Unveiling Rare Patterns: Enhancing Interpretability and Discovering Unexpected Insights in Data

DISSERTATION

zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von M.Sc. Sadeq Hussein Saleh Darrab

geb. am 01.01.1981 in Hajah

Gutachterinnen/Gutachter

Prof. Dr. rer. nat. habil. Gunter Saake
Prof. Dr. Simeon Simoff
Prof. Dr. Mohamed A. A. Al-qaness

Magdeburg, den 22.04.2025

Darrab, Sadeq Hussein Saleh:
Unveiling Rare Patterns: Enhancing Interpretability and Discovering Unexpected Insights in Data
Dissertation, Otto-von-Guericke University Magdeburg, 2025.

Abstract

In data analysis, rare pattern mining is essential for uncovering valuable insights by identifying uncommon patterns that often escape traditional methods. Despite notable advancements in deep learning, challenges related to explainability and interpretability persist, particularly when it comes to identifying and understanding rare, but impactful patterns. Rare pattern mining, especially association rule mining, offers an interpretable approach that provides insights that are crucial for informed decision-making.

This doctoral dissertation addresses key challenges in rare pattern mining through four main contributions: (1) developing an efficient method for discovering rare patterns, (2) discovering the concise representation of rare patterns to reduce redundancy, (3) unveiling interesting patterns by filtering out irrelevant and noisy patterns, and (4) demonstrating practical applicability through a case study focused on interpretability. First, we introduce a novel depth-first search approach to overcome the limitations of traditional methods, achieving substantial improvements in speed and memory efficiency, particularly in sparse datasets, and thus establish a new standard for rare pattern extraction. To address redundant pattern generation, we propose a method for identifying maximal rare patterns, providing a concise output that enhances analysis efficiency. In addition, we introduce a model that isolates interesting patterns by effectively filtering out irrelevant and noisy data, ensuring that only impactful patterns are highlighted, thereby enhancing interpretability and reducing information overload.

The practical implications of these contributions are demonstrated in a healthcare case study focused on heart disease, where interpretability and transparency are essential. By analyzing patient data, our methods not only reveal critical risk factors but also identify vulnerability patterns in asymptomatic individuals, enabling early intervention and improved health outcomes. This case study underscores the value of model transparency and interpretability, particularly in high-stakes applications.

Through these contributions, this thesis makes significant advances in the efficiency, relevance, and interpretability of rare pattern mining. The proposed methods provide robust, actionable insights across various domains, providing a foundation for more effective data-driven decision-making.

Zusammenfassung

In der Datenanalyse spielt die Identifikation seltener Muster eine entscheidende Rolle bei der Gewinnung wertvoller Erkenntnisse, da solche ungewöhnliche Muster häufig von herkömmlichen Methoden übersehen werden. Trotz bemerkenswerter Fortschritte im Bereich des Deep Learning bestehen weiterhin Herausforderungen im Hinblick auf Erklärbarkeit und Interpretierbarkeit, insbesondere bei der Identifikation und dem Verständnis seltener, aber bedeutender Muster. Das Mining seltener Muster, insbesondere das Assoziationsregel-Mining, bietet einen interpretierbaren Ansatz, der Einblicke liefert, die für fundierte Entscheidungsprozesse unerlässlich sind.

Diese Doktorarbeit befasst sich mit den zentralen Herausforderungen im Mining seltener Muster durch vier wesentliche Beiträge: (1) die Entwicklung einer effizienten Methode zur Entdeckung seltener Muster, (2) die Bestimmung einer kompakten Darstellung seltener Muster zur Reduzierung von Redundanz, (3) die Aufdeckung interessanter Muster durch das Herausfiltern irrelevanter und störender Muster und (4) die Demonstration der praktischen Anwendbarkeit durch eine Fallstudie mit Fokus auf Interpretierbarkeit. Zunächst wird ein neuartiger Tiefensuchansatz vorgestellt, der die Einschränkungen traditioneller Methoden überwindet und deutliche Verbesserungen in Bezug auf Geschwindigkeit und Speichereffizienz, insbesondere in dünn besetzten Datensätzen, erzielt, wodurch ein neuer Standard für die Extraktion seltener Muster gesetzt wird. Um die Erzeugung redundanter Muster zu vermeiden, schlagen wir eine Methode zur Identifikation maximal seltener Muster vor, die eine kompakte Ausgabe liefert und die Analyseeffizienz erhöht. Darüber hinaus führen wir ein Modell ein, das interessante Muster isoliert, indem es irrelevante und störende Daten effektiv herausfiltert, sodass nur wirkungsvolle Muster hervorgehoben werden, wodurch die Interpretierbarkeit verbessert und die Informationsüberlastung reduziert wird.

Die praktischen Implikationen dieser Beiträge werden in einer Fallstudie im Gesundheitswesen, die sich auf Herzerkrankungen konzentriert, veranschaulicht, bei der Interpretierbarkeit und Transparenz wesentlich sind. Durch die Analyse von Patientendaten enthüllen unsere Methoden nicht nur kritische Risikofaktoren, sondern identifizieren auch spezielle Faktoren bei asymptomatischen Personen, was eine frühzeitige Intervention und verbesserte Gesundheitsprognosen ermöglicht. Diese Fallstudie unterstreicht den Wert von Modelltransparenz und Interpretierbarkeit, insbesondere in Anwendungen mit hohen Anforderungen.

Durch diese Beiträge macht diese Arbeit bedeutende Fortschritte in der Effizienz, Relevanz und Interpretierbarkeit des Mining seltener Muster. Die vorgeschlagenen Methoden bieten robuste, umsetzbare Einblicke in verschiedenen Bereiche und bilden eine Grundlage für eine effektivere, datengetriebene Entscheidungsfindung.

Contents

List of Figures

List of Tables

1. Introduction

Data mining is a crucial process for uncovering patterns and gaining insights from datasets. It plays a significant role in fields such as data analysis, machine learning, and business intelligence because it enables the discovery of connections, associations, correlations, and trends that would otherwise remain hidden. A key technique in data mining is pattern mining, which involves identifying and extracting recurring patterns or structures within datasets to obtain meaningful, interesting, and actionable knowledge [Luna et al., 2019]. Depending on the data being analyzed, this process can uncover various types of patterns, including sequential patterns, frequent patterns, and graphs.

In recent years, deep learning (DL) has gained significant attention because of its ability to learn complex patterns from vast datasets. Despite their impressive performance, DL models have been criticized for their lack of explainability and interpretability [Rudin, 2019]. These models typically operate as "black boxes," making it difficult for practitioners and domain experts to understand the rationale behind their decisions or derive actionable insights from underlying patterns. This lack of transparency poses critical challenges in fields such as healthcare, finance, and security, where trust in and understanding of the decision-making process is critical.

In contrast, Association Rule Mining (ARM) [Troncoso-García et al., 2023] provides a more transparent and interpretable approach to knowledge discovery. ARM uncovers relationships between variables and reveals not only frequent patterns but also rare patterns that may be of particular significance. Rare pattern mining within ARM is particularly valuable because it focuses on identifying less common yet highly impactful patterns that are often overlooked by traditional frequent pattern mining and black-box DL models. These rare patterns can provide critical insights into exceptional behaviors or anomalies such as early signs of disease, fraudulent transactions, and emerging trends in large datasets.

The ability to identify rare patterns is vital because these infrequent occurrences often represent anomalies or outliers that can have significant implications. ARM provides clear and interpretable results, making it particularly suitable for applications in which

understanding the rationale behind discovered patterns is essential. By capturing both frequent and rare patterns, ARM addresses some of the key limitations of DL models and provides valuable, actionable insights that enhance decision-making in complex, data-rich environments.

Thus, the shift toward ARM, with a particular emphasis on rare pattern mining, presents a transparent and powerful alternative for knowledge discovery. This approach not only mitigates the limitations associated with DL models but also enables the extraction of valuable, interpretable insights that are crucial in fields where transparency and trust in the decision-making process are crucial.

1.1 Frequent pattern mining

Frequent pattern mining (FPM) is a widely utilized pattern mining technique aimed at discovering recurring patterns and connections within datasets. This method can be employed to detect associations or correlations between various items in a dataset, such as products purchased by customers, symptoms experienced by patients, or keywords found in documents. One of the primary challenges of FPM is to efficiently uncover frequent patterns, which are subsets of items that co-occur in a dataset with a frequency exceeding a predefined minimum support threshold. To address this issue, various algorithms for FPM, such as Apriori, FP-growth, Eclat, and PrefixSpan, have been developed, which employ different strategies to reduce the search space and improve performance [Luna et al., 2019].

Although FPM is a widely used method for identifying frequent patterns, it faces several challenges that hinder its usefulness and applicability. One of the challenges in pattern mining is the potential for a large number of patterns to emerge when a low minimum support threshold is used, including rare interesting patterns. Therefore, when the threshold is extremely low, the analysis becomes computationally intractable. However, setting the constraint too high can limit the analysis to a small number of common patterns, potentially missing more subtle and valuable rare patterns. Therefore, identifying the optimal balance between constraints and extraction is critical for real-world applications that require the discovery of meaningful and surprising patterns. This is particularly relevant in areas such as fraud detection, insurance and healthcare analytics, and fault detection in safety-critical systems, where rare patterns may indicate critical events.

Another challenge associated with FPM lies in its tendency to generate patterns that mainly align with expected phenomena, leading to a focus on widely recognized occurrences. The drawback of this tendency becomes evident when it may result in overlooking patterns that occur less frequently but offer valuable insights and meaningful information. In certain situations, it is more efficient and desirable to emphasize the exploration and extraction of less common patterns, as they allow for the identification of outliers, anomalies, or hidden correlations that may be essential for a thorough understanding of the underlying data. Hence, common (frequent) patterns may already be well known and therefore cannot provide new insights. To achieve a more comprehensive and meaningful analysis of datasets, greater emphasis should be placed on less frequently occurring patterns that generate new and interesting insights rather than focusing solely on well-established occurrences of FPM.

1.2 Rare pattern mining

Rare pattern mining is a sophisticated method that is increasingly used to address these challenges. Rare pattern mining has emerged as a pivotal methodology in the field of data mining, distinguished by its specialized focus on uncovering infrequent yet substantial patterns within datasets. Unlike traditional mining techniques that predominantly target frequent occurrences, rare pattern mining seeks to identify events or patterns that may be sparse but possess considerable importance. The main objective of rare pattern mining is to uncover hidden rare events, outliers, or exceptional associations that might otherwise remain undetected by conventional mining methods [Akdas et al., 2024; Darrab et al., 2021b; Gui et al., 2024; Liu et al., 2023].

Rare patterns do not occur frequently and are crucial for a number of applications due to their unique capabilities, such as fraud detection, medicine, anomaly detection, and security. In fraud detection, rare pattern mining can be used to identify fraudulent transactions or activities that deviate from normal behavior or patterns. For example, rare pattern mining can detect credit card fraud by identifying unusual combinations of items or amounts purchased by cardholders. Healthcare leverages this methodology to diagnose rare diseases and ensure a comprehensive understanding of diverse medical conditions. For example, rare pattern mining can detect drug reactions by identifying rare associations between drugs and symptoms. Rare pattern mining can be used to detect anomalies or outliers in data, which may indicate errors, faults, or malicious attacks. For example, rare pattern mining can detect network intrusions by identifying uncommon patterns in network traffic or packets. Beyond these, safety-critical systems find rare pattern mining essential for detecting faults and uncommon events, contributing to improved system reliability. The versatility of rare pattern mining is underscored by its applicability in domains where uncovering unusual associations is vital for informed decision making [Akdas et al., 2024; Borah and Nath, 2020; Darrab et al., 2021b].

This innovative approach targets infrequent patterns with the potential to produce significant outcomes, making it an effective tool for understanding and adapting to the intricacies of diverse domains. Rare pattern mining offers an in-depth understanding of patterns by overcoming the limitations of conventional frequent pattern mining, thereby enriching insights and fostering more effective outcomes. Throughout this thesis, we employ the expressions ”infrequent patterns” and ”rare patterns” interchangeably.

1.3 Challenges of rare pattern mining

In contrast to conventional frequent pattern mining, rare pattern mining focuses on less common events, allowing for a more in-depth understanding of datasets and the discovery of unknown relationships that might otherwise go undetected. Consequently, specialized algorithms and evaluation metrics are required to address the unique characteristics of these infrequent patterns. The extraction of infrequent patterns from data is a complex process because of the large size and noisy nature of data, which can hinder the identification of useful patterns and relationships [Borah and Nath, 2019; Darrab et al., 2021b]. Considering the complexities inherent in

mining rare patterns, several key challenges must be addressed to ensure the efficacy and relevance of the extracted insights. This thesis focuses on various challenges, such as performance, redundancy, interpretability, and interestingness.

- Performance: This remains a foremost concern, given the computational demands involved in identifying rare patterns within vast datasets. Because of the sparsity and irregularity characteristics of rare patterns, traditional pattern mining algorithms may have difficulty handling them efficiently, which necessitates the development of novel approaches capable of mining rare patterns efficiently.
- Redundancy: The issue of redundancy is a significant obstacle to the effective mining of rare patterns. These results can be cluttered and obscured by redundant patterns that convey redundant information or provide little additional insight. Therefore, it is imperative for rare pattern mining endeavors to be optimized by filtering redundant patterns and prioritizing those that contain unique or actionable insights.
- Interestingness: Identifying statistically significant patterns is not sufficient; it is critical to identify patterns that are both statistically significant and unexpected. By incorporating measures of interestingness into the mining process, we can prioritize patterns that offer the greatest potential to yield novel insights or drive meaningful actions.
- Interpretability: It emerges as another critical challenge in the context of rare pattern mining. As pattern complexity increases, it becomes more difficult to interpret its implications and derive actionable insights. Ensuring that the extracted patterns are interpretable to domain experts is essential for facilitating informed decision making and deriving value from the mining process.

For rare pattern mining to succeed, we must address these diverse challenges and enable it to exploit its full potential to uncover valuable information from rarely occurring data.

1.4 Goal of this thesis

This thesis aims to provide efficient solutions to the challenges associated with rare pattern mining. The focus of this thesis is to identify the limitations of current methods in generating interesting rare patterns to prevent negative outcomes in a variety of areas. This thesis aims to develop effective methods for discovering rare patterns while considering issues such as the generation of a large number of rules, redundancy, interpretability, and level of interest. To examine rare patterns in a real-world problem, the proposed methods are applied to a case study analysis, specifically focusing on heart disease within the healthcare sector. This thesis aims to answer the following questions through pattern mining:

1. What are the current methods available for mining rare patterns, and how do they differ in their effectiveness?
2. How can we generate rare patterns efficiently to optimize processing time and resources?

3. How can we focus on identifying meaningful rare patterns while minimizing redundancy?
4. Among the many unpredictable rare patterns, how can we filter out the most interesting and unexpected insights?
5. How can rare pattern mining be effectively applied in a real-world case study to demonstrate its practical importance and validate its impact?

1.5 Main contributions

In this subsection, we summarize the contributions made in response to the five questions outlined in the previous subsection.

For Question 1, we conducted a literature review [Darrab et al., 2021b] and published a survey that answers: *What are the current methods available for mining rare patterns, and how do they differ in their effectiveness?*

This section highlights our contributions to the challenges addressed in this thesis, focusing on the discovery of new, unexpected, and meaningful patterns in response to Questions 2 through 5.

1.5.1 Developing an efficient method for discovering rare patterns

The task of extracting unusual patterns from data has been the subject of extensive research. Existing methods can be divided into two categories based on their original approach: apriori-based [Agrawal and Srikant, 1994] and FP-based [Wang et al., 2002]. Both of these methods have limitations when addressing rare-pattern problems. Apriori-based methods are time consuming and require a large amount of memory because of the need to rescan the entire dataset and use a generate-test approach. To overcome this limitation, tree-based methods have been proposed for mining rare patterns using a depth-first approach. However, these methods are difficult to scale when mining sparse datasets, which commonly contain unusual patterns. Therefore, designing an algorithm to mine interesting rare patterns from sparse datasets remains a significant challenge. To address this challenge, we propose a novel pre-post rare method for rare pattern mining called the Rare Pre-Post(RPP) algorithm. The experimental results demonstrate that our RPP algorithm outperforms state-of-the-art methods for rare pattern mining [Darrab et al., 2020]. The details of this contribution and the proposed method are discussed in Chapter 4.

1.5.2 Discovering the concise representation of rare patterns

Although our proposed algorithm, which was discussed in the previous subsection, is designed to efficiently generate rare patterns, one drawback of the existing methods for rare pattern mining is that they often produce a large number of redundant patterns. This poses problems in terms of improving the efficiency and effectiveness of the process. Rare pattern mining is highly sensitive to the support threshold, because a low value can result in an excessive number of patterns. Mining a large number of patterns can overwhelm decision makers, and many of the generated patterns

may be redundant, making it necessary to aggregate them for subsequent analysis. This problem is particularly severe when dealing with rare events, as it is crucial to identify them as early as possible to avoid unfavorable outcomes. To address this issue, it is desirable to extract a concise representation of rare patterns to provide more meaningful results. Generating a concise representation of rare patterns often involves summarizing the entire set of patterns, which can be significantly smaller than the complete set of patterns. In addition, discovering concise representations is often faster than discovering a complete set of rare patterns. Although several methods have been proposed to extract maximal and closed frequent patterns, there is no known method for mining concise representations of rare patterns. Therefore, designing a method to identify condensed rare patterns without redundant patterns is an important research topic. To address this challenge, we developed an algorithm called MaxRI, which can discover maximal rare patterns [Darrab et al., 2021a]. We provide details of this contribution and the method proposed in Chapter 5.

1.5.3 Interestingness of patterns

Pattern mining is crucial in data analysis to unearth valuable insights from datasets. However, sifting through vast amounts of data to identify meaningful patterns is challenging. Rare pattern mining has emerged as a solution that focuses on extracting non-conforming patterns that yield actionable insights. Existing methods, although abundant, often flood decision makers with an overwhelming number of rules complicate the analysis. For instance, a pattern with d items can generate $2^d - 1$ rules, which necessitates extensive downstream analysis. To streamline this process, it is imperative to extract fewer yet more meaningful rules. A clustering-based approach utilizing DBSCAN [Bui-Thi et al., 2020] was proposed to identify unexpected rare rules. However, this method has notable limitations, including suboptimal performance due to its reliance on the Apriori algorithm, potential oversight of nested cluster structures, and sensitivity to DBSCAN's hyperparameters.

To address these challenges, we propose OPECUR, an efficient model based on OPTICS clustering of ECLAT-generated unexpected rules. The experimental results demonstrate OPECUR's superiority of OPECUR in terms of the F1-score, AUC, and speed compared with DBSCAN-based methods, providing faster and more insightful rule recovery [Darrab et al., 2022a].

Clustering models such as DBSCAN and OPECUR provide end users with manageable rule sets. However, these models suffer from limitations, including performance degradation and the need for parameter adjustment. In addition, they failed to discover a complete set of rare patterns. To address this challenge, we proposed UCRP-miner: Mining patterns that matter, a model that utilizes frequent patterns to identify unexpected rules. This approach allows for efficient retrieval of complete set of unexpected rules and produces actionable rules that are both new and interesting. To evaluate the effectiveness of our model, we conducted experiments using real-life datasets and found that it outperformed the state-of-the-art model in terms of both time and accuracy, generated usable patterns, and produced non-redundant rules that significantly reduced the effort required to generate interesting patterns [Darrab et al., 2022b]. Chapters 6 and 7 provide a detailed explanation of the contributions and the proposed methodologies.

1.5.4 A case study with interpretability focus

Heart disease affects a significant proportion of the population and is a leading cause of mortality worldwide. Data mining methods, such as logistic regression, neural networks, and random forests [Singh and Kumar, 2020], have been utilized to detect and predict heart diseases by analyzing patient data, uncovering patterns, and forecasting outcomes. However, these methods often lack interpretability, which makes it difficult to understand the decision-making process of the model. It is crucial that the model's predictions are explainable to provide insights into its reasoning and enhance trust.

Association rule mining offers a promising approach to address interpretability by identifying the relationships between variables and the risk of heart disease. This method can reveal influential factors and their relationships with model-predicted outcomes, which aids in understanding heart disease risks. However, traditional association rule mining faces challenges, including generating an overwhelming number of rules, many of which may be uninformative or irrelevant, and potentially overlooking rare yet significant patterns with low support.

To address these limitations, we propose Exploring the Predictive Factors of Heart Disease using Association Rule Mining (EPFHD-RARMING), a model designed to generate only relevant, interpretable rules. This approach highlights key factors and their relationships, supporting the early detection of cardiovascular diseases. Our model provides a deeper understanding of heart disease etiology based on experimental findings, uncovering both common and rare patterns, including those with low support, which may be crucial for early diagnosis [Darrab et al., 2024]. The details of our contribution and methodology are explained in Chapter 8.

1.6 Corresponding publications

This dissertation is based on a collection of peer-reviewed articles presented at conferences and workshops and published in journals. The articles were ordered according to their appearance in this dissertation.

1. Sadeq Darrab, David Broneske, and Gunter Saake. *RPP algorithm: A method for discovering interesting rare itemsets*, volume 1234 CCIS. Springer Singapore, 2020.
2. Sadeq Darrab, David Broneske, and Gunter Saake. *MaxRI: A method for discovering maximal rare itemsets.* In *2021 4th International Conference on Data Science and Information Technology*, pp. 334–341, 2021.
3. Sadeq Darrab, David Broneske, and Gunter Saake. *Modern applications and challenges for rare itemset mining. International Journal of Machine Learning (IJML)*, 11(3):208–218, 2021.
4. Sadeq Darrab, Priyamvada Bhardwaj, David Broneske, and Gunter Saake. *OPECUR: An enhanced clustering-based model for discovering unexpected rules.* At the *International Conference on Advanced Data Mining and Applications*, pages 29–41. Springer, 2022.

5. Sadeq Darrab, David Broneske, and Gunter Saake. *UCRP-Miner: Mining patterns that matter*. At the *2022 5th International Conference on Data Science and Information Technology (DSIT)*, pages 1–7. IEEE, 2022.
6. Sadeq Darrab, David Broneske, and Gunter Saake. *Exploring the predictive factors of heart disease using rare association rule mining. Scientific Reports*, 14(1):18178, 2024.

1.7 Outline of this thesis

The remainder of this dissertation is structured as follows. Chapter 2 provides a comprehensive background on rare pattern mining. Chapter 3 presents a survey of the existing rare pattern mining approaches. In Chapter 4, we address the first challenge of this thesis and propose an efficient algorithm to identify rare patterns. Chapter 5 addresses the second challenge, focusing on reducing redundancy in rare pattern mining. Chapters 6 and 7 explore the third challenge, identifying unexpected and interesting rules. In Chapter 8, we apply these findings to the healthcare domain, particularly in the discovery of rare patterns associated with heart diseases. Finally, Chapter 9 concludes the thesis and outlines the future research directions.

2. Background

This chapter explores the development and significance of rare pattern mining, focusing on Association Rule Mining (ARM) and clustering as foundational unsupervised techniques. These methods are crucial for rare pattern mining because they enable the discovery of previously undetected structures and relationships within data, without relying on predetermined labels or targets. A thorough understanding of these foundational principles is essential for recognizing their importance and role in contemporary data analysis [Hassija et al., 2024].

Association Rule Mining (ARM) is a vital area within data mining that focuses on uncovering subtle, nontrivial, and significant patterns from extensive datasets. The primary objective of ARM is to identify the relationships between variables in large databases, forming the basis for a wide range of applications, including market basket analysis, bioinformatics, and recommendation systems [Fournier-Viger et al., 2017; Zhang and Wu, 2011]. Traditionally, ARM has concentrated on frequent patterns, that is, patterns that commonly co-occur within a dataset above a certain threshold. However, this focus on frequency often overlooks rare patterns, which despite their infrequent occurrence, can offer profound insights [Borah and Nath, 2017; Lu et al., 2020].

Rare pattern mining is a specialized subfield of ARM that focuses on identifying infrequent patterns and recognizing their potential to reveal unforeseen connections and provide substantial value across a wide range of domains, including healthcare, retail, and banking [Ninoria and Thakur, 2020; Shrivastava and Johari, 2016]. Mining rare patterns presents particular challenges due to their low support values, meaning they do not frequently appear in datasets. However, as noted by Lu et al. [2020], these rare patterns represent novel and valuable knowledge.

Although traditional frequent pattern mining has been extensively studied and applied, rare pattern mining has emerged as a crucial complement, enabling a more comprehensive understanding of data by uncovering the full spectrum of associations. This approach addresses the limitations of focusing solely on frequent patterns by emphasizing the significance of rare occurrences, which can lead to actionable insights

across various fields, as demonstrated by Borah and Nath [2017]; Ninoria and Thakur [2020]; Shrivastava and Johari [2016]. The continued development of algorithms and techniques for rare pattern mining is essential for advancing the field of data mining, ensuring that both frequent and infrequent phenomena are captured and analyzed, as highlighted by Lu et al. [2020].

Rare pattern mining has received considerable attention in recent years because of its wide-ranging implications across a variety of sectors. It is increasingly important for informed decision making and the extraction of practical insights to identify uncommon yet valuable patterns. An overview of rare pattern mining is provided in this chapter, setting the stage for a more detailed examination of ARM, clustering, and the specific challenges and opportunities within rare pattern mining in the following sections.

2.1 Problem description

In this section, we explore the tasks of frequent and rare pattern mining, along with related concepts integral to these processes. We discuss the fundamental ideas behind pattern mining, particularly focusing on the identification of frequent and rare patterns, and the extraction of interesting association rules between patterns within transactional datasets. To elucidate these concepts, we present the following motivating example:

Motivating Example

Consider the transaction dataset provided in Table 2.1, where the minimum support threshold is set to 50%.

Table 2.1: A toy transaction dataset

id	Transaction
t1	a, b, c, d
t2	b, d
t3	a, b, c, e
t4	c, d, e
t5	a, b, c

Figure 2.1 shows a comprehensive subset lattice corresponding to the five items listed in Table 2.1. This lattice captures all the possible patterns that can be extracted from the toy transaction dataset. The structure of the lattice is organized such that each level contains patterns of the same length, with the lattice culminating at the top with the null set. The diagram also explicitly annotates the frequencies of each pattern, as they occur within the dataset. This systematic representation facilitates a clearer understanding of the distribution and relationships of the patterns within the given transactional data.

In this example, frequent patterns appear in at least 50% of the transactions as per the minimum support threshold. Patterns such as {a, b, c}, which appear in three of the five transactions, are considered frequent. In contrast, patterns such as {e},

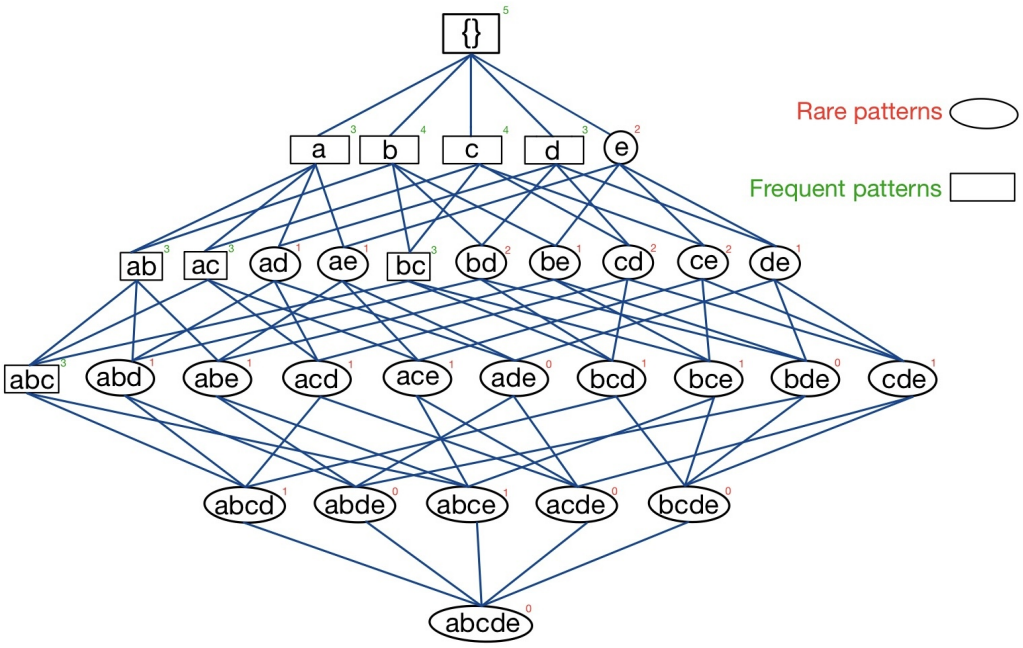


Figure 2.1: A lattice structure illustrating a hierarchical arrangement of patterns and their corresponding frequencies. In this lattice, square-shaped elements represent frequent patterns, while oval-shaped ones signify rare patterns.

which appear less frequently, are categorized as rare. Identifying both frequent and rare patterns provides valuable insights into the underlying data structure, enabling the discovery of significant associations that might otherwise go unnoticed.

The lattice structure in Figure 2.1 effectively demonstrates the organization and visualization of frequent and rare patterns, making it easier to analyze and interpret the data. By examining this lattice, one can observe how patterns are interconnected and how their frequencies vary, which is essential for understanding pattern mining dynamics in transactional datasets. This understanding lays the groundwork for further exploration of the methods and challenges involved in frequent and rare pattern mining.

Definition 2.1: Support of a Pattern

Support of a Pattern: The support of a pattern X in a dataset D is the number of transactions within D that contain X. It quantifies the absolute frequency of the occurrence of a pattern in a dataset [Luna et al., 2019]. Mathematically, it can be expressed as

$$\sup(X) = |\{t \in D \mid X \subseteq t\}|.$$

where $|\cdot|$ denotes the cardinality of the set.

Following our motivating example presented in Table 2.1, which is illustrated in the lattice shown in Figure 2.1, the numbers above each node represent the pattern support. For example, 'a: 3', 'ac: 3', and so forth.

Definition 2.2: Relative Support of a Pattern

The relative support for a pattern X in a dataset D is the proportion of transactions within D that contain X relative to the total number of transactions in D. This measure provides an understanding of the pattern's prevalence relative to the dataset size [Luna et al., 2019]. It is given by:

$$relSup(X) = \frac{sup(X)}{|D|},$$

where $|D|$ denotes the total number of transactions in the dataset.

For example, the relative support of pattern bc is $\frac{3}{5}$, where five represents the total number of transactions in Table 2.1.

Definition 2.3: Frequent Patterns

A pattern X whose support $sup(X)$ satisfies a user-specified support threshold $minSup$ is called a *frequent pattern* [Luna et al., 2019]. Formally, this can be expressed as

$$sup(X) \geq minSup,$$

Patterns such as 'a: 3', 'c: 4', 'ac: 3', and 'ab: 3' are considered frequent patterns, as their support is no less than the threshold $minSup$ of 50%.

2.2 Compact representations of patterns

Mining all the patterns in the previous section can result in a large number of patterns. To overcome these challenges, various methods have been developed to uncover condensed patterns. An interesting aspect of generating these patterns is to generate a few patterns that compress the entirety of the generated patterns. These patterns are called closed and maximal frequent patters [Wu et al., 2021].

Definition 2.4: Closed Frequent Patterns

A closed frequent pattern is a frequent pattern for which no proper superset has the same support as the patterns itself. A pattern C is a closed frequent pattern if it is frequent and there are no patterns C' such that $C \subset C'$ and $\sup(C) = \sup(C')$ [Rodríguez-González et al., 2018]:

$$\forall C' \supset C, \quad \sup(C') > \sup(C) \quad \text{or} \quad \sup(C') < \sup(C)$$

These are subsets of frequent patterns that do not have a larger set with identical frequency, representing complete and non-redundant item sets. For example, as illustrated in Figure 2.1, the pattern 'ac: 3' is not a closed pattern because it has a superset, 'abc: 3', with the same support level. In contrast, a closed pattern like 'b: 4' does not have any superset sharing the same support, making it a unique and closed frequent item set within the dataset. Identifying such patterns is crucial for

minimizing redundancy in the data mining process and ensuring that only the most significant sets are considered. Closed patterns are important because they generate complete information without loss. This is because subsets of patterns with the same support can be derived from their superset, which shares the same support level. By focusing on closed patterns, we ensure that all relevant information is captured efficiently while also simplifying the dataset by eliminating redundant patterns. This efficiency is critical in data mining, where the goal is to extract the most meaningful insights from large datasets without overlooking significant details.

Definition 2.5: Maximal Frequent Patterns

A maximal frequent pattern is a frequent pattern for which no proper superset is frequent. It aims to identify the largest sets of patterns that frequently appear together while ensuring that none of their supersets meet the minimum support threshold [Yang, 2004].
A pattern M is a maximal frequent pattern if it is frequent, and there is no patterns M' such that $M \subset M'$ and M' is frequent:

$$\forall M' \supset M, \sup(M') < \text{minsup}$$

Maximal frequent patterns are subsets of frequent patterns that do not have a larger set in which they are frequent. Such patterns are also known as maximal frequent patterns. They comprise complete and non-redundant patterns. For example, in Figure 2.1, 'abc: 3' is a maximal frequent pattern since it has no frequent superset; its superset 'abcd: 1' is infrequent as its support drops below 3. The maximal pattern set is the largest set of frequent patterns. These patterns represent the most concise representations of the data and lie at the boundary where rare patterns may occur.

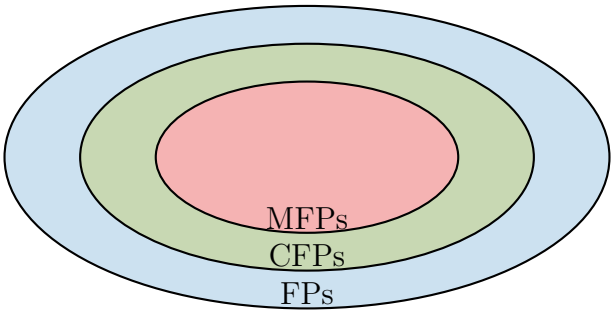
To understand the structure and organization of patterns, it is essential to understand the relationships between Frequent Patterns (FPs), Closed Frequent Patterns (CFPs), and Maximal Frequent Patterns (MFPs). Given a dataset D and minimum support threshold σ, the relationships can be defined as follows:

$$\text{MFPs} \subseteq \text{CFPs} \subseteq \text{FPs}$$

where:

- $\text{FPs} = \{I \mid s(I) \geq \sigma\}$ represents the set of all patterns that are frequent, meeting or exceeding the minimum support threshold.
- $\text{CFPs} = \{I \in \text{FPs} \mid \nexists J \supset I, s(J) = s(I)\}$ includes those patterns that are closed, meaning no superset of I has the same support as I.
- $\text{MFPs} = \{I \in \text{CFPs} \mid \nexists J \supset I, s(J) \geq \sigma\}$ contains those patterns that are maximal, as no superset of these patterns is frequent.

Visual representation can enhance our understanding of hierarchical structures. The following diagram illustrates the relationships between these subsets:



Frequent, closed frequent, and maximal frequent patterns represent well-known phenomena and situations that commonly occur within datasets and typically generate no surprise. These patterns are often predictable and, while useful, do not offer new insights into the less understood aspects of the data.

Focusing on rare patterns can reveal unexpected insights and more meaningful results. This matters greatly in situations where detecting rare patterns early can avoid negative outcomes. In healthcare, for example, identifying unusual patient reactions to treatment allows timely intervention.

Exploring these rare patterns requires innovative data analysis techniques that can distinguish between noise and genuinely insightful anomalies. By focusing on the nuances of these less frequent occurrences, we can uncover valuable insights that remain hidden when analyzing only the most common patterns. This approach not only enriches our understanding, but also enhances decision-making processes, especially in high-stakes environments where the cost of missing such patterns is significant.

2.3 Rare patterns

Frequent pattern mining is pivotal for identifying patterns within a dataset that meets a specific support threshold. As illustrated in Figure 2.1 and detailed in Table 2.1, with $n = 5$, there is the potential to generate $2^n = 32$ patterns, including null patterns. Of these, only 10 patterns, highlighted with rectangular shapes, were classified as frequent because they met or exceeded a support threshold of 3. The remaining 22 patterns, depicted with circular shapes, were initially disregarded as infrequent and were potentially treated as noise. However, these patterns can offer critical insights, particularly in areas where rare events are key indicators.

In practical domains, such as security monitoring, healthcare analytics, system failure detection, and credit card fraud prevention, recognizing these infrequent patterns is crucial for identifying potential threats or anomalies. For example, consider frequent patterns such as 'abc: 3', 'ab: 3', and 'ac: 3' where the numbers indicate their support, while 'abcd: 2' is excluded from frequent mining. This exclusion prompts an inquiry into whether this pattern is merely noise or if it could indicate underlying insights. The pattern 'abcd' might represent a sequence of events or transactions that, although infrequent, could reveal unique behavior or an unforeseen threat.

Analyzing infrequent (rare) patterns is vital for gaining a comprehensive understanding of complex systems and datasets and ensuring their proactive management.

Identifying these overlooked patterns can lead to the discovery of hidden insights, risk mitigation, and enhancement of decision-making processes in various fields [Bhatt and Patel, 2015; Borah and Nath, 2020; Gui et al., 2024; Kataria et al., 2019; Liu et al., 1999; Lu et al., 2020; Szathmary et al., 2007; Tsang et al., 2011; Vanamala et al., 2021].

Understanding the different types of patterns is crucial for effective analysis. Here, we define three key types of patterns: rare patterns, interesting rare patterns, and maximal rare patterns, each playing a unique role in uncovering insights from the data, particularly when frequent pattern analysis may overlook significant but less common phenomena.

Definition 2.6: Rare Patterns

Rare patterns are patterns X that do not meet the minimum support threshold minSup. Despite their infrequent occurrence in a dataset, these patterns may provide valuable insights [Darrab et al., 2021b]. Mathematically, they are expressed as

$$\sup(X) < \text{minSup}.$$

Similar to the example illustrated in the 2.1, patterns such as 'bd: 2', 'cd: 2', and 'abcd: 1' are examples of rare patterns, as their support levels fall below the minimum support threshold minSup = 3.

Definition 2.7: Interesting Rare Patterns

A rare pattern X [Darrab et al., 2021b] is considered interesting if it satisfies both the condition of rarity $\text{Sup}(X) < \text{minSup}$ and a minimum support threshold minRare, ensuring that the pattern is substantial enough to warrant further analysis:

$$\text{Sup}(X) < \text{minSup} \wedge \text{Sup}(X) \geq \text{minRare}$$

where $\text{Sup}(X)$ represents the support of the item set X, minSup defines the upper limit for maintaining rarity, and minRare sets the lower boundary of interest. Suppose minRare = 1, then patterns such as 'bd: 2', 'cd: 2', 'ade:1' are considered interesting rare patterns.

Definition 2.8: Maximal Rare Patterns

A Maximal Rare Pattern X is termed a maximal rare pattern if it qualifies as an interesting rare pattern, as per Definition 2.3, and there is no other interesting rare pattern Y such that $X \subset Y$. These patterns represent the largest of the rare patterns that maintain their properties of interest without being overshadowed by a larger pattern. The complete set of maximal rare patterns in the motivating example is {'abcd: 1', 'acde: 1'}.

2.4 Challenges in mining interesting rare patterns

Identifying interesting rare patterns [Darrab et al., 2021b; Koh and Ravana, 2016; Lu et al., 2020; Tsang et al., 2011] within large datasets is a complex task that presents several substantial challenges, as discussed in the introduction chapter. Although most of these patterns may be uninteresting or simply represent noise within the data, a small subset can hold considerable significance. The successful identification of these rare but valuable patterns is crucial for deriving meaningful insights for practical applications. For clarity, in this section and the entire thesis, the terms *rare patterns*, *infrequent patterns*, and *low-support patterns* are used interchangeably to refer to patterns with low support.

- **Volume of Patterns**: The combinatorial nature of pattern generation in large datasets leads to an exponential increase in the number of possibilities, resulting in a combinatorial explosion. This vast search space complicates the process of distinguishing meaningful patterns from an overwhelming number of potential candidates, making the identification of truly valuable but rare patterns exceedingly challenging.
- **Low support**: Patterns identified during the mining process often exhibit very low support, meaning they occur infrequently within the dataset. Traditional data mining approaches tend to disregard these low-support patterns, assuming that their rarity diminishes their significance. However, infrequent patterns can still offer critical insights, especially in domains where rare events are of significant interest, such as fraud detection or rare disease identification.
- **Distinguishing noise from significant patterns**: A core challenge in rare pattern mining is differentiating between noise and genuinely valuable patterns. Noise can easily overshadow significant rare patterns, making it difficult to focus on data that holds the most potential for actionable insights. The ability to filter out noise while retaining important patterns is crucial for the success of rare pattern mining.
- **Effective identification of rare patterns**: Identifying rare patterns that are both meaningful and actionable requires advanced analytical techniques and often a deep understanding of the domain. Standard techniques may fail to detect these patterns, necessitating the use of more sophisticated methods that can not only identify rare patterns, but also evaluate their relevance within a specific context.

Addressing these challenges necessitates the application of advanced data mining techniques specifically designed to efficiently navigate the expansive search space and isolate patterns of true significance. Techniques such as lowering the support thresholds, utilizing supplementary measures such as confidence and lift, and integrating domain-specific expertise into the mining process are imperative. These methodologies not only reduce the complexity of the search space but also increase the probability of uncovering rare patterns that may signify atypical yet crucial phenomena. The insights derived from these patterns are particularly valuable in fields where the timely identification of rare events is paramount, such as in healthcare for the detection of rare diseases, finance for the identification of fraudulent activities, or cybersecurity for the recognition of anomalies.

Although the detection of meaningful rare patterns is inherently challenging, their potential benefits are substantial. The successful identification and exploitation of these rare, yet insightful patterns can lead to significant advancements in the understanding and management of complex systems and environments. Upon identification of these patterns, analysis of the relationships between them is a critical step in the data mining process. This procedure, known as association rule mining, is pivotal for uncovering underlying associations within large datasets.

The following section will explore association rule mining in greater detail, with a particular emphasis on the importance of recognizing not only frequent patterns, but also often-overlooked rare patterns. The identification of rare patterns is essential because they can reveal critical insights that might otherwise remain undiscovered, such as anomalies in security data or rare medical conditions. These insights are indispensable for informed decision-making and for enhancing predictive accuracy.

2.5 Association rule mining

Association rule mining, as defined by Agrawal and Srikant [1994], is a foundational unsupervised learning technique designed to uncover hidden patterns within a dataset. This technique employs "if-then" logic, known as association rules, where each rule comprises two components: an antecedent (the "if" part) and a consequent (the "then" part), both of which are sets of items. For instance, in the heart disease dataset utilized in our research, an example of an association rule might be: if 'asymptomatic', 'fasting blood sugar' = 1, and 'male', then heart disease. This suggests that patients with 'asymptomatic' chest pain, fasting blood sugar level of 1, and male sex are more likely to develop heart disease. Association rule mining consists of two steps.

- **Identifying interesting patterns**: Patterns are sets of items that appear together in a dataset and are deemed interesting if they satisfy a predefined threshold constraint.
- **Generating association rules**: Once interesting patterns are identified, rules are generated by dividing these patterns into an antecedent and a consequent. These rules are then evaluated using metrics such as support, confidence, and lift.

A review of the fundamental concepts and definitions of association rule mining is necessary to fully understand the scope and significance of our proposed work. The following formal definition is provided by Agrawal and Srikant [1994]:

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n unique items and let $DB = \{T_1, T_2, \ldots, T_m\}$ represent a set of m transactions that collectively form the dataset. Each transaction $T_i \subseteq I$ includes one or more items from I. An association rule is an implication of form $X \to Y$, where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \emptyset$. Here, X is the antecedent of the rule and Y is the consequent.

The quality of an association rule is commonly assessed using the following two metrics.

- **Support (Supp)**: This metric measures the frequency or proportion of transactions that contain both X and Y. It is defined as:

$$\text{Supp}(X \to Y) = \frac{\sigma(X \cup Y)}{m}$$

  where $\sigma(X \cup Y)$ denotes the number of transactions that include both X and Y.

- **Confidence (Conf)**: This metric measures the conditional probability or strength of the rule, defined as:

$$\text{Conf}(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

  where $\sigma(X)$ is the number of transactions containing X.

A high support value indicates that the rule is generally applicable across the dataset, whereas a high confidence value suggests that the rule is reliable.

Next, we review some key definitions related to association rule mining (ARM) to better understand and interpret the results and implications of our proposed approaches:

Definition 2.9: Strong Association Rule

An association rule $X \to Y$ is considered strong if its support (Supp) and confidence (Conf) measures meet specified minimum thresholds (minSup and minConf, respectively) Aggarwal [2015].

Definition 2.10: Unexpected Association Rule

An association rule (rare) $X \to Y$ is unexpected with respect to a known (frequent) rule $A \to B$ if the following conditions are met [Bui-Thi et al., 2020; Darrab et al., 2024]:

1. **Antecedent Similarity**: The antecedents of the rules (i.e., A and X) exhibit statistical significance within the dataset and demonstrate high similarity, surpassing a predefined similarity threshold.
2. **Consequence Exclusivity**: The consequences of the rules (i.e., B and Y) are mutually exclusive or oppositely related.

To illustrate these concepts, consider the following example from a heart disease dataset. Let us assume that we derive two rules from the data.

The first rule (A) suggests that a combination of factors—such as 'high heart rate', 'male', 'normal fasting blood sugar', and 'no exercise-induced angina'—typically indicates a lack of heart disease (Y= 'no'). This rule is considered strong if it satisfies certain thresholds for support (minSup) and confidence (minConf).

Now, consider another rule (X) that includes an additional factor—'high old-peak'—which is a less common symptom. This rule indicates that, with this rare

combination of factors, there is a higher likelihood of heart disease (Y= 'yes'). Despite being rare, the statistical significance and similarity to frequent rule (A) make this association unexpected.

Moreover, the consequences of these rules—reduced risk of heart disease for rule A and the presence of heart disease for rule X—are oppositely related, meeting the condition for consequence exclusivity. This unexpected finding suggests a need for further investigation by medical professionals.

The conventional support-confidence model for generating frequent patterns has gained widespread popularity due to its simplicity. Raw frequency counts and conditional probabilities are crucial for supporting claims and for determining confidence levels. However, as highlighted in Aggarwal [2015], the frequency of patterns does not always correlate with the most interesting patterns.

We address this limitation by evaluating the generated rules using additional metrics such as lift, leverage, and conviction [Tew et al., 2014]. By incorporating these statistical measures into our analysis, we can identify rules that are not only frequent but also meaningful. The definitions of these metrics are as follows:

- **Lift** Tew et al. [2014]: This metric measures how much more likely the antecedent and consequent of a rule are to occur together than would be expected if they were statistically independent. Mathematically, it is defined as the ratio of the observed support of the rule to the expected support if the antecedent and consequent are independent.

$$\text{Lift}(X \to Y) = \frac{\text{Supp}(X \to Y)}{\text{Supp}(X) \cdot \text{Supp}(Y)}$$

- **Leverage** Tew et al. [2014]: Leverage quantifies the difference between the observed support of a rule and the expected support if the antecedent and consequent were independent. It is computed as:

$$\text{Leverage}(X \to Y) = \text{Supp}(X \to Y) - (\text{Supp}(X) \cdot \text{Supp}(Y))$$

- **Conviction** Tew et al. [2014]: Conviction measures the degree to which the consequent of a rule is dependent on the antecedent. It is interpreted as the ratio of the expected frequency that the antecedent occurs without the consequent if they are independent to the observed frequency of the antecedent occurring without the consequent:

$$\text{Conviction}(X \to Y) = \frac{1 - \text{Supp}(Y)}{1 - \text{Conf}(X \to Y)}$$

These metrics provide deeper insights into the strength and reliability of association rules, beyond what support and confidence alone can offer. By incorporating these advanced measures into our analysis, we enhance our ability to discern truly significant patterns from those that are frequent, but may not necessarily be meaningful in practical applications.

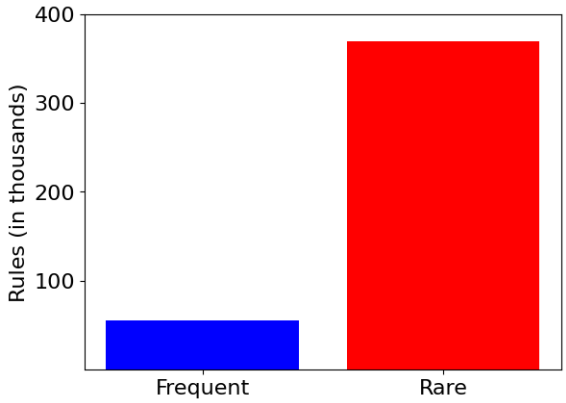


Figure 2.2: A comparison of the number of frequent and rare rules generated from the heart disease dataset.

Contrary to the assertions of numerous studies suggesting that the generation of association rules from patterns is a straightforward process, in actuality, the task is considerably complex. The primary challenge does not reside in rule generation itself, but rather in ensuring the production of meaningful rules. The potential number of rules derived from the patterns can be substantial, rendering the analysis both computationally expensive and impractical. This issue becomes particularly pronounced when attempting to identify interesting rules among the rare ones. For instance, in our heart disease dataset, which comprises only 1,190 transactions, a substantial 448,981 rare and frequent rules are generated, as illustrated in Figure 2.2. Our proposed solution addresses this challenge by focusing solely on the generation of interesting and unexpected rules that can assist clinicians in assessing the likelihood of heart disease based on the presented symptoms.

As we investigate the complexities of association rule mining, a significant challenge emerges: the generation of an immense number of rules. This issue becomes increasingly problematic when addressing rare patterns because each rare pattern with length n can potentially generate 2^n - 2 rules [Zaki, 2000a]. This exponential growth in the number of possible rules exacerbates the difficulty of analyzing vast datasets to obtain valuable insights.

The management of a large number of potential rules presents a fundamental challenge, particularly regarding rare patterns. Although rare patterns are rare, they have substantial impacts despite their infrequency. The proliferation of the derived rules associated with these patterns can quickly become unmanageable, making the identification of valuable insights similar to the search for needles in a haystack.

Addressing this challenge is the primary objective of this thesis. We aim to refine the rule-generation process to ensure that only the most salient and potentially impactful rules are considered. This approach involves sophisticated filtering techniques and criteria that prioritize relevance and significance over quantities. By focusing on generating a manageable set of interesting rules, we enhance the practical utility of our findings and ensure that the rules we consider are those most likely to offer genuine insights.

Mining association rules from large datasets presents several significant challenges, including the substantial volume of potential rules, complexity in identifying meaningful patterns, and heterogeneous distribution of data points across the rule space. To address these challenges, researchers have increasingly adopted advanced analytical strategies that facilitate effective management and extraction of valuable insights from complex datasets. Among these strategies, clustering-based methods have demonstrated particular efficacy in reducing complexity and enhancing rule manageability by aggregating similar data points. Consequently, these methods have become indispensable tools for processing large datasets and deriving actionable insights.

2.6 Clustering-based methods for mining patterns

Density-based clustering approaches are most commonly employed to manage the overwhelming number of rules generated from rare patterns. These methods identify clusters of similar rules based on their densities in feature space. Two prominent algorithms in this context are Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [Ester et al., 1996] and Ordering Points To Identify the Clustering Structure (OPTICS) [Ankerst et al., 1999], which are defined and discussed below.

Definition 2.11: DBSCAN

DBSCAN identifies clusters based on the density of data points in the feature space. It groups points that are closely packed together, while marking outliers that lie alone in low-density regions. The algorithm requires the following two parameters.

- **eps (ϵ)**: The maximum distance between two points for one to be considered as in the neighborhood of the other.
- **minPts**: The minimum number of points required to form a dense region (a cluster).

DBSCAN is particularly useful in datasets where the clusters are irregular or intertwined, as it does not assume any prespecified cluster shape. However, its performance is highly dependent on the distance threshold ϵ, which can limit its effectiveness in datasets with varying densities, potentially causing it to miss significant clusters or to create too many small clusters.

Definition 2.12: OPTICS

OPTICS extends the capabilities of DBSCAN to better handle data with varying density levels. Unlike DBSCAN, OPTICS does not require a global distance threshold (ϵ) for all the points. Instead, it uses a method to vary this parameter, allowing it to identify meaningful clusters in the data that would otherwise be too sparse when using a single threshold:

- **Core Distance**: For a point in the dataset, the core distance is the smallest distance such that the point is a core point, with minPts within this distance.

- **Reachability Distance**: Defined for a point p to point o as the maximum of the core distance of p and the Euclidean distance between p and o.

This approach enables OPTICS to create an ordered list of points representing the structure of the data based on reachability distances, which facilitates the identification of clusters with varying densities. It effectively addresses the shortcomings of DBSCAN by adapting it to changes in data density, making it more flexible and capable of revealing smaller and more significant clusters that DBSCAN might overlook.

DBSCAN and OPTICS are two well-known clustering algorithms that are renowned for their effectiveness in handling data-clustering tasks by focusing on the density of data points. A recent approach utilizing a clustering-based model was proposed to mine unexpected rare rules using the DBSCAN clustering technique [Bui-Thi et al., 2020]. This model, after clustering association rules, uses DBSCAN to determine whether the rules are either noise or unexpected based on a contradiction check.

Although DBSCAN is effective in reducing the number of rules for analysis and concentrating on those that are densely grouped and potentially informative, this approach has several notable limitations.

1. **Performance issues:** DBSCAN's effectiveness is somewhat limited by its reliance on the Apriori algorithm for rule mining, which may not be the most efficient method, especially for large datasets.
2. **Missed opportunities:** The algorithm may overlook interesting, unexpected rules due to its inability to detect nested cluster structures. This limitation means that subtle yet potentially significant patterns can be missed.
3. **Parameter sensitivity:** The outcomes of DBSCAN heavily depend on the choice of hyperparameters, such as ϵ and *minPts*. Proper tuning is crucial because inappropriate values can lead to suboptimal clustering, affecting both the detection of genuine clusters and the identification of noise.

The "noise" identified by OPTICS, as illustrated in Figure 2.3, may not merely represent outliers but could signify rare, critical instances or patterns within the dataset. This feature of OPTICS is particularly valuable in association rule mining, where these noise points might represent rare yet crucial events such as fraud in financial transactions, anomalies in network security, or unusual patient responses in healthcare. By effectively isolating and analyzing these points, OPTICS enables the generation of nuanced, rare, and actionable rules from what might initially appear as noise, providing significant insights and opportunities for intervention in various applications.

Although existing models address some of the challenges associated with generating large numbers of rules and extracting unexpected patterns, they exhibit several critical limitations. First, the performance of these models in terms of both time and memory is significantly impeded by their reliance on traditional methods aimed at recovering the complete set of patterns with low support. Although these approaches

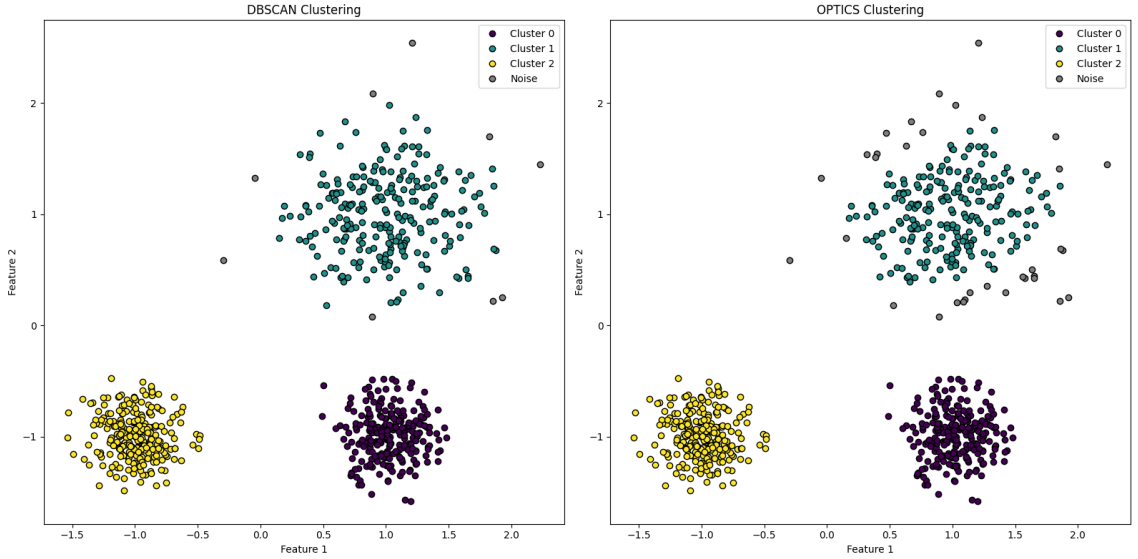


Figure 2.3: Comparative Performance of DBSCAN and OPTICS on the 'make_blobs' Dataset from Scikitlearn.

are intended to capture both frequent and rare patterns, they are inefficient and resource intensive [Singh et al., 2014]. Second, clustering-based models often generate redundant patterns, which diminishes the clarity and utility of the results. Third, parameter tuning in clustering-based models, particularly for parameters such as ϵ and $minPts$, is inherently challenging. Suboptimal parameter settings can lead to failure in identifying the desired patterns. Finally, although some techniques have succeeded in reducing the number of patterns by orders of magnitude, they still fail to produce a comprehensive set of interesting patterns, particularly when the focus is on recovering the most promising, anomalous, or unexpected patterns in real-world applications.

These limitations highlight the need for more efficient and effective models to overcome these challenges. To address this need, we propose an unexpected closed rare-pattern miner (UCRP-Miner) [Darrab et al., 2022b], a model specifically designed to efficiently extract a comprehensive set of unexpected patterns. UCRP-Miner leverages frequent patterns, conceptualized as well-established co-occurring phenomena or beliefs, to generate unexpected patterns. This approach directly addresses the limitations of clustering-based methods by focusing on the extraction of meaningful rules without generating an excessive number of candidate patterns, thereby enhancing the efficiency and relevance of the mining process.

This novel approach surpasses conventional clustering techniques by focusing on extracting significant rules while reducing excessive and redundant patterns. The model effectively isolates the most informative and actionable patterns in large datasets, especially rare occurrences often overlooked by existing methods.

To validate the effectiveness of our model, we conducted comprehensive case studies including those published in a peer-reviewed journal [Darrab et al., 2024]. This study emphasizes the practical implications of mining rare patterns in the context of heart disease, demonstrating the capacity of the model to identify patients who are

currently asymptomatic but may be at risk of developing heart disease in the future. Such insights are crucial for preventive healthcare and offer a proactive approach to managing potential health risks.

In conclusion, the advancements presented in this thesis contribute significantly to the field of data mining, particularly to the analysis of rare patterns. The methodologies developed show substantial potential for uncovering critical insights across various domains, including healthcare, finance, and cybersecurity.

3. Related Work on Rare Pattern Mining

In Chapter 2, we discuss the foundational concepts and definitions that serve as the basis of this dissertation, with a particular focus on association rule- and clustering-based techniques. These methodologies are crucial for identifying significant patterns within data, especially in the context of rare pattern mining. In this chapter, we present a comprehensive review of existing approaches that address the challenges associated with rare pattern mining. Much of the material presented here is derived from our previously published survey [Darrab et al., 2021b], which has been thoroughly revised and updated to incorporate recent advancements, including findings from our own research. Our aim is to provide an exhaustive and current overview, contributing to the broader discourse on advanced data mining techniques and their application in discovering rare patterns.

3.1 Introduction

Data mining is a sophisticated process aimed at uncovering hidden patterns and latent knowledge from large datasets, which once extracted should be interpretable to support human decision making. The core tasks of data mining can be broadly categorized into three primary areas: clustering, classification, and association rule mining (ARM) [Han et al., 2022]. Although clustering and classification have been extensively studied, ARM remains an area of active research due to its evolving nature and diverse applications. ARM identifies significant relationships between groups of items that frequently co-occur within a dataset. For instance, in a retail dataset, the association {beer $\rightarrow$ chips} suggests that customers who purchase beer are more likely to purchase chips. Such associations provide valuable insights for business decision making, enabling managers to implement strategies such as targeted promotions to boost sales.

One of the primary advantages of ARM is its inherent explainability and ease of understanding. Unlike more complex models [Hassija et al., 2024], such as deep

learning, which often function as ”black boxes,” ARM provides clear, interpretable rules that directly describe the relationships in the data, making it easier for domain experts and decision-makers to grasp the rationale behind the insights. This feature of ARM addresses the critical challenge of explainability and interpretability in advanced models, particularly in deep learning, where the inner workings of models can be difficult to understand. Consequently, ARM is particularly valuable in fields where interpretability is essential for human decision making. Consequently, ARM has gained significant attention across various fields, including bioinformatics, network traffic analysis, medical diagnosis, and market basket analysis [Fournier-Viger et al., 2017].

Association rule mining (ARM) consists of two critical steps: (1) frequent pattern mining (FPM), which identifies frequent patterns, and (2) extracting association rules from these patterns. Most studies [Agrawal and Srikant, 1994; Han et al., 2000; Pei et al., 2001; Zaki, 2000b] have focused on FPM because of its computational intensity, as generating association rules is straightforward. FPM aims to discover patterns that frequently co-occur within a dataset, thereby providing valuable insights for decision-making. Although frequent patterns offer useful information, they often represent predictable or well-established phenomena. Consequently, recent research has increasingly shifted toward the discovery of rare patterns, which may reveal less obvious but potentially valuable insights.

Rare pattern mining (RPM) was introduced to address this challenge [Borah and Nath, 2020; Gui et al., 2024; Kataria et al., 2019; Lu et al., 2020; Szathmary et al., 2007; Tsang et al., 2011; Vanamala et al., 2021], with the goal of discovering patterns that occur infrequently but hold significant value, particularly in fields such as medical research, fraud detection, and security. Conventional approaches face the ”rare item problem,” where reducing the minimum support threshold to capture rare patterns results in an excessive number of frequent patterns, leading to substantial computational inefficiencies. Conversely, increasing the minimum support threshold risks missing valuable rare patterns. This ongoing challenge remains unresolved [Bhatt and Patel, 2015; Darrab et al., 2021b; Koh and Ravana, 2016].

In recent years, rare pattern mining (RPM) has received increasing attention due to its critical role in real-world applications such as medical research, DNA analysis, and homeland security. For instance, identifying rare patterns of passenger behavior at airports can facilitate the detection of suspicious activities and help mitigate potential security threats [Troiano and Scibelli, 2014]. Despite their increasing importance, conventional RPM methodologies continue to face challenges related to scalability and performance.

Several comprehensive surveys on Rare Pattern Mining (RPM), such as those by [Darrab et al., 2021b; Koh and Ravana, 2016], offer in-depth reviews of the field. Despite significant advancements in RPM, key challenges persist, including efficient extraction of rare patterns, addressing redundancy issues, minimization of noise in identifying meaningful patterns, and emphasis on the novelty of mined patterns.

The primary contributions of this chapter are as follows:

- A thorough and critical review of current methodologies for rare pattern mining, accompanied by a detailed comparison and evaluation of existing techniques in the field.
- The introduction of recent advancements and innovative extensions to established rare pattern mining methods, with an emphasis on their significance and potential to impact both research and practice.
- A comprehensive discussion on the practical applications of rare patterns across multiple domains, illustrating their utility in addressing complex real-world challenges.
- The identification of key challenges associated with rare pattern mining, along with informed suggestions for future research directions aimed at overcoming these challenges and advancing the state of the art in the field.

The remainder of this chapter is organized as follows. Section 3.1 discusses the limitations of frequent pattern mining when applied to rare patterns, identifies their inadequacies, and establishes a foundation for exploring more specialized approaches. Section 3.2 analyzes the Breadth-First Search (BFS) and Depth-First Search (DFS) methodologies in the context of rare pattern mining, emphasizing how these traversal algorithms aid in efficiently identifying rare patterns. Section 3.3 examines techniques for mining frequent patterns that also encompass rare patterns, and addresses the intersection between frequent and rare pattern mining strategies. Section 3.4 provides an in-depth examination of rare pattern mining methodologies, focusing on techniques specifically designed for the identification and management of rare patterns. Section 3.5 introduces advanced techniques in rare pattern mining, highlighting recent innovations that address the challenges related to scalability and computational efficiency. Section 3.6 explores the practical applications of rare pattern mining and demonstrates its significance through case studies across various domains. Section 3.7 identifies the current research opportunities and challenges within the field and offers insights into unresolved issues and potential directions for future research. Section 3.8 focuses on the specific challenges addressed in this thesis, detailing how the proposed methods contribute to advancements in this field. Finally, Section 3.9 presents a summary of key points.

3.2 Shortcomings of frequent pattern mining for rare patterns

Frequent pattern mining [Agrawal and Srikant, 1994; Han et al., 2000] is a fundamental technique in data mining that is widely used to identify frequently occurring patterns within datasets, based on a user-defined minimum support threshold. This approach has proven effective in various domains, including market basket analysis, web usage mining, and bioinformatics, where frequent patterns provide valuable insights for decision-making processes, such as product recommendations, customer behavior analysis, and inventory management.

Despite their success in uncovering frequent patterns, traditional frequent pattern mining algorithms struggle to detect rare patterns that occur infrequently but may

carry significant importance. This inherent bias toward frequent patterns presents a major limitation, as rare patterns often reveal critical insights into applications such as anomaly detection, fraud prevention, and medical research. For instance, in market basket analysis, a rare combination of items can indicate niche customer preferences that can be exploited in targeted marketing strategies. In medical research, rare patterns may represent subtle correlations, leading to novel discoveries or early disease indicators. However, adjusting the support threshold to capture these rare patterns often results in an exponential increase in the number of frequent patterns, which not only increases computational complexity, but also creates challenges in result interpretation. However, increasing the support threshold to control complexity may inadvertently exclude valuable rare patterns from the analysis, further compounding this issue.

Rare patterns, although infrequent, it signifies outliers, anomalies, or emerging trends, which are elements that are often more significant than the more commonly occurring patterns that dominate traditional analyses. Detecting these rare patterns requires specialized techniques that go beyond conventional frequent pattern mining algorithms because these methods must balance the trade-off between computational efficiency and the need to uncover rare but highly relevant insights.

In the following subsections, we explore advanced methods developed to address the limitations of frequent pattern mining, with a particular focus on approaches that are capable of identifying rare patterns. These techniques aim to reduce bias towards frequent patterns, enabling the discovery of rare patterns that are of substantial importance in various real-world applications.

3.3 Structural insights for effective rare pattern mining

Rare pattern mining is a crucial task in association rule learning, where the chosen traversal strategy can significantly influence the efficiency of mining algorithms. A breadth-first search (BFS) is a commonly used approach that explores all nodes at the current depth level before progressing to the next level. This strategy is particularly effective for datasets where the desired patterns are located closer to the root of the search tree. However, BFS often requires multiple scans of the dataset and candidate pattern generation, which can be both time-consuming and memory intensive, particularly for larger datasets.

To address the limitations of the Breadth-First Search (BFS), Depth-First Search (DFS) was implemented. DFS explores each branch of the search tree to its maximum depth before backtracking, rendering it more memory-efficient by storing only the current path and a limited portion of the tree. In contrast to BFS, DFS typically requires a maximum of two scans of a dataset, which enhances its efficiency in specific tasks. However, the DFS presents several challenges. In algorithms such as FP-Growth, the construction of conditional trees for rare patterns can be both computationally intensive and memory demanding, particularly when processing sparse datasets in which infrequent patterns are dispersed throughout the search space.

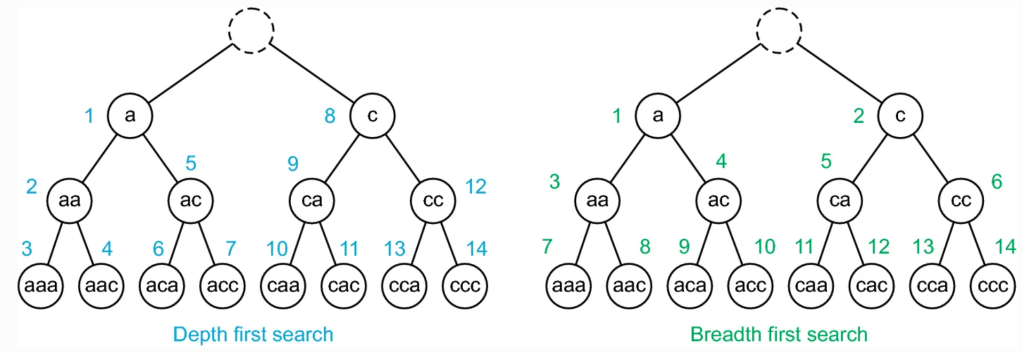


Figure 3.1: A schematic representation of BFS and DFS in pattern mining, starting from the root (the empty set) and progressing through successive levels of complexity, from 1-patterns to 3-patterns [Titarenko et al., 2019].

Figure 3.1 illustrates the differences between the two approaches. On the left, DFS is depicted, where the nodes are explored deeply before backtracking. The BFS is shown on the right, where nodes at the same depth level are explored before moving deeper into the tree.

3.4 Mining frequent patterns including rare ones

Traditional frequent pattern mining (FPM) algorithms often face challenges in capturing rare patterns because they typically rely on a single or fixed minimum support threshold. This approach fails to account for the fact that items in real-world datasets often appear at vastly different frequencies. Consequently, rare but potentially significant patterns may be overlooked. To address this limitation, extensive research has been conducted on mining frequent patterns, including rare patterns, by introducing multiple minimum support thresholds [Chen et al., 2014; Gupta and Chandra, 2020; Telikani et al., 2020; Xu and Dong, 2013]. These methods assign varying minimum support thresholds to individual items based on their frequency, thereby enabling the discovery of correlations between frequent and rare items in a more nuanced manner. Thus, patterns are considered interesting if their support satisfies or exceeds the minimum support threshold for the rarest item in the pattern.

Two common algorithmic strategies are employed for mining frequent patterns that include rare patterns: breadth-first and depth-first searches. In the following sections, we first examine breadth-first search algorithms and their applications in mining both frequent and rare patterns, followed by a discussion of depth-first search algorithms.

3.4.1 Breadth-first search methods

A key breadth-first search algorithm for mining frequent patterns, including rare ones under multiple minimum support thresholds, is the MSapriori algorithm introduced by [Liu et al., 1999]. This algorithm extends the original Apriori algorithm [Agrawal and Srikant, 1994] by assigning a unique minimum item support (MIS) value to each item, allowing it to capture patterns involving both frequent and rare items. MSapriori begins by generating candidate 1-patterns and retains only those with a

support greater than or equal to their individual MIS. It then generates candidate 2-patterns from the frequent 1-patterns, considering those patterns whose support meets the lowest MIS among their constituent items. This process continues iteratively until no additional candidate patterns are generated. The resulting patterns are considered interesting if their support exceeds the lowest MIS value of the items in the pattern.

Several enhancements to MSapriori have been proposed to improve its performance [Kiran and Reddy, 2009; Lee et al., 2005; Xu and Dong, 2013]. For example, [Xu and Dong, 2013] introduced the MSB_apriori method, which, like MSapriori, assigns a distinct MIS value to each item. A pattern is considered frequent if it satisfies the lowest MIS among all its items. Unlike MSapriori, which introduces specific modifications to the Apriori steps, MSB_apriori follows the basic Apriori procedure in two phases: (1) mining all potential patterns using a unified minimum support and (2) filtering these patterns based on multiple minimum supports to identify desired patterns. A pattern X is classified as frequent if its support $\sup(X)$ is greater than or equal to the MIS of its least frequent item.

In [Kiran and Reddy, 2009], the IMSApriori algorithm was proposed, which enhances MSapriori by incorporating a Support Difference (SD) to adjust the minimum support for items. This SD allows for small deviations in item frequency while keeping rare patterns relevant. The MIS for each item is defined as

$$\mathrm{MIS}(i) = \begin{cases} S(i) - \mathrm{SD} & \text{if } S(i) - \mathrm{SD} > \mathrm{LS} \\ \mathrm{LS} & \text{otherwise} \end{cases} \tag{3.1}$$

where $S(i)$ is the support of item i and LS is the lowest minimum support. The SD is computed as:

$$\mathrm{SD} = \lambda(1 - \alpha) \tag{3.2}$$

where λ is the maximum support among all items, and α is a user-defined parameter between 0 and 1. This approach reduces the number of frequent patterns involving rare items while preserving their relevance.

Other algorithms have extended Apriori to mine both frequent and rare patterns using multiple minimum support thresholds [Bansal et al., 2013; Lee et al., 2005]. Lee et al. [2005] proposed an algorithm that identifies all frequent 1-patterns, including rare ones, by comparing each item's support to its predefined MIS. For k-patterns to be considered interesting, their support is compared to the maximum MIS values for the items in the pattern. Similarly, Bansal et al. [2013] introduced a method to adjust the MIS of each item i using the following equation:

$$\mathrm{MIS}(i) = \begin{cases} \beta S(i) & \text{if } \beta S(i) > \mathrm{LS} \\ S(i) & \text{otherwise} \end{cases} \tag{3.3}$$

where β is a user-specified parameter between zero and one, $S(i)$ is the support of item i, and LS is the least minimum support threshold. This equation ensures that

frequent patterns involving rare items are discovered while pruning frequent patterns composed primarily of common items.

While these methods address the rare item problem by identifying both frequent and rare patterns, they depend on a breadth-first search strategy that employs a candidate generation-and-test approach. This method is computationally expensive in terms of time and memory, particularly for long patterns.

3.4.2 Depth-first search methods for mining patterns

Depth-first search algorithms have been developed to overcome the limitations of apriori-like algorithms, particularly their inefficiency in handling large datasets due to repeated scans and pattern generation. A prominent depth-first approach is the multiple-item support tree (MIS-Tree) proposed by Hu and Chen [2006]. The MIS-tree structure addresses the inefficiencies of generate-and-test methods, such as those used in MSapriori, by storing critical information about frequent patterns, including rare items, in a compact tree structure. This significantly reduces the number of dataset scans required, offering improved performance over traditional breadth-first methods.

Building on the MIS tree concept, the CFP-growth algorithm was introduced in Hu and Chen [2006]. This algorithm enhances the FP-growth approach by incorporating an initial dataset scan to construct an MIS-Tree, followed by a pruning and merging phase to eliminate non-contributory items. This pruning process results in a more compact MIS tree, facilitating the efficient mining of both frequent and rare patterns. However, CFP-growth has certain limitations, notably the initial inclusion of non-contributory items that continue to be examined until their respective conditional pattern bases are fully processed, often without yielding valuable patterns.

To address these shortcomings, an improved version called CFP-growth++ has been proposed [Kiran and Reddy, 2011]. This enhanced algorithm utilizes the lowest minimum support threshold (LMS) for pruning, as opposed to the least item support (MIN) used in CFP-growth. In addition, it introduces a mechanism for removing leaf nodes that cannot generate valuable patterns and applies a conditional closure property, enabling the discovery of all relevant patterns without requiring full exhaustion of the conditional pattern base. These improvements reduce both the execution time and search space, thereby significantly enhancing the efficiency of the algorithm.

Further advancements include the MS-FP-growth algorithm [Taktak and Slimani, 2014], which employs variable minimum support thresholds at different pattern levels and dynamically adjusts the support values. This flexibility allows the algorithm to increase or decrease the threshold depending on the level of the patterns (K-patterns), thereby enabling the identification of rare patterns at specific levels with tailored support values.

Another significant development involves the use of statistical methods to determine item support values as demonstrated in Chen et al. [2014]. By applying the central limit theorem, they calculated the mean item support (MIS) values to refine the pattern generation. The formula for calculating the mean support is as follows:

$$\mu(i_j) = \frac{1}{n}\sum_{j=1}^{n} \sup(i_j) \tag{3.4}$$

$$\mathrm{MIS}(i_j) = \mu(i_j) - \sqrt{\frac{1}{n}\sum_{j=1}^{n} \left(\sup(i_j) - \mu(i_j)\right)^2} \tag{3.5}$$

where $\mu(i_j)$ represents the average frequency of items at the same level, allowing for the automated adjustment of support thresholds to optimize the mining process for both frequent and rare patterns.

Additionally, [Darrab and Ergenç, 2016; Wang and Chang, 2019] extended the FP-growth algorithm to support mining under multiple minimum supports without requiring tree reconstruction, thus preserving efficiency. Similarly, the MISeclat [Darrab and Ergenc, 2017] algorithm employs different data structures and support mechanisms to improve rare pattern mining. These advancements have demonstrated the ongoing evolution of mining algorithms to better handle the complexities of diverse and large datasets.

Although these methods successfully uncover both frequent and rare patterns by assigning distinct minimum support thresholds to each item, they present certain challenges. Although this strategy facilitates the discovery of rare patterns, it also substantially increases the number of generated patterns, thereby intensifying computational demands and prolonging analysis time. Furthermore, determining an appropriate minimum support for each individual item introduces additional complexity. Given these limitations, there is an increasing need for approaches tailored specifically to focus on mining rare patterns alone, thereby reducing the computational overhead and simplifying the analysis. The following section examines methods developed to meet this need by concentrating exclusively on rare pattern discovery.

3.5 Rare pattern mining

As discussed in the previous section, traditional frequent pattern mining algorithms that attempt to mine both frequent and rare patterns face significant challenges, particularly when determining an appropriate minimum support threshold. One major limitation of these methods is the implicit assumption that all items in the dataset exhibit similar frequency behavior, which is rarely the case in practice. Item frequencies often vary greatly depending on their role or value within specific contexts. For example, in retail, items such as bread are purchased more frequently than higher-margin items such as saucepans, which are purchased less frequently but are essential to profitability. This disparity highlights the ”rare item problem,” where valuable patterns involving infrequent items are often overlooked because they fail to meet a predefined frequency threshold [Darrab et al., 2021b; Liu et al., 1999; Lu et al., 2020; Selvarani and Jeyakarthic, 2021].

Mining both frequent and rare patterns simultaneously tends to produce an overwhelming number of patterns, many of which represent well-known phenomena and

provide limited insights. Frequent patterns typically capture common behaviors that are already understood, offering little unexpectedness. In contrast, rare patterns, despite their low frequency, can be of greater significance, particularly in fields where identifying anomalies or uncovering uncommon associations is crucial. For example, in airport security, rare patterns may signal unusual behaviors that could indicate potential threats. In telecommunications, rare events may predict equipment failure, and in medical diagnostics, identifying rare genetic markers can lead to breakthroughs in disease understanding. Similarly, rare transaction patterns in credit card data can help detect fraudulent activities that traditional frequent pattern mining techniques may miss [Adda et al., 2007; Bhatt and Patel, 2015; Borah and Nath, 2020; Koh and Ravana, 2016; Szathmary et al., 2012].

Recognizing the importance of rare patterns, recent research has shifted towards the development of specialized techniques for rare pattern mining [Borah and Nath, 2019; Darrab et al., 2021b]. These methods explicitly focus on identifying rare patterns rather than mining both frequent and rare patterns together, thereby reducing the explosion of redundant patterns and concentrating on uncovering meaningful, often unexpected insights.

Rare pattern mining techniques [Akdas et al., 2024] are specifically designed to capture patterns that occur infrequently but hold significant values. These methods overcome the limitations of traditional algorithms by employing advanced search strategies that are typically categorized into breadth-first and depth-first approaches. These strategies enhance the efficiency of search space exploration by filtering out the noise generated by frequent patterns, while focusing on the discovery of rare yet valuable insights. In the following sections, we examine these methods in detail and explore their practical applications across various domains.

3.5.1 Breadth-first search methods

The breadth-first search (BFS) in rare pattern mining employs a level-wise exploration strategy, as illustrated on the right side of Figure 3.1. The method begins with simple patterns (e.g., 1-patterns) and progresses to more complex patterns (e.g., 2-patterns, 3-pattetc.). In each iteration, candidate patterns are generated and evaluated against a predefined minimum support threshold. The process continues until no further patterns are generated, ensuring a comprehensive exploration of the pattern space and identifying both frequent and rare patterns, without prematurely excluding infrequent patterns.

A notable contribution to BFS in rare pattern mining is the work of [Sadhasivam and Angamuthu, 2011], that introduced two automated support thresholds: AvgSup and MedianSup. These thresholds adjust for item frequency variations and enhance rare pattern detection. AvgSup represents the average support of all unique items in the dataset, while MedianSup is calculated from the median between the highest and lowest supports. Based on these thresholds, patterns are classified into three categories: Most_Interesting_Group (MiG) for patterns with support above AvgSup, Somewhat_Interesting_Group (SiG) for patterns between MedianSup and AvgSup, and Rare_Interesting_Group (RiG) for patterns below both thresholds.

The RiG category is crucial in rare pattern mining as it targets infrequent yet valuable patterns, often overlooked by traditional frequent pattern mining approaches.

This stratified approach ensures a more thorough exploration of the pattern space, capturing both frequent and rare patterns.

In another study, [Szathmary et al., 2007] introduced the MRG-Exp and ARIMA methods to improve rare pattern discovery. These methods distinguish between Minimal Generators (MGs), Minimal Rare Generators (MRGs), and Minimal Zero Generators (MZGs). MRG-Exp exploits the properties of MRGs to generate new patterns from the bottom-up, while ARIMA uses MRGs to limit the exploration space, reducing computational overhead while maintaining precision.

Similarly, [Koh and Rountree, 2005] advanced the traditional Apriori algorithm with Apriori-Inverse, which employs dual-thresholds (maxsup and minsup) to capture rare patterns effectively. Through this modification, patterns that fall between these thresholds can be identified, thereby overcoming the limitations of the traditional frequent pattern mining.

[Troiano et al., 2009] proposed AfRIM, a top-down algorithm that starts with the largest patterns and works downward, particularly effective in datasets where significant patterns encompass smaller rare ones. In addition, [Tummala et al., 2018] refined clustering techniques to group similar patterns, thereby improving the efficiency of mining both frequent and rare patterns in large datasets. As the data volume increases, traditional methods face scalability challenges. To mitigate this, [Padillo et al., 2017] introduced Apriori-Inverse-MR, a MapReduce-based algorithm designed for big data. By leveraging distributed processing, this approach efficiently discovers rare patterns and mines association rules, optimizing performance in large datasets.

Despite the effectiveness of BFS in exploring the search space, its computational cost in terms of memory usage and repeated database scans presents significant challenges. The generation of numerous candidate patterns, many of which do not lead to meaningful insights, further contributes to resource consumption. These limitations underscore the need for more efficient strategies to minimize unnecessary candidate generation while prioritizing rare but valuable patterns.

In the next subsection, we explore DFS as a more resource-efficient solution to BFS. By focusing on specific branches of the pattern space, DFS reduces memory and computational demands, accelerating the identification of rare patterns.

3.5.2 Depth-first search methods

To address the limitations of breadth-first search approaches, [Han et al., 2000] introduced a depth-first approach that eliminates the need for multiple scans and avoids the generate-and-test mechanism, providing a more efficient method for mining rare patterns.

Building on this idea, [Tsang et al., 2011] introduced RP-Tree, the first tree-based algorithm specifically designed for rare pattern mining. Similar to the FP-growth algorithm [Han et al., 2000], RP-Tree employs a tree structure to store key information for the mining process. The algorithm performs two scans: the first to count the support of individual items and the second to construct the RP-Tree. This tree includes only transactions that contain at least one rare item. By focusing exclusively

on rare items, RP-Tree efficiently reduces computational overhead, making rare pattern discovery more effective.

To further optimize rare pattern mining, [Bhatt and Patel, 2015] introduced the Maximum Constraint RP-Tree (MCRP-Tree). This enhanced version assigns varying Minimum Item Support (MIS) values to items based on their frequency, ensuring that a pattern is only considered interesting if it meets the highest MIS value of its constituent items. Similar to RP-Tree, MCRP-Tree excludes transactions with only frequent items, refining the mining process to focus on rare but significant patterns. This method not only improves computational efficiency but also minimizes the generation of redundant or unimportant patterns, making it highly effective for discovering actionable insights.

In response to the challenges of large-scale datasets, recent innovations have incorporated distributed computing platforms such as Apache Spark. For instance, [Liu et al., 2016] developed a spark-based extension of RP-Tree for big data environments. This approach partitions the dataset into frequent and rare vertical segments to efficiently count candidate pattern support. As with RP-Tree, it discards patterns that lack rare items, optimizing performance for large datasets. Similarly, [Liu and Pan, 2018] introduced SRAM, an algorithm designed to mine rare patterns in wireless networks. By analyzing Network Performance Counters (NPCs), SRAM identifies rare patterns that contribute to Key Quality Indicator (KQI) degradation, helping to pinpoint the root causes of abnormal events in wireless networks.

A number of depth-first search strategies, especially tree-based structures like RP-Trees and MCRP-Trees, have been shown to reduce the computational burden associated with rare pattern mining. When combined with the scalability of modern big data platforms, these methods are well equipped to handle the growing complexity of datasets across a wide range of domains, enabling efficient discovery of rare yet valuable patterns.

However, although RP-Tree and its variants have delivered promising results, they encounter challenges in mining rare patterns in sparse datasets, where such patterns are often predominant. RP-Tree's inefficiency in these cases highlights the need for a new structure optimized for rare pattern mining in sparse data environments.

In this dissertation, we propose a novel structure specifically designed to enhance the mining of rare patterns in sparse datasets. Our approach aims to overcome the limitations of current methods, such as RP-Tree, providing a more efficient solution for discovering rare patterns in complex, sparsely populated data environments.

3.6 Advanced techniques in rare pattern mining

The rare pattern mining methods mentioned in the previous section primarily rely on frequency as the main metric for identifying interesting patterns. However, in practical scenarios, factors such as utility, data fuzziness, continuous data streams, and item weights play significant roles in extracting meaningful patterns. Consequently, several advanced techniques have been developed that incorporate dimensions beyond frequency to better capture rare but valuable insights. This section explores four advanced approaches to rare pattern mining: High Utility Rare Pattern Mining, Fuzzy

Rare Pattern Mining, Rare Pattern Mining in Data Streams, and Rare Weighted Pattern Mining.

3.6.1 High utility rare pattern mining

Identifying rare patterns that also provide high utility is crucial in contexts where utility, rather than frequency alone, drives decision making. To address this gap, researchers [Arunkumar et al., 2020; Chan et al., 2003; Goyal et al., 2015] introduced the concept of utility as a metric to assess the value of a pattern.

In many scenarios, identifying rare patterns with high utility yields valuable insight. For example, a sales manager might prioritize infrequent purchase patterns that generate substantial profits over frequently purchased, low-margin products. To support such use cases, [Pillai et al., 2013] proposed the High Utility Rare Itemset (HURI) algorithm, which consists of two main steps:

- Identifying rare patterns that do not satisfy the maximum support threshold.
- Selecting high utility rare patterns whose utility values meet or exceed a predefined minimum utility threshold.

Additionally, [Ryang et al., 2014] introduced the MHU-Growth (Multiple Item Supports with High Utility Growth) method, which incorporates both utility factors and multiple minimum support levels. This technique utilizes the MHU-tree, constructed in a single scan of the dataset, to store transactional and utility-related information. It employs four pruning conditions to reduce the search space and candidate patterns, thereby improving the efficiency of high-utility rare pattern mining.

Although research on utility-based rare pattern mining remains limited [Arunkumar et al., 2020], these approaches highlight the importance of integrating utility metrics into rare pattern discovery. This is particularly relevant in industries such as retail, finance, and healthcare, where utility-driven insights can significantly impact decision making and profitability.

3.6.2 Fuzzy rare pattern mining

Fuzzy rare pattern mining [Cui et al., 2022] addresses the limitations of traditional rare pattern mining by incorporating fuzzy logic into the mining process, enabling the analysis of data that is inherently imprecise or vague. This technique was first introduced by Kuok et al. [1998], who developed a method for discovering quantitative frequent patterns by assigning fuzzy membership values to items within transactions. Fuzzy pattern mining can handle uncertainties and ambiguities in real-world data by allowing varying degrees of membership rather than binary inclusion.

Fuzzy rare pattern mining is particularly effective when the data are not strictly binary or when it is difficult to categorize the data precisely. To identify uncommon learning behaviors, it may be necessary to process data that are imprecise or difficult to quantify in the field of education. Although common student behaviors are relatively straightforward to capture, rare or atypical behaviors may require more

nuanced detection techniques. Studies such as [Chen and Chang, 2016; Cui et al., 2022; Weng, 2011] have successfully integrated fuzzy set theory with data mining techniques to uncover rare patterns in educational datasets. These studies have provided valuable insights into outlier behaviors, such as uncommon learning patterns, which are likely to go undetected using conventional mining methods.

3.6.3 Rare pattern mining in data streams

The continuous generation of vast volumes of data in fields such as telecommunications, sensor networks, and online platforms presents unique challenges for traditional rare pattern mining algorithms. These algorithms are typically designed for static datasets and often require multiple scans of the data, making them unsuitable for mining patterns from dynamic, continuous data streams.

To address these challenges, rare pattern mining techniques specifically tailored to data streams have been developed. These methods aim to process transactions in real-time, capturing approximate sets of interesting patterns without the need for re-scanning the data. For example, [Huang et al., 2012] introduced the Streaming Rare Pattern Tree (SRP-tree) algorithm, which processes data in a single pass using a sliding window technique. The SRP-tree efficiently maintains essential information for mining while a connection table tracks items within the window. Once an item's support falls below the minimum threshold, the SRP-tree generates all subsets of infrequent items, making it a highly efficient solution for real-time rare pattern mining in continuous data streams.

Rare pattern mining in data streams is crucial in applications where timely detection of unusual patterns is essential, such as network monitoring, fraud detection, and real-time sensor data analysis.

3.6.4 Rare weighted pattern mining

Rare weighted pattern mining assigns weights to items based on their local significance within transactions. This approach enables the discovery of patterns that, although infrequent, may carry substantial importance due to the assigned weights, reflecting the varying significance of items in different contexts.

An early method for weighted pattern mining was proposed by Wang et al. [2000], introducing the concept of Weighted Association Rules (WAR). This method consists of two main steps:

- Generate frequent patterns using traditional approaches.
- Apply the weight parameter during rule generation to create weighted association rules, ensuring that the significance of item quantities is considered in the discovered patterns.

Although the discovery of rare weighted patterns is a less explored area, it has gained attention due to its potential to provide deeper insights. For instance, [Cagliero and Garza, 2013] introduced an algorithm to mine infrequent weighted patterns using two key measures:

- **IWI-support-min**: Measures a pattern based on the least interesting item's weight within the transaction.
- **IWI-support-max**: Measures a pattern based on the most interesting item's weight.

These measures help identify a meaningful subset of rare weighted patterns, which are particularly valuable in domains such as retail, healthcare, and targeted marketing, where identifying high-weight rare patterns can lead to more informed decision-making and more precise targeting of resources.

3.7 Applications of rare pattern mining

Rare pattern mining has gained considerable attention for its capacity to uncover infrequent yet highly valuable insights across various fields. Many real-world applications require the identification of rare events or associations, as they often carry critical, actionable information. The ability to detect such rare patterns has proven instrumental in solving complex problems, improving decision-making, and identifying anomalies that might otherwise go unnoticed. In this section, we explore key domains in which rare pattern mining has been effectively applied, highlighting its impact on enhancing operational efficiency, risk management, and predictive accuracy across diverse industries.

3.7.1 Fraud detection

Anomaly detection is one of the most prominent applications of rare pattern mining, particularly in fields such as cybersecurity, fraud detection, and network management. Anomalies, by definition, represent rare events that can have significant consequences if left undetected. Rare pattern mining is a powerful tool for identifying these anomalies, enabling early detection and mitigation.

In fraud detection, rare pattern mining has been extensively applied to identify fraudulent activities in financial transactions. For example, algorithms developed by Awoyemi et al. [2017]; Seeja and Zareapoor [2014] have successfully utilized rare pattern mining to uncover infrequent but suspicious credit card transactions, enhancing the accuracy and efficiency of fraud detection systems. Rare transaction patterns that deviate from a customer's usual spending behavior can be flagged for further investigation, allowing financial institutions to reduce the risk of financial losses.

In network management, rare pattern mining has proven invaluable for detecting performance anomalies in wireless networks. The Spark-based Rare Association Rule Mining (SRAM) approach proposed by Liu et al. [2016] linked network performance counters (NPCs) to key quality indicators (KQIs), allowing network service providers to rapidly identify the root causes of performance degradation. This method significantly improves network reliability and service quality by addressing rare but critical system faults. Similarly, [He et al., 2004] introduced the Frequent Pattern Outlier Factor (FPOF), which employs rare pattern detection to identify outliers in transactional data, further enhancing anomaly detection across diverse domains.

In insurance, rare pattern mining used to detect fraudulent claims by identifying unusual patterns that deviate significantly from typical claims. These applications demonstrate the capability of rare pattern mining to not only mitigate financial losses but also strengthen security measures across multiple sectors.

3.7.2 Medical diagnostics

In the healthcare sector, the ability to detect rare patterns within large datasets, such as those generated by electrocardiograms (ECG), positron emission tomography (PET), and magnetic resonance imaging (MRI), holds significant potential for early disease diagnosis and prevention. Rare pattern mining facilitates the identification of subtle and infrequent correlations that may indicate early onset of diseases, thereby enabling timely intervention. For instance, early detection of rare arrhythmic patterns in ECG data can prevent severe cardiac events. Recent studies [Borah and Nath, 2018; Piri et al., 2018] have demonstrated the efficacy of rare pattern mining in uncovering significant yet infrequent correlations between patient attributes and diseases such as breast cancer and hepatitis, thereby enhancing early detection and intervention strategies.

Furthermore, [Darrab et al., 2024] employed machine learning techniques enhanced by rare pattern mining to predict the onset of heart disease prior to the appearance of clinical symptoms. This early prediction strategy demonstrates the potential of rare pattern mining to support preventative healthcare through timely risk assessment and targeted intervention.

3.7.3 Network security

Network security is another domain in which rare pattern mining is invaluable. Malicious attacks, such as denial-of-service (DoS) attacks, are relatively rare events compared with regular network traffic. Detecting these rare attacks is critical for maintaining the security and integrity of computer networks. Rare pattern mining techniques can be employed to analyze network traffic data and identify unusual patterns that may indicate ongoing attacks or security breaches [Huang et al., 2012].

In telecommunications and sensor networks, rare pattern mining is similarly applied to detect rare events such as equipment failures, anomalies in data transmission, or unusual sensor readings. Although these events are infrequent, they often signal critical issues that require immediate attention, such as infrastructure failure or abnormal environmental conditions.

For example, in wireless networks, rare pattern mining can be used to monitor network performance indicators and detect early signs of equipment malfunction or service quality degradation [Liu and Pan, 2018]. Similarly, in sensor networks deployed for industrial or environmental monitoring, rare pattern mining facilitates the detection of irregular sensor readings, which may indicate hazardous conditions or malfunctioning equipment.

Furthermore, rare sequences of network packets or atypical communication patterns can serve as early indicators of potential threats. By focusing on these rare patterns, network administrators can take proactive measures to prevent or mitigate

cyberattack. In addition, rare pattern mining enables the identification of emerging threats that may not conform to known attack patterns, thereby enhancing the overall resilience and effectiveness of cybersecurity systems.

3.7.4 Education systems

In educational systems, rare pattern mining has proven instrumental in uncovering hidden but significant student behaviors within large-scale datasets, offering educators valuable insight into student engagement and performance. For example, [Romero et al., 2010] developed a method for mining rare patterns from student interactions on online learning platforms such as Moodle, identifying atypical behaviors that may signify academic disengagement or learning difficulties. These insights enable timely interventions aimed at improving student outcomes.

Moreover, [Weng, 2011] introduced a fuzzy apriori-based rare pattern mining (FARIM) approach to detect rare learning behaviors in quantitative educational data. This method provides a robust framework for identifying unique challenges encountered by specific student groups, thereby supporting more personalized educational strategies.

Through the application of rare pattern mining, educational systems can more effectively identify and address uncommon but critical student behaviors, ultimately enhancing personalized learning and educational outcomes.

3.8 Research opportunities and challenges

Many algorithms have facilitated the search for interesting rare patterns. However, significant challenges remain in improving the efficiency, accuracy, and applicability of rare pattern mining for real-world applications. In this section, we discuss some of the major challenges and areas that require further exploration in rare pattern mining. These challenges provide promising opportunities for future research [Borah and Nath, 2020; Darrab et al., 2021b; Gui et al., 2024].

3.8.1 Efficient discovery of rare patterns

It has been previously noted that rare pattern mining is computationally costly, particularly in terms of memory consumption and execution time. Despite the effectiveness of existing algorithms, many require multiple dataset scans and complex computations, particularly when dealing with large datasets. To improve efficiency, it is imperative to develop more sophisticated algorithms to minimize resource consumption without compromising accuracy. This may require the development of new data structures, improvement of search space pruning techniques, and introduction of more effective constraints.

3.8.2 Concise representation of rare patterns to avoid redundancy

The generation of redundant patterns is a major challenge in rare pattern mining. An overwhelming number of rare patterns may obscure important insights because many are not useful or meaningful. To address this issue, there is a need to develop more

sophisticated methods that focus on mining only meaningful and non-redundant rare patterns. Techniques such as mining rare closed patterns and maximal rare patterns have been explored to reduce redundancy, but there is still room for improvement. Future research should investigate advanced methods for filtering rare patterns to ensure that only those with practical significance are retained.

3.8.3 Focusing on interesting rare patterns

Another significant challenge is differentiating genuinely interesting rare patterns from infrequent patterns. Not all rare patterns hold value and mining all rare occurrences can result in the discovery of irrelevant or trivial patterns. A critical research direction involves developing new measures of "interestingness" to better identify the most valuable rare patterns. These measures can incorporate domain-specific insights, anomaly detection principles, or integration of utility and weight into the mining process. By prioritizing truly interesting rare patterns, the results become more actionable and applicable to real-world scenarios.

3.8.4 Handling noise in rare pattern mining

Datasets often contain noise, which can significantly affect the quality of mined patterns. In the context of rare pattern mining, noise can lead to the discovery of patterns that do not represent true correlations, thereby reducing the quality of results. Development of techniques that can effectively handle noise is a key challenge. These techniques may include robust statistical methods, noise filtering, or integration of noise tolerance directly into mining algorithms. Future research could focus on creating noise-resilient algorithms that can uncover meaningful rare patterns in noisy datasets.

3.8.5 Enhancing explainability and interpretability

Despite remarkable advancements in deep learning (DL) and machine learning (ML) in recent years, the discovery of rare association rules remains crucial in applications where rare patterns hold significant value. Although DL and ML models have demonstrated exceptional performance in various tasks, their inherent complexity often results in "black box" models that lack transparency, making it challenging to elucidate or interpret the underlying decision-making processes. This limitation is particularly problematic in fields where explainability and interpretability are critical, such as healthcare, cybersecurity, and finance.

Rare association rules offer a complementary approach that ensures greater transparency and comprehensibility in the analysis of rare but impactful patterns. Unlike complex DL and ML models, rare association rules provide clear, interpretable relationships between infrequent patterns and their associated outcomes. These rules offer actionable insights, rendering them more suitable for real-world applications, where understanding the rationale behind a decision is as important as the decision itself.

For instance, in medical diagnostics, rare association rules can reveal subtle correlations between rare symptoms and diseases, which can subsequently be used to inform

personalized treatment plans. Similarly, in cybersecurity, rare association rules can aid in identifying atypical sequences of events that signal potential threats, enabling security teams to implement proactive measures. While DL and ML can uncover complex patterns, the interpretability of rare association rules ensures that these insights can be effectively communicated and acted upon by domain experts.

The challenge lies in developing hybrid approaches that incorporate explainability and interpretability provided by rare association rules. Future research should focus on algorithms that seamlessly integrate these methods, enabling the discovery of rare patterns that are not only accurate but also easily understood and actionable. By ensuring that these rules are interpretable and explainable, researchers can bridge the gap between cutting-edge machine learning technologies and practical, real-world applications where rare patterns are critical.

3.8.6 Case studies in rare pattern mining

Although rare pattern mining has shown significant promise in theoretical research, there is a pressing need for empirical case studies to demonstrate its practical applicability. Case studies are essential to illustrate how rare pattern mining can effectively address critical challenges in domains such as healthcare, cybersecurity, and finance. By providing concrete examples of rare but impactful patterns, case studies can showcase the utility of these techniques in solving real-world problems and offer insights into how rare pattern mining methods can be refined and adapted for specific applications.

For instance, in healthcare, case studies can highlight the early detection of rare symptoms or unusual patient behaviors, leading to improved diagnosis and treatment outcomes. In cybersecurity, case studies may demonstrate how rare pattern mining can be used to identify anomalous network activities that signal potential threats. Additionally, in finance, case studies can explore how rare pattern mining aids in detecting fraudulent transactions. These examples underscore the importance of grounding theoretical advancements in practical applications, reinforcing the value of rare pattern mining in addressing complex real-world problems.

3.8.7 Addressing the scalability challenge

As data continues to grow in volume and complexity, rare pattern mining must be adapted to meet the demands of big data. Scalability is a significant challenge, as traditional algorithms may not be able to handle the massive datasets generated in fields such as genomics, financial analysis, and social media. Research on distributed and parallelized mining algorithms, which leverage modern computing frameworks such as MapReduce and Spark, is crucial for improving the scalability of rare pattern mining. These approaches can help process large-scale data efficiently and uncover rare patterns in real time, thus providing timely insights in fast-paced environments.

3.8.8 Diversity of data

Data come in various forms and formats, such as spatio-temporal, sequential, and continuous data. Each of these data types presents unique challenges for rare

pattern mining. There is a need to develop methods that can handle diverse data formats while considering factors such as the quantity, utility, weight, and dynamic nature of data. Research could focus on adapting rare pattern mining techniques to work effectively across different data types, ensuring that meaningful patterns are discovered even in complex datasets.

3.9 Challenges addressed in this dissertation

Despite advancements in rare pattern mining, methods still face several limitations, particularly in effectively identifying rare yet valuable patterns. These limitations often result in inefficiencies and the generation of redundant patterns that obscure meaningful insight. This dissertation addresses these challenges by developing efficient methods for mining rare patterns, with a focus on reducing redundancy and enhancing the interestingness of the discovered patterns. Our approach is motivated by the need for more targeted and relevant rare pattern discovery in complex, evolving datasets, with a specific application in heart disease, one of the leading causes of death globally.

A key focus of this thesis is the development of techniques that ensure the discovery of non-redundant, interesting rare patterns. In the context of heart disease, where identifying subtle yet critical patterns could significantly improve early diagnosis and treatment, our methods are designed to capture patterns that are both informative and actionable. Additionally, we explore the use of rare association rules to enhance the explainability of the patterns, ensuring they are not only detected but also interpretable by healthcare professionals.

The primary challenges addressed in this dissertation encompass most of the limitations discussed in the previous section, particularly in Sections 3.8.1 through 3.8.6. The remaining limitations, which are covered in Sections 3.8.7 and 3.8.8, are beyond the scope of this study and will be considered in future work. The major challenges addressed in this thesis are as follows.

- **Efficient mining of rare patterns**: This challenge addresses the limitation discussed in Subsection 3.8.1 of the previous section. Timely discovery of rare patterns is essential for various applications. This dissertation tackles the challenge of improving the efficiency of rare pattern mining by introducing a novel technique that optimizes the discovery process and reduces both the computation time and resource usage.
- **Reducing redundancy in rare pattern mining**: In this challenge, we deal with the the limitation discussed in Subsection 3.8.2 of the previous section. One of the main challenges in rare pattern mining is the proliferation of redundant patterns, which complicates the mining process and diminishes practical utility. This dissertation focuses on minimizing redundancy by developing methods that efficiently prune irrelevant patterns while retaining the most meaningful ones.
- **Interestingness of patterns**: This challenge addresses the limitations discussed in Subsections 3.8.3 and 3.8.4 of the previous section. Rare patterns are highly susceptible to noise, which can obscure valuable insights and result in irrelevant

patterns. This thesis tackles this issue by implementing noise-filtering measures that enhance the quality of discovered patterns, ensuring that only the most significant and relevant patterns are retained.

- **Interpretability in critical applications: A case study on heart disease**: In this challenge, we focus on limitations in Subsections 3.8.5 and 3.8.6 in the previous section. To illustrate the practical applicability of our proposed methods, we conduct a case study focused on heart disease in the healthcare sector. This study demonstrated the effectiveness of the proposed models in addressing real-world challenges, offering actionable insights that support early detection and treatment, ultimately enhancing health outcomes. Using rare association rules in heart disease analysis enhances interpretability by uncovering subtle patterns and relationships, making insights more transparent and easier to understand for healthcare practitioners focused on early detection and intervention.

3.10 Chapter summary

This chapter provides a comprehensive overview of rare pattern mining and its significance, challenges, and applications across various domains. It begins with an examination of the limitations of traditional frequent pattern mining methods in capturing valuable rare patterns, leading to the exploration of specialized rare pattern mining algorithms developed to address contemporary data challenges, including scalability, computational efficiency, and dynamic dataset management.

Key applications of rare pattern mining are presented, illustrating its utility in fields such as medical diagnosis, education, and cybersecurity. These examples demonstrate how rare pattern identification can enable early disease detection, improve student performance monitoring, and enhance security threat mitigation.

The chapter also addresses several critical research challenges, such as developing more efficient algorithms, minimizing redundant patterns and avoiding noise. A central focus remains on improving methods for assessing the interestingness of rare patterns, enabling the distinction between valuable insights and trivial findings, particularly in high-dimensional datasets.

Furthermore, the discussion highlights the critical importance of scalability and adaptability in rare pattern mining to effectively manage heterogeneous data types, including temporal, spatial, and sequential data. This underscores the necessity for robust algorithmic frameworks capable of addressing the inherent complexities associated with large-scale data environments.

While substantial progress has been achieved in this domain, several fundamental challenges and open research questions persist. Addressing these issues remains essential to advancing the field and unlocking the full potential of rare pattern mining for generating meaningful insights and practical innovations across diverse application areas.

4. Mining Rare Patterns Efficiently

This chapter addresses the first challenge by proposing a novel method for mining rare patterns. In the previous chapter, we reviewed related work and examined various methods for rare pattern mining, emphasizing tree-based approaches as superior in performance. Despite being considered the most efficient method for rare pattern generation, the RP-growth algorithm [Tsang et al., 2011] exhibits performance limitations on sparse datasets and requires substantial computational time and memory resources, particularly when mining with low user-defined thresholds.

To address these limitations, we introduce the Rare Pre Post (RPP) algorithm [Darrab et al., 2020], whcih is a more efficient approach for rare pattern mining. RPP overcomes the generation of unnecessary candidate patterns and eliminates the requirement for constructing conditional trees, thereby significantly reducing the computational overhead. This method employs a novel data structure, N-list, to further optimize the mining process. Experimental evaluations conducted on both sparse and dense datasets demonstrate that the proposed method, RPP, outperforms RP-growth, offering an improvement in performance of approximately an order of magnitude.

4.1 Introduction

Rare pattern mining (RPM) is the discovery of infrequent patterns that provides valuable insights despite their rarity. RPM identifies less obvious patterns in datasets that are significant, as opposed to frequent pattern mining (FPM), which identifies common patterns in datasets. As an example, RPM can significantly improve patient safety and treatment outcomes by detecting rare complications or side effects earlier, which can significantly enhance patient safety and treatment outcomes.

RPM is also significant in domains such as predicting telecommunication equipment failures [Bhatt and Patel, 2015], detecting fraudulent credit card transactions [Weiss, 2004], and monitoring adverse drug reactions [Ji et al., 2012]. The early identification of rare yet consequential events in these fields facilitates more effective decision-making and timely interventions. For instance, recognizing rare patterns in transaction data

can help prevent significant financial losses by uncovering fraudulent activities that deviate from normal behavior.

However, RPM often depends on a user-defined support threshold to filter out uninteresting patterns, which introduces a fundamental trade-off. Lowering the support threshold captures rare patterns but generates an overwhelming number of patterns, many of which are irrelevant, thereby increasing computational costs and complicating the analysis. Conversely, raising the support threshold to reduce computational overhead excludes potentially valuable rare patterns, diminishing the effectiveness of the mining process. This challenge, known as the rare item problem [Borah and Nath, 2020; Darrab et al., 2021b], underscores the difficulty in balancing comprehensive pattern discovery with computational efficiency.

To address these challenges, various RPM approaches have been proposed and categorized according to their search strategies: breadth-first search (BFS) and depth-first search (DFS) [Darrab et al., 2021b]. BFS methods, such as the Apriori algorithm [Agrawal and Srikant, 1994], systematically explore the search space but suffer from drawbacks such as generating excessively large candidate sets and requiring multiple data scans, making them unsuitable for RPM, especially with large datasets.

In response to these limitations, DFS-based methods have been developed, utilizing more sophisticated data structures such as the FP-tree [Han et al., 2000]. A notable advancement is the RP-growth algorithm [Tsang et al., 2011], which reduces the need for repeated data scans and enhances memory efficiency by compressing the dataset into an FP-tree. This compression allows for faster pattern mining without generating excessive candidate sets.

Despite these improvements, RP-growth still faces challenges. Specifically, it tends to generate numerous unnecessary conditional trees, significantly increasing the search time and memory usage, particularly in sparse datasets where rare patterns are more widespread. Sparse datasets pose substantial difficulties for RPM algorithms due to the uneven distribution of patterns, with rare patterns often obscured by noise. Studies have shown that the performance of existing RPM methods degrades significantly in such environments [Borah and Nath, 2019] because they struggle to efficiently navigate sparse data and isolate patterns of interest.

The purpose of this chapter is to address these shortcomings by introducing the Rare Pre Post(RPP) algorithm, a novel approach for efficiently discovering rare patterns. The RPP algorithm leverages the **N-list** data structure [Deng et al., 2012], which has demonstrated its potential to enhance pattern mining efficiency in other contexts. The primary innovation of the RPP algorithm is its ability to bypass the generation of unnecessary candidate patterns and avoid the construction of conditional trees, thereby mitigating the inefficiencies of previous approaches.

The key contributions of the proposed method are as follows:

- **RPPC-tree Construction**: We propose the **RPPC-tree**, a new data structure that captures essential information for rare pattern extraction. By focusing on transactions containing rare items, the RPPC-tree effectively reduces the dataset to its most relevant components, streamlining the mining process.

- **RN-list Creation**: We introduce the **RN-list**, a novel representation of significant rare items that forms the backbone of the mining process. This data structure enables efficient intersection operations and facilitate the comprehensive identification of rare patterns with minimal computational overhead.
- **Efficient Mining Process**: The **RPP algorithm** utilizes RN-lists to systematically intersect them, generating rare patterns while significantly reducing computational costs in terms of both time and memory. This approach ensures that no relevant patterns are missed, offering a highly efficient and scalable solution for rare pattern mining.

The RPP algorithm presents a promising approach for addressing the challenges of rare pattern mining, especially in sparse datasets where existing methods often fall short. By enhancing both efficiency and accuracy, the proposed algorithm provides a method for more effective real-world applications of rare pattern mining.

In the following sections, we present the proposed method, experimental results, and conclusions. To avoid redundancy, related work is not revisited here, as it has been thoroughly covered in the introduction and previous chapters.

4.2 Proposed approach: RPP algorithm

In Chapter 3, we reviewed various approaches for rare pattern mining, including level-wise methods such as Apriori-Inverse and Apriori-Rare [Darrab et al., 2021b; Liu et al., 1999; Lu et al., 2020; Selvarani and Jeyakarthic, 2021]. These methods use a bottom-up traversal strategy but are computationally intensive due to multiple dataset scans. Depth-wise methods, such as RP-growth, CFP-growth, and mis-eclat, improve efficiency by addressing the limitations of level-wise methods, specifically repeated scans and candidate generation [Borah and Nath, 2020; Darrab et al., 2021b]. Among these, the RP-tree structure has shown superior performance with promising results. However, challenges remain, particularly with algorithms such as RP-growth, which face inefficiencies in sparse datasets because of the repeated construction of conditional trees.

To address these limitations, we propose the RPP algorithm, described in Algorithm 1 and illustrated in Figure 4.1. The RPP algorithm eliminates the need for conditional tree construction, candidate generation, and multiple dataset scans by leveraging RN-lists, as introduced in [Deng et al., 2012]. This approach significantly reduces computational overhead while ensuring accurate rare pattern extraction. Inspired by the method in [Deng et al., 2012], which targets frequent patterns, our approach focuses specifically on rare patterns. By overcoming the inefficiencies of the RP-tree, particularly in sparse datasets where rare patterns are more prevalent, our method is highly effective for mining rare patterns. Moreover, by completely avoiding conditional tree construction, greater efficiency is achieved.

4.2.1 Step-by-step process of the proposed approach

The purpose of this section is to demonstrate the functionality of the proposed RPP method through an illustrative example that clarifies its three essential phases:

Algorithm 1 The Proposed RPP Algorithm for Mining Rare Patterns

Input: Dataset D of transactions, $minSup$, $maxSup$
Output: Set of rare patterns $Rare\_Patterns$

procedure RPP
    Step 1: Construct RPPC-tree
    Initialize an empty RPPC-tree
    for each transaction T in dataset D **do**
        if T contains at least one rare item **then**
            Add T to the RPPC-tree
            Set current node to root
            for each item x in T **do**
                if node for x exists as a child of current node **then**
                    Increment the count of node x
                else
                    Create new node for item x, set count = 1
                    Assign pre-order value
                end if
                Move current node to child node representing x
            end for
        end if
    end for
    Perform post-order traversal to assign post-order values

    Step 2: Generate RPP-Codes and Construct RN-lists
    Initialize empty lists for RPP-codes and RN-lists
    for each node X in RPPC-tree (pre-order traversal) **do**
        Generate RPP-code for X: $\{(X_{\text{pre-order}}, X_{\text{post-order}}), \text{count}\}$
        Store RPP-code associated with each item
    end for
    Sort RPP-codes in ascending order of pre-order values
    Construct RN-lists from sorted RPP-codes

    Step 3: Mine Rare Patterns using RN-lists
    Initialize $Rare\_Patterns \leftarrow \emptyset$
    for each pair of RN-lists (RN_list_X, RN_list_Y) **do**
        $RN\_Intersection \leftarrow RN\_list_X \cap RN\_list_Y$
        if $RN\_Intersection \neq \emptyset$ **then**
            Compute $support(RN\_Intersection)$
            if $minSup \leq support(RN\_Intersection) < maxSup$ **then**
                $Rare\_Patterns \leftarrow Rare\_Patterns \cup \{RN\_Intersection\}$
            end if
        end if
    end for
    Repeat intersections iteratively for increasing pattern length k
    Terminate when no further intersections can occur
end procedure

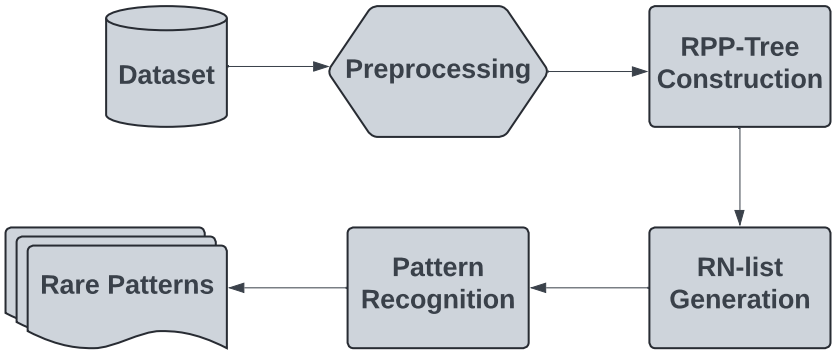


Figure 4.1: The RPP method workflow

construction of the RPPC-tree, creation of RN-lists for items, and generation of rare patterns.Thee phases are described in detail in the following sections. subsections.

Motivating Example: Given the transaction dataset *DB* in Table 4.1, let the maximum support threshold (*maxSup*) and rare support threshold (*minSup*) be 4 and 2, respectively. The task of rare pattern mining is to extract a set of all rare patterns with support not less than *minSup* and no more than *maxSup*.

Table 4.1: A simple dataset

TID	Items	Ordered Items
1	a, b, c, d	b, c, a, d
2	b, d	b, d
3	a, b, c, e	b, c, a, e
4	c, d, e, h	c, d, e
5	a, b, c, g	b, c, a

4.2.2 Construction of the RPPC-tree

The initial phase of the RPP method consists of the construction of the RPPC-tree, which stores critical information from the dataset, as illustrated in Figure 4.2 and in lines 1-19 of Algorithm 1. The tree is constructed using two sequential scans of the dataset:

- **First Scan:** The support of 1-items is calculated, and items with support less than *minSup* are discarded.
- **Second Scan:** Transactions containing at least one rare item (items with support between *minSup* and *maxSup*) are added to the RPPC-tree.

The RPPC-tree consists of a root node labeled “null” and child nodes representing patterns. Each node contains the following fields:

- **item-name**: The name of the item represented by the node.

- **count**: The number of transactions that reach this node.
- **pre-order and post-order**: The positions of the node during pre-order and post-order traversal of the tree.

The RPPC-tree differs from the traditional RP-tree by omitting node links and incorporating pre-order and post-order fields for each node. This structure is used to form the NL-list, which is discussed in the following subsections. Once the RN-lists are generated, the RPPC-tree is discarded.

In our motivating example, as shown in Table 4.1, the RPPC-tree is constructed by scanning the dataset and removing items with support less than $minSup = 2$. For example, items {g, h} are discarded. The remaining items are sorted in descending support order and are used to construct the RPPC-tree during the second scan. Figure 4.2 shows the resulting RPPC-tree, where each node contains the item, count, and pre-post rank. For example, node c has RPP-code {((2, 3): 3)}, meaning c has a count of 3 and a pre-post rank of (2, 3).

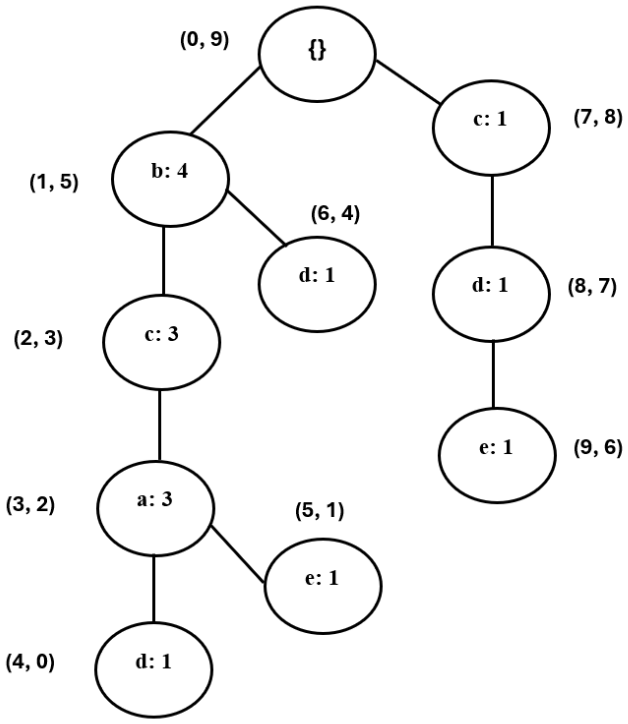


Figure 4.2: The RPPC-tree constructed from the transactions in Table 4.1.

4.2.3 Generating RN-lists of items

The second phase of the RPP method aims to generate RN-lists from the RPPC-tree, as illustrated in Figure 4.1 and lines 20–27 in Algorithm 1. An RN-list comprises RPP-codes for each item in a tree. The RN-list of item X encompasses all RPP-codes of nodes representing X, arranged in ascending order of their pre-order values. The support of X is computed as the sum of the counts in its corresponding RN-list.

Based on the RPPC-tree shown in Figure 4.2, the RN-lists for the interesting 1-patterns can be generated. For instance, the RN-list of item e contains two RPP-codes: {((5, 1): 1), ((9, 6): 1)}. The support of e is calculated as $1 + 1 = 2$. Table 4.2 shows the RN-lists of all interesting items from the motivating example.

Table 4.2: RN-lists of interesting rare items

Item	RPP-codes	Support
b	{(1, 5): 4}	4
c	{(2, 3): 3, (7, 8): 1}	4
a	{(3, 2): 3}	3
d	{(4, 0): 1, (6, 4): 1, (8, 7): 1}	3
e	{(5, 1): 1, (9, 6): 1}	2

4.2.4 Generation of rare patterns

The final phase, as shown in Figure 4.1 and lines 28 through 40 in Algorithm 1, involves generating rare patterns by comparing the RN-lists of two items, X and Y. Pattern XY can be generated if X is an ancestor of Y, which is determined when X.pre-order $<$ Y.pre-order and X.post-order $>$ Y.post-order.

Consider generating a rare pattern {ce}. The RN-list of c is {((2, 3): 3), ((7, 8): 1)}, and the RN-list of e is {((5, 1): 1), ((9, 6): 1)}. The comparison proceeds as follows.

- The RPP-code of c {((2, 3): 3)} is compared with {((5, 1): 1)}, and since $2 < 5$ and $3 > 1$, the ancestor-descendant relation holds. We add {((2, 3): 1)} to the RN-list of {ce}.
- The next RPP-code of c {((7, 8): 1)} is compared with {((9, 6): 1)}, and since $7 < 9$ and $8 > 6$, the ancestor-descendant relation holds again. We add {((7, 8): 1)} to the RN-list of {ce}.

The final RN-list of {ce} is {((2, 3): 1), ((7, 8): 1)}, and its support is $1 + 1 = 2$. Because $2 \leq \text{Sup}(ce) = 2 < 4$, {ce} is a valid rare pattern.

By repeating this process, we can generate the following rare patterns for our motivating example:

$$\{a : 3, e : 2, d : 2, ba : 3, bd : 2, ca : 3, cd : 2, ce : 2, bca : 3\}.$$

We generate rare patterns only from rare items {a, d, e} because items with support greater than *maxSup* are excluded from consideration.

4.3 Experimental results

To evaluate the performance of the RRP algorithm, we compared it with the most efficient algorithm for mining rare patterns, RP-growth [Deng et al., 2012]. Several experiments were conducted on four real-world datasets: Mushroom, Retail, Pumsb, and Kosarak. Both sparse datasets (Kosarak and Retail) and dense datasets (Mushroom and Pumsb) were used in the evaluation process. The characteristics of the

datasets are summarized in Table 4.3. The last column in Table 4.3 shows the density of the datasets, calculated as the ratio of the average transaction size to the total number of distinct items.

For each dataset, the number of transactions, number of distinct items, and average transaction size are denoted by # of Trans, # of Items, and AvgTrans, respectively. All datasets were downloaded from the FIMI repository [FIMI Repository]. The experiments were conducted on a Windows 10, 64-bit operating system with an Intel Core i7-7700HQ CPU at 2.80 GHz, 16 GB of RAM, and a 1 TB hard disk. Both algorithms were implemented in Java to ensure a consistent implementation environment. The source code for the RPgrowth algorithm was obtained from [Fournier-Viger et al., 2016].

Table 4.3: Characteristics of the Datasets

Dataset	Size (MB)	# of Items	# of Trans	AvgTrans	minSup (%)	maxSup	Density
Mushroom	19.3	119	8,124	23	{0.1, 0.2, ..., 0.9}	0.01	19.3
Retail	4.2	16,470	88,126	10.3	{0.1, 0.2, ..., 1}	0.1	0.006
Pumsb	16.3	2,113	49,046	74	{52.5, 55, ..., 70}	0.8	3.5
Kosarak	30.5	41,271	990,002	8.1	{0.1, 0.2, ..., 0.9}	0.01	0.002

Table 4.3 provides key information about each dataset, including size in megabytes (MB), the number of distinct items, the number of transactions, the average number of items per transaction (AvgTrans), and the support thresholds (*minSup* and *maxSup*) used. Additionally, the density of each dataset is provided, which is a critical factor for assessing performance, especially when comparing sparse and dense datasets.

The primary performance metrics evaluated were runtime, memory consumption, and scalability. These metrics were analyzed across both sparse and dense datasets to assess the scalability and efficiency of the RPP algorithm in comparison with the RP-growth method.

4.3.1 Execution time

To assess the performance of the RPP algorithm, we compared it to the RP-growth algorithm across all datasets listed in Table 4.3. Each experiment was conducted using two support thresholds: *maxSup* (maximum support) and *minSup* (minimum rare support). While *maxSup* remained fixed, *minSup* was varied as indicated in Table 4.3.

The goal of these experiments was to extract rare patterns with support values less than *maxSup* but greater than or equal to *minSup*. Figures 4.3a through 4.3d show the runtime performances of the RPP and RP-growth algorithms across all datasets. In these figures, the X-axis represents *minSup* values, and the Y-axis shows execution time in seconds. The results clearly indicate that the RPP algorithm consistently outperformed RP-growth across all datasets in terms of execution time.

The enhanced efficiency of the RPP algorithm can be attributed to its use of RN-lists, which significantly reduces the computational overhead involved in generating rare patterns. In contrast, RP-growth incurs higher overhead due to the construction of

conditional trees for each rare item. This difference becomes particularly noticeable at lower *minSup* values, where the number of rare patterns increases significantly.

For example, when mining the Kosarak dataset with a *minSup* threshold of 0.1%, 759,391 rare patterns are generated, whereas with a *minSup* of 0.9%, only 32 rare patterns are produced. Despite the significant increase in rare patterns at lower *minSup* values, the RPP algorithm maintains high efficiency and achieves significantly faster execution times compared to RP-growth

As *minSup* increases, the performance gap between RPP and RP-growth narrows. This is expected, as higher *minSup* values reduce the number of rare patterns, resulting in more comparable performance between the two algorithms.

Overall, the results demonstrate that the RPP algorithm is significantly more efficient, particularly when dealing with large datasets and low *minSup* values, where the number of rare patterns is substantial.

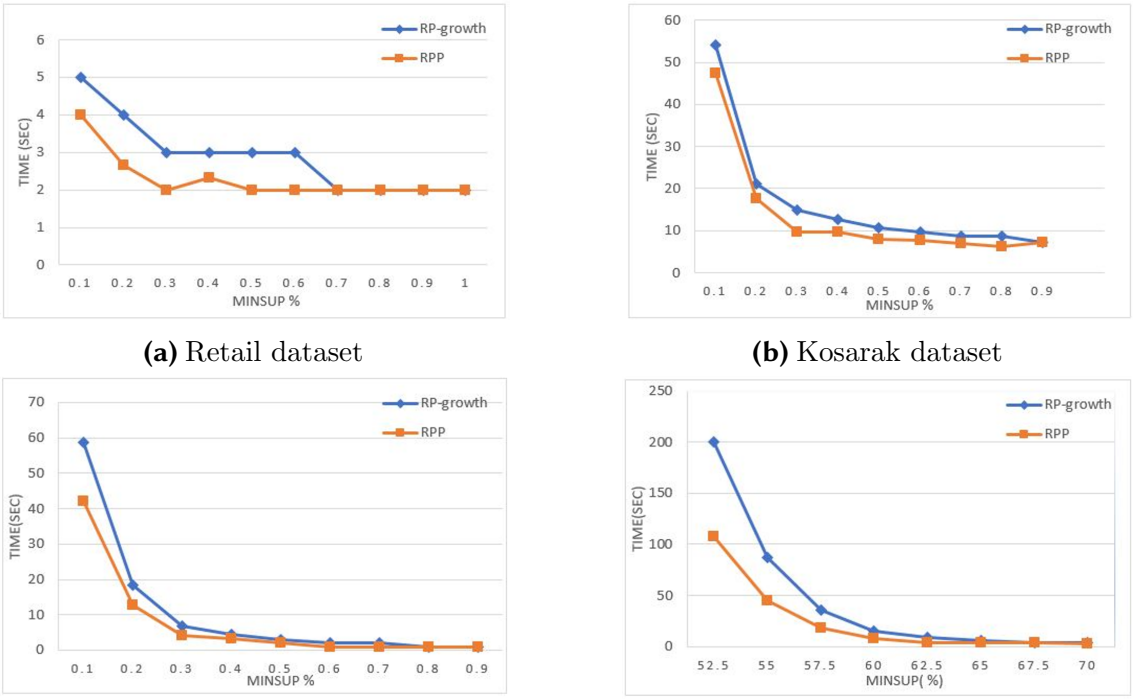
Our analysis compares the performance of the RPP and RP-growth algorithms across four datasets, as shown in Figure 4.3. The runtime shows significant variation depending on the dataset characteristics. The RPP algorithm provides strong performance in both sparse and dense datasets. For instance, in the sparse datasets Retail and Kosarak, as illustrated in Figures 4.3a and 4.3b, the RPP algorithm consistently outperforms the RP-growth algorithm in terms of runtime efficiency.

Interestingly, despite being optimized for sparse datasets, where rare patterns are more prevalent, the RPP algorithm also performs competitively in dense datasets. In the Mushroom and PUMSB datasets, shown in Figures 4.3c and 4.3d, the RPP algorithm achieves better runtime performance compared to RP-growth, even though tree-based methods tend to work well with dense patterns due to their compressed nature. This ability of the RPP algorithm to maintain superior runtime performance, regardless of dataset density, underscores its efficiency and adaptability across varying data characteristics.

4.3.2 Memory consumption

To evaluate memory consumption, we utilized the same support thresholds (*maxSup* and *minSup*) as outlined in Table 4.3. Comprehensive experiments were conducted using the datasets listed in Table 4.3. Figures 4.4a - 4.4d illustrate the memory usage of the RPP and RP-growth algorithms across varying *minSup* values. As in the runtime experiments, the *maxSup* threshold was fixed, while the *minSup* threshold was varied. In these figures, the X-axis represents the *minSup* values, and the Y-axis represents the memory consumption for both algorithms.

The RPP algorithm demonstrates greater memory efficiency on sparse datasets. Figures 4.4a and 4.4b show that, on sparse datasets like Retail and Kosarak, RPP generally consumes less memory than RP-growth. Sparse datasets contain fewer rare patterns, but the search space is large. RPP effectively reduces this search space by utilizing RN-lists to directly extract rare patterns without constructing conditional trees. In contrast, RP-growth generates a large RP-tree on sparse datasets, leading to the creation of numerous conditional trees, which significantly increases memory usage. In the Retail dataset, RPP consumes less memory when *minSup* exceeds



(a) Retail dataset

(b) Kosarak dataset

(c) Mushroom dataset

(d) PUMSB dataset

Figure 4.3: Runtime comparisons between RPP and RP-growth across four datasets

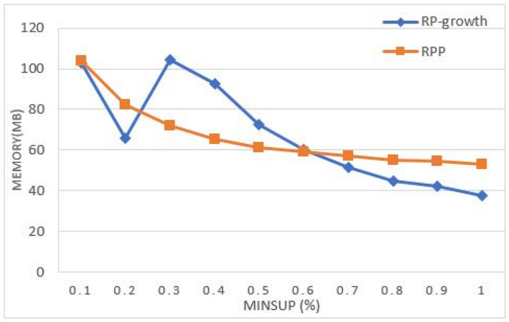
0.3%, although it marginally surpasses RP-growth's memory usage at higher *minSup* values (above 0.6%).

For dense datasets, such as Mushroom and PUMSB, as shown in Figures 4.4c and 4.4d, RPP consumes more memory than RP-growth. Since most items in dense datasets frequently co-occur, the RP-tree remains compact, allowing RP-growth to generate rare patterns without the overhead associated with multiple conditional trees. In contrast, RPP consumes more memory due to additional data structures, specifically the maintenance of pre-order and post-order information for each node in the RPPC-tree and the larger RN-lists, as more items must be processed during mining. Additionally, dense datasets often have larger average transaction sizes, further contributing to the increased memory usage in RPP.

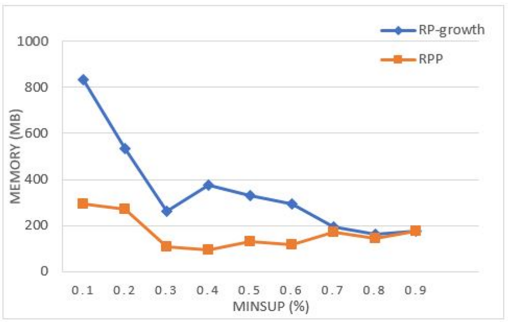
Overall, while the RPP algorithm performs efficiently in terms of memory usage for sparse datasets, it requires more memory than RP-growth when mining dense datasets due to its handling of more extensive data structures.

4.3.3 Scalability

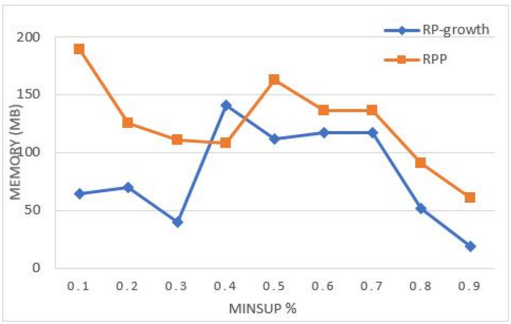
To evaluate the scalability of the proposed RPP algorithm compared with RP-growth, we conducted experiments on the largest dataset, Kosarak, which contains approximately 1 million transactions. The dataset was evenly divided into ten parts, and for each experiment, 10% of the dataset was incrementally added to the previous accumulative parts. Figures 4.5a and 4.5b present the experimental results, highlighting the scalability of both algorithms in terms of execution time and memory consumption.



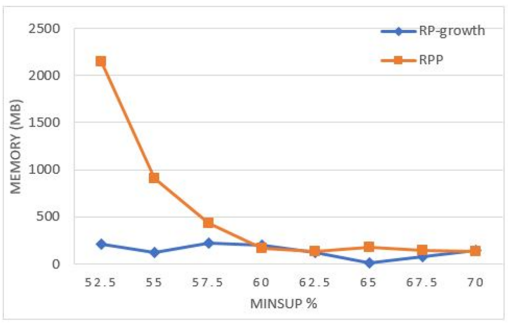
(a) Retail dataset



(b) Kosarak dataset



(c) Mushroom dataset



(d) PUMSB dataset

Figure 4.4: Memory consumption comparisons between RPP and RP-growth across four datasets

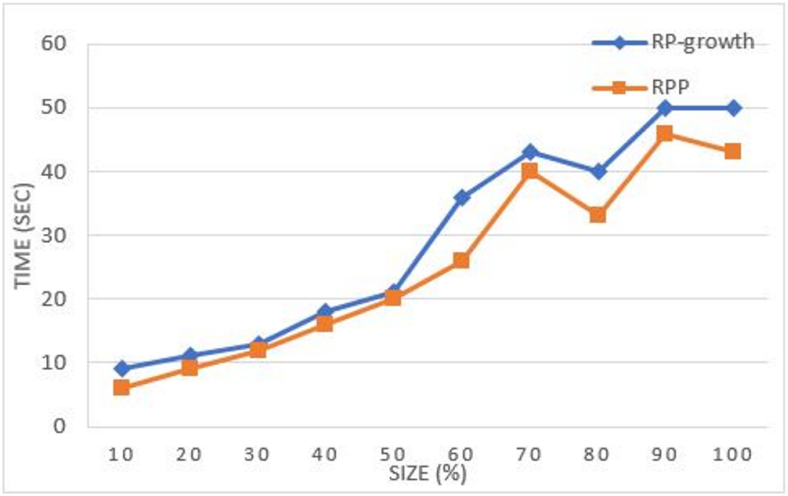
The results indicate that the RPP algorithm scales more efficiently than the RP-growth as the dataset size increases. RPP requires less time and memory due to its use of RN-lists during the mining process, allowing it to focus exclusively on rare patterns. In contrast, RP-growth must traverse a much larger search space and generate a substantial number of conditional trees, leading to significantly higher execution time and memory usage.

As shown in Figure 4.5a, the RPP algorithm consistently outperforms RP-growth in terms of execution time across all dataset sizes. This efficiency is attributed to RPP's ability to narrow down the mining process by directly leveraging RN-lists, thereby reducing computational overhead. Conversely, RP-growth needs to repeatedly construct conditional trees, which becomes increasingly computationally expensive as the dataset size increases.

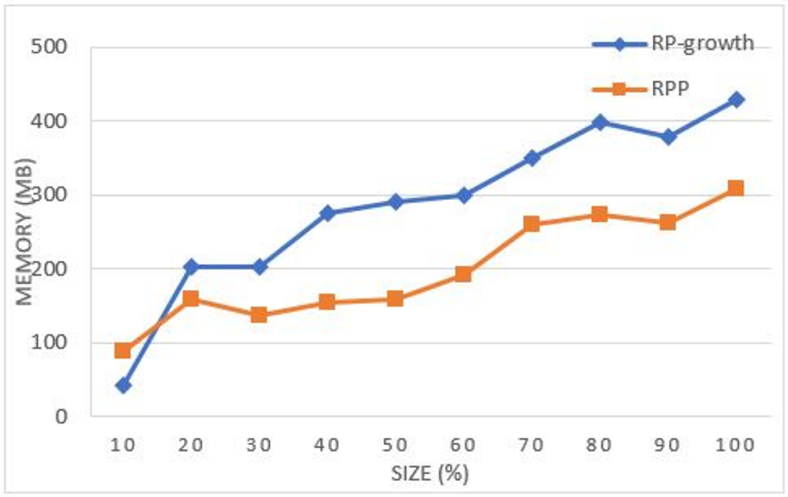
In terms of memory consumption, the RPP algorithm demonstrates better scalability than RP-growth when processing sparse datasets, requiring less memory as dataset size increases, as illustrated in Figure 4.5b. However, as noted earlier, RPP tends to consume more memory on dense datasets due to the need to manage pre-order and post-order information in the RPPC-tree, as well as larger RN-lists. Nevertheless, the scalability advantage of RPP remains evident as dataset size increases, particularly in sparse datasets such as Kosarak.

4.3.4 Discussion

The experimental results demonstrate that the RPP algorithm consistently outperforms RP-growth in terms of execution time, particularly for sparse datasets with



(a) Kosarak dataset - Time Scalability



(b) Kosarak dataset - Memory Scalability

Figure 4.5: Scalability comparisons between RPP and RP-growth for the Kosarak dataset in terms of time and memory

low *minSup* values, where the number of rare patterns significantly increases. The advantage of RPP stems from its use of RN-lists, which eliminates the need for conditional tree generation, thereby reducing computational overhead and accelerating the mining process. In contrast, RP-growth encounters difficulties due to larger search spaces and the repetitive construction of conditional trees, leading to slower execution times, especially at lower *minSup* values. However, as *minSup* increases and the number of rare patterns decreases, the performance gap between the two algorithms shrinks.

In terms of memory consumption, the behavior of the two algorithms varies based on dataset characteristics. RPP demonstrates superior memory efficiency in sparse datasets such as Retail and Kosarak, as it avoids the creation of large conditional trees, which are a significant memory burden for RP-growth. However, in dense datasets such as Mushroom and PUMSB, RP-growth proves more memory-efficient due to the compact nature of the RP-tree, minimizing the need for additional conditional trees.

In contrast, RPP incurs higher memory costs in dense datasets due to the overhead of maintaining pre-order and post-order information in the RPPC-tree, along with the larger RN-lists required to handle the frequent co-occurrence of items and larger transaction sizes typical of dense datasets.

When considering scalability, RPP exhibits clear advantages over RP-growth, particularly in large datasets such as Kosarak. As dataset size increases, RPP scales efficiently in both execution time and memory usage, benefiting from its reliance on RN-lists for direct rare pattern extraction. In contrast, RP-growth suffers from scalability limitations as the dataset grows, requiring substantial memory and computational resources to manage the increasing number of conditional trees and expanding search space. Though RP-growth performs adequately on smaller datasets, its efficiency declines sharply with larger datasets, positioning RPP as a more scalable and effective solution for large-scale rare pattern mining.

4.4 Chapter summary

The experimental results highlight the RPP algorithm's strengths and weaknesses in comparison to RP-growth for rare pattern mining. RPP consistently outperforms RP-growth in execution time, particularly for sparse datasets with low minimum support thresholds (*minSup*), by leveraging RN-lists, which eliminates the need for conditional tree generation and reduces computational overhead. Memory consumption varies based on dataset density: RPP is more memory efficient for sparse datasets as it avoids the burden of constructing large conditional trees, whereas RP-growth is more efficient in dense datasets because of its compact RP-tree structure. Additionally, RPP demonstrates superior scalability, handling larger datasets with lower memory and time requirements compared to RP-growth, which struggles with increased dataset size due to its reliance on extensive conditional tree construction. Overall, RPP presents a more scalable and effective approach for large-scale rare pattern mining, especially in scenarios with sparse datasets and low support thresholds.

5. Efficient Discovery of Compact Rare Patterns

In the previous chapter, we explored the RPP algorithm and its ability to identify rare patterns within datasets. Despite its effectiveness, RPP algorithm generates an exhaustive number of patterns, many of them are redundant. This redundancy increases computational complexity, consumes more time and memory, and ultimately limits the scalability and practicality of the algorithm, especially in large datasets or time-sensitive applications.

To address these inefficiencies, this chapter introduces the Maximal Rare Pattern (MaxRI) algorithm, which refines rare pattern mining by focusing on maximal rare patterns—the longest rare patterns in a dataset. These patterns act as concise representatives of the full rare pattern set, significantly reducing redundancy and improving the efficiency and interpretability.

The proposed MaxRI algorithm narrows the search space, leading to faster processing and lower memory usage, making it particularly valuable in real-world scenarios with limited computational resources. In addition, we present the Recover Rare Patterns (RRI) method, which allows users to extract rare patterns of any specified length from the maximal rare patterns, providing flexibility without unnecessary complexity.

Together, MaxRI and RRI offer a more efficient approach to rare pattern mining, outperforming traditional methods in terms of speed and memory usage. These algorithms are particularly suited for applications such as anomaly detection, where the discovery of rare yet meaningful patterns is essential.

5.1 Introduction

Frequent pattern mining has been extensively studied in the field of data mining, resulting in the development of numerous methods aimed at generating condensed representations of frequent patterns such as closed and maximal patterns [Gouda and Zaki, 2001; Luna et al., 2019]. These methods efficiently reduce redundancy

and improve scalability and interpretability. In contrast, rare pattern mining, which seeks to uncover infrequent but significant patterns, has received comparatively less attention.

Although several methods have been proposed to address rare pattern mining [Darrab et al., 2020; Koh and Rountree, 2005; Tsang et al., 2011], many suffer from the generation of overly large and redundant sets of patterns. Rare pattern mining is particularly sensitive to the choice of minimum support thresholds; setting the threshold too low can result in an overwhelming number of patterns, particularly in dense datasets. This challenge is especially critical in domains where the early detection of rare events is vital, such as anomaly detection and predictive diagnostics. Thus, the development of approaches that generate concise, non-redundant representations of rare patterns is an important and timely research problem.

In frequent pattern mining, methods for identifying closed and maximal patterns have been well established [Burdick et al., 2005; Gouda and Zaki, 2001; Lu et al., 2020; Mi, 2022; Wu et al., 2022], and offer compact representations of frequent phenomena. However, to our knowledge, no corresponding methods have been developed for mining concise representations of rare patterns, such as maximal rare patterns. This represents a significant gap in literature. Maximal rare patterns represent subsets of any other rare patterns, provide a non-redundant, concise representation of rare events. Mining these patterns would reduce computational overhead and improve interpretability, making them particularly useful in fields where rare events are more insightful than frequent ones, such as fraud detection and medical diagnosis.

To address these challenges, this chapter introduces the Maximal Rare Itemset (MaxRI) algorithm, which aims to recover a compressed and non-redundant representation of rare patterns. By extracting maximal rare patterns, the algorithm reduces the computational time and memory usage, making the results more manageable and facilitating expert analysis. This approach addresses the inefficiencies of existing methods by eliminating redundancy and enhancing the overall performance.

Furthermore, to support deeper analysis, we propose a Recovering Rare Itemsets (RRI) procedure. This method enables users to extract rare patterns of a specified length k from a set of maximal rare patterns, providing flexibility for targeted exploration without the need to generate a full set of patterns. This added capability significantly enhances the practical applicability of the proposed methods

The key contributions of this chapter are as follows.

- Introduction of the MaxRI algorithm for discovering non-redundant maximal rare patterns, overcoming limitations of existing approaches.
- Efficient extraction of maximal rare patterns using the MRI-tree structure.
- Extensive experimental validation on real-world dense datasets, showing that MaxRI significantly outperforms state-of-the-art algorithms in both time and memory efficiency.
- Introduction of the RRI algorithm for retrieving rare patterns of specified lengths, enabling targeted exploration.

The remainder of this chapter is organized as follows. Section 5.2 introduces the MaxRI algorithm, along with a motivating example to illustrate its functionality. Section 5.3 presents the experimental results and provides insights into the performance of the algorithm. Finally, Section 5.4 summarizes the key findings and concludes the chapter.

5.2 Proposed approach: MaxRI algorithm

In this section, we describe the proposed approach, called the Maximal Rare Itemset (MaxRI) algorithm, which is designed to mine a concise representation of rare patterns, specifically focusing on maximal rare patterns. The goal of MaxRI is to extract representative rare patterns that reduce runtime, minimize memory consumption, and facilitate expert analysis. In addition, we introduce the RRI algorithm, which enables users to retrieve interesting rare patterns of a specified length k from the concise representation generated by MaxRI.

The MaxRI algorithm utilizes an FP-tree structure [Borgelt, 2005] to identify a comprehensive set of long rare patterns. The process begins with a preprocessing phase that eliminates unpromising items (i.e., items with support below the minSup threshold). Subsequently, the dataset is compressed into a compact structure termed the Maximal Rare Itemset Tree (MRI-tree). Finally, the algorithm extracts representative rare patterns directly without employing a candidate-test approach or conditional trees. The detailed process is illustrated in Figure 5.1 and outlined in Algorithm 2. To better understand the operation of the MaxRI algorithm, we explain its three phases using a motivating example.

Motivating Example: Given the transaction dataset DB shown in Table 5.1, the maximum support threshold (maxSup) is set to 4, and the rare support threshold (minSup) is set to 2. The goal of maximal rare pattern mining is to extract a set of maximal rare patterns whose support is greater than or equal to minSup, but less than maxSup.

Table 5.1: Original dataset

TID	Items	Sorted Items
1	1, 3, 4	3, 1, 4
2	2, 3, 5	2, 3, 5
3	1, 2, 3, 5	2, 3, 5, 1
4	2, 5	2, 5
5	1, 2, 3, 5	2, 3, 5, 1

Table 5.2: Support of 1-items

Item	Support
5	4
3	4
2	4
1	3
4	1

5.2.1 Preprocessing phase

This step in the proposed algorithm, as illustrated in Figure 5.1 and described in lines 3 to 8 of Algorithm 2, involves scanning the dataset to compute the support count of each 1-item. Unpromising items, whose support does not meet the minimum threshold, are removed from the dataset. The remaining transactions are then sorted in descending order based on item support. The resulting dataset is shown in the

Table 5.3: Tidset of items

Item	Tidset
1	T1, T3, T5
2	T2, T3, T4, T5
3	T1, T2, T3, T5
4	T1
5	T2, T3, T4, T5

right column of Table 5.1, and the corresponding support values of the 1-patterns are presented in Table 5.2.

In addition, this phase generates a Tidset (a set of transaction IDs) to facilitate the calculation of support for rare subset patterns derived from maximal rare patterns, as demonstrated in Table 5.3. This step is critical for efficiently managing the mining process, ensuring that only relevant patterns are retained for further analysis. Tidset is used in the RRI algorithm to extract rare patterns of length k from the concise representation generated by MaxRI.

5.2.2 Construction of the MRI-tree

In this phase, the second step of the MaxRI algorithm is executed, as shown in the workflow in Figure 5.1 and detailed in lines 1–5 of Algorithm 2. The MRI-tree is constructed using the transactions in the right column of the preprocessed dataset in Table 5.2. Similar to the FP-tree construction process, the dataset is scanned again to build the MRI-tree. Initially, the tree contains a root node labeled `null`.

The transactions in the right column of Table 5.2 are then inserted into the MRI-tree in descending order of their support. If an inserted transaction shares a prefix with previously inserted transactions, the count of all nodes along the shared path is incremented by one. Otherwise, new nodes are created and initialized with a count of one. Each node in the tree contains the following information: the item's name, count, children, parent, and link to other nodes with the same item name.

For example, consider node (2:4) in the tree, where numbers 2 and 4 represent the item's name and its occurrence on path {1532}, respectively. Node (2:4) has two children (3 and 5), and its parent is the root of the tree. The compact MRI tree for our motivating example is shown in Figure 5.1.

To facilitate efficient traversal of the tree, a rare header table is created to store only rare items that are eligible for generating maximal rare patterns. By reordering the items in descending order of their support, the resulting compact tree ensures that the most relevant rare items are placed at the bottom of the tree.

5.2.3 Mining process for the MaxRI algorithm

This phase represents the final phase of the MaxRI algorithm, as illustrated in Figure 5.1, with detailed steps outlined in lines 6-39 of Algorithm 2. The mining process follows a bottom-up approach, starting from the lowest item in the rare header table. For each rare item, the MaxRI algorithm retrieves the corresponding

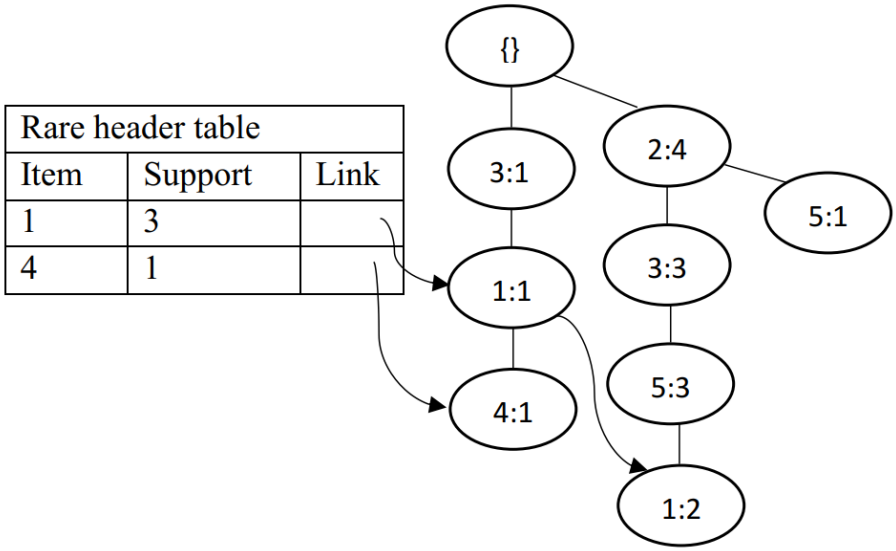


Figure 5.1: The compact MRI-tree after adding all transactions.

paths for item i from the compact MRI tree. The support count of a retrieved path is set as the occurrence of item i within that path.

It is important to note that the support count of the prefix items cannot exceed that of the suffix item, as the items are inserted in descending order of their support. This ensures that the mining process effectively identifies the most relevant maximal rare patterns.

Condition 1: Criteria for adding patterns to the result set

A pattern X is added to the result set $SetMRI$ **only if** both of the following conditions are satisfied:

- The support of X lies within the minimum and maximum thresholds:

$$\text{minSup} \leq \text{Support}(X) \leq \text{maxSup}.$$

- There is no pattern Z already in path P such that $X \subset Z$ (i.e., X is not a subset of any existing pattern in $SetMRI$).

Let SetMRI be a set of retrieved maximal rare patterns. For each item i in the rare header table, the mining process can be summarized as follows:

Case 1: Single Path X

- If only one path X is retrieved for an item i, first check whether **Condition 1** is satisfied. If it is, add this path to the result set, SetMRI. Otherwise, skip mining for this item and continue with the remaining items in the rare header table.

Case 2: Multiple Paths X and Y

- If multiple paths X and Y are retrieved, choose the longest path. Let X be the longest.

    1. **For the longest path** X: Add X to the result set $SetMRI$, similar to Case 1. This ensures that the longest maximal rare pattern is added as early as possible.
    2. **For path subsets**: If two paths $Y \subset Z$ are found, add Z (the superpattern) along with its support to the result set $SetMRI$, provided it satisfies **Condition 1**.
    3. **For identical paths**: If two paths are identical, select one and update its support by summing the supports of both paths and add it along with its support to the result set $SetMRI$, provided it satisfies **Condition 1**. The updated support for X is:

$$X.\text{count} \leftarrow X.\text{count} + Y.\text{count}.$$

    4. **For independent paths**: For any path K that is not a subset of any other path, add it to the result set $SetMRI$ if it satisfies **Condition 1**.

- This process is repeated for the remaining items in the rare header table.

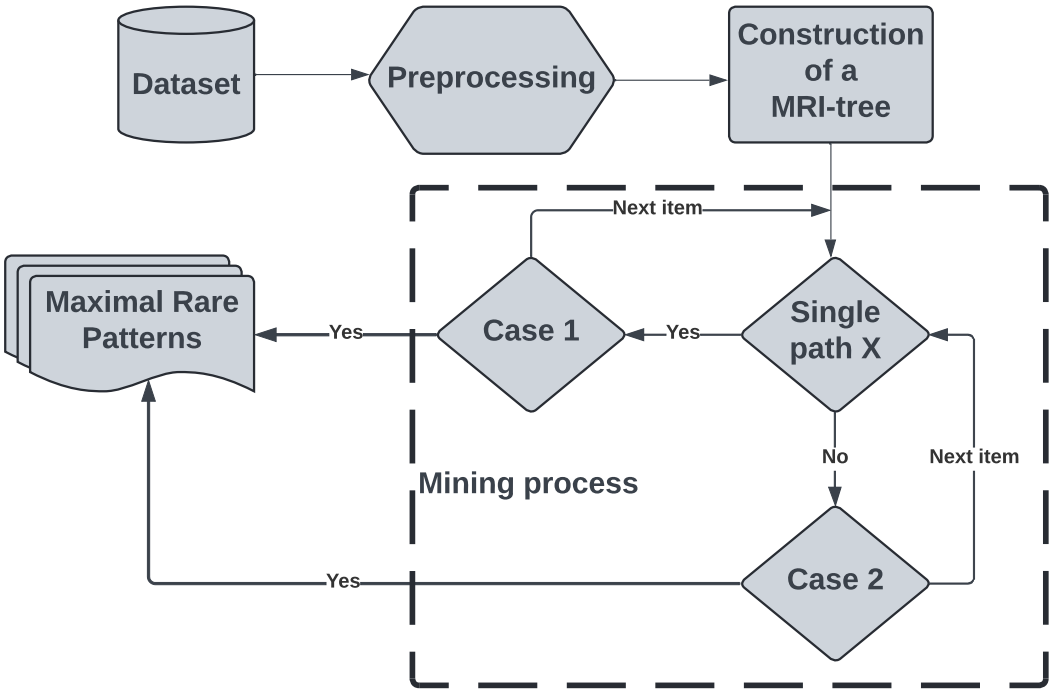


Figure 5.2: Workflow of the MaxRI algorithm

To illustrate the mining process of the MaxRI algorithm, let us consider a motivating example. Given the compact MRI tree and rare header table shown in Figure 5.1, the task is to discover the complete set of maximal rare patterns.

Algorithm 2 MaxRI: Mining Maximal Rare Patterns

Input: Dataset D, $minSup$, $maxSup$
Output: Set of maximal rare patterns $SetMRI$

procedure MaxRI(D, $minSup$, $maxSup$)
  $SetMRI \leftarrow \{\}$
  ▷ Phase 1: Preprocessing
  preprocess dataset, remove items $<minSup$, generate Tidsets
  construct MRI-tree from preprocessed dataset
  build rare header table ordered by descending support

  function Valid(X) ▷ Check pattern validity
    return support(X)$\geq minSup$ **and** support(X)$<maxSup$ **and** no super-pattern of X in $SetMRI$
  end function
  for each item i in header table **do**
    retrieve paths containing item i
    if single path X **then**
      if Valid(X) **then**
        add X to $SetMRI$
      end if
    else
      select longest path X
      if Valid(X) **then**
        add X to $SetMRI$
      end if
      for each pair (Y, Z), $Y \subset Z$ **do**
        if Valid(Z) **then**
          add Z to $SetMRI$
        end if
      end for
      for identical paths (X, X') **do**
        support(X) $\leftarrow$ support(X) + support(X')
        if Valid(X) **then**
          add X to $SetMRI$
        end if
      end for
      for remaining paths K **do**
        if Valid(K) **then**
          add K to $SetMRI$
        end if
      end for
    end if
  end for
  return $SetMRI$
end procedure

Algorithm 3 RRI Algorithm for Refining Rare Patterns

Input: Set of maximal rare patterns $SetMRI$, Tidsets, subset length k
Output: Set of refined rare patterns $SetRRI$

```
1: procedure RRI(SetMRI, Tidsets, k)
2:     SetRRI ← {}                          ▷ Initialize the set of refined rare patterns
3:     for each maximal rare pattern X in SetMRI do
4:         if length(X) ≥ k then    ▷ Check if maximal pattern length is at least k
5:             for each subset Y of X where length(Y) = k do
6:                 Tidset_Y ← intersect Tidsets of items in Y
7:                 calculate support(Y) from Tidset_Y
8:                 if support(Y) satisfies minSup and maxSup then
9:                     add Y to SetRRI
10:                end if
11:            end for
12:        end if
13:    end for
14:    return SetRRI
15: end procedure
```

The MaxRI algorithm begins with the lowest rare item, 4, from the rare header table, as in Figure 5.1. For item 4, there is a single prefix path from the root to suffix item 4, which is represented as $\{1:1, 3:1\}$. Here, the number following the colon indicates the count of the respective item on the path. The only pattern that can be generated from item 4 is $\{413:1\}$, because the support of $\{413\}$, $Sup(413) = 0.20$, is less than the maximum support threshold ($maxSup = 0.80$) and greater than or equal to the minimum support threshold ($minSup = 0.20$). Thus, the maximal rare pattern $\{413:1\}$ is added to the set of representative rare patterns, SetMRI.

Next, for rare item 1, there are two prefix paths from the root to suffix item 1: $\{5:3, 3:3, 2:4\}$ and $\{3:1\}$. The longest maximal rare pattern in this case is $\{1532\}$, with a support count equal to the relative support count of item 1 across these paths, which is 0.40. Consequently, $\{1532\}$ with support 0.40 is added to SetMRI, as it is not a subset of any maximal rare pattern already in SetMRI. On the other hand, path $\{31\}$ is discarded because $\{1532\}$, already in SetMRI, contains $\{31\}$ as a subset.

Thus, the concise set of maximal rare patterns is $\{413:0.20, 1532:0.40\}$. These patterns provide an efficient and interpretable representation of the entire set of rare patterns. The resulting rare patterns are discovered with computational efficiency and can be easily understood by domain experts.

In contrast, if we apply traditional methods, such as the RP-Tree or RPP algorithms, to the same dataset with identical $maxSup$ and $minSup$ thresholds, the complete set of rare patterns is as follows: $\{(4:0.20), (41:0.20), (43:0.20), (413:0.20), (1:0.60), (15:0.40), (13:0.40), (12:0.40), (153:0.40), (152:0.40), (132:0.40), (1532:0.40), (53:0.60), (32:0.60), (532:0.60)\}$.

Generating a complete set of rare patterns through traditional methods significantly degrades the performance, making both the discovery and analysis of these patterns computationally expensive.

5.2.4 Recovering k-length rare patterns from maximal rare patterns

This phase is classified as post-processing, in which we extract the complete set of rare patterns from the maximal rare patterns generated in the preceding step. Additionally, it allows for the selection of patterns based on desired lengths. The steps and workflow of this procedure are presented in Algorithm 3. While the generated maximal rare patterns are easily interpretable, their subset patterns support is not retained upon generation. In practice, an expert may require certain subsets of these maximal rare patterns for further analysis. To address this, we propose the Recover Rare Patterns from Maximal Rare Patterns (RRI) algorithm, which retrieves interesting subset rare patterns of length k from a set of maximal rare patterns.

The RRI algorithm operates as follows: it takes as input the Tidset of items (as shown in Table 5.3), which is generated during the preprocessing phase, the desired length k of the rare patterns, and the set of maximal rare patterns with lengths greater than or equal to k. The output of the RRI algorithm is the complete set of rare k-patterns. For each maximal rare pattern X, the RRI algorithm intersects the Tidset of the rare pattern with Tidsets of length $k-1$ of the other items in X.

To demonstrate the operation of the RRI algorithm, consider the Tidsets presented in Table 5.3. Let the desired subset rare pattern length be 3. Assume there is one maximal rare pattern, $\{1532\}$, with a length greater than 3. The Tidset of item 1 is then intersected with the Tidsets of 2-subsets from the remaining items (i.e., $\{53\}, \{52\}, \{32\}$). The result of intersecting the Tidset of 1 with that of $\{53\}$ is $\{T3, T5\}$, indicating that Items 1, 5, and 3 co-occur in transactions $T3$ and $T5$. The relative support of pattern $\{153\}$ is computed as:

$$\frac{|Tidset(153)|}{|Total\ Transactions|} = \frac{2}{5} = 0.40$$

This process is repeated for Rare Item 1 with the remaining subsets. The resulting subset of rare patterns for rare item 1 is $\{153 : 0.40, 152 : 0.40, 132 : 0.40\}$. This process terminates when no further rare patterns are produced. The final result is $\{153 : 0.40, 152 : 0.40, 132 : 0.40, 532 : 0.60\}$.

5.3 Experimental results

In this section, we present an experimental evaluation of the proposed MaxRI and RRI methods for mining representative rare patterns. The experiments were conducted on two real-world datasets: Mushroom and Accidents [FIMI Repository]. These datasets were selected due to their complexity and high density. Table 5.4 provides a summary of the key characteristics of these datasets, where #Trans denotes the number of transactions and Avg represents the average transaction length.

The experiments were performed on a system operating Windows 10 (64-bit), equipped with an Intel Core i7-7700HQ CPU at 2.80 GHz, 16 GB of RAM, and a 1 TB hard drive. This configuration was chosen to represent a typical environment for algorithm benchmarking in terms of computational resources. The performance of the proposed algorithms was evaluated against well-known algorithms, namely RP-growth and RPP [Darrab et al., 2020; Deng et al., 2012], both of which are well-established in the field of rare pattern mining for their efficiency.

All algorithms were implemented in Java, providing a uniform execution environment for performance evaluation. The comparison focused on two key metrics: runtime efficiency and memory consumption. The results, which measure both execution time and peak memory usage, demonstrate the superior performance of MaxRI and RRI, with significant improvements.

Table 5.4: Characteristics of the datasets used in the experiments

Name	#Trans	#Items	Avg Transaction Length
Mushroom	8,416	119	23
Accidents	340,183	468	33.8

5.3.1 Execution time

To evaluate the execution time of the proposed methods, MaxRI and RRI, we compared their performance with two algorithms, RPP and RP-growth, using the datasets described in Table 5.4. To limit the number of generated rare patterns, the maximum support threshold (`maxSup`) was set to 10%, and the minimum support threshold (`minSup`) was varied from 0.1% to 1% across all experiments. This range was selected to explore how different levels of rare patterns affect performance. The rare patterns considered in these experiments have support values below `maxSup` and greater than or equal to `minSup`. In each graph, the X-axis represents different `minSup` values, while the Y-axis corresponds to the execution time in seconds.

For the Mushroom dataset, Figure 5.3 presents the runtime of the proposed algorithms, MaxRI and RRI, compared with RPP and RP-growth. As shown, MaxRI and RRI consistently outperformed state-of-the-art methods across all `minSup` values. The performance improvement is significant, with the proposed methods achieving up to 1000 times faster execution than traditional approaches for mining rare patterns. This improvement is largely due to the ability of MaxRI to effectively reduce the search space by avoiding the generation of candidate patterns and the use of projected conditional trees. Instead, MaxRI directly identifies representative rare patterns (i.e., long rare patterns) without intermediate steps, leading to faster runtimes. RRI further optimizes performance by focusing on discovering specific subsets of rare patterns, leveraging Tidsets and maximal rare patterns to perform targeted intersections between items.

For the Accidents dataset, Figure 5.4 illustrates the runtime of MaxRI and RRI. Interestingly, RP-growth and RPP are not shown in the graph because they failed to complete their execution within a two-hour limit. Even after adjusting the `minSup` threshold to 1%, the competitor algorithms took approximately 1000 seconds to

complete, while generating millions of rare patterns. In contrast, MaxRI identified only 91 representative rare patterns under the same conditions, demonstrating its ability to drastically reduce the number of patterns while maintaining interpretability and relevance. This result highlights the advantage of focusing on representative rare patterns, which provide more meaningful insights and faster responses to abnormal events, while also being more manageable for post-processing.

From Figure 5.4, it is clear that the execution times for MaxRI and RRI remain low, with runtimes capped at approximately 25 and 43 seconds, respectively. This represents a significant improvement over traditional methods, further confirming the scalability and efficiency of the proposed approaches.

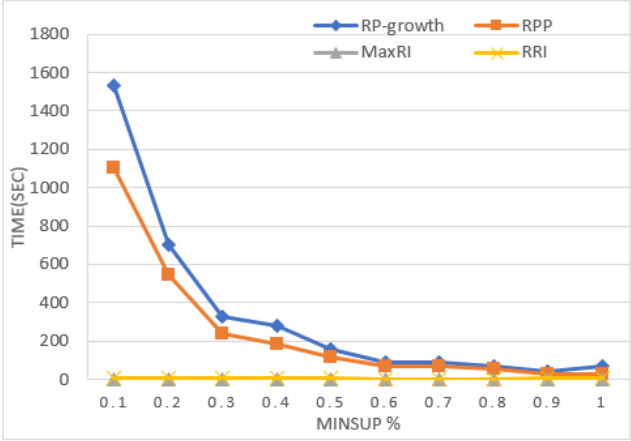


Figure 5.3: Execution time comparison on the Mushroom dataset.

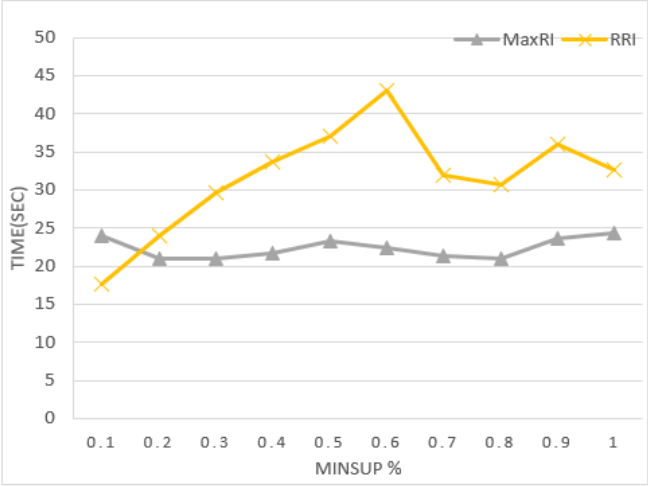


Figure 5.4: Execution time comparison the Accidents dataset.

5.3.2 Memory consumption

The memory consumption of the proposed methods, MaxRI and RRI, was evaluated and compared with algorithms, RPP and RP-growth, using the datasets presented in Table 5.4. The same experimental setup as the previous execution time experiment

was applied here, with the X-axis representing different `minSup` values and the Y-axis denoting the memory cost.

For the Mushroom dataset, Figure 5.5 presents the memory consumption of the proposed algorithms, MaxRI and RRI, in comparison with RPP and RP-growth. As shown, the proposed methods, MaxRI and RRI, consistently consume less memory than the RPP and RP-growth algorithms. The RRI algorithm consumes slightly more memory than MaxRI, primarily because RRI needs to retain the Tidset of items in memory during the intersection process. In contrast, both RP-growth and RPP are memory-intensive, because they generate a large number of rare patterns. Additionally, RP-growth requires memory for conditional trees, which further increases memory usage. The RPP algorithm is particularly memory-hungry, as it must maintain pre-order and post-order nodes in the tree, as well as RN-lists of items for intersection operations.

For the Accidents dataset, the memory consumption of the proposed methods is depicted in Figure 5.6. As noted earlier in the execution time analysis, the memory costs of the RP-growth and RPP algorithms are not displayed, as these methods were unable to complete their execution within the two-hour runtime limit. Figure 5.6 demonstrates that, similar to the Mushroom dataset, the RRI algorithm consumes more memory than MaxRI due to the necessity of storing Tidsets for intersection operations.

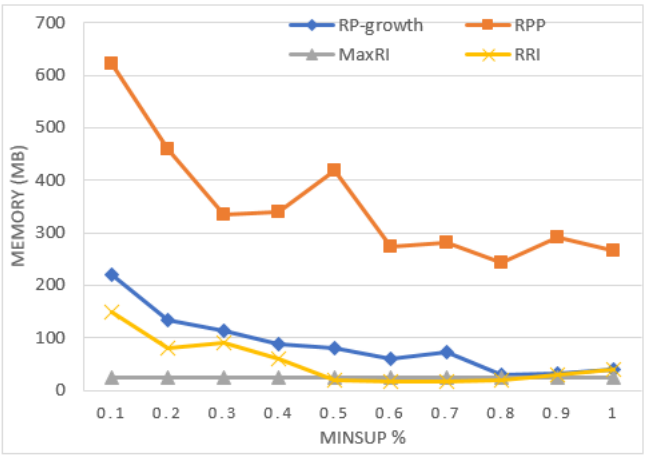


Figure 5.5: Memory cost comparison for the Mushroom dataset.

5.3.3 Discussion

The empirical results demonstrate that rare pattern mining from dense datasets yields an unmanageable number of rare patterns. The large volume of these patterns complicates downstream analysis, and the performance of traditional methods, such as RP-growth and RPP algorithms, incurs significant costs in terms of both runtime and memory usage. Our proposed methods effectively address these limitations by generating representative rare patterns more efficiently in terms of both time and memory consumption.

As shown in Figures 5.3–5.6, the runtime and memory efficiency of the proposed methods are markedly superior, especially for highly dense datasets such as Mushroom

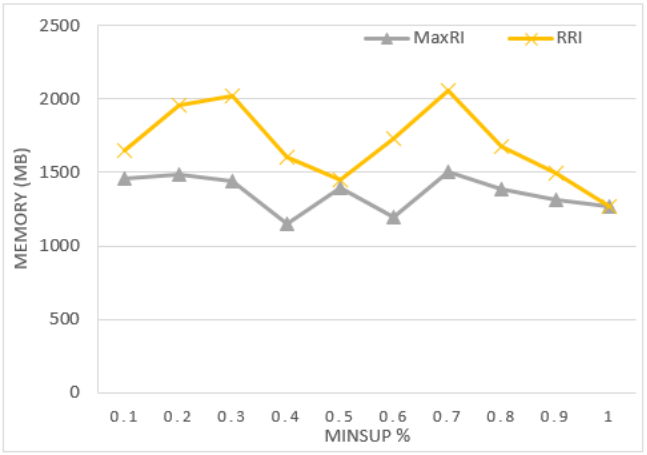


Figure 5.6: Memory cost comparison for the Accidents dataset.

and Accidents. These results indicate that our algorithms outperform state-of-the-art approaches, particularly in time and memory costs. The MaxRI algorithm excels by recovering a concise set of representative rare patterns, specifically the maximal rare patterns, which retain all the necessary information without redundancy.

In the context of rare patterns, maximal rare patterns are information-preserving because the rare item (i.e., the item with the lowest support count in each maximal rare pattern) dictates the support of any subset pattern containing that item. For instance, in the motivating example, rare item 1 leads to the generation of the maximal rare pattern $\{1532 : 0.40\}$. Any subset of this maximal rare pattern that contains rare item 1 will have a support of 0.40. This ensures that, from the resulting representative rare patterns, we can derive meaningful insights without producing a vast number of redundant and irrelevant patterns.

Furthermore, the proposed methods demonstrate a significant speed advantage, being approximately 1000 times faster than the traditional algorithms compared in this study. This dramatic improvement highlights the effectiveness of MaxRI and RRI in reducing both computational complexity and memory overhead.

5.4 Chapter summary

In this chapter, we addressed the challenges of rare pattern mining in dense datasets, specifically focusing on the limitations of traditional methods such as RP-growth and RPP. These methods often generate an excessive number of rare patterns, resulting in substantial computational overhead in both memory usage and runtime. This inefficiency arises because traditional approaches struggle to prune the search space effectively, particularly for dense datasets containing numerous infrequent patterns. To overcome these challenges, we propose the MaxRI algorithm, which efficiently retrieves representative rare patterns while reducing redundancy. Additionally, the RRI algorithm was introduced to extract meaningful subset rare patterns from representative patterns generated by MaxRI. Our experimental results demonstrated substantial improvements in runtime and memory efficiency, with performance gains of up to 1000 times faster than those achieved by traditional methods.

With the generation of representative rare patterns that are now optimized, the next logical progression is to derive meaningful association rules from these patterns. In the following chapters, we focus on generating rare association rules that are interpretable, unexpected, and actionable. These rare rules are of particular interest because of their potential to reveal valuable insights, support decision-making processes, and uncover hidden but significant relationships within data.

6. Discovering Unexpected Rules

In previous chapters, we addressed *rare pattern generation* as a foundational step in discovering association rules. This chapter shifts the focus to the generation of *rare association rules* that quantify the relationships between these patterns. Such rules are increasingly relevant in critical domains, including medicine, fraud detection, and malware analysis, where they uncover unexpected yet valuable insights often missed by more common patterns. For these insights to be actionable, the generated rules must not only be rare but also exhibit high utility, requiring both high user confidence and models that are interpretable with minimal tuning complexity.

A key challenge in rare pattern mining is *data imbalance*, as seen in medical datasets where the proportion of ill patients is significantly smaller than that of healthy individuals. This imbalance complicates the discovery of rare patterns. Although recent clustering models have attempted to address this issue, their performance often suffers in terms of time efficiency and accuracy.

In this chapter, following the identification of rare patterns, we address the next critical step: generating association rules that are both unexpected and interesting. We propose an efficient model to assess the quality of these rare rules. By applying our model to three real-world medical datasets, we demonstrate that it outperforms existing models in terms of both speed and precision, yielding more accurate results. This ensures that users are presented with only the most relevant and compact rules, thereby reducing the effort required for post-processing and making insights more actionable.

6.1 Introduction

Pattern mining is a well-established technique for identifying patterns that occur together within a dataset. These patterns often form association rules, which are typically represented by $X \rightarrow Y$, where X and Y are sets of patterns. For instance, in market basket analysis, the rule *milk* $\rightarrow$ *bread* suggests that customers who purchase milk are also likely to buy bread. This process, known as Association Rule Mining

(ARM), has numerous applications across domains such as retail, medical diagnosis, and web usage mining [Agrawal and Srikant, 1994; Ahmed and Barkat Ullah, 2018].

A significant advantage of ARM is its interpretability, which provides clear and actionable insights for decision-making. However, a major limitation is the large number of rules generated. For a pattern with d items, the number of possible rules can reach $2^d - 2$ [Zaki, 2000a], which leads to a flood of redundant, noisy, and irrelevant rules. This challenge is particularly pronounced for rare patterns, which, despite their infrequency, can still generate numerous rules, many of which are irrelevant or noisy. Consequently, identifying the most unexpected and insightful rules from a large set is challenging.

Rare association rules possess considerable value because they uncover hidden insights that are often overlooked when focusing on frequent patterns. Despite their rarity, these rules can provide critical insights into fields such as fraud detection, medical research, and network security. However, the primary challenge lies in distinguishing meaningful rules from irrelevant or redundant ones. Effective filtering of these less useful rules is crucial for retaining only those that provide novel and actionable insights, ensuring that the results are both valuable and applicable for real-world applications.

Clustering-based techniques have been introduced to address the issue of rule explosion in ARM. These methods group similar rules to reduce the overall rule set and improve interpretability [Gupta et al., 1999; Lent et al., 1997; Toivonen et al., 1995]. One prominent approach, DBSCAN, clusters rules based on their similarity, thereby facilitating the interpretation of results [Bui-Thi et al., 2020]. However, DBSCAN and similar techniques primarily focus on frequent patterns and often neglect rare association rules. Rare-pattern mining, also known as unexpected rule mining, seeks to bridge this gap by identifying rare but valuable association rules [Borah and Nath, 2019; Darrab et al., 2021b]. Despite advances in this field, current state-of-the-art methods such as DBSCAN have several limitations. These shortcomings are high computational costs, sensitivity to parameter settings (e.g., ϵ and *minPts* in DBSCAN), and reliance on single-class metrics, such as the F1-score, which can lead to misleading results in imbalanced datasets [Jeni et al., 2013].

Consequently, more efficient and interpretable models for rule generation are required, particularly when addressing rare patterns. These models should address both computational complexity and interpretability to ensure that the insights generated are actionable and valuable in real-world applications.

To address these limitations, we propose a novel model called *OPECUR* (OPTICS-based Clustering of ECLAT-Generated Unexpected Rules), which is designed to enhance the discovery of rare and unexpected association rules. Our primary contributions are as follows:

- We utilize the FP-growth and ECLAT algorithms to efficiently generate a comprehensive set of association rules, mitigating the time and memory overhead associated with Apriori-based methods.
- We implement the OPTICS algorithm for clustering, a density-based method that minimizes the need for extensive parameter tuning. Unlike DBSCAN,

OPTICS can identify clusters of varying densities, facilitating the discovery of more accurate and unexpected association rules.

- We evaluate the quality of the generated rules using three machine learning classifiers: Support Vector Machines (SVM), Random Forest (RF), and a Neural Network-based Multi-Layer Perceptron (MLP).
- Our experimental results demonstrate that *OPECUR* consistently outperforms state-of-the-art DBSCAN-based models in terms of F1-score and Area Under the Curve (AUC). Furthermore, *OPECUR* generates unexpected rules more efficiently by identifying a larger number of interesting and novel rules.

The structure of this chapter is organized as follows: Section 6.2 introduces the proposed model in detail, Section 6.3 presents the experimental results, and Section 6.4 concludes the chapter with a summary of key findings.

6.2 Proposed method: OPECUR Model

In this chapter, we introduce the Optimized Clustering for Unexpected Rare Rules (OPECUR) model, an advanced clustering framework for effectively discovering unexpected and valuable rare association rules. The OPECUR model workflow illustrated in Figure 6.1 outlines an approach that surpasses current state-of-the-art methods in terms of both accuracy and efficiency.

The workflow begins by generating a comprehensive set of association rules from the dataset with a low minimum support threshold to ensure that even rare but insightful patterns are captured. Once generated, the rules are transformed into feature vectors based on item correlations within the dataset. A contradiction check function is then applied to the noise points identified in the clustering phase, which is a critical step for isolating the final set of unexpected association rules, setting OPECUR apart from comparable methods.

The OPECUR model comprises two primary phases: association rule generation and clustering, as shown in Figure 6.1. In the first phase, efficient algorithms such as FP-growth and ECLAT generate a complete set of patterns, optimizing time and computational resources. The second phase employs the OPTICS clustering technique [Ankerst et al., 1999] to group rules and detect outliers as potentially unexpected patterns. This technique effectively distinguishes dense clusters and isolates unexpected rule patterns from outliers, thereby enhancing the discovery of rare and valuable rules.

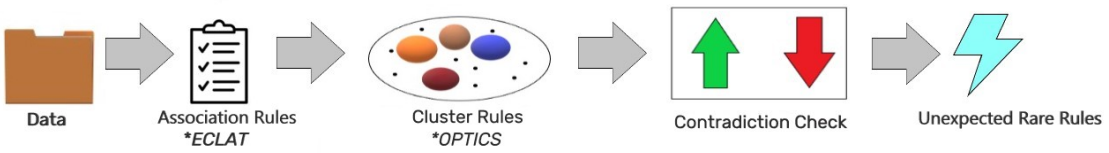


Figure 6.1: OPECUR workflow for generating unexpected rare rules.

6.2.1 Generating association rules

The state-of-the-art approach [Bui-Thi et al., 2020] utilizes the Apriori algorithm to generate association rules. However, Apriori's requirement for generating numerous candidate sets and performing multiple dataset passes results in significant computational inefficiencies. Our proposed model, OPECUR, overcomes these limitations by employing more efficient and scalable FP-growth and ECLAT algorithms.

FP-growth [Han et al., 2004] constructs a compact tree structure, FP-Tree, which captures essential information for mining without the overhead of candidate generation and testing. This algorithm requires at most two scans of the dataset: the initial scan constructs the FP-Tree by incrementally adding transactions and filtering them to exclude irrelevant items. Once the tree is constructed, the mining process utilizes a divide-and-conquer approach to enhance the efficiency.

The ECLAT algorithm [Zaki et al., 1997] employs a depth-first search strategy to generate a comprehensive set of patterns. By operating in a vertical data format, ECLAT avoids multiple scans and calculates pattern support through set intersections, significantly improving computational performance.

The use of FP-growth and ECLAT is advantageous over apriori-based methods, which are commonly used in comparative models. Furthermore, to ensure the discovery of rare patterns, we set a low support threshold, an approach that is viable given the manageable size of the dataset.

6.2.2 Clustering-based approach

In our proposed model, the second phase focuses on a clustering-based approach for identifying interesting rare rules. In this phase, similar association rules are grouped using the density-based clustering algorithm *OPTICS* [Ankerst et al., 1999]. The OPECUR algorithm sorts data points based on the reachability distance, defined as the maximum distance from a point p to another point and p's core distance. Here, the core distance is the distance to the n^{th} nearest neighbor, where n is a user-defined parameter *minPts*. This method ensures that points within the reachability distance are clustered.

One advantage of using the reachability distance is the reduced need to pre-define an *eps* value, as this is automatically derived from *minPts*. This flexibility enables OPTICS to effectively handle datasets with varying cluster densities and identify nested clusters. As shown in Figure 6.2 and 6.3, the OPECUR model generates a greater number of clusters than DBSCAN. Because OPTICS dynamically adjusts *eps*, it can discover clusters that may be missed by DBSCAN, which relies on a fixed *eps* parameter. In both figures, the red points represent noise points that are not part of any cluster.

To extract unexpected rules, we focused on analyzing these noise points because they often contain hidden and interesting patterns. To determine whether a noise rule is unexpected, we use a process similar to that in [Bui-Thi et al., 2020] with an important modification. In our approach, the OPECUR algorithm first generates clusters, and the contradiction check function is applied only afterward to the noise points. This contrasts with the method in [Bui-Thi et al., 2020], in which the contradiction check

is embedded within the clustering process. By separating these steps, our method reduces the parameter tuning complexity and prevents additional constraints from influencing the clustering outcome. Consequently, the OPECUR model improves the detection of unexpected rules while minimizing the risk of overlooking important patterns.

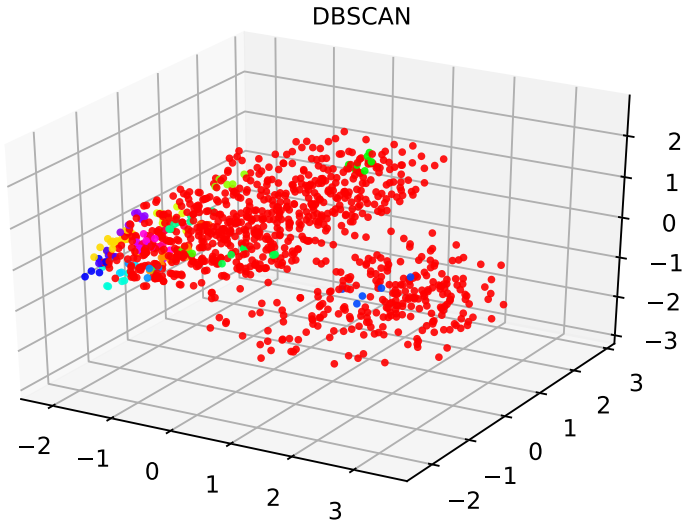


Figure 6.2: Clusters in the Breast Cancer Dataset using DBSCAN

If all these conditions are satisfied, the noise rule $X \rightarrow Y$ is identified as an unexpected rule that contradicts the clustered rule $X' \rightarrow Y'$.

1. $Y \neq Y'$
2. High cosine similarity between X and X'
3. High confidence for both rules
4. The rule $X \rightarrow Y$ is identified as a noise point

If all these conditions are met, the noise rule $X \rightarrow Y$ is flagged as an unexpected rule that contradicts the clustered rule $X' \rightarrow Y'$.

6.3 Experimental evaluation

In this section, the performance of our proposed model is compared with that of the state-of-the-art model introduced in [Bui-Thi et al., 2020]. In the following, we introduce the evaluation setup and then explain the results of our experiments.

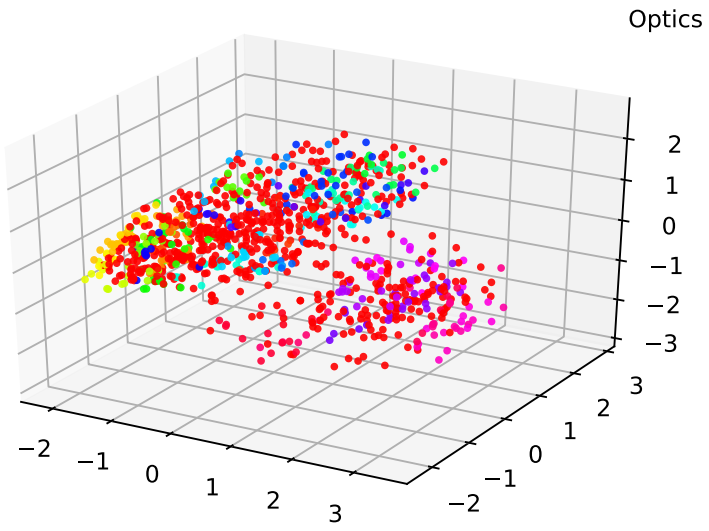


Figure 6.3: Clusters in the Breast Cancer Dataset using OPECUR

6.3.1 Experimental setup

To evaluate our proposed model, OPECUR, we compared it with the state-of-the-art model introduced in [Bui-Thi et al., 2020], using three real-world medical datasets: Breast Cancer, Cleveland Heart Disease, and Hepatitis, all obtained from the UCI repository [Kelly et al.]. Table 6.1 provides an overview of the main characteristics of datasets. The Cleveland Heart Disease dataset has five target classes for prediction: 0 (absence) and 1, 2, 3, 4 (presence) [1]. Both the Hepatitis and Cleveland datasets contain a mix of real-valued and categorical attributes, whereas the Breast Cancer and Hepatitis datasets exhibit imbalanced distributions, with minority classes comprising 29% and 25% of instances, respectively. This skewed distribution poses a significant challenge for extracting representative and meaningful rare rules from these datasets.

To assess the effectiveness of our approach, we conducted the following experiments:

- **Scalability**: In the first experiment, we compare the time required by OPECUR and the DBSCAN-based model to generate the complete set of rules, providing insights into scalability.
- **Clustering Quality**: In the second experiment, we evaluate the quality of clustered association rules produced by OPECUR in comparison to those from the DBSCAN-based model.
- **Rule Quality for Classification**: Lastly, we examine the quality of rare rules generated by our clustering model, OPECUR, against those produced by DBSCAN_based model. We trained a classifier using these rules to assess their utility for decision-making, specifically in identifying individuals as healthy or ill.

[1] For simplicity, we convert this to a binary classification: 0 (absence) and 1 (presence).

All experiments were conducted on Google Colab with 12GB RAM, allowing us to evaluate the scalability of our models in a constrained environment. The experimental results are presented in the following subsections.

Table 6.1: Dataset Details

Dataset	Instances	Classes	Attributes	% Minority Class
Breast Cancer	286	2	9	0.29
Hepatitis	155	2	20	0.25
Cleveland	303	2	14	0.44

*All datasets are obtained from the UCI Machine Learning Repository.

6.3.2 Experiment 1: execution time comparison

This experiment evaluated the execution time of our proposed OPECUR model in comparison to a state-of-the-art DBSCAN-based model for generating a complete set of association rules. While the DBSCAN-based model utilizes the Apriori algorithm, we employ FP-growth and ECLAT algorithms to generate the same rule set. For each experiment, the minimum support threshold (*minSup*) was varied from 0.01 to 0.4 to ensure consistency in the output rules.

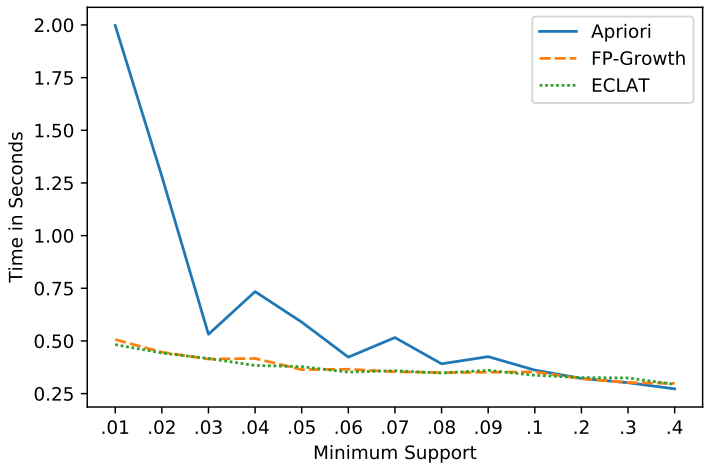
Figures 6.4a to 6.4c illustrate the performance of each algorithm across all datasets. As shown in the graphs, FP-growth and ECLAT significantly reduce execution time compared to Apriori. In particular, even at low *minSup* values, FP-growth and ECLAT efficiently retrieve patterns, underscoring their effectiveness in generating comprehensive rule sets that include rare and meaningful patterns.

Based on these results, our algorithm adopts ECLAT for rule generation because it demonstrates superior performance when mining patterns at low support thresholds, which is essential to effectively uncovering rare patterns.

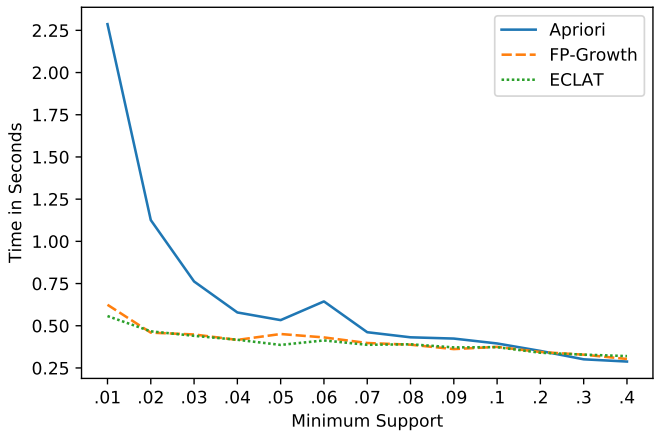
6.3.3 Experiment 2: clustering process comparison

In this experiment, we compared the clustering results of our proposed model, OPECUR, with those of the state-of-the-art DBSCAN-based model. To ensure a fair comparison, we set the parameters consistent with those used in the DBSCAN-based model [Bui-Thi et al., 2020]: specifically, *minPts* is set to 10, with *delta1* and *delta2* (contradiction check parameters) set to 1 and -1, respectively. Table 6.2 presents the clustering results for OPECUR and DBSCAN. The results indicate that OPECUR generates more unexpected rare rules than DBSCAN, and successfully identifies interesting rare patterns that DBSCAN misses. This improvement is due to OPECUR's automatic calculation of the *minEps* parameter, which enables it to detect more clusters and yield fewer noise rules than DBSCAN. These findings suggest that OPECUR is more effective at identifying unexpected rare rules that may be overlooked by DBSCAN; however, evaluating the meaningfulness of these rules still needs more investigation.

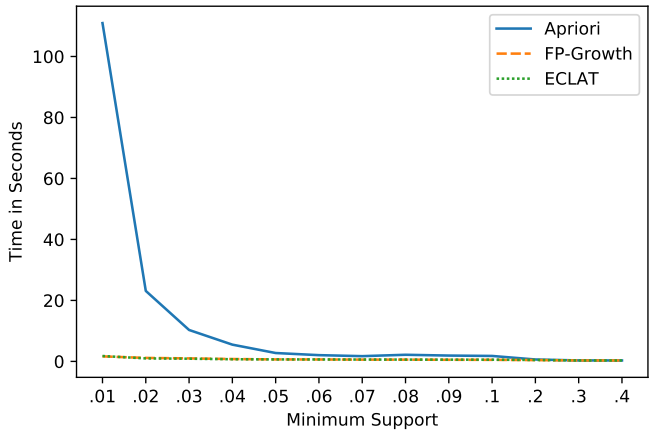
To evaluate the quality of unexpected rare rules generated by OPECUR, we adopted an evaluation strategy similar to that in [Bui-Thi et al., 2020], which measures the impact of rare rules using machine learning classifiers. Their approach assumes that a rule is meaningful if it enhances the decision boundary of a classifier and, thus, its performance. We modified this approach as follows:



(a) Breast Cancer Dataset



(b) Cleveland Heart Disease Dataset



(c) Hepatitis Dataset

Figure 6.4: Runtime Performance Comparison Across Datasets

- **Cross-Validation**: Instead of using an independent hold-out method, we apply 3-fold cross-validation on the datasets. For each iteration, rules from two folds are used for training, while the remaining fold serves as the test set. This process is repeated until each fold has served as an independent test set, and the average scores are reported across all folds to provide a robust estimate of the generalization error.
- **Classification Task**: We employ three classifiers, Random Forest (RF), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP), implemented in the Sklearn library [Pedregosa et al., 2011] with default settings. The models are evaluated using the F1 and AUC metrics. The F1 score balances precision and recall, whereas the Area Under the Curve (AUC) offers insights into the balance between true-positive and false-positive rates.

Table 6.2: Comparison of Clustering Algorithms

Algorithm	Noise	Rare	Clusters
Breast Cancer Dataset			
DBSCAN	748	3	20
OPECUR	386	23	54
Cleveland Dataset			
DBSCAN	17,269	415	4
OPECUR	15,291	1,650	17
Hepatitis Dataset			
DBSCAN	11,925	0	3
OPECUR	7,289	4	10

* All values are based on minpoints = 10 for all datasets. This table describes the number of clusters, noise points, and rare rules generated by each algorithm for each dataset.

Figures 6.5 and 6.6 show the performance of our proposed model, OPECUR, compared with the DBSCAN-based model across all three datasets. For both evaluation metrics, our model, OPECUR, demonstrated consistently better performance than the DBSCAN-based model [Bui-Thi et al., 2020] on all datasets. This improvement stems from OPECUR's ability to identify a greater number of clusters while generating fewer noise rules that satisfy the contradiction check. This is a clear advantage over the DBSCAN model, which can only identify a limited subset of actual clusters. Consequently, rules that should ideally be classified as noise are still assigned to a cluster by DBSCAN and are thus considered frequent rules. Consequently, our model generates a higher number of unexpected rare rules and achieves superior performance in terms of the F1 and AUC scores.

Notably, the OPECUR model achieved its highest performance on the hepatitis dataset. This is attributed to the large number of attributes in the dataset, which enhances the detection of correlations within these attributes. Furthermore, an in-depth analysis of the rules and clusters reveals that the DBSCAN model struggles with this dataset, as it fails to identify nested clusters within the hepatitis data.

6.3.4 Experiment 3: evaluation of unexpected rules

In this experiment, we evaluated the unexpected rules generated by our model using two key criteria. First, we apply the contradiction check approach used in the comparative model [Bui-Thi et al., 2020] with identical parameter settings. This

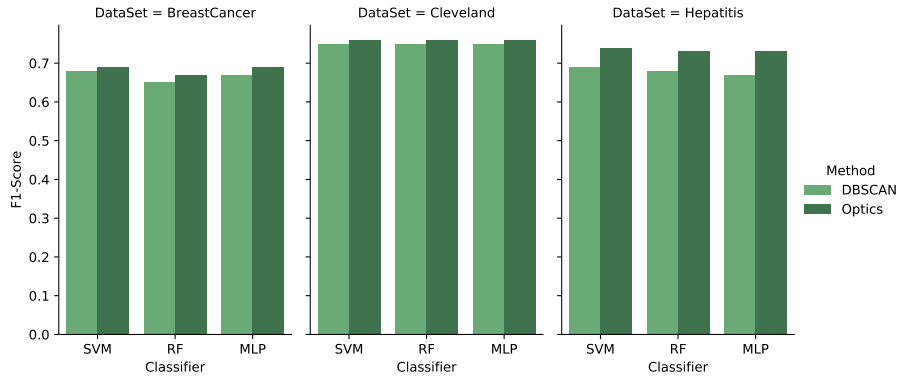


Figure 6.5: F1 Score

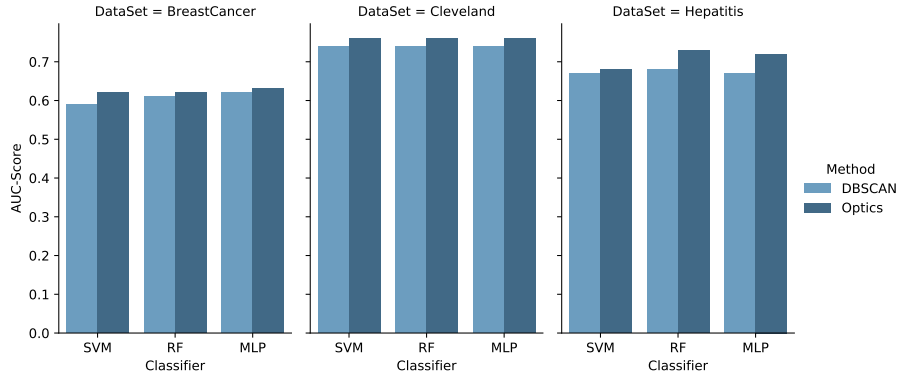


Figure 6.6: AUC Score

F1-Score and AUC-Score of SVM, RF, and MLP classifiers using rules derived from DBSCAN- and Optics-based models. "DBSCAN" and "Optics" refer to models built on rules extracted from the outputs of the respective clustering models, not the original clustering algorithms.

evaluation demonstrates that our OPECUR model produces more insightful and unexpected association rules. For example, the rule **'age=50-59', 'breast=left', 'deg-malig=3', 'irradiat=no', 'menopause=ge40', 'tumor size=30-34' → class=yes** is identified by OPECUR as unexpected. This rule contradicts the following two rules:

- **'age=50-59', 'breast=left', 'irradiat=no', 'menopause=ge40' → class=no**
- **'breast=left', 'irradiat=no', 'menopause=ge40' → class=no**

Second, we compared the unexpected rules generated by our model with those identified by the rare-pre-post-order (RPP) algorithm [Darrab et al., 2020], which was recently introduced to discover the complete set of rare rules. Our results indicate that the unexpected association rules generated by OPECUR are a meaningful subset of the rare rules found by RPP, providing a more concise and targeted set of insights.

For instance, RPP generates the following rare rule: **'tumor size=30-34', 'inv-nodes=3-5', 'node-caps=no', 'menopause=ge40', 'deg-malig=3', 'irradiat=no' → class=yes**. This rule indicates an increased risk of cancer recurrence among patients characterized by a tumor size of 30–34 mm, involvement of 3–5 lymph nodes, absence of node caps, malignancy grade 3, postmenopausal status, and no prior radiation

therapy. Similarly, OPECUR identifies the rule **'menopause=ge40', 'inv-nodes=3-5', 'node-caps=no', 'irradiat=no'** → **class=yes** as rare. These findings highlight that our proposed model captures insights similar to RPP while offering a more concise rule set.

Hence, the unexpected association rules generated by OPECUR are informative and meaningful. They represent a refined subset of rare patterns previously identified by the RPP algorithm [Darrab et al., 2020], providing clearer insight into the underlying structure of the data.

6.4 Chapter summary

The discovery of unexpected (rare) rules has gained increasing attention because of its potential to uncover valuable hidden knowledge in domains such as medical diagnosis, fault detection, and fraud prevention. However, existing state-of-the-art models often suffer from degraded performance and fail to capture a complete set of unexpected association rules. To address these challenges, we developed a clustering-based model, OPECUR, that efficiently identifies unexpected association rules from real-world datasets. Our model surpasses the many limitations of previous approaches by generating a comprehensive set of unexpected rules and autonomously setting the parameters during the clustering process. This functionality enables the model to identify a wider range of clusters, yielding more unexpected rules that enhance the decision boundaries of machine learning classifiers.

The unexpected rules generated by OPECUR were evaluated based on criteria such as contradiction checks, classifier performance, and comparisons with the RPP algorithm. The experimental results indicate that our model is scalable and capable of generating reliable and insightful unexpected rules. However, although OPECUR demonstrates strong capabilities in discovering a diverse set of interesting rules, the clustering-based approach may still limit its ability to identify a truly complete set of unexpected rules. This potential limitation will be the focus of the following chapter, in which we explore strategies to mitigate this effect and further enhance rule discovery.

7. Exploring Meaningful and Unexpected Patterns

In the previous chapter, we introduced methods for discovering interesting rare rules using clustering-based techniques. Although effective, these clustering methods have inherent limitations that often result in an incomplete set of unexpected rules. This incompleteness restricts the discovery of potentially valuable insights within the data, as certain rare patterns that deviate significantly from the norm remain undetected.

In this chapter, we propose a novel model designed to address these limitations by capturing a comprehensive set of interesting rare rules. The core concept of our approach is grounded in utilizing frequent patterns as a baseline or set of "beliefs." Frequent patterns represent common co-occurrences within the dataset, encapsulating known and predictable behaviors. By treating these frequent patterns as normative behaviors, we aim to uncover rare rules that diverge from this baseline, highlighting unexpected and potentially insightful deviations. This approach enables the identification of a more complete and non-redundant set of rare rules, thereby significantly enhancing the depth of insight derived from pattern mining. The content of this chapter is based on our paper published in [Darrab et al., 2022b].

7.1 Introduction

In recent years, deep learning (DL) models have achieved impressive success in pattern recognition and predictive analytics due to their high accuracy and scalability with large datasets. Despite these strengths, DL models frequently operate as "black boxes," lacking the interpretability needed for critical applications involving rare or infrequent phenomena. In contrast, data mining techniques, particularly association rule mining (ARM), offer a more transparent approach by revealing significant relationships between patterns. This interpretability makes ARM particularly valuable in domains where explainability and actionable insights are essential, such as fraud detection, disease diagnosis [Altaf et al., 2017], and road traffic accident prediction [Joshi et al., 2020].

To address the need for comprehensive rule discovery, a range of methods have been developed within ARM [Aljehani and Alotaibi, 2024; Darrab and Ergenç, 2016; Kamepalli and Bandaru, 2019]. Algorithms such as Apriori and FP-growth [Singh et al., 2014], along with numerous extensions, are widely used to generate association rules. However, although these methods effectively capture a complete set of patterns, they frequently produce an excessive number of rules. This overabundance not only presents scalability challenges but also complicates analysis, particularly when the objective is to identify rare and unexpected rules. The resulting surplus of rules can obscure the critical rare patterns that are most valuable for decision makers, such as atypical transaction behaviors in fraud detection or indicators of rare diseases in medical diagnostics.

To overcome this challenge, clustering-based methods have been introduced to reduce the number of generated rules by grouping similar rules into clusters [Dahbi et al., 2016; Lent et al., 1997; Toivonen et al., 1995]. For example, the approach in [Toivonen et al., 1995] prunes undesirable association rules while grouping others into clusters [Lent et al., 1997]. Although clustering-based models improve efficiency by focusing on frequent events, they may overlook less frequent but potentially more valuable patterns. For real-world applications, especially in fields such as fraud detection and medical diagnosis, capturing these infrequent patterns, such as unusual spending behaviors or rare disease patterns, can provide crucial insights.

Efforts to address the need to mine unexpected rare rules have led to the development of models such as DBSCAN and OPECUR [Bui-Thi et al., 2020; Darrab et al., 2022a], which employ clustering to capture rare events. These models operate by generating a comprehensive set of patterns with a low support threshold and then clustering similar rules. However, despite their utility, these approaches have several limitations. First, their reliance on a low support threshold results in a performance bottleneck because generating the full set of patterns (including rare ones) is computationally intensive. Second, rare patterns are often clustered with frequent patterns based on distance similarity, which can result in missed rare rules that are critical for specific applications. Moreover, these models are highly sensitive to hyperparameter settings; improper tuning of parameters, such as ϵ and $minPts$ can lead to suboptimal results, further complicating the detection of unexpected patterns.

In real-world applications, accurately identifying the most promising rare patterns, often representing anomalous or unexpected behaviors, is crucial; however, current clustering-based methods are unable to achieve this goal comprehensively. These limitations underscore the need for an efficient model that can capture the entire set of interesting rare rules without excessive redundancy. To address this gap, we propose the Unexpected Closed Rare Pattern Miner (UCRP-miner), a novel model designed to effectively retrieve unexpected patterns. The UCRP-miner considers frequent patterns as a foundation or set of "beliefs" (representing expected patterns), from which it identifies rare patterns that deviate from these norms, offering a structured way to extract insightful deviations. This approach ensures that both frequent and rare patterns are captured while minimizing redundancy and enhancing interpretability.

The principal contributions of this chapter are as follows:

- To address the challenges of redundancy and performance degradation, we propose a method that recovers a compact representation of patterns, specifically through closed rare patterns.
- We use frequent patterns as a baseline ”belief” set to identify unexpected rules. By comparing closed rare patterns with frequent ones, we flag rare patterns that deviate from these baseline ”beliefs” as potentially insightful, uncovering novel findings.
- Experimental results demonstrate that UCRP-miner significantly outperforms existing methods, delivering precise and insightful results with improved efficiency.

The remainder of this chapter is organized as follows. Section 7.2 describes the proposed model. Section 7.3 provides an analysis of the experimental results. Finally, Section 7.4 concludes the chapter.

7.2 Proposed method: UCRP-Miner

In this chapter, we introduce UCRP-Miner, a novel approach designed to uncover patterns that yield new, unknown, and unexpected insights within datasets. As illustrated in Figure 7.1, UCRP-Miner first preprocesses the data and prepares it for the mining process to ensure efficiency and accuracy. Following this, a complete set of closed (both frequent and rare) patterns is generated. Among these, rare patterns exhibiting high similarity to common (frequent) patterns based on a predefined threshold are identified as candidates for interesting patterns.

The UCRP-Miner framework is structured into two main phases: identifying closed patterns and extracting interesting patterns. In the first phase, the model generates closed patterns that serve as a compact representation of the dataset, ensuring minimal information loss. In the second phase, UCRP-Miner identifies interesting patterns by leveraging frequent closed patterns as a baseline (or set of beliefs). This comparison isolates rare patterns that deviate significantly from expected behaviors, thereby capturing meaningful and actionable insights that are likely to be of high interest to users.

Through these phases, UCRP-Miner efficiently retrieves rare patterns that are unexpected relative to frequent patterns, thereby enhancing interpretability and focusing on the most insightful deviations. The following subsections detail the steps involved in extracting these valuable patterns from datasets.

7.2.1 Preprocessing phase

In this phase, we removed useless transactions from a given dataset. A transaction is considered in the mining process if none of the following conditions occur.

- A transaction includes only one item.
- A transaction is a duplicate of another.

Thus, we consider only meaningful transactions in the preprocessing phase, and those that do not generate knowledge while mining interesting patterns are discarded.

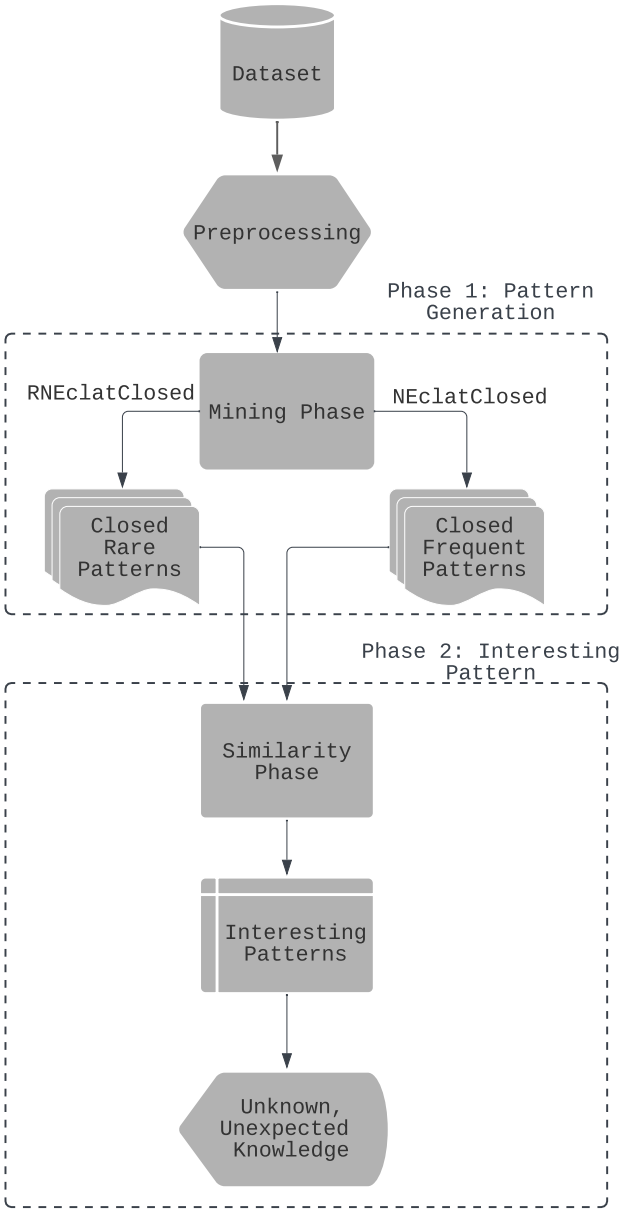


Figure 7.1: UCRP-Miner workflow for mining meaningful patterns

7.2.2 Generation of patterns

A significant challenge in frequent pattern mining, known as pattern explosion, arises from the overwhelming number of generated patterns, which often exceed the dataset size by orders of magnitude. Many of these patterns are redundant and provide repetitive information. Various methods, such as closed or maximal frequent patterns, have been developed to reduce the pattern count. Despite these attempts, frequent patterns remain too extensive for expert analysis, emphasizing patterns that are predictable and commonly occurring, often overlooking less common, yet insightful patterns. Thus, discovering rare, unexpected patterns that diverge from the general trends (frequent ones) within the data is valuable, as they may reveal previously unexplored phenomena.

Existing state-of-the-art models, such as those presented by Bui-Thi et al. [2020]; Darrab et al. [2022a] apply clustering techniques to generate unexpected rules. They utilized traditional pattern-mining methods (e.g., Apriori, FP-growth, and Eclat) at lowered support thresholds to capture rare patterns, generating a complete set of rules before clustering them. Subsequently, noisy rules are filtered out by comparing the clustered rules, retaining those that satisfy the constraints (e.g., contradiction checking and similarity) as unexpected rules. However, these methods have certain limitations. First, they generate an extensive number of patterns and rules, often orders of magnitude greater than the dataset size, leading to high time and memory consumption, particularly with large datasets. Second, clustering-based methods may inadvertently group rare patterns with high similarities, limiting the discovery of truly unexpected patterns.

Our proposed frequent-based model, UCRP-Miner, addresses these issues by implementing faster, more scalable, and efficient techniques to generate a concise representation of both frequent and rare patterns, thereby reducing redundancy in the pattern set.

UCRP-Miner avoids generating redundant patterns by creating a compact representation rather than producing the entire set of patterns. Representing frequent patterns in a condensed form, such as closed patterns, significantly reduces the number of patterns needing extraction. Closed patterns provide a lossless and compact representation of all complete patterns, making them ideal for our purposes. Therefore, in UCRP-Miner, we focus on generating closed patterns.

Figure 7.1 illustrates the UCRP-Miner workflow, where the method generates two primary types of patterns: frequent and rare closed patterns. To create frequent closed patterns, we used an Eclat-based approach called NEclatClosed [Aryabarzan and Minaei-Bidgoli, 2021], which leverages a depth-first search strategy. This method operates in a vertical data format, which eliminates the need for multiple dataset scans using intersections to determine pattern support counts.

For mining rare closed patterns, we introduce the RNEclatClosed method, an adaptation of the NEclatClosed algorithm with the following key distinctions: 1) in contrast to NEclatClosed, RNEclatClosed orders items by decreasing support; 2) it employs two constraints, MaxSup and MinSup, which serve as user-defined thresholds to delineate frequent and rare patterns; and 3) items with support values below MinSup are excluded to ensure that they do not influence the generation of rare closed patterns. Consequently, UCRP-Miner utilizes both frequent and rare patterns to produce insightful and interesting patterns, as detailed in the subsequent section.

7.2.3 Interesting patterns

In the second stage of our UCRP-Miner model, we identify interesting patterns that can be used to generate meaningful and unexpected patterns. Figure 7.1 illustrates how this stage builds on the patterns generated in the first phase to produce insightful patterns. This model classifies patterns into frequent (common) and infrequent (rare) closed patterns. Frequent patterns identified through NEclatClosed represent typical phenomena and generally reflect known and expected knowledge. These

frequent patterns serve as a foundational set of beliefs against which we identify deviations—rare patterns that potentially lead to novel insights.

In UCRP-Miner, interesting patterns are those that deviate from established norms, offering knowledge that was not previously considered. To discover these patterns, we compared the foundational beliefs (frequent closed patterns) with the set of closed rare patterns generated by RNEclatClosed. Each rare pattern is evaluated against the frequent patterns, disregarding noise, to identify deviations from beliefs as follows.

Consider two patterns, P and P', where P is a frequent closed pattern and P' is a rare closed pattern. A rare pattern P' is deemed interesting if it satisfies the following conditions:

- P' is a closed rare pattern such that $MinSup \leq Sup(P') < MaxSup$.
- P is a closed frequent pattern where $Sup(P) \geq MaxSup$.
- P' and P have a high similarity, meeting a user-defined similarity threshold, such as a cosine measure.

The identified interesting patterns are then used to generate unexpected rules, providing valuable insights beyond the expected patterns found in the dataset.

7.3 Experimental evaluation

The performance of our proposed model, UCRP-miner, is compared in this section with the state-of-the-art models introduced in [Bui-Thi et al., 2020; Darrab et al., 2022a]. Our next step is to describe the evaluation setup and discuss the results of the experiments that we conducted.

7.3.1 Datasets and experimental setup

To evaluate our proposed model, UCRP-miner, we compared it with the state-of-the-art models, DBSCAN and OPECUR, introduced in [Bui-Thi et al., 2020; Darrab et al., 2022a] by utilizing four real-life datasets. The datasets are BMSWebView2, Kddcup99, Mushrooms, and Accidents from the UCI repository [Kelly et al.]. The datasets, BMSWebView2, Kddcup99, Mushrooms, and accidents represent click-stream data, a wide variety of intrusions simulated in a military network environment, data of different physical activities, and traffic accident data, respectively. In Table 7.1, #Trans, AT,#Items, MaxSup, and MinSup indicate the number of transactions, average transactions, number of unique items, user-defined minimum support threshold for mining frequent patterns, and user-defined minimum threshold for mining rare patterns, respectively.

To evaluate the efficiency and effectiveness of the proposed method, three experiments were conducted. The first is to compare closed frequent patterns with infrequent ones. The second experiment compared the interesting patterns produced by the UCRP-miner model with those produced by clustering-based models. NEclatClosed's source code is available on Java [Fournier-Viger et al., 2016]. The experiments were all run on Google Colab with a limited amount of RAM (12GB) to evaluate the scalability and ease of use of the models. The results are discussed in the following subsections.

Table 7.1: Dataset Details

Dataset	#Trans	AT	#Items	MinSup	MaxSup
BMSWebView2	77,512	4.62	3,340	0.0005	0.1
Kddcup99	1,000,000	16	135	0.2	0.4
Mushrooms	8,564	23	119	0.01	0.1
Accidents	340,183	33	468	0.2	0.4

All datasets were taken from the UCI Machine Learning Repository.

7.3.2 Experiment 1: pattern generation

This experiment evaluated the output generated by the proposed UCRP-Miner model, distinguishing between the two types of patterns. The patterns produced by NEclatClosed represent known and expected knowledge and are hence regarded as frequent patterns or beliefs. Conversely, infrequent patterns with low support are generated by the proposed extension, RNEclatClosed. Because many patterns are redundant and can often be inferred from other patterns, we focus on closed patterns to provide a condensed and concise representation, thereby reducing redundancy. Closed frequent patterns (CFPs) and closed rare patterns (CRPs) are shown in Figure 7.2. For all datasets in Table 7.1, CFPs are patterns with support values exceeding the MaxSup threshold, whereas CRPs satisfy the MinSup threshold without surpassing MaxSup.

Figure 7.2 (a)–(c) displays the CFPs and CRPs derived from the NEclatClosed and RNEclatClosed algorithms across all datasets. Information in CFPs pertains to frequently occurring patterns, representing predictable and well-known phenomena. In contrast, CRPs consist of patterns with lower support, which may reveal unexpected and potentially valuable insights. The figure illustrates that in most datasets, the number of CRPs exceeds that of CFPs. This is attributed to the fact that mining with a lower support threshold generates more patterns than mining with a higher support threshold. Rare patterns (CRPs) are of particular interest in research, as they can uncover non-obvious details. Although rare, patterns may contain noise and often reveal unforeseen and intriguing information.

Identifying patterns within CRPs that differ from the norm (CFPs) is essential for generating a comprehensive set of interesting patterns. The following experiment, discussed in Experiment 2, explored these types of patterns in greater depth.

7.3.3 Experiment 2: interesting patterns

Our experiment compared the results of our proposed model, UCRP-miner, with the state-of-the-art clustering models, OPECUR and DBSCAN, to generate interesting patterns. The goal is to find interesting patterns contained in rare ones that are similar to a set of beliefs (frequent patterns) but have lower support. We used cosine similarity to quantify the similarity between rare and frequent patterns. The similarity thresholds ranged from 40% to 80%. Figure 7.3 illustrates that only a few sets of rare patterns are considered interesting in all the datasets, with a similarity threshold of 40%. Figure 7.4 shows the meaningful patterns generated for all datasets. The number of valuable patterns decreased as the similarity threshold increased.

We evaluated the quality of the unexpected patterns generated by our proposed UCRP-miner model by comparing them with clustering models [Bui-Thi et al., 2020;

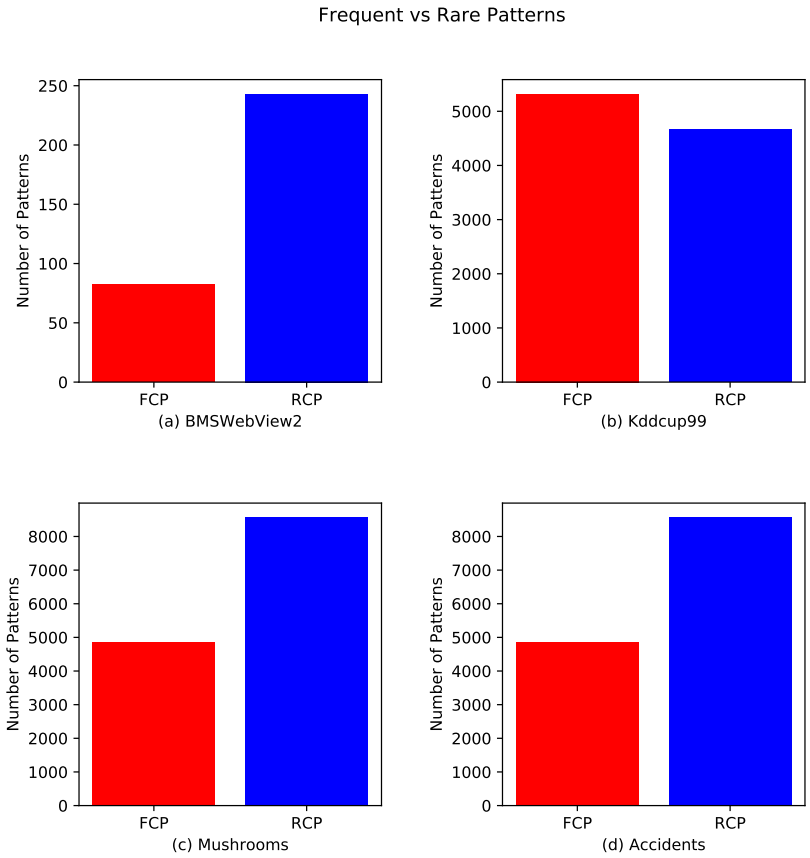


Figure 7.2: Closed (frequent and rare) patterns

Darrab et al., 2022a]. Figure 7.5 shows the interesting patterns produced by our new model, UCRP-miner, compared with state-of-the-art models [Bui-Thi et al., 2020; Darrab et al., 2022a]. For all datasets in Table 7.1, the graph shows that our model generates a complete set of interesting patterns, whereas the clustering-based model does not. Models based on clustering do not produce all the interesting patterns because they calculate the distance between rare and frequent patterns and group those with a high degree of similarity together. Because clustering considers noise patterns as the source of unexpected patterns, interesting patterns are not found in clustering-based models, as they may reside together with the frequent ones in the same cluster. For example, if we consider two patterns: a frequent pattern [a,b,c] with high support and a rare pattern [a, b, c, d] with low support, our model will suggest the rare pattern as interesting because it is rare and likely to be similar to the frequent pattern. On the other hand, rare patterns may still be assigned to a cluster by clustering and, thus, be considered frequent. According to the proposed model, UCRP-miner, this is a clear advantage compared to clustering models, which identify only a limited number of interesting patterns from noise. We use mushrooms as an example from the real dataset to highlight our model's ability to detect interesting patterns that lead to generating novel knowledge from our experimental results. Let us take a frequent pattern, FCP = {'121', '38', '36', '94', '90', '128', '23', '57', '31', '104', '1', '56', '71', '67', '41', '97': 864} which we consider as a belief since it is common with high relative frequency support = 0.10. Alternatively, let us take the rare pattern with low relative support = 0.01, RCP = {'121', '38', '36', '94', '90',

'128', '23', '57', '31', '104', '1', '56', '71', '67', '41', '97', **'51', '107'**: 108}. To calculate the cosine similarity between these patterns, we convert them to 0s or 1s based on their occurrences. Hence, FCP and RCP will be {1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, **0, 0**} and {1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, **1, 1**}, respectively. In our model, UCRP-miner, we can detect that RCP is an interesting pattern because it deviates from FCP and has 94% similarity to FCP. For further investigation, these types of patterns will be introduced to domain experts. The importance of detecting these patterns arises from the proposed model's ability to successfully discover interesting patterns with several characteristics, including little support and deviation from normal behavior (set of beliefs), while producing unexpected, new, or novel knowledge. The state of are models [Bui-Thi et al., 2020; Darrab et al., 2022a] fail to produce such patterns because RCP and FCP may reside in the same cluster. Figure 7.5 also shows that cluster-based models fail to generate interesting patterns from the accident dataset. The accident dataset is very dense; therefore, there are no noise patterns that can generate interesting patterns. Thus, the experimental results indicate that our proposed model can detect more interesting patterns than state-of-the-art clustering-based models. Consequently, our proposed model can create a complete set of patterns that matter.

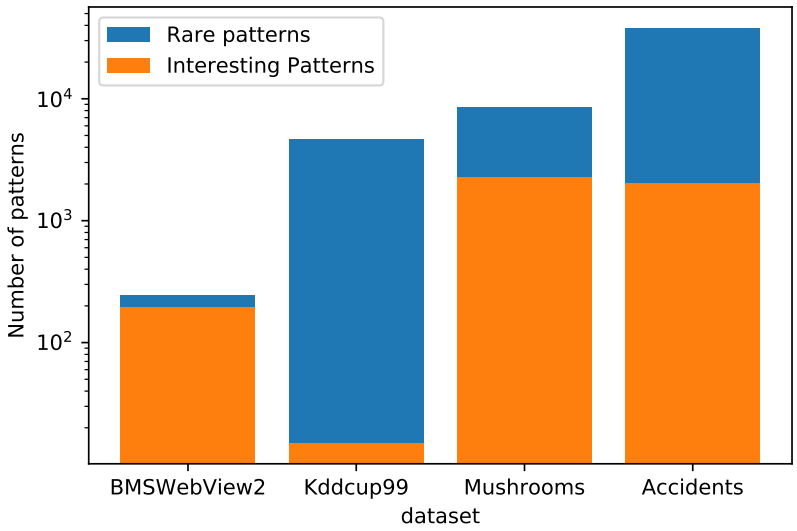


Figure 7.3: Interesting patterns versus all rare patterns for 40% similarity

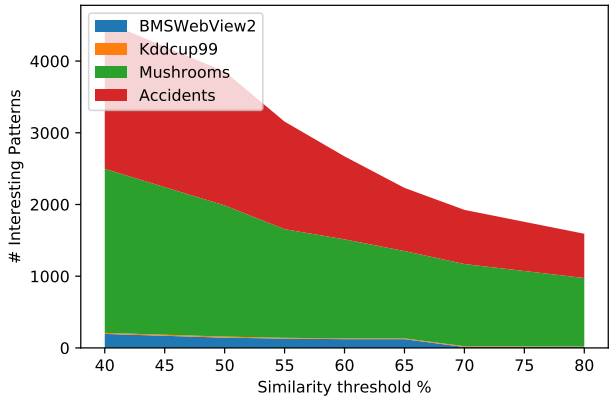


Figure 7.4: Interesting patterns for various similarity thresholds

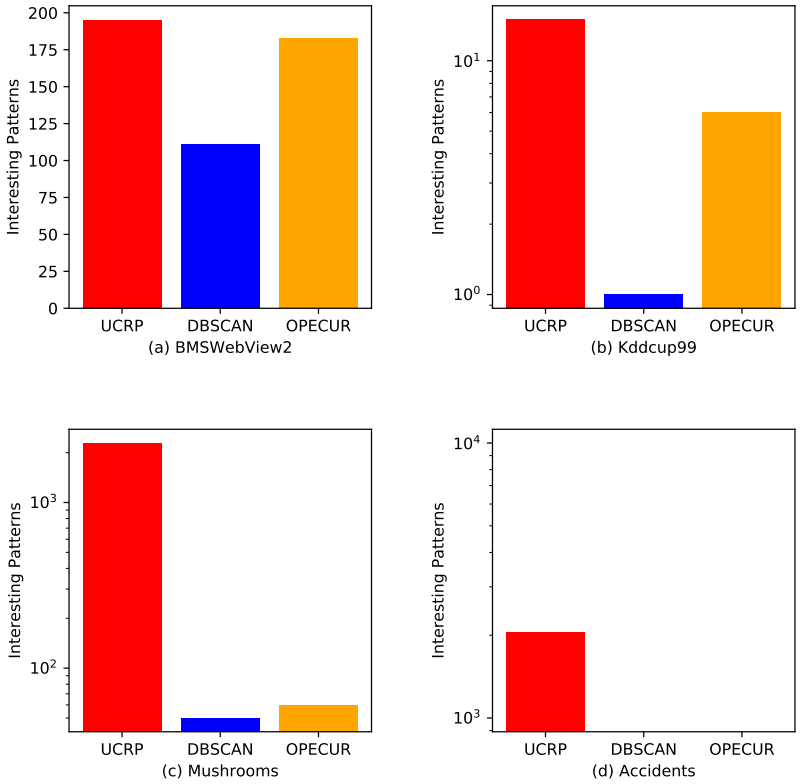


Figure 7.5: Interesting patterns generated by the proposed model, UCRP-miner, and the state-of-the-art models DBSCAN and OPECUR

7.4 Chapter summary

The discovery of interesting rare patterns has received considerable attention for its ability to uncover hidden insights within data. Although various methods have been developed to identify these patterns, existing clustering-based models often struggle to reliably generate a meaningful and diverse set of patterns and sometimes fail to capture the full scope needed for comprehensive analysis.

To address these limitations, we developed UCRP-Miner, a novel model designed to retrieve a comprehensive set of meaningful patterns from real-world datasets. By focusing on closed patterns, UCRP-Miner effectively reduces redundancy and captures deviations from the expected behavior, resulting in a concise and informative representation of the data. This approach makes UCRP-Miner particularly valuable for applications such as anomaly detection in healthcare, industrial damage monitoring, traffic analysis, and criminal investigations.

The experimental results demonstrate the effectiveness of UCRP-Miner in generating novel and unexpected patterns, highlighting its potential for uncovering valuable insights. However, certain limitations remain, particularly regarding the applicability of the model to real-world scenarios to demonstrate its full efficiency. In the following chapter, we address these limitations through a case study in the health sector, specifically focusing on heart disease. This case study provides a foundation for evaluating UCRP-Miner in a real-world context and directing future improvements.

8. Discovering Hidden Risk Factors for Heart Disease Using Rare Association Rule Mining: A Case Study

In previous chapters, we addressed the primary challenges of this dissertation, including generating rare patterns and deriving interesting rare rules. This chapter focuses on the applicability of the proposed methods and highlights the importance of rare pattern mining in the health sector. Our model demonstrates its capacity to uncover rare, unexpected rules that tackle explainability and interpretability challenges in predictive modeling, with a specific emphasis on cardiovascular disease, a leading cause of mortality worldwide. Although predictive models such as logistic regression, neural networks, and random forests have been effective, they often lack transparency and interpretability [Hassija et al., 2024]. In this chapter, we introduce Exploring the Predictive Factors of Heart Disease using Association Rule Mining (EPFHD-RARMING), an innovative approach that employs rare association rule mining to enhance the understanding and prediction of heart disease. By uncovering unexpected rules, EPFHD-RARMING identifies critical factors that contribute to heart disease, detects high-risk patterns even in asymptomatic individuals, and facilitates early intervention.

This case study underscores the value of rare pattern mining in revealing impactful relationships that might otherwise remain hidden. The objective of this method extends beyond rule identification, aiming to discover surprising and meaningful patterns that offer actionable insights for healthcare professionals. Effectively integrated with established feature engineering techniques, EPFHD-RARMING enhances practical utility, enabling medical professionals to manage patient care more proactively. This chapter illustrates the effectiveness of EPFHD-RARMING in providing deeper insights into heart disease, and offers significant advancements in medical analytics and patient outcomes. Moreover, its applicability extends beyond health

care, demonstrating the importance of identifying rare and meaningful patterns in other fields.

8.1 Introduction

According to the World Health Organization (WHO), heart disease is a significant global health concern [World Health Organization, 2021]. Annually, approximately 17.9 million people lose their lives due to cardiovascular disease (CVDs), which is the leading cause of mortality. In 2017, approximately 10.6 million new cases of coronary heart disease were reported worldwide, resulting in the loss of 8.9 million lives. The WHO estimates that by 2030, CVD death rates will increase to 23.66 million due to heart disease. This pandemic has substantial economic implications. Medical expenses linked to heart disease are expected to increase by 41% in the United States, from 126.2 $ billion in 2010 to 177.5 $ billion by 2040 [Bhatt et al., 2023].

The diagnosis of certain medical conditions can be challenging for physicians, necessitating both expeditious and accurate assessments. In the case of heart diseases, the utilization of computerized technologies is imperative to assist doctors in providing more precise and timely diagnoses [Abdelhamid et al., 2023]. Various machine learning techniques have been proposed to support the early detection and diagnosis of heart diseases, including random forests, logistic regression, support vector machines, and IOT networks [Arumugam et al., 2023; Jindal et al., 2021; Yashudas et al., 2024]. However, these models are often difficult to interpret, which can impede the understanding of the underlying rationale for their predictions, leading to a lack of confidence in the outcomes for both clinicians and patients.

Association rule mining (ARM) [Luna et al., 2019] is a widely recognized and highly interpretable data mining technique that reveals hidden patterns and correlations among various factors. Its prominence, ease of interpretation, and ability to extract valuable knowledge make it an excellent tool for real-world applications, such as market basket analysis and web traffic analysis. Despite their potential, ARM have not been widely adopted in the field of medicine. This is unfortunate because association rules can identify every pattern in a given dataset, which is highly beneficial for clinical data analysis. Using association rules, clinicians can expeditiously and automatically make well-informed diagnoses, extract valuable information, and develop essential knowledge bases. Despite the advantages of ARM [Brin et al., 1997], it presents several challenges. A significant challenge is the generation of numerous irrelevant and repetitive rules. Moreover, the most interesting rules often have low support values and are referred to as rare rules. Low support thresholds can result in an overwhelming number of rules, which complicates their management and analysis. Consequently, appropriate methods are necessary to determine the usefulness of the rules and identify the most relevant ones.

To address these challenges, we propose EPFHD-RARMIN, a model designed to identify factors contributing to heart disease while minimizing the generation of redundant or uninformative rules. The model focuses on producing only relevant and meaningful rules by integrating both frequent and rare patterns. Frequent patterns capture established associations consistent with prior knowledge or common trends. In contrast, rare patterns highlight deviations that emerge when additional features

are considered but exhibit significantly lower support. By combining both types, EPFHD-RARMIN model enables the discovery of rare rules that challenge dominant patterns, offering deeper insights into predictive factors and associated symptoms of heart disease.

Feature selection has gained extensive attention in recent years due to its significant role in identifying the most important features for model predictions [Chen et al., 2024a,b]. However, focusing solely on features and their impact on model predictions neglects the importance of determining the patterns associated with these features that lead to predictions. Therefore, in our proposed model, we emphasize not only feature selection but also the patterns that may indicate the development of heart disease when these features (symptoms, in our case study of heart disease) are present. To the best of our knowledge, this study is the first attempt to use simple yet powerful rule-mining algorithms to extract symptoms and identify patterns indicative of future heart diseases. The rules generated by our model have the potential to assist clinicians in making informed decisions for the early detection and treatment of heart diseases. Our primary objective in this study is to generate rules that are both insightful and applicable for predicting heart disease, thereby enhancing the explainability and transparency of predictive models. The main contributions of this chapter are summarized as follows:

- **Innovative rule extraction:** Our model, EPFHD-RARMING, specifically addresses the challenge of traditional association rule mining, which often produces an excessive number of low-support rules. It extracts a meaningful set of association rules from this extensive rules, focusing on those that are truly insightful and relevant, thus mitigating the common issue of the overwhelming quality of rule quantities.
- **Critical factor identification:** Our model is highly effective in uncovering pivotal factors and symptoms of heart disease, utilizing advanced analytics to prioritize the most significant variables associated with cardiovascular risks, thereby enhancing early detection and intervention strategies.
- **Predictive vulnerability analysis:** This approach diverges from conventional models by identifying not only conditions directly linked to heart disease but also seemingly healthy states that may predispose individuals to future health risks. This predictive analysis of vulnerability provides a more comprehensive and nuanced understanding of potential health trajectories.
- **Comprehensive data exploration through unsupervised learning:** Utilizing the unsupervised tool of Association Rule Mining (ARM), our methodology offers a more thorough exploration of datasets to identify overlooked patterns and factors. This comprehensive analysis aids in understanding the complex interactions between variables and heart disease. Furthermore, the rule-based approach enhances interpretability and usability, particularly in clinical settings, making the findings accessible and actionable to medical professionals.

8.2 Related work

Despite the significant challenges presented by heart disease, which remains the leading cause of death worldwide, machine learning techniques have greatly assisted

in the analysis of clinical data. These techniques make use of the vast amount of readily available healthcare data and have become powerful decision-making and forecasting tools.

Various studies have explored the potential of machine learning for predicting heart disease. In [Motarwar et al., 2020], the Random Forest algorithm emerged as the most accurate method for predicting heart disease. Another study [Mohan et al., 2019] proposed an innovative approach that combined various features and classification techniques to enhance prediction accuracy. In [Katarya and Meena, 2021], machine learning techniques for heart disease prediction were reviewed, revealing a variety of data mining strategies with varying degrees of effectiveness and accuracy. Similarly, a study [Marimuthu et al., 2018] performed a comprehensive review of different machine learning techniques, including artificial neural networks, decision trees, fuzzy logic, K-nearest neighbors, naïve Bayes, and support vector machines, in the context of heart disease prediction.

Furthermore, extensive research has been conducted to predict and evaluate the risk factors associated with heart disease. In [Jindal et al., 2021], various machine learning algorithms, such as logistic regression and KNN, were used to predict and classify patients with heart disease. Another study [Yang et al., 2023] utilized an optimized LightGBM classifier with improved hyperparameters and a focal loss function optimized using OPTUNA. This model, evaluated using CVD data from the Framingham Heart Institute, achieved an AUC value of 97.8%, outperforming other comparative models in terms of accuracy.

A novel Recommendation System for CVD Prediction Using an IoT Network (DEEP-CARDIO) was proposed in another study [Yashudas et al., 2024], which offers prior diagnosis, treatment, and dietary recommendations for cardiac diseases. This system collected data from four biosensors (ECG, pressure, pulse, and glucose) and processed them using an Arduino controller. The BiGRU attention model diagnosed and classified CVD into five categories and achieved an overall accuracy of 99.90%. Furthermore, the QMBC technique, which employs the Quine McCluskey method to derive the Minimum Boolean expression for a target feature, was introduced [Kapila et al., 2023]. By combining the predictions from the seven classifiers, the ensemble model forms a comprehensive dataset to apply the minimum Boolean equation with an 80:20 train-to-test ratio. The proposed QMBC model demonstrated superior performance compared to current state-of-the-art models and previously suggested methods, indicating its potential for improved cardiovascular disease prediction.

Although many machine learning techniques have been proposed for the early detection and diagnosis of heart disease, clinicians often struggle to trust these models because of their lack of interpretability. This difficulty in understanding the basis of the predictions compromises the reliability and acceptance of the models. To address this issue, it is essential to focus on developing transparent and interpretable models that enable clinicians and patients to comprehend underlying mechanisms and gain confidence in their predictions. Several studies have investigated the utilization of rule-based methods, particularly association rule mining, in the domain of heart disease detection.

A novel methodology and algorithm for mining distributed medical data sources using association rules, specifically focusing on predicting heart diseases, were presented in [Khedr et al., 2021]. Another study [Sonet et al., 2017] utilized association rule mining to uncover concealed patterns related to frequently occurring heart diseases in the Bangladeshi population. Associative classification mining was employed in another study [Lakshmi and Reddy, 2015] to construct a classifier using rules of high interest for an accurate heart disease prediction. An enhanced association rule mining approach for detecting coronary artery disease using a heart disease dataset was introduced in a previous study [Yadav et al., 2014].

While current methods focus on improving prediction accuracy and identifying factors that contribute to cardiac disease, several limitations persist. One significant challenge is managing unlabeled data, which is crucial for developing robust and comprehensive models. Additionally, these approaches often fail to explore the relationships between various symptoms and heart disease-causing factors, potentially overlooking critical indicators.

Many recent studies [Fournier-Viger et al., 2017] have generated rules based on frequent patterns, resulting in predictable and well-known outcomes. Despite their utility, these studies often produce an overwhelming number of rules, making analysis and interpretation costly. To address this limitation, we introduce a novel modeling approach designed to generate a limited number of insightful and interesting rules, thereby enhancing both the efficiency and effectiveness of rule analysis.

Association rule mining, particularly for rare patterns, is essential for making decisions regarding heart disease. In this work, we propose a novel method that not only identifies factors leading to heart disease but also uncovers patterns that may indicate future disease development. Our approach uses frequent patterns as a foundation for discovering interesting patterns associated with heart diseases. We developed a model to identify these patterns and their potential to lead to heart disease when combined with specific risk factors.

8.3 Dataset: heart disease

This section provides an overview of the heart disease datasets used in this study. Our approach aims to identify predictive factors for heart disease and analyze patterns in healthy individuals who may be at risk of developing the condition in the future. The dataset obtained from IEEE DataPort [Siddhartha, 2020] was constructed by combining several popular heart disease datasets to create a comprehensive resource that was previously unavailable to researchers. This newly assembled dataset contains 1,190 instances and 12 common features, making it the largest heart disease dataset currently available for research. The data were curated from five sources: Cleveland, Hungarian, Swiss, Long Beach VA, and Statlog (Heart). By integrating these datasets, we aim to advance machine learning and data mining applications related to heart disease. This consolidated dataset enables researchers to develop more accurate and effective methods for the early detection and prevention of heart disease.

Tables 8.1 and 8.2 present the dataset characteristics. Table 8.1 summarizes several key characteristics and Table 8.2 provides a description of the nominal attributes.

Table 8.1: Heart Disease Dataset Characteristics

S.No.	Attribute	Description	Unit	Data Type
1	Age	Age	Years	Numeric
2	Sex	Gender	1 = Male, 0 = Female	Binary
3	Chest Pain Type	Type of chest pain	1, 2, 3, 4	Nominal
4	Resting Blood Pressure	Resting blood pressure	mm Hg	Numeric
5	Serum Cholesterol	Serum cholesterol level	mg/dl	Numeric
6	Fasting Blood Sugar	Fasting Blood Sugar (>120 mg/dL)	1 = High, 0 = Normal	Binary
7	Resting ECG Results	Resting electrocardiogram results	0, 1, 2	Nominal
8	Maximum Heart Rate Achieved	Maximum heart rate achieved	{min=60, max=202}	Numeric
9	Exercise Induced Angina	Exercise-induced angina	0 = No, 1 = Yes	Binary
10	Oldpeak	ST depression induced by exercise relative to rest	depression	Numeric
11	Slope of the Peak Exercise ST Segment	Slope of the peak exercise ST segment	1, 2, 3	Nominal
12	Class	Diagnosis of heart disease	0 = No, 1 = Yes	Binary

Table 8.2: Description of Nominal Attributes

Attribute	Description
Sex	1 = Male 0 = Female
Chest Pain Type	1: Typical angina 2: Atypical angina 3: Non-anginal pain 4: Asymptomatic
Fasting Blood Sugar	(Fasting blood sugar > 120 mg/dl) 1 = True, 0 = False
Resting ECG Results	0: Normal 1: Having ST-T wave abnormality (> 0.05 mV) 2: Showing left ventricular hypertrophy by Estes' criteria
Exercise Induced Angina	1 = Yes 0 = No
Slope of the Peak Exercise ST Segment	1: Upsloping 2: Flat 3: Downsloping
Class	1 = Heart disease 0 = Normal

To ensure the highest level of data quality and consistency, a rigorous preprocessing pipeline was developed, which included several crucial steps, such as handling missing values and standardizing the representation of data. Our dataset did not contain any missing or null values, and a value of zero was found only once in an instance where the **St slope** was 0. Because it did not contribute to pattern or rule generation, we removed it, resulting in 1189 transactions. However, because our proposed method involves unsupervised techniques rather than classification tasks, it is crucial to perform preprocessing and feature selection tailored to our objectives. The processes are described in detail in the following sections.

8.4 The proposed model: EPFHD-RARMING

Our proposed model, EPFHD-RARMING, builds on previous work [Darrab et al., 2022b] and aims to generate rules that aid in the early detection of heart disease and predict the factors contributing to its development. This was achieved through a three-phase process specifically designed as a case study of heart disease. Our approach allows for the identification of rare but significant associations that traditional methods often overlook, providing deeper insights into the factors leading to conditions such as heart disease. The workflow of this model is illustrated in Figure 8.1.

Utilizing our method to enhance rule-based machine learning in medical datasets, we developed and implemented the *Mine Interesting Rules* algorithm 4. This algorithm systematically mines interesting rules from a dataset in three main phases, ensuring comprehensive analysis and interpretation.

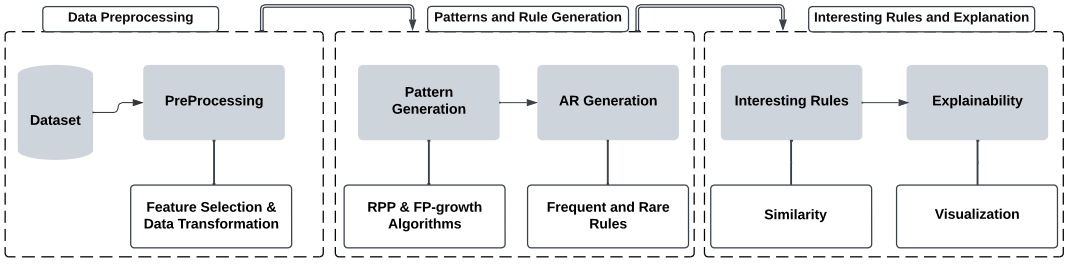


Figure 8.1: EPFHD-RARMING model for detecting heart disease risk factors

Algorithm 4 Mine Interesting Rules

Input: DB (dataset), minSup (minimum support), minRare (minimum rareness), simT (similarity threshold)
Output: List of interesting rules
ds_cleaned ← clean_data(DB) ▷ Clean the dataset
ds_transformed ← transform_data(ds_cleaned) ▷ Perform Feature Selection and Data Transformation
FPs ← find_patterns(ds_transformed, minS) ▷ Find Frequent Patterns
f_rules ← generate_rules(FPs, metrics=[confidence, leverage, lift, conviction]) ▷ Generate Frequent Rules
f_rules_yes ← filter_rules(f_rules, consequent="Yes")
f_rules_no ← filter_rules(f_rules, consequent="No")
r_patterns ← find_patterns(ds_transformed, maxSup=minSup, minSup=minRare) ▷ Find Rare Patterns
r_rules ← generate_rules(r_patterns, metrics=[confidence, leverage, lift, conviction]) ▷ Generate Rare Rules
r_rules_yes ← filter_rules(r_rules, consequent="Yes")
r_rules_no ← filter_rules(r_rules, consequent="No")
interesting_rules ← []
for each rule r_rule in r_rules_yes **do**
    for each rule f_rule in f_rules_no **do**
        similarity ← calculate_similarity(r_rule.antecedents, f_rule.antecedents)
        if similarity > simT **and** r_rule.consequent ≠ f_rule.consequent **then**
            interesting_rules.append((r_rule, f_rule))
        end if
    end for
end for
return interesting_rules

8.4.1 Algorithm for mining interesting rules in heart disease prediction

In this subsection, we describe the proposed algorithm and demonstrate its functionality.

Algorithm 4 outlines the process of mining interesting rules from a dataset in three main phases. In the data preparation and cleaning phase, the dataset (ds) is prepared to handle missing values, outliers, and noise, followed by feature selection and data transformation to make it suitable for association rule mining. In lines 1-2, the algorithm starts by defining the inputs (ds, $minSup$, $minRare$, $simT$) and the expected output, which is a list of interesting rules, where ds stands for dataset, $minSup$ for minimum support threshold, $minRare$ for minimum rare support, and $simT$ for similarity threshold. In line 3, the dataset was cleaned to handle missing values, outliers, and noise, thus ensuring the data quality for further analysis. After data cleaning, the dataset undergoes feature selection and transformation to enhance its suitability for Association Rule Mining (ARM), facilitating the discovery of meaningful patterns. From lines 5 to 12, the second phase, pattern discovery and rule extraction, was performed. In line 5, the FPs are identified using a minimum support threshold ($minSup$). From lines 6 to 8, frequent rules are generated and filtered based on specified metrics and categorized into two types of rules based on their consequent values (”Yes” for heart disease and ”No” for healthy). Similarly, rare patterns were found using both the minimum support ($minSup$) and minimum rareness ($minRare$) thresholds, and the corresponding rare rules were generated and filtered from lines 9 to 12. In the final phase, lines 13 to 22, insightful rule identification and interpretation, and interesting rules are identified by comparing rare and frequent rules. The similarity between the antecedents of rare rules (with a ”Yes” consequent) and frequent rules (with a ”No” consequent) was calculated. If the similarity exceeds a specified threshold ($simT$) and the consequences differ, the pair of rules is considered interesting and is added to the list of interesting rules. Ultimately, the algorithm returns a list of rules of interest. Following these steps, the algorithm effectively cleans and transforms the data, discovers frequent and rare patterns, generates and filters rules, and identifies rules of interest for further analysis.

In the following subsections, a thorough explanation of each phase of the model is provided, including data preparation and transformation, pattern discovery and rule extraction, and insightful rule identification and interpretation phases. The ultimate goal of this model is to provide a comprehensive understanding of its operation, with the aim of aiding the early detection of heart disease and predicting the factors that contribute to its development.

8.4.2 Data preparation and transformation phase

In this subsection, we focus on the preprocessing phase, which involves converting the dataset from a supervised classification task to an unsupervised association rule mining task. The preprocessing phase is critical for preparing a heart disease dataset for the mining process. We discuss the two primary steps of the preprocessing phase in detail, as outlined in the first phase of our workflow and described in lines 1 and 2 of the proposed algorithm 4.

- Selection of features: In this step, we employ multiple techniques to identify the most relevant factors contributing to heart disease. This process involves selecting a subset of the most informative features for the mining process. Selecting appropriate features can enhance the quality and efficiency of subsequent mining processes. Consequently, the feature selection process helped to identify the most important attributes related to heart disease.
- Dataset transformation: The heart disease dataset must be transformed into a suitable format for mining association rules. The preferred format for mining association rules is a Boolean transactional representation, where each instance is represented as a set of items and each item represents a selected feature. The value for each item was either present (1) or absent (0). This transformation prepares a dataset for further mining to generate association rules.

The details of the above steps for the heart disease dataset used in this study are as follows.

8.4.2.1 Selection of features

The cardiac disease dataset comprises 12 features, as presented in Table 8.1. We implemented a reliable feature selection process involving five distinct approaches to identify the crucial features that contribute to cardiac disease. To comprehensively understand this condition, it is imperative to include all relevant features that impact cardiac disease. Using the following selection methods, we derived a final set of 10 of the 12 features in the heart disease dataset. To select the most significant feature for the mining process, we utilized the following scikit-learn feature selection methods:

- Feature selection using the chi-squared statistic.
- Feature selection using ANOVA F-value.
- Features selected through mutual information.
- Features selected via Recursive Feature Elimination with logistic regression.
- Features selected based on random forest feature importance.

Figure 8.2 illustrates the significance of the selected features using all approaches employed in this study. According to the graph, the features 'ST slope' and 'oldpeak' appear to have the greatest influence on the outcome variable, as they are included by all approaches. 'Max heart rate', 'exercise angina', and 'chest pain type' occupy a secondary position in terms of importance, being favored by four out of five applied methods. The selection of 'cholesterol' is made by three of the five methods, while the selection of 'age' and 'sex' is made by two methods, and the selection of 'fasting blood sugar' is made by a single method.

All of these features were incorporated into our approach to comprehensively address the majority of important factors. Therefore, the following features were chosen for this proposed model: **{'ST slope', 'age', 'chest pain type', 'cholesterol', 'exercise angina', 'fasting blood sugar', 'max heart rate', 'oldpeak', and 'sex'}**. These features represent the union of the top six features of each feature selection method.

In addition, the class feature 'target' was included. Consequently, 10 out of the 12 features were used in our study.

By implementing rigorous feature selection methods and conducting a comprehensive analysis of all selected features, this study aims to achieve a thorough understanding of the key factors influencing the prevalence of heart disease.

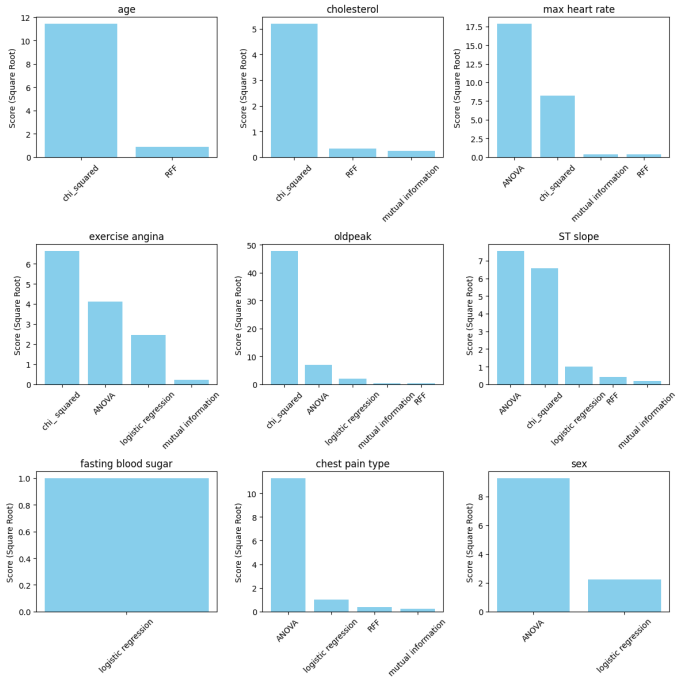


Figure 8.2: Significance of Selected Features Using Different Approaches

8.4.2.2 Dataset transformation

For the mining process to be effective, all features in the dataset related to heart disease must be presented in a binary format. There are four continuous attributes, namely 'age', 'cholesterol', 'max heart rate', and 'oldpeak', which must be discretized to convert continuous data into discrete categories or bins. This process is performed as follows for these four features:

- The 'age' feature is divided into three bins: 'young', 'middle-aged', and 'elderly'. The bin edges are specified as [0, 30, 60, np.inf], where 'np.inf' represents infinity.
- The 'cholesterol' feature is discretized into three bins: 'chollow', 'cholnormal', and 'cholhigh'. The bin edges are defined as [-1, 200, 240, np.inf].
- The 'max heart rate' feature is discretized into three bins: 'heartratelow', 'heartratenormal', and 'heartratehigh'. The bin edges were specified as [0, 100, 160, np.inf].
- The 'oldpeak' feature is discretized into three bins: 'oldpeaklow', 'oldpeakmoderate', and 'oldpeakhigh'. The bin edges are specified as [- np.inf, 1.0, 2.0, np.inf].

Discretizing these continuous variables into discrete categories simplifies the data representation and facilitates subsequent analysis during the mining process. To

	ST_slope0	ST_slope1	ST_slope2	ST_slope3	asymptomatic	atypicalangina	cholhigh	chollow	cholnormal	elderly	...
0	False	True	False	False	False	True	True	False	False	False	...
1	False	False	True	False	False	False	False	True	False	False	...
2	False	True	False	False	False	True	True	False	False	False	...
3	False	False	True	False	True	False	False	False	True	False	...
4	False	True	False	False	False	False	False	True	False	False	...

5 rows × 28 columns

Figure 8.3: Dataset after preprocessing phase

prepare the dataset for analysis, the data must be transformed into a binary format, either [0,1] or true/false. The `TransactionEncoder()` method was employed for one-hot encoding, resulting in a final dataset consisting of 1189 rows and 28 dimensions. A representation of the first five rows of the preprocessed dataset is shown in Figure 8.3.

8.4.3 Pattern discovery

In the pattern discovery phase, a formalized approach is employed to identify significant patterns that are subsequently used to generate association rules. This process involves the exploration of both frequent and rare patterns within a dataset using specialized algorithms designed for each pattern type.

To uncover FPs, the FP-growth algorithm [Han et al., 2000] was employed, which is renowned for its efficiency in discovering patterns that meet a predefined support threshold. This enables the detection of frequent patterns that occur regularly within a dataset and possesses a significant value. These patterns often correspond to well-established phenomena or widely anticipated information.

For the discovery of rare patterns, the Rare Pre-Post (RPP) algorithm [Darrab et al., 2020] was applied, which facilitates the identification of patterns that occur infrequently but offer unique and valuable insights. Rare patterns, although less common, can provide important information that is often overlooked by more traditional frequent pattern mining approaches.

A comprehensive set of patterns is generated by integrating FP-growth for frequent pattern mining with the RPP algorithm for rare pattern mining. This dual approach allows for the generation of a complete set of association rules encompassing both frequent and rare patterns. Consequently, meaningful insights and valuable knowledge can be extracted from the data to enhance our understanding of the underlying patterns.

8.4.4 Rule generation

Following the generation of frequent and rare patterns, we can derive association rules. The analysis of these rules provides valuable insights into the relationships and dependencies between different items and attributes within a dataset. By examining these rules, we can gain a comprehensive understanding of the underlying patterns and associations in the data.

Our model generates two types of rules: frequent and rare. Frequent rules represent beliefs or associations within a dataset that are considered significant, and we identify

rules that are highly supported and satisfy several statistical metrics. In this study, we were particularly interested in rules in which the consequent represents healthy patients without heart disease. Health attributes or factors associated with good health and absence of heart disease were revealed by these frequent rules.

In contrast, rare patterns lead to rules with low support but still possess statistical significance based on the metrics we employed. Examination of these rare rules provides insights into the unique factors associated with heart disease, which is of particular interest. In this study, we used rare rules to represent non-healthy patients, specifically those with heart disease.

It is possible to gain a comprehensive understanding of the associations and dependencies present in the data by considering both frequent and rare rules. By doing so, we can determine which attributes or factors contribute to good health, as well as those that indicate heart disease. Combining frequent and rare rules allows us to extract valuable knowledge from the data and make informed decisions based on the patterns found in the data.

8.4.5 Insightful rule identification and interpretation

The final phase of our proposed model, detailed in lines 13–22 of Algorithm 4, focuses on generating and interpreting interesting rules. This important phase emphasizes identifying the features that cause heart disease and uncovering interesting rules. By setting a set of beliefs—in our case, frequent rules that represent healthy patients with no heart disease—we aimed to identify those with heart disease whose characteristics differ slightly from those of healthy patients.

8.4.5.1 Interesting rules

By identifying rare rules with low support that satisfy all statistical metrics based on the background information provided, we can further refine the set of interesting association rules. In this study, we aimed to identify rare rules that deviate from the common (frequent) rules that represent healthy individuals. For this phase, we consider the following factors to determine the interestingness and unexpected nature of the rule.

- **Similarity of antecedents**: We determine the similarity between the rare rule's antecedents and the known frequent rule's antecedents using a similarity measure, such as the Jaccard similarity approach.
- **Contrasting consequences**: We compare the frequent and rare rules to see if they have contrasting consequences. In this case, the consequences of the two rules should oppose each other to be considered interesting.
- **Low support**: The rare rule, denoted as Rrule, must satisfy all predefined metrics. In particular, its support should be low, indicating that it deviates from the normal rule (Frule).

Incorporating these conditions allows us to filter and prioritize rare rules that lack support, deviate from normal patterns, exhibit similar antecedents, and contrast consequences with frequent rules. The refined rules provide valuable insights into the underlying patterns and deviations from the norm of the dataset, allowing us to further understand exceptional cases and unexpected associations.

8.4.5.2 Explainability

The application of association rule mining, a rule-based data-mining technique, is essential because of its interpretability and ease of understanding. Therefore, we summarized the interesting rules generated by our model, focusing on the factors that contribute to heart disease. Our proposed model includes comprehensive documentation supported by tables, examples, and illustrations to enhance the clarity and interpretation of association rules. By identifying the factors within the rare rules that deviate from our established beliefs, represented by the frequent rules, we gained deeper insight into the specific factors contributing to heart disease. This understanding can significantly aid in deciphering the causes of heart disease and ultimately support its prevention and treatment.

8.5 Experimental results

In this section, we present the outcomes of our proposed model for generating rules that are not only interesting and concise, but also highly valuable. We provide a detailed explanation of the results, demonstrating the effectiveness and efficiency of our model in producing rules that possess the desired characteristics without generating a large number of rules. In the following subsections, we present the experimental results of the proposed EPFHD-RARMING model.

8.5.1 Experimental setup

The experiments were conducted on Google Colab using the following commonly used parameters and constraints: When mining both frequent and rare patterns for pattern generation, it is essential to adhere to the following constraints:

- To obtain frequent patterns, a minimum support threshold, $minSup$, of 0.01 is established in order to identify frequent patterns. This signifies that a pattern is classified as frequent only when it occurs in no fewer than 0.01 of the instances within the dataset.
- Rare patterns are identified by focusing on patterns with support below this minimum support, $minSup$, and above the minimum support, $minRar$, of 0.001, denoted as $minRar = 0.001$. Therefore, we aim to identify rare patterns with support of less than $minSup$ and support equal to or greater than $minRar$.

Regarding rule generation, it is necessary to adhere to several conditions to determine the criteria for compelling rules within the proposed model. These conditions are applicable to both frequent and rare situations. Thus, only the rules that met these stringent criteria were recognized in our proposed model.

Minimum support of rules: Frequent rules must have a minimum support of $minSup$. Similarly, for rare rules, we consider rules with support less than $minSup$ but still exceeding $minRar$.

Metric requirements: To be considered a strong rule, whether frequent or infrequent, the below popular metrics must be met.

Table 8.3: Column Name Mapping

No	Original Name	Shortened Name	No	Original Name	Shortened Name
1	asymptomatic	asym	15	heart_rate_high	hrhigh
2	atypical_angina	atangina	16	heart_rate_low	hrlow
3	cholhigh	hcol	17	heart_rate_normal	hrnoraml
4	chollow	lcol	18	middle-aged	maged
5	cholnormal	ncol	19	non_anginal_pain	napain
6	downsloping	dsloping	20	oldpeakhigh	peakhigh
7	elderly	elderly	21	oldpeaklow	peaklow
8	exercise_angina0	exangina0	22	oldpeakmoderate	peakmoderate
9	exercise_angina1	exangina1	23	typical_angina	tangina
10	fasting_blood_sugar0	fbsugar0	24	upsloping	usloping
11	fasting_blood_sugar1	fbsugar1	25	young	young
12	female	F	26	heart_disease	yes
13	man	M	27	no_heart_disease	no
14	flat	flat			

- **Confidence**: Confidence score must exceed 0.80.
- **Lift**: Lift must be greater than 1.
- **Leverage**: Leverage should be greater than 0.
- **Conviction**: Conviction should be greater than 1.

In addition, for brevity, we have replaced the names of the columns with abbreviations, as shown in Table 8.3. By equalizing full column names with their abbreviated counterparts (e.g., 'asymptomatic' to 'asym' and 'heart_disease' to 'yes'), the mapping provides clarity and brevity in data representation.

8.5.2 Patterns generation

To identify frequent patterns, we utilized the FP-growth algorithm presented in [Han et al., 2000]. Furthermore, we employ the RPP algorithm [Darrab et al., 2020] to detect rare patterns that may result in unexpected outcomes. Figure 8.4 shows the results of our case study dataset for heart diseases. The graph shows a significant number of rare patterns due to the use of very low support levels. There are 81,632 patterns in this phase of pattern generation, of which 22,178 are frequent and 59,454 are rare.

Following the generation of both frequent and rare patterns, we derived the association rules. Our approach is significant for identifying the most valuable and noteworthy rules while filtering out the majority of less-relevant rules produced by rare patterns. The subsequent section explains the process of uncovering these intriguing and insightful rules from a comprehensive set of patterns.

8.5.3 Rule generation

After the generation of frequent and rare patterns, as shown in Figure 8.4, the subsequent step involves the formulation of an exhaustive set of rules. As shown

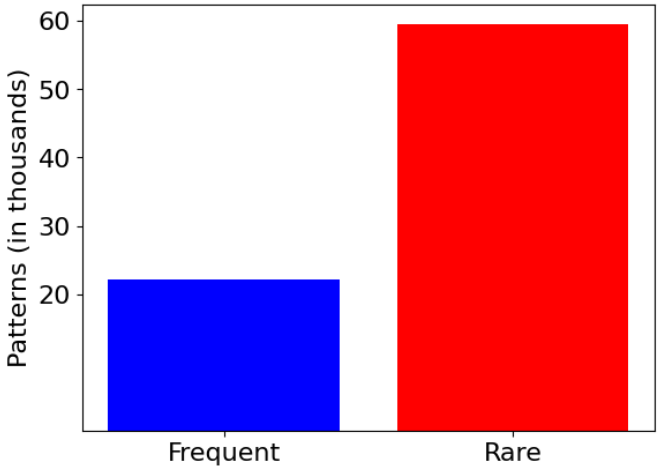


Figure 8.4: A comparison of the number of frequent and rare patterns generated from the heart disease dataset.

in Figure 2.2, the number of rules generated from frequent patterns is considerably high, amounting to 55,307, where the antecedent support is equal to or exceeds the specified minimum support threshold $minSup = 0.01$. In addition, a considerably larger number of rules, totaling 389,531, were derived from rare patterns, where their support fell below the $minSup$ threshold. This substantial number of rules highlights a significant limitation within the domain of association rule mining and underscores the need for a methodology that enables effortless identification of insightful association rules. Consequently, the purpose and challenges of this study revolve around developing a methodology that enables the identification of rules leading to the discovery of factors that contribute to the detection of heart disease. Within the scope of this study, we concentrated on rules that are relevant to this objective.

To mitigate the growth of rules and address the aforementioned challenge, extensive exploration of rule generation has been conducted. To facilitate understanding, the rules were organized based on their outcomes, specifically distinguishing between those that indicated the presence or absence of heart disease. Consequently, our attention is solely directed towards rules that relate to the presence or absence of heart disease, as it aligns with our primary objective. As shown in Figure 8.5, these rules can be classified into four categories.

- Frequent rules leading to the occurrence of heart disease.
- Frequent rules that emphasize health in the absence of heart disease.
- Rare rules indicating the presence of heart disease.
- Rare rules suggesting the absence of heart disease.

It is important to emphasize that all the rules under consideration are deemed reliable, as they fulfill all the requisite conditions within the experimental configuration. Consequently, we can classify the rules into the following four categories:

- **Type 1**: There are 2,624 rules that are frequent and have the consequence ”no heart disease”. Here is an example for such kind of rules: {**’maged’, ’fbsugar0’,**

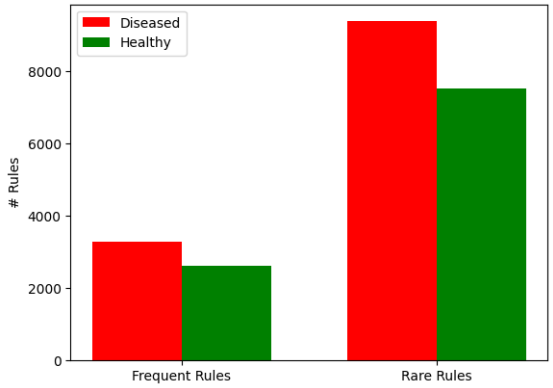


Figure 8.5: A comparison of the number of frequent and rare rules generated from the heart disease dataset.

'usloping', 'peaklow' ⇒ **'No'**}. The rule's various evaluation metrics are as follows: **support (0.24), confidence (0.86), lift (1.81), leverage (0.11), and conviction (3.73)**.

- **Type 2**: There are 3,293 frequent rules with "heart disease" as their consequent. Here is an example of such kind of rule: {**hrnoraml', 'exangina1', 'asym', 'M'** ⇒ **'Yes'**}. The rule's various evaluation metrics are as follows: **support (0.21), confidence (0.92), lift (1.7), leverage (0.08), conviction (6.02)**.
- **Type 3**: A total of 7,530 rules are rare and have "no heart disease" as an outcome. For example, the rule: **{'fbsugar0', 'tangina', 'hrnoraml', 'exangina0', 'dsloping', 'M'}** ⇒ **'No'** indicates there is no presence of heart disease. The rule's various evaluation metrics are as follows: **support (0.001), confidence (1.0), lift (2.11), leverage (0.0008), conviction (Infinity)**.
- **Type 4**: A total of 9,381 rare rules indicate the presence of "heart disease". In the case of this type of rule, for example, {**'asym', 'exangina1', 'hrnoraml', 'flat', 'ncol', 'elderly', 'M'** ⇒ **'Yes'**}. The rule's various evaluation metrics are as follows: **support (0.009), confidence (1.0), lift (1.89), leverage (0.004), conviction (Infinity)** .

8.5.3.1 Type 1 and 2 (frequent rules)

As shown in Figure 8.5, Type 1 rules consist of 2,624 frequent rules associated with the consequence "no heart disease." These rules represent healthy patients and indicate that the symptoms exhibited by patients do not indicate the presence of a heart disease. In contrast, Type 2 rules indicate frequent rules with high support that represent patients with heart disease. These rules are representative of a well-established phenomenon, in which frequent rules exhibit a high level of support and express specific expectations.

The insights gained from these rule types are widely recognized and can be readily interpreted by domain experts. Numerous studies have extensively examined this category of rules [Fournier-Viger et al., 2017], leading us to regard them as a set of prevailing beliefs because they encapsulate the most common patterns.

Our analysis focuses on Type 1 rules, which we use as the foundation for identifying unexpected and intriguing rules. We elaborate on this endeavor in the following

section titled "Interesting Rules," which constitutes the principal contribution of our novel model, EPFHD-RARMING.

8.5.3.2 Type 3 and 4 (rare rules)

Typically, traditional approaches primarily emphasize the analysis of rules that fall into Categories 1 and 2, whereas rules of Types 3 and 4 are often overlooked when using conventional approaches. However, it is essential to recognize that these rules within Categories 3 and 4 have the potential to yield interesting results. Consequently, the identification of interesting rules among this plethora of rules poses a substantial challenge, particularly when seeking unexpected and significant rules related to the occurrence of heart disease. As part of our research endeavor, we conducted a comprehensive examination of these rules to rectify analytical oversight. Our primary focus remains on rules of type 4, which constitute a set of 9,381 rare rules indicative of the presence of "heart disease." Considering these rules, we are looking for factors that contribute to heart disease. In contrast, we chose to omit the rules of Type 3, which typically indicate the presence of healthy patients and display low levels of support in our dataset. An extensive analysis of these rules would result in excessive costs, without providing substantial insights into our primary objective of identifying patients with heart disease.

8.5.4 Interesting rules

The primary objective of this subsection is to identify and analyze rare patterns that contrast with those observed in patients without cardiac conditions. A unique aspect of these patterns is the similarity in their antecedents, while their consequences differ, often contradicting one another and resulting in unexpected outcomes. Accordingly, we examined both Type 4 rules, which represent rare patterns associated with heart disease, and Type 1 rules, which reflect frequent patterns indicative of healthy patients.

To determine the rules of interest, we employed objective metrics such as lift, confidence, leverage, and conviction, as defined in Definitions 2 and 3. These rules, whether frequent or rare, must satisfy these metrics to be considered strong rules and demonstrate their objective interest.

Moreover, we explore rare rules that deviate from the normal rules (i.e., frequent rules without heart disease) due to symptoms that reduce the support of the rules. To identify such rules, we used the Jaccard metric and set the similarity threshold to 0.80. This allowed us to identify patterns that are associated with the absence of heart disease and become rare in the presence of heart disease when another factor is introduced. Hence, we identified a total of only 163 interesting rules using the proposed model. Analyzing these rules can provide valuable insights for medical experts, particularly in identifying symptoms that may be indicative of heart disease.

Let us analyze two specific rules, denoted as "frequent" and "rare". The first rule, represented as **'heartrate = normal', 'oldpeak = high', 'exercise angina = 0', 'fasting blood sugar = 0', 'cholesterol = high', 'sex = female' ==> 'Yes' (heart disease)**, is classified as a **rare rule**. Conversely, the second rule, expressed as **'heartrate = normal', 'exercise angina = 0', 'fasting blood sugar = 0', 'cholesterol = high', 'sex=**

female' ==> 'No' (no heart disease), is categorized as a **frequent rule**. The two rules demonstrated a substantial degree of similarity (0.83). This similarity indicates that when a seemingly healthy patient exhibits specific characteristics, including a normal heart rate, absence of exercise-induced angina, normal fasting blood sugar levels, high cholesterol levels, and female sex, a flag is raised, suggesting a potential risk of heart disease, particularly if the old peak value is high. In other words, it is important to note that a patient may develop heart disease if their oldpeak value becomes high if they have symptoms, as indicated in the frequent rule.

The visualization presented in Figure 8.6 illustrates two critical rules derived from the heart disease dataset. These rules highlight the importance of identifying rare but significant patterns that can drastically alter prediction outcomes. The first rule, with a consequent 'No heart disease,' has antecedents that include 'Oldpeak = low,' 'Middle age,' 'High heart rate,' 'Low cholesterol,' and 'Fasting blood sugar = 0.' This rule has a support value of 0.02 and a confidence of 0.82, indicating that it is relatively common and reliable in predicting the absence of heart disease under these conditions.

The second rule, which includes the additional antecedent 'Asymptomatic,' changes the prediction to 'Heart disease.' Despite having a lower support value of 0.004, this rule boasts a higher confidence value of 0.83. The transformation from a 'No heart disease' to a 'Heart disease' consequent upon adding the 'Asymptomatic' condition underscores the critical nature of this rare pattern. The similarity in antecedents between these two rules, differing only by the presence of 'Asymptomatic,' makes the second rule particularly intriguing and significant for heart disease detection.

This analysis highlights how our proposed model effectively identifies interesting rare rules by examining this example and demonstrating the primary results of our study. By focusing on such rare but valuable rules, healthcare professionals can better identify and manage patients who might otherwise be overlooked because of the rarity of these conditions. This approach not only enhances the accuracy of heart disease predictions but also contributes to a more nuanced understanding of the various factors involved. The 'Asymptomatic' feature, when combined with other symptoms, can change the risk assessment from no heart disease to high risk, emphasizing its role in medical diagnostics.

Our EPFHD-RARMING model successfully extracted 163 relevant rules from a large number of rules. These rules can provide valuable insights for medical experts in the investigation of symptoms that may indicate cardiovascular disease.

8.5.5 Explanation and interpretation of interesting rules

This section provides a thorough and comprehensive explanation of the intriguing rules generated, ensuring that they are clearly communicated and understood by the end user. In total, 163 interesting rules were identified using the proposed model, and their visual representations are shown in Figure 8.7. It is worth noting that the graph indicates a high similarity between frequent and rare rules, with a similarity score exceeding 0.80. Our focus is on rare rules that deviate from frequent rules by introducing additional symptoms, resulting in the formation of new rules with lower support but yielding more unexpected insights. For example, the labeled rules in

the graph showcase both frequent and rare rules that diverge from them. Figure 8.7 shows the relationship between frequent and rare rules in terms of their support and similarity. The plot uses three dimensions to represent the key metrics:

- **X-axis (frequent rule support):** This axis represents the support values of frequent rules, indicating how often these rules occur within the dataset.
- **Y-axis (rare rule support):** This axis represents the support values of rare rules, showing how often these less common rules appear within the dataset.
- **Z-axis (Jaccard similarity)** This axis represents the Jaccard similarity between the antecedents of frequent and rare rules. A higher similarity value indicates a greater overlap between the sets of conditions that define the rules.

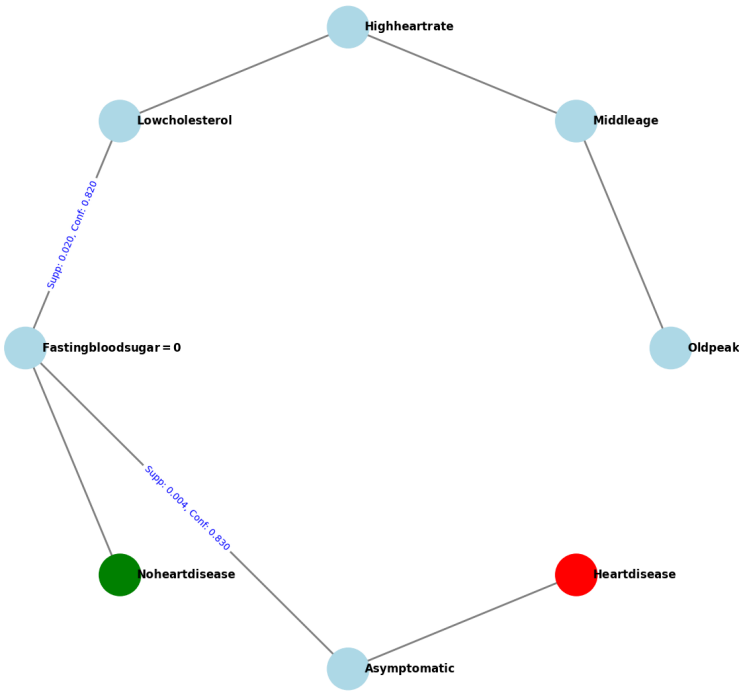


Figure 8.6: Rule visualization showing how the 'asym' feature changes prediction from 'No' to 'Yes' for heart disease. The model highlights rare rules by changes in support and confidence.

The points in the plot are color-coded according to their similarity values, with a color bar on the side serving as a reference for the similarity scale. The plot's interactive features allow users to scroll over individual points to view detailed information, including the rule pair ID, antecedents, consequences, support, confidence for both frequent and rare rules, and the similarity value for each rule pair.

In the highlighted example, point represents a pair of rules with the following details.

- **Frequent rule:** {Antecedents: 'Middle age', 'Male', 'Fast blood suger =0, 'Upsloping ST slope', 'Non-anginal pain', 'Exercise-induced angina = 0'}, Consequent: 'No heart disease', Support: 0.048, and Confidence: 0.89.

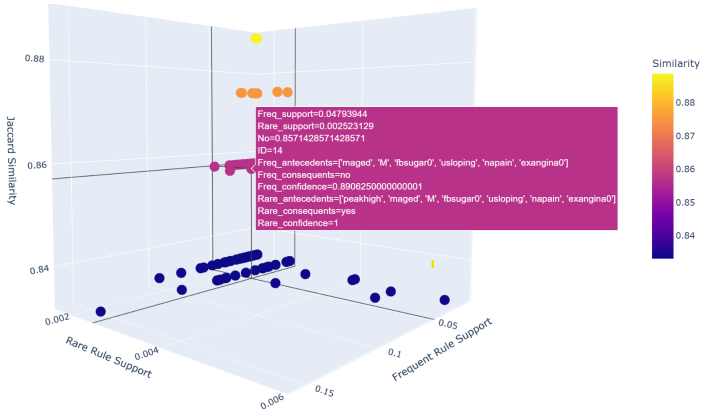


Figure 8.7: 163 interesting rules plotted in 3D

- **Rare rule:** {Antecedents: 'Middle age', 'Male', 'Fast blood suger =0, 'Upsloping ST slope', 'Non-anginal pain', 'Exercise-induced angina = 0', 'OldPeak high'}, Consequent: 'heart disease', Support: 0.003, and Confidence: 1.
- **Similarity (Jaccard):** 0.857.

This specific pair of rules is significant because the addition of the antecedent 'OldPeak High' in the rare rule changes the consequent from 'No heart disease' to 'Heart disease.' Despite the rare rule having a much lower support value, the high Jaccard similarity (0.86) with the frequent rule indicates that the conditions for both rules are very similar. This insight is crucial as it highlights how a slight change in conditions can alter the outcome, emphasizing the importance of considering rare rules in heart disease prediction and diagnosis. The confidence levels of both rules also suggest their reliability, making them valuable for further analysis and application in medical diagnostics.

Table 8.4 presents the most interesting rules based on similarity measures. It is important to note that all these rules are frequent and correspond to healthy patients, becoming rare and indicative of heart disease when an additional symptom is included. A possible explanation for the occurrence of interesting rare rules in the dataset is that adding another factor or symptom to the frequent rules reduces their support and makes them rare.

Let us consider rule number 7 in Table 8.4 to illustrate how interesting rules are generated. In the absence of the red symptom **oldpeak with a high value**, the frequent rule with the factors **'middle-aged', 'high heart rate', 'male', 'fasting blood sugar = 0', 'upsloping ST slope', 'no exercise-induced angina' ==> 'no heart disease'** suggests that individuals with these factors are generally free from heart disease. However, when a new rule, a rare one, is formed by including a high 'oldpeak' value, the generated rule **'high oldpeak', 'middle-aged', 'high heart rate', 'male', 'fasting blood sugar = 0', 'upsloping ST slope', 'no exercise-induced angina' ==> 'heart disease'** identifies patients at risk of heart disease. Although the support of the frequent rule is 0.06 out of 1189, indicating that approximately 71 patients with these factors are healthy, the support of this new rare rule decreases to 0.002, implying that only two patients with these factors actually have heart disease.

Table 8.4: Top 10 most interesting rare rules that

No	Uncommon rules (heart disease)
1	{'peaklow', 'maged', 'hrhigh', 'M', 'lcol', 'fbsugar0', 'usloping', 'exangina0', 'asym' }
2	{'ncol', 'maged', 'hrhigh', 'fbsugar0', 'usloping', 'napain', 'exangina0', 'peakhigh'}
3	{'hrnoraml', 'peaklow', 'maged', 'lcol', 'fbsugar0', 'napain', 'exangina0', 'flat'}
4	{'atangina', 'hcol', 'maged', 'F', 'usloping', 'hrnoraml', 'exangina0', 'fbsugar1'}
5	{'hrnoraml', 'peaklow', 'maged', 'M', 'lcol', 'napain', 'exangina0', 'flat'}
6	{'atangina', 'hcol', 'peaklow', 'maged', 'F', 'usloping', 'exangina0', 'fbsugar1'}
7	{'maged', 'hrhigh', 'M', 'fbsugar0', 'usloping', 'napain', 'exangina0', 'peakhigh'}
8	{'atangina', 'hcol', 'peaklow', 'maged', 'F', 'fbsugar0', 'hrnoraml', 'flat'}
9	{'peaklow', 'maged', 'hrhigh', 'M', 'lcol', 'usloping', 'exangina0', 'asym'}
10	{'atangina', 'hcol', 'peaklow', 'maged', 'F', 'hrnoraml', 'exangina0', 'fbsugar1'}

The following section provides an in-depth analysis of the factors that contribute to the development of heart disease. Our findings underscore the significance of these factors and their impact on the prediction of heart disease, further validating the proposed EPFHD-RARMING model. This analysis not only offers deeper insights into the relationship between these factors and heart disease but also enhances the interpretability and explainability of our findings.

It is imperative to emphasize that the rules outlined in Tables 8.5 through 8.10 pertain to healthy individuals, as they represent the most frequent rules associated with no heart disease (set of beliefs) and high support when the crucial "red feature" is absent. This "red feature" serves as the pivotal element that transforms these common rules—characterized by their high support and absence of cardiac disease—into uncommon rules with reduced support when cardiac disease is present. In the following subsections, we provide a thorough examination of these contributing factors.

8.5.5.1 ST depression induced by exercise relative to rest (Oldpeak)

The significance of the old peak value in the examination of noteworthy rules cannot be overstated, as it was present in 69 of the 163 rules under investigation. The transformation from a state of good health to one of rarity with heart disease is strongly indicative of heart disease, as the transition occurs when the old peak is high and is combined with these 69 common rules.

Our model places a high emphasis on the importance of a high peak value because it is associated with nearly 40% of the interesting rules. This suggests a strong connection between a high peak value and an increased risk of heart disease. This information is valuable in identifying healthy rules that include these factors and serves as a significant alarm for potential heart disease in patients with high peak values.

The presence of a high oldpeak value (indicated as 'peakhigh') is a critical factor that triggers the transition from frequent, benign patterns to rare, high-risk patterns. This shift highlights the importance of closely monitoring old peak values, as they can serve as early warning signs for the onset of cardiovascular disease. The transformation from frequent to rare rules not only underscores the predictive power of the old peak

but also demonstrates the utility of our model in identifying these crucial changes in health status.

In Table 8.5, we present the top 10 rules that demonstrate the transformation of frequent rules (common in healthy patients) into rare rules because of the high old-peak values, signifying the development of heart disease. These rules illustrate the significant impact of Oldpeak on cardiovascular risk and the value of our model in uncovering these patterns.

To illustrate how unexpected rare rules are generated, we consider Rule 1 in Table 8.5. This rule indicates that if the cholesterol level is within the normal range, the patient's age falls within the middle range, there is a high heart rate, fasting blood sugar levels are normal, the ST depression on the Resting Electrocardiogram presents an upsloping pattern, the chest pain type is non-anginal pain, and exercise-induced angina is absent, then a diagnosis of heart disease is made when the "Oldpeak" value is high.

This example underscores the importance of analyzing various factors in conjunction with the old peak to make early and accurate diagnoses. Effective treatments aimed at reducing mortality associated with cardiovascular diseases can be developed by understanding these contributing factors. The availability of such frequent rules, particularly for individuals at high risk of developing heart disease when "Oldpeak" (ST depression induced by exercise relative to rest), is highly important. Hence, high oldpeak values often indicate ischemia or reduced blood flow to the heart muscle, which is a critical factor in the development of cardiovascular disease. By identifying patterns in which high oldpeak values correlate with other risk factors, healthcare providers can develop more targeted interventions to manage and mitigate these risks.

The insights provided by our model highlight the need for a comprehensive evaluation that considers the interplay between multiple factors. By identifying these critical patterns, healthcare professionals can better assess and manage at-risk patients, ultimately improving outcomes and reducing the burden of cardiovascular diseases.

Overall, the presence of a high oldpeak value as a significant marker in our model underscores the importance of detailed cardiovascular assessment and proactive management strategies. The rules identified by our model provide a roadmap for clinicians to follow, ensuring that at-risk patients receive the necessary care to prevent the progression of heart disease. This proactive approach can lead to earlier interventions, better patient outcomes, and a reduction in the overall incidence of cardiovascular events.

Please note that this explanation is applicable to the remaining rules found in Table 8.5, emphasizing the broad applicability and significance of our findings across different patient profiles.

8.5.5.2 The Slope of the peak exercise ST segment (ST Slope)

The experimental results of the proposed model demonstrate that the ST Slope plays a significant role in the onset of cardiovascular disease, particularly when its value is ('flat'). According to the results, 28 out of 163 significant rare rules indicate

Table 8.5: Top 10 Rare Heart Disease Rules with High Oldpeak Values

No	Uncommon rules (heart disease)
1	{'ncol', 'maged', 'hrhigh', 'fbsugar0', 'usloping', 'napain', 'exangina0', 'peakhigh' }
2	{'maged', 'hrhigh', 'M', 'fbsugar0', 'usloping', 'napain', 'exangina0', 'peakhigh' }
3	{'ncol', 'maged', 'M', 'fbsugar0', 'usloping', 'napain', 'exangina0', 'peakhigh' }
4	{'ncol', 'M', 'fbsugar0', 'usloping', 'napain', 'exangina0', 'peakhigh' }
5	{'maged', 'hrhigh', 'M', 'usloping', 'napain', 'exangina0', 'peakhigh' }
6	{'maged', 'hrhigh', 'M', 'fbsugar0', 'usloping', 'exangina0', 'peakhigh' }
7	{'maged', 'hrhigh', 'M', 'fbsugar0', 'usloping', 'napain', 'peakhigh' }
8	{'maged', 'M', 'lcol', 'usloping', 'napain', 'exangina0', 'peakhigh' }
9	{'ncol', 'maged', 'hrhigh', 'M', 'usloping', 'napain', 'peakhigh' }
10	{'maged', 'M', 'fbsugar0', 'usloping', 'ncol', 'exangina0', 'peakhigh' }

cardiovascular disease when their ST Slope is 'flat', compared to healthy patients (frequent rules without cardiovascular disease). This transition from frequent to rare rules signifies an increased likelihood of cardiovascular disease, indicating a critical shift in health status.

The flat ST Slope is particularly noteworthy because it reflects a significant modification of the underlying factors or conditions. A flat ST Slope during peak exercise typically indicates an abnormal response to physical stress, which can be a precursor to serious cardiovascular issues. The occurrence of rarity, along with specific rule attributes such as a flat ST Slope, may act as a strong marker of cardiovascular risk. Further investigation of the causes and consequences of this transformation on health outcomes is required.

The top 10 interesting rules, displayed in Table 8.6, illustrate how this factor determines the rules that lead to cardiovascular disease. These rules deviate from norms as their support falls and are often missed during frequent pattern mining. For example, a rule might indicate that a patient with normal cholesterol levels and no other significant symptoms, when combined with a flat ST Slope, suddenly falls into a high-risk category for heart disease.

Our proposed model identifies these critical deviations and uncovers hidden patterns that are not apparent in traditional analyses. The identification of a flat ST Slope as a significant risk factor for cardiovascular disease highlights the importance of this symptom in clinical assessments. By recognizing the importance of a flat ST Slope, healthcare professionals can better assess the risk of cardiovascular disease in patients who might otherwise appear healthy.

As a result, the presence of a flat ST Slope is a vital factor in our model for detecting rare but significant rules that indicate cardiovascular disease. This insight underscores the importance of considering the ST Slope in comprehensive cardiovascular risk assessments. The exceptional rules identified by our model, as shown in Table 8.6, provide a deeper understanding of the factors contributing to cardiovascular risk. These findings emphasize the necessity for thorough evaluations that include the ST Slope, enabling early detection and improved management of heart disease.

Table 8.6: Top 10 Rare Heart Disease Rules: (ST Slope) with Flat Values

No	Uncommon rules (heart disease)
1	{'hrnoraml', 'peaklow', 'maged', 'lcol', 'fbsugar0', 'napain', 'exangina0', 'flat'}
2	{'hrnoraml', 'peaklow', 'maged', 'M', 'lcol', 'napain', 'exangina0', 'flat'}
3	{'atangina', 'hcol', 'peaklow', 'maged', 'F', 'fbsugar0', 'hrnoraml', 'flat'}
4	{'atangina', 'hcol', 'peaklow', 'maged', 'M', 'fbsugar0', 'hrnoram', 'flat'}
5	{'atangina', 'hcol', 'peaklow', 'maged', 'fbsugar0', 'hrnoraml', 'exangina0', 'flat'}
6	{'atangina', 'hcol', 'peaklow', 'M', 'hrnoraml', 'exangina0', 'flat'}
7	{'peaklow', 'M', 'lcol', 'napain', 'hrnoraml', 'exangina0', 'flat'}
8	{'atangina', 'hcol', 'peaklow', 'F', 'fbsugar0', 'hrnoraml', 'flat'}
9	{'atangina', 'hcol', 'peaklow', 'maged', 'hrnoraml', 'exangina0', 'flat'}
10	{'atangina', 'hcol', 'peaklow', 'maged', 'fbsugar0', 'hrnoraml', 'flat'}

8.5.5.3 Type of chest pain: asymptomatic

The type of chest pain experienced by the patients was a significant factor identified using our proposed model. Our experimental results demonstrate that chest pain plays a crucial role in the development of heart disease. Specifically, among the 163 significantly rare rules in otherwise healthy patients, 23 have been found to be at risk for heart disease when chest pain is ('asymptomatic (asym)'). This indicates that the presence of asymptomatic chest pain, when combined with other common health indicators, significantly alters patient's health status, suggesting a high likelihood of heart disease.

The presence of asymptomatic chest pain is particularly concerning because it often goes unnoticed by patients, delaying diagnosis and treatment. Our findings underscore the importance of identifying subtle yet critical symptoms. The transition from frequent to rare rules signifies a substantial shift in health status, in which the addition of asymptomatic chest pain to an otherwise benign condition increases the risk of heart disease.

The top 10 interesting rules, shown in Table 8.7, illustrate how this factor deviates from the norms and leads to rules that express patients with heart disease, despite their rarity. These rules highlight the critical nature of asymptomatic chest pain as a determinant of cardiovascular disease onset. For instance, a rule might indicate that a middle-aged individual with normal cholesterol levels and no other significant symptoms, when combined with asymptomatic chest pain, suddenly falls into the high-risk category for heart disease.

Our proposed model is effective in identifying the important factors that contribute to the development of heart disease. By focusing on the presence of asymptomatic chest pain, our model uncovered hidden patterns that are not evident in traditional analyses. This insight is invaluable for early detection and intervention, as it identifies patients who might otherwise be overlooked due to the absence of more obvious symptoms.

Therefore, the identification of asymptomatic chest pain as a significant risk factor for heart disease was a major finding of the proposed model. The exceptional rules identified by our model, as shown in Table 8.7, provide a deeper understanding of the factors contributing to cardiovascular risk. These rules emphasize the importance of

Table 8.7: Top 10 Rare Heart Disease Rules: Chest pain type as asymptomatic

No	Uncommon rules (heart disease)
1	{'peaklow', 'maged', 'hrhigh', 'M', 'lcol', 'fbsugar0', 'usloping', 'exangina0', 'asym'}
2	{'peaklow', 'maged', 'hrhigh', 'M', 'lcol', 'usloping', 'exangina0', 'asym'}
3	{'maged', 'hrhigh', 'M', 'lcol', 'fbsugar0', 'usloping', 'exangina0', 'asym'}
4	{'hcol', 'peaklow', 'elderly', 'F', 'fbsugar0', 'usloping', 'exangina0', 'asym'}
5	{'peaklow', 'hrhigh', 'M', 'lcol', 'fbsugar0', 'usloping', 'exangina0', 'asym'}
6	{'hrhigh', 'M', 'lcol', 'fbsugar0', 'usloping', 'exangina0', 'asym'}
7	{'peaklow', 'hrhigh', 'M', 'lcol', 'usloping', 'exangina0', 'asym'}
8	{'maged', 'hrhigh', 'M', 'fbsugar1', 'usloping', 'exangina0', 'asym'}
9	{'hcol', 'elderly', 'F', 'fbsugar0', 'usloping', 'exangina0', 'asym'}
10	{'hcol', 'peaklow', 'elderly', 'F', 'fbsugar0', 'exangina0', 'asym'}

thorough clinical assessments, including the evaluation of subtle symptoms, such as asymptomatic chest pain. By incorporating these insights, healthcare professionals can improve early diagnosis and treatment, ultimately reducing the incidence and severity of heart disease.

8.5.5.4 Max heart rate

The importance of the maximum heart rate in identifying unusual rules that diverge from typical frequent rules, which serve as standard beliefs, cannot be overstated. Our findings reveal that when the maximum heart rate is low, 12 out of 163 significantly rare rules have been linked to cardiovascular disease in otherwise healthy individuals. These deviations from the norm occur when the maximum heart rate ('hrlow') is low. Consequently, these 12 unusual rules represent deviations from the expected patterns and are indicative of cardiovascular disease in patients.

Our proposed model successfully uncovered these unique rules, highlighting critical contributors to the onset of cardiovascular disease, despite their rarity. It is essential to note that these rules apply to women. Furthermore, six of these distinctive rules are relevant to elderly women, specifically rules 1, 2, 8, 9, 11, and 12, as shown in Table 8.8. This suggests that elderly women with a low maximum heart rate are at a particularly heightened risk of developing cardiovascular disease, emphasizing the need for targeted interventions and monitoring of this demographic.

In contrast, when the maximum heart rate was high, our model did not identify any distinct rules directly linking this symptom to cardiovascular disease. However, it is crucial to emphasize that a high maximum heart rate is associated with 59 interesting and rare rules. Although not directly caused by a high maximum heart rate, these rules are linked to other significant factors that contribute to the development of cardiovascular disease. These factors include a high peak, asymptomatic chest pain, and various other symptoms. This indicates that while a high maximum heart rate alone may not be a direct indicator, its presence along with other risk factors can significantly increase the likelihood of cardiovascular disease.

This dual insight, which highlights the critical role of both low and high maximum heart rates in different contexts, demonstrates the robustness of the proposed model. This underscores the importance of considering the maximum heart rate in comprehensive cardiovascular risk assessments. By identifying these rare but critical rules,

Table 8.8: Rare Heart Disease Rules: Max heart rate as low

No	Uncommon rules (heart disease)
1	{'hcol', 'elderly', 'F', 'fbsugar0', 'napain', 'exangina0', 'hrlow'}
2	{'peakmoderate', 'elderly', 'F', 'fbsugar0', 'flat', 'exangina0', 'hrlow'}
3	{'F', 'fbsugar0', 'flat', 'napain', 'exangina0', 'hrlow'}
4	{'peakmoderate', 'F', 'fbsugar0', 'flat', 'exangina0', 'hrlow'}
5	{peakmoderate', 'hcol', 'F', 'fbsugar0', 'exangina0', 'hrlow'}
6	{'peakmoderate', 'F', 'fbsugar0', 'napain', 'exangina0', 'hrlow'}
7	{'hcol', 'F', 'fbsugar0', 'napain', 'exangina0', 'hrlow'}
8	{'hcol', 'elderly', 'F', 'fbsugar0', 'napain', 'hrlow'}
9	{'hcol', 'elderly', 'F', 'napain', 'exangina0', 'hrlow'}
10	{'peaklow', 'maged', 'fbsugar1', 'ncol', 'exangina0', 'hrlow'}
11	{'elderly', 'F', 'fbsugar0', 'napain', 'exangina0', 'hrlow'}
12	{'peakmoderate', 'elderly', 'F', 'fbsugar0', 'exangina0', 'hrlow'}

our model provides valuable information that can aid in early diagnosis and targeted intervention, ultimately contributing to better patient outcomes and personalized healthcare strategies.

In summary, the presence of a low maximum heart rate is a vital factor in our model for detecting rare but significant rules that indicate cardiovascular disease, particularly in women and elderly women. Conversely, a high maximum heart rate, although not directly causal, is associated with other risk factors that collectively indicate an increased risk of heart disease. These insights from our model emphasize the importance of a holistic approach to cardiovascular risk assessment, considering various interrelated factors to improve the accuracy and effectiveness of disease prediction and management.

8.5.5.5 Exercise-induced angina

Our research highlights the critical role of exercise-induced angina in identifying exceptional and atypical rules that deviate from established norms. Exercise-induced angina, indicated by a value of 1, is a condition in which chest pain occurs during physical activity due to reduced blood flow to the heart. This factor was proven to be significant in our study.

Our findings indicate that when ('exercise-induced angina ('exangina1') is present, 13 of the 163 rare rules exhibit a strong association with cardiovascular diseases in otherwise healthy individuals. These 13 unusual rules, detailed in Table 8.9, contrast sharply with conventional norms and are effective in identifying patients with cardiovascular disease. These rules indicate that the presence of exercise-induced angina, combined with other factors, significantly alters a patient's health status, leading to a higher risk of cardiovascular disease.

A detailed analysis of these rules revealed that the presence of exercise-induced angina, when combined with other health indicators, serves as a critical marker for cardiovascular disease. This demonstrates the power of our innovative model to uncover important health insights that may be missed by conventional analysis. Despite their rarity, these rules provide valuable information for the early diagnosis and prevention of heart disease.

Table 8.9: Rare Heart Disease Rules: exercise-induced angina is present

No	Uncommon rules (heart disease)
1	{'hcol', 'peaklow', 'maged', 'fbsugar0', 'usloping', 'hrnoraml', 'exangina1'}
2	{'peakmoderate', 'elderly', 'F', 'fbsugar0', 'flat', 'hrnoraml', 'exangina1'}
3	{'atangina', 'hcol', 'peaklow', 'maged', 'fbsugar0', 'hrnoraml', 'exangina1'}
4	{'peaklow', 'M', 'elderly', 'fbsugar0', 'usloping', 'ncol', 'exangina1'}
5	{'peaklow', 'maged', 'hrhigh', 'lcol', 'fbsugar0', 'usloping', 'exangina1'}
6	{'maged', 'F', 'fbsugar0', 'flat', 'napain', 'exangina1'}
7	{'peakmoderate', 'elderly', 'F', 'fbsugar0', 'hrnoraml', 'exangina1'}
8	{'peaklow', 'M', 'elderly', 'usloping', 'ncol', 'exangina1'}
9	{'peaklow', 'elderly', 'fbsugar0', 'usloping', 'ncol', 'exangina1'}
10	{'peaklow', 'maged', 'hrhigh', 'lcol', 'fbsugar0', 'exangina1'}
11	{'peaklow', 'hrhigh', 'lcol', 'fbsugar0', 'usloping', 'exangina1'}
12	{'maged', 'M', 'fbsugar1', 'usloping', 'ncol', 'exangina1'}
13	{'peaklow', 'maged', 'hrhigh', 'lcol', 'usloping' ,'exangina1'}

In contrast, our study found that when exercise-induced angina was absent (denoted by a value of 0), no exceptionally rare rules were generated. This absence indicates that the lack of exercise-induced angina does not contribute to significant deviations from the norm and thus does not highlight any unusual patterns or risk factors for cardiovascular disease.

Thus, the presence of exercise-induced angina is a vital factor in our model for detecting rare but critical rules that point to cardiovascular disease. This insight underscores the importance of exercise-induced angina in clinical assessments and highlights its role in the early detection and management of heart disease. The exceptional rules identified by our model, as shown in Table 8.9, provide a deeper understanding of the factors contributing to cardiovascular risk.

8.5.5.6 Presence of fasting blood sugar

The experimental results revealed the crucial function of ('fasting blood sugar ('fbsugar1') in identifying exceptional and unconventional rules that diverge from established norms, especially those associated with frequently occurring rules without heart disease. Notably, all these rare rules apply to women, as shown in the top 10 interesting rules illustrated in Table 8.10. Our findings demonstrate that when fasting blood sugar is present (indicated by a value of 1), 23 of 163 rare rules exhibit a considerable association with cardiovascular disease in otherwise healthy women.

The presence of elevated levels of fasting blood sugar often coincides with other risk factors such as high cholesterol and angina, particularly in women. This correlation underscores the heightened risk of heart disease in the presence of these factors. For example, women with high cholesterol levels and positive fasting blood sugar test results are at a significantly increased risk, especially if they also experience symptoms, such as angina or exercise-induced angina. This highlights the multifaceted nature of cardiovascular risk, in which the interaction between multiple factors compounded the overall risk.

Our analysis showed that these rare rules deviate significantly from conventional norms, effectively identifying female patients at risk for cardiovascular disease. This

Table 8.10: Rare Heart Disease Rules: fast blood suger is present

No	Uncommon rules (heart disease)
1	{'atangina', 'hcol', 'peaklow', 'maged', 'F', 'usloping', 'hrnoraml', 'exangina0', 'fbsugar1'}
2	{'atangina', 'hcol', 'maged', 'F', 'usloping', 'hrnoraml', 'exangina0', 'fbsugar1'}
3	{'atangina', 'hcol', 'peaklow', 'maged', 'F', 'usloping', 'exangina0', 'fbsugar1'}
4	{ 'atangina', 'hcol', 'peaklow', 'maged', 'F', 'hrnoraml', 'exangina0', 'fbsugar1'}
5	{'hcol', 'peaklow', 'maged', 'F', 'usloping', 'hrnoraml', 'exangina0', 'fbsugar1'}
6	{'atangina', 'hcol', 'peaklow', 'F', 'usloping', 'hrnoraml', 'exangina0', 'fbsugar1'}
7	{'atangina', 'hcol', 'F', 'usloping', 'hrnoraml', 'exangina0', 'fbsugar1'}
8	{'atangina', 'hcol', 'peaklow', 'F', 'usloping', 'exangina0', 'fbsugar1'}
9	{'hcol', 'peaklow', 'maged', 'F', 'hrnoraml', 'exangina0', 'fbsugar1'}
10	{'atangina', 'hcol', 'maged', 'F', 'hrnoraml', 'exangina0', 'fbsugar1'}

deviation from frequent patterns signifies a substantial change in health status, indicating a critical shift towards disease when fasting blood sugar levels are high. The presence of high fasting blood sugar, as highlighted by our novel model, is a crucial determinant in uncovering the pivotal factors that contribute to the onset of cardiovascular disease in women. Despite their rarity and deviation from conventional norms, these rules provide essential insights into early diagnosis and intervention.

Conversely, the absence of a positive fasting blood sugar test result (denoted by a value of 0) did not generate any exceptional rare rules. This suggests that normal fasting blood sugar levels do not significantly contribute to deviations from the norm, thereby not highlighting any unusual patterns or risk factors for cardiovascular disease. The absence of this factor indicates a lower-risk profile aligned with conventional medical understanding.

In summary, the presence of fasting blood sugar is a vital factor in our model for detecting rare but critical rules that point to cardiovascular disease in women. This insight underscores the importance of considering fasting blood sugar levels in clinical assessments and highlights their role in the early detection and management of heart disease. The exceptional rules identified by our model provide a deeper understanding of the factors contributing to cardiovascular risk, particularly in female patients.

8.6 Discussion

In our EPFHD-RARMING model, we aim to uncover and highlight rare rules that contradict expectations, thereby leading to remarkable discoveries. Our novel method is successful because it can extract interesting rules from a large number of rules. Within this model, we employ well-established frequent rules as our grounding truth, representing widely accepted beliefs due to their high frequency of co-occurrence.

Our model identifies specific factors that account for the transformation of common rules into rare ones, even with low support. These findings are particularly noteworthy, as demonstrated in our experiments. Several significant factors play a crucial role in the development of heart disease, including ST depression induced by exercise relative to rest (Oldpeak), slope of the peak exercise ST segment (ST Slope), asymptomatic chest pain, low heart rate, exercise-induced angina, and fasting blood sugar. Figure 8.8

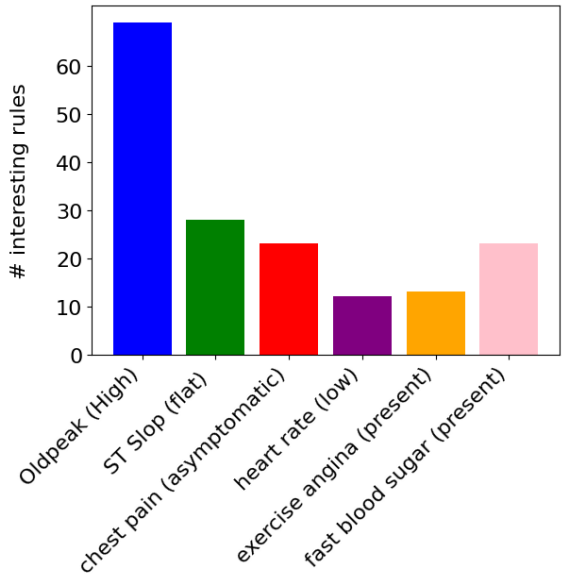


Figure 8.8: Factors contributing to the generation of interesting rules with heart disease as an outcome.

illustrates the factors associated with the generation of unexpected rules which lead to heart disease. The graph shows that a high Oldpeak is the most prominent indicator, followed by a flat ST segment slope and asymptomatic chest pain. Detailed explanations of these six factors are provided in Subsections 8.5.5.1 through 8.5.5.6.

The EPFHD-RARMING model not only unveils these previously unknown associations but also illuminates the intricate interplay of these factors, providing valuable insights into the development of heart disease in otherwise healthy individuals.

The factors identified by our model were further validated by applying multiple feature selection algorithms that consistently identified the same variables as critical contributors. The convergence of methodologies across multiple approaches demonstrates the reliability and robustness of the factors identified in our model. According to our model, the factors contributing to heart disease include **'oldpeak', 'ST slope', 'chest pain type', 'max heart rate', 'exercise angina', and 'fasting blood sugar'**. Figure 8.2 illustrates the key features that contribute to the prediction of heart disease, as identified by multiple feature selection methods. According to the graph, the factors generated by the proposed solution are extremely important. Additionally, a recent study Ozcan and Peker [2023] confirmed the significance of these factors, further attesting to their substantial impact on predictive modeling. This external corroboration strongly supports the accuracy and relevance of the proposed solution.

Our results were compared with those of a recent study Ozcan and Peker [2023] that used a classification and regression tree algorithm for heart disease prediction, focusing primarily on model accuracy. Although their approach identified key risk factors through supervised learning, our novel method leveraged rare rules to analyze unsupervised datasets, emphasizing interpretability and explainability. Unlike the supervised approach, our model uncovered detailed patterns and provided a comprehensive view of the factors leading to heart disease, making our findings more actionable for healthcare professionals. Additionally, our model identifies patterns that may indicate future heart disease development, aiding early detection

and intervention. Our holistic and unsupervised approach makes our method highly adaptable to various domains and offers a more comprehensive understanding of cardiovascular health.

Notably, our model utilizes an unsupervised method, specifically Association Rule Mining (ARM), which enhances the credibility of our findings. The unsupervised nature of the proposed approach underscores the independence and objectivity of the model, allowing it to uncover patterns without predefined labels. This aspect of our model makes it highly adaptable and applicable to various domains beyond health, such as finance, marketing, and other fields in which identifying rare patterns is crucial.

The alignment of our model's findings with established feature selection algorithms, together with supporting evidence from recent studies, provides substantial evidence of the accuracy and correctness of our proposed solution. The utilization of an unsupervised method further emphasizes the ability of the model to autonomously identify and validate crucial factors in the absence of labeled data. In contrast to conventional approaches for identifying factors that play a major role in prediction, our proposed model effectively identifies a diverse set of notable rules, which can be summarized as follows:

- **Frequent rules relating to heart disease:** These rules closely reflect those derived from traditional methodologies, representing well-established rules associated with patients affected by heart disease.
- **Frequent rules facilitating early detection of heart disease:** Among the vast number of rules, our proposed model, EPFHD-RARMING, identified 163 interesting frequent rules that represent healthiness. The identification of these frequent rules that deviate to rare and interesting patterns upon the occurrence of one of the critical factors (such as 'oldpeak', 'ST slope', 'chest pain type', 'max heart rate', 'exercise angina', and 'fasting blood sugar') helps medical experts detect patients who may be at risk of developing heart disease. These vulnerable frequent patterns aid in the early determination of potential heart disease development.
- **Identifying risk factors:** Our model has been successful in identifying risk factors that contribute to the onset of heart disease.

The results of our model should be further investigated by domain experts and tested using additional datasets to fully ascertain its effectiveness and importance. Such validation will help establish its generalizability and potential for broader applications. Thus, we can ensure that the insights provided by our model are robust and reliable, paving the way for its application in real-world scenarios.

Consequently, our proposed EPFHD-RARMING model effectively identifies and prioritizes rare association rules characterized by low support, distinct differences from common rules, shared antecedents, and contrasting outcomes compared to prevalent rules. The identified rules offer critical insights into exceptional cases and unexpected associations within the dataset, thus enhancing our understanding of deviations from typical patterns. Unlike conventional methods, which often overlook

essential risk indicators and fail to capture complex factor relationships, our model leverages the unsupervised ARM approach to conduct a comprehensive analysis. This innovative methodology yields novel insights into cardiovascular health dynamics, identifying patterns indicative of both healthy conditions and potential risks. Hence, our approach represents a substantial advancement, contributing significantly to the deeper comprehension of the intricate mechanisms underlying heart disease.

8.7 Chapter summary

This paper presents EPFHD-RARMING, an unsupervised model designed to identify the key factors leading to heart disease. Unlike traditional supervised methods, our approach uses rare association rule mining to improve the efficiency and specificity of identifying predictive indicators. By defining frequent rules as foundational beliefs, we isolated rare but significant rules that shed light on the onset of heart disease. In addition, our method identifies sensitive frequent rules that correspond to symptoms present in healthy individuals who may develop heart disease if the factors identified by our model are triggered. This predictive capability allows for early intervention. Our approach overcomes the limitations of traditional association rule mining, which often produces a large volume of rules. Instead, EPFHD-RARMING selects a manageable set of 163 rules from thousands, focusing on rare, divergent factors with low support and highlighting unique patterns and deviations within the dataset. This approach confirms the utility of our model in detecting key contributors to heart disease and enhances our understanding of exceptional and unforeseen cases in medical data.

9. Conclusion and Future Work

This section concludes by summarizing the core contributions of our study and providing a detailed outline of potential future research directions. We highlight the significance of our findings and discuss their broader implications in the field of rare pattern mining, specifically focusing on interpretability, performance, redundancy, and interestingness. These contributions offer substantial insight into their potential applicability across diverse domains. Future research will aim to refine our proposed methods to efficiently manage large-scale and complex datasets, integrate them with advanced machine learning techniques, and explore new application areas where rare pattern mining could yield actionable and interpretable insights, particularly in high-stakes real-world contexts.

9.1 Conclusion

In recent years, deep learning achieved remarkable advancements in various fields. However, the interpretability and explainability of these models remain challenging, especially in high-stakes domains, such as healthcare and finance, where understanding a model's decision-making rationale is crucial. To address these challenges, we employed rare pattern mining as a complementary approach, providing insights that are often overlooked by traditional deep learning models. Below, we detail the key contributions of this research, each designed to address specific limitations within rare pattern mining and interpretability.

9.1.1 Efficient discovery of rare patterns

A primary contribution is the development of an algorithm called the Rare Pre Post(RPP) algorithm [Darrab et al., 2020], which efficiently discovers rare patterns and outperforms existing methods, particularly in sparse datasets where rare patterns may otherwise be undetectable. The RPP algorithm addresses the computational challenges of identifying rare patterns and optimizing the retrieval process to recover patterns that are often missed by conventional state-of-the-art techniques. By targeting the unique challenges presented by sparsity, RPP contributes to the field by enhancing the reliability and efficiency of rare pattern mining and by setting a new benchmark for the identification of rare yet meaningful patterns.

9.1.2 Concise representation of rare patterns

Generating a concise representation of rare patterns often involves summarizing the entire set, creating a representation that is significantly smaller yet equally informative. Although methods exist for mining maximal and closed frequent patterns, no method is available for generating concise representations of rare patterns using maximal rare patterns. To address this, we developed the Maximal Rare Itemsets (MaxRI) algorithm [Darrab et al., 2021a] that identifies condensed rare patterns by eliminating redundant and irrelevant information. MaxRI offers a concise view of rare patterns, ensuring a faster discovery process, while reducing the complexity of the results. This contribution establishes a foundation for future research on condensed rare pattern mining, providing a practical tool for applications in which interpretability and data efficiency are critical.

9.1.3 Identification of interesting rare patterns

To reduce noise and irrelevant patterns, we designed methods to discover interesting rare patterns by emphasizing the actionable and interpretable results. This contribution includes the following two models.

- **OPECUR**: An efficient model based on OPTICS clustering of ECLAT algorithm to generate unexpected rules [Darrab et al., 2022a]. The OPECUR model significantly outperformed the existing DBSCAN-based methods in terms of F1-score, AUC, and speed. This performance gain allows quicker and more insightful rule discovery, providing end users with a manageable and useful set of rules.
- **UCRP-Miner**: we introduce UCRP-Miner to retrieve a complete set of unexpected rules [Darrab et al., 2022b]. UCRP-Miner employs frequent patterns to identify unexpected rules, thereby generating actionable insights that are both novel and meaningful. By testing UCRP-Miner on real-world datasets, we demonstrated its ability to outperform the state-of-the-art models in terms of time efficiency and accuracy. In addition, UCRP-Miner produces non-redundant usable patterns, significantly reducing the effort required to discover actionable rules.

These methods collectively provide a robust framework for identifying rare patterns that are relevant and manageable, thereby addressing the key challenges of performance degradation and parameter sensitivity.

9.1.4 Predictive factors of heart disease: a case study on heart disease

A notable application of our method is in the healthcare domain, particularly in identifying factors associated with heart disease. We developed a specialized model, EPFHD-RARMING [Darrab et al., 2024], designed to identify both common and rare factors that may contribute to the onset of heart disease. This model generates interpretable rules that highlight critical factors, support early diagnosis,

and potentially improve health care outcomes. By uncovering patterns with low support, EPFHD-RARMING contributes to a deeper understanding of cardiovascular disease causes, revealing patterns that may otherwise remain hidden yet are essential for preventive care.

9.1.5 Summary

In summary, our dissertation presents several significant contributions that bridge the gap between rare pattern mining and its practical applications in real-world contexts. By addressing issues related to performance, redundancy, and interestingness, our contributions provide actionable insights that can be applied to critical decision-making processes, particularly in domains where interpretability is essential. The patterns identified are not only interpretable and explainable but also contribute valuable, actionable knowledge that enhances the decision-making process across various crucial applications.

9.2 Future Work

Our future work will aim to expand the capabilities of rare pattern mining across several critical areas, thereby improving both its scalability and applicability to diverse data structures and domains. Below, we outline specific directions for future research.

9.2.1 Scalability of rare pattern mining in big data

The primary objective of future research is to adapt rare pattern mining algorithms to handle the vast amounts of data generated in fields such as social media and healthcare. To achieve this, we will utilize the proposed algorithms, RPP and MaxRI, in distributed systems [Kumar and Mohbey, 2022], such as Apache Flink, Spark, and Hadoop. These systems are equipped to manage large-scale data, providing a scalable solution that can efficiently mine rare patterns in high-dimensional datasets, while maintaining performance and interpretability. This approach is intended to enhance the ability of rare pattern mining to process big data, making it applicable in real-world crucial environments.

9.2.2 Mining rare patterns in complex data types

Rare pattern mining can be extended to more complex data types, particularly sequential [Huang et al., 2024] and stream data [Kumar and Mohbey, 2022], where rare patterns hold significant values. Sequential data, often found in bioinformatics, and stream data, common in real-time monitoring applications, pose unique challenges due to the dynamic relationships and temporal dependencies between data points. Developing novel methods that address these challenges will allow rare pattern mining to capture critical insights within these datasets while preserving interpretability. This direction will enable the discovery of meaningful patterns in fields where the structure of the data is as crucial as the data itself.

9.2.3 Integrating causality for enhanced rare pattern discovery

To further improve the discovery and interpretability of rare patterns, we plan to integrate rare pattern mining with causality analysis. By incorporating causal relationships, we can go beyond simple correlations to reveal patterns that suggest potential cause-and-effect dynamics, enhancing the significance and applicability of the results [Nogueira et al., 2022]. This combined approach will yield more meaningful and actionable insights, especially in fields such as healthcare, finance, and the social sciences, where understanding causal mechanisms is essential. In addition, integrating causal inference and machine learning will refine our methodologies, making them suitable for large-scale, complex datasets. This approach will allow rare pattern mining to deliver insights that are not only interpretable but also causally relevant, thus supporting informed decision-making in critical scenarios.

9.2.4 Expanding applicability across diverse domains

Rare pattern mining offers significant potential across a variety of domains where detecting uncommon patterns can yield crucial insights [Akdas et al., 2024; Darrab et al., 2024]. For instance, in healthcare, rare pattern detection can serve as an early warning system for diseases by identifying unusual combinations of symptoms or genetic markers. Beyond healthcare, rare pattern mining can be instrumental in high-risk domains such as finance, where anomalies in transaction data may indicate fraudulent activity.

Each domain presents distinct challenges and demands tailored methodologies; however, the shared advantage of uncovering significant yet infrequent patterns can improve preventative measures, bolster risk management, and enable more precise interventions. For example, in cybersecurity, identifying unusual patterns in network traffic can reveal potential cyber-attacks by highlighting anomalies in logs, network flows, or user behavior. Similarly, in manufacturing, the detection of rare patterns in sensor data or production logs can signal equipment malfunctions or product defects, facilitating proactive maintenance and quality control to prevent costly issues.

9.2.5 Importance of privacy preserving rare pattern mining

The future of research is significantly dependent on ensuring privacy in rare pattern mining [Gui et al., 2024], particularly as data collection expands into sensitive domains such as healthcare, finance, and personal behavior analysis. Although traditional data mining techniques are robust, they frequently compromise sensitive information when identifying and analyzing infrequent patterns, potentially revealing distinctive individual characteristics. Consequently, it is imperative to implement privacy-preserving methodologies in rare pattern mining to ensure data confidentiality and compliance with privacy regulations. Future research should focus on developing advanced cryptographic and anonymization techniques to protect individual privacy while maintaining the accuracy and interpretability of rare pattern discoveries. By integrating privacy measures into rare pattern mining, researchers can extend its application to privacy-sensitive domains without compromising ethical standards or data security, thereby fostering greater trust and facilitating wider adoption in critical sectors such as healthcare, cybersecurity, and personalized services. This approach will enable stakeholders to leverage the full potential of rare pattern insights while adhering to robust data protection standards.

Bibliography

Abdelaziz A Abdelhamid, Marwa M Eid, Mostafa Abotaleb, and SK Towfek. Identification of cardiovascular disease risk factors among diabetes patients using ontological data mining techniques. *Journal of Artificial Intelligence and Metaheuristics*, 4(2):45–53, 2023. (cited on Page 102)

Mehdi Adda, Lei Wu, and Yi Feng. Rare itemset mining. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pages 73–80. IEEE, 2007. (cited on Page 37)

Charu C Aggarwal. *Data mining: the textbook*, volume 1. Springer, 2015. (cited on Page 22 and 23)

Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. *Proceedings of 20th International Conference on Very Large Data Bases, VLDB*, 1215:487–499, 1994. (cited on Page 9, 21, 30, 31, 33, 50, and 78)

Mohiuddin Ahmed and Abu S.S.M. Barkat Ullah. Infrequent pattern mining in smart healthcare environment using data summarization. *Journal of Supercomputing*, 74 (10):5041–5059, 2018. (cited on Page 78)

Devrim Naz Akdas, Derya Birant, and Pelin Yildirim Taser. Erim: An ensemble of rare itemset mining and its application in the automotive industry. *Expert Systems*, 41(6):e13122, 2024. (cited on Page 7, 37, and 136)

Shahad S Aljehani and Youseef A Alotaibi. Preserving privacy in association rule mining using metaheuristic-based algorithms: A systematic literature review. *IEEE Access*, 2024. (cited on Page 90)

Wasif Altaf, Muhammad Shahbaz, and Aziz Guergachi. Applications of association rule mining in health informatics: a survey. *Artificial Intelligence Review*, 47(3): 313–340, 2017. ISSN 15737462. (cited on Page 89)

Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2): 49–60, 1999. (cited on Page 25, 79, and 80)

K Arumugam, Mohd Naved, Priyanka P Shinde, Orlando Leiva-Chauca, Antonio Huaman-Osorio, and Tatiana Gonzales-Yanac. Multiple disease prediction using machine learning algorithms. *Materials Today: Proceedings*, 80:3682–3685, 2023. (cited on Page 102)

MS Arunkumar, P Suresh, and C Gunavathi. High utility infrequent itemset mining using a customized ant colony algorithm. *International Journal of Parallel Programming*, 48(5):833–849, 2020. (cited on Page 40)

Nader Aryabarzan and Behrouz Minaei-Bidgoli. Neclatclosed: A vertical algorithm for mining frequent closed itemsets. *Expert Systems with Applications*, 174:114738, 2021. (cited on Page 93)

John O Awoyemi, Adebayo O Adetunmbi, and Samuel A Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 International Conference on Computing Networking and Informatics (ICCNI)*, pages 1–9. IEEE, 2017. (cited on Page 42)

Anubha Bansal, Neelima Baghel, and Shruti Tiwari. An novel approach to mine rare association rules based on multiple minimum support approach. *International Journal of Advanced Electrical and Electronics Engineering*, 10:75–80, 2013. (cited on Page 34)

Chintan M Bhatt, Parth Patel, Tarang Ghetia, and Pier Luigi Mazzeo. Effective heart disease prediction using machine learning techniques. *Algorithms*, 16(2):88, 2023. (cited on Page 102)

Urvi Y Bhatt and Pratik A Patel. An effective approach to mine rare items using maximum constraint. In *2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO)*, pages 1–6. IEEE, 2015. (cited on Page 19, 30, 37, 39, and 49)

Anindita Borah and Bhabesh Nath. Rare association rule mining: a systematic review. *International Journal of Knowledge Engineering and Data Mining*, 4(3-4): 204–258, 2017. (cited on Page 13 and 14)

Anindita Borah and Bhabesh Nath. Identifying risk factors for adverse diseases using dynamic rare association rule mining. *Expert Systems with Applications*, 113: 233–263, 2018. (cited on Page 43)

Anindita Borah and Bhabesh Nath. Rare pattern mining: challenges and future perspectives. *Complex & Intelligent Systems*, 5:1–23, 2019. (cited on Page 7, 37, 50, and 78)

Anindita Borah and Bhabesh Nath. Rare association rule mining from incremental databases. *Pattern Analysis and Applications*, 23(1):113–134, 2020. (cited on Page 7, 19, 30, 37, 44, 50, and 51)

Christian Borgelt. An implementation of the fp-growth algorithm. In *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, pages 1–5, 2005. (cited on Page 65)

Sergey Brin, Rajeev Motwani, and Craig Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, pages 265–276, 1997. (cited on Page 102)

Danh Bui-Thi, Pieter Meysman, and Kris Laukens. Clustering association rules to build beliefs and discover unexpected patterns. *Applied Intelligence*, 50(6): 1943–1954, 2020. (cited on Page 10, 22, 26, 78, 80, 81, 82, 83, 85, 90, 93, 94, 95, 96, and 97)

Douglas Burdick, Manuel Calimlim, Jason Flannick, Johannes Gehrke, and Tomi Yiu. Mafia: A maximal frequent itemset algorithm. *IEEE transactions on knowledge and data engineering*, 17(11):1490–1504, 2005. (cited on Page 64)

Luca Cagliero and Paolo Garza. Infrequent weighted itemset mining using frequent pattern growth. *IEEE Transactions on Knowledge and Data Engineering*, 26(4): 903–915, 2013. (cited on Page 41)

Raymond Chan, Qiang Yang, and Yi-Dong Shen. Mining high utility itemsets. In *Third IEEE International Conference on Data Mining*, pages 19–19. IEEE Computer Society, 2003. (cited on Page 40)

Chen-Tung Chen and Kai-Yi Chang. A study on the rare itemsets of students' learning effectiveness by using fuzzy data mining. In *2016 International Conference on Advanced Materials for Science and Engineering (ICAMSE)*, pages 703–706. IEEE, 2016. (cited on Page 41)

Yi-Chun Chen, Grace Lin, Ya-Hui Chan, and Meng-Jung Shih. Mining frequent patterns with multiple item support thresholds in tourism information databases. In *Technologies and Applications of Artificial Intelligence: 19th International Conference, TAAI 2014, Taipei, Taiwan, November 21-23, 2014. Proceedings*, pages 89–98. Springer, 2014. (cited on Page 33 and 35)

Yuantao Chen, Runlong Xia, Kai Yang, and Ke Zou. Dnnam: Image inpainting algorithm via deep neural networks and attention mechanism. *Applied Soft Computing*, 154:111392, 2024a. ISSN 1568-4946. (cited on Page 103)

Yuantao Chen, Runlong Xia, Kai Yang, and Ke Zou. Micu: Image super-resolution via multi-level information compensation and u-net. *Expert Systems with Applications*, 245:123111, 2024b. ISSN 0957-4174. (cited on Page 103)

Yanling Cui, Wensheng Gan, Hong Lin, and Weimin Zheng. Fri-miner: fuzzy rare itemset mining. *Applied Intelligence*, pages 1–16, 2022. (cited on Page 40 and 41)

Azzeddine Dahbi, Mohamed Mouhir, Youssef Balouki, and Taoufiq Gadi. Classification of association rules based on K-means algorithm. *Colloquium in Information Science and Technology, CIST*, 0:300–305, 2016. (cited on Page 90)

Sadeq Darrab and Belgin Ergenç. Frequent pattern mining under multiple support thresholds. *WSEAS TRANSACTIONS on COMPUTER RESEARCH*, 10(11): 1–10, 2016. (cited on Page 36 and 90)

Sadeq Darrab and Belgin Ergenc. Vertical pattern mining algorithm for multiple support thresholds. *Procedia computer science*, 112:417–426, 2017. (cited on Page 36)

Sadeq Darrab, David Broneske, and Gunter Saake. Rpp algorithm: a method for discovering interesting rare itemsets. In *Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings 5*, pages 14–25. Springer, 2020. (cited on Page 9, 49, 64, 72, 86, 87, 111, 114, and 133)

Sadeq Darrab, David Broneske, and Gunter Saake. Maxri: A method for discovering maximal rare itemsets. In *2021 4th International Conference on Data Science and Information Technology*, pages 334–341, 2021a. (cited on Page 10 and 134)

Sadeq Darrab, David Broneske, and Gunter Saake. Modern applications and challenges for rare itemset mining. *International Journal of Machine Learning (IJML)*, 11(3):208–218, 2021b. (cited on Page 7, 9, 19, 20, 29, 30, 36, 37, 44, 50, 51, and 78)

Sadeq Darrab, Priyamvada Bhardwaj, David Broneske, and Gunter Saake. Opecur: An enhanced clustering-based model for discovering unexpected rules. In *International Conference on Advanced Data Mining and Applications*, pages 29–41. Springer, 2022a. (cited on Page 10, 90, 93, 94, 96, 97, and 134)

Sadeq Darrab, David Broneske, and Gunter Saake. Ucrp-miner: Mining patterns that matter. In *2022 5th International Conference on Data Science and Information Technology (DSIT)*, pages 1–7. IEEE, 2022b. (cited on Page 10, 27, 89, 106, and 134)

Sadeq Darrab, David Broneske, and Gunter Saake. Exploring the predictive factors of heart disease using rare association rule mining. *Scientific Reports*, 14(1):18178, 2024. (cited on Page 11, 22, 27, 43, 134, and 136)

ZhiHong Deng, ZhongHui Wang, and JiaJian Jiang. A new algorithm for fast mining frequent itemsets using n-lists. *Science China Information Sciences*, 55:2008–2030, 2012. (cited on Page 50, 51, 55, and 72)

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996. (cited on Page 25)

FIMI Repository. Frequent itemset mining dataset repository. http://fimi.uantwerpen.be/data/. Accessed: 2021. (cited on Page 56 and 71)

Philippe Fournier-Viger, Chun-Wei Lin, Alberto Gomariz, Tarek Gueniche, Amir Soltani, Zhi-Hong Deng, and Hoc-Tri Lam. The spmf open-source data mining library version 2. In *Proceedings of the 19th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2016) Part III*, volume 9853 of *Lecture Notes in Computer Science*, pages 36–40. Springer, 2016. (cited on Page 56 and 94)

Philippe Fournier-Viger, Jerry Chun-Wei Lin, Bay Vo, Tin Truong Chi, Ji Zhang, and Hoai Bac Le. A survey of itemset mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(4):e1207, 2017. (cited on Page 13, 30, 105, and 116)

Karam Gouda and Mohammed Javeed Zaki. Efficiently mining maximal frequent itemsets. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 163–170. IEEE, 2001. (cited on Page 63 and 64)

Vikram Goyal, Siddharth Dawar, and Ashish Sureka. High utility rare itemset mining over transaction databases. In *Databases in Networked Information Systems: 10th International Workshop, DNIS 2015, Aizu-Wakamatsu, Japan, March 23-25, 2015. Proceedings 10*, pages 27–40. Springer, 2015. (cited on Page 40)

Yijie Gui, Wensheng Gan, Yongdong Wu, and S Yu Philip. Privacy preserving rare itemset mining. *Information Sciences*, page 120262, 2024. (cited on Page 7, 19, 30, 44, and 136)

Gunjan K. Gupta, Alexander Strehl, and Joydeep Ghosh. Distance based clustering of association rules. *Intelligent Engineering Systems Through Artificial Neural Networks*, 9:759–764, 1999. (cited on Page 78)

Manoj Kumar Gupta and Pravin Chandra. A comprehensive survey of data mining. *International Journal of Information Technology*, 12(4):1243–1257, 2020. (cited on Page 33)

Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. *ACM sigmod record*, 29(2):1–12, 2000. (cited on Page 30, 31, 38, 50, 111, and 114)

Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8(1):53–87, 2004. (cited on Page 80)

Jiawei Han, Jian Pei, and Hanghang Tong. *Data mining: concepts and techniques*. Morgan kaufmann, 2022. (cited on Page 29)

Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74, 2024. (cited on Page 13, 29, and 101)

Zengyou He, Xiaofei Xu, Joshua Zhexue Huang, and Shengchun Deng. A frequent pattern discovery method for outlier detection. In *Advances in Web-Age Information Management: 5th International Conference, WAIM 2004, Dalian, China, July 15-17, 2004 5*, pages 726–732. Springer, 2004. (cited on Page 42)

Ya-Han Hu and Yen-Liang Chen. Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism. *Decision support systems*, 42(1):1–24, 2006. (cited on Page 35)

David Huang, Yun Sing Koh, and Gillian Dobbie. Rare pattern mining on data streams. In *Data Warehousing and Knowledge Discovery: 14th International Conference, DaWaK 2012, Vienna, Austria, September 3-6, 2012. Proceedings 14*, pages 303–314. Springer, 2012. (cited on Page 41 and 43)

Gengsen Huang, Wensheng Gan, and Philip S Yu. Taspm: Targeted sequential pattern mining. *ACM Transactions on Knowledge Discovery from Data*, 18(5): 1–18, 2024. (cited on Page 135)

László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 245–251. IEEE, 2013. (cited on Page 78)

Yanqing Ji, Hao Ying, John Tran, Peter Dews, Ayman Mansour, and R Michael Massanari. A method for mining infrequent causal associations and its application in finding adverse drug reaction signal pairs. *IEEE transactions on Knowledge and Data Engineering*, 25(4):721–733, 2012. (cited on Page 49)

Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, and Preeti Nagrath. Heart disease prediction using machine learning algorithms. In *IOP Conference Series: Materials Science and Engineering*, volume 1022, page 012072. IOP Publishing, 2021. (cited on Page 102 and 104)

Sisir Joshi, Abeer Alsadoon, S. M.N.Arosha Senanayake, P. W.C. Prasad, Shiaw Yin Yong, Amr Elchouemi, and Trung Hung Vo. *Pattern mining predictor system for road accidents*, volume 1287. Springer International Publishing, 2020. (cited on Page 89)

Sujatha Kamepalli and Srinivasa Bandaru. Weighted based frequent and infrequent pattern mining model for real-time e-commerce databases. *Advances in Modelling and Analysis B*, 62(2-4):53–60, 2019. ISSN 12404543. doi: 10.18280/ama$_$b.622-404. (cited on Page 90)

Ramdas Kapila, Thirumalaisamy Ragunathan, Sumalatha Saleti, T. Jaya Lakshmi, and Mohd Wazih Ahmad. Heart disease prediction using novel quine mccluskey binary classifier (qmbc). *IEEE Access*, 11:64324–64347, 2023. (cited on Page 104)

Mohak Kataria, C Oswald, and B Sivaselvan. A novel rare itemset mining algorithm based on recursive elimination. In *Software Engineering: Proceedings of CSI 2015*, pages 221–233. Springer, 2019. (cited on Page 19 and 30)

Rahul Katarya and Sunit Kumar Meena. Machine learning techniques for heart disease prediction: a comparative study and analysis. *Health and Technology*, 11(1):87–97, 2021. (cited on Page 104)

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The uci machine learning repository. https://archive.ics.uci.edu. Accessed: 2022. (cited on Page 82 and 94)

Ahmed M Khedr, Zaher Al Aghbari, Amal Al Ali, and Mariam Eljamil. An efficient association rule mining from distributed medical databases for predicting heart diseases. *IEEE Access*, 9:15320–15333, 2021. (cited on Page 105)

Rage Uday Kiran and P Krishna Reddy. An improved multiple minimum support based approach to mine rare association rules. In *2009 IEEE Symposium on Computational Intelligence and Data Mining*, pages 340–347. IEEE, 2009. (cited on Page 34)

Rage Uday Kiran and P Krishna Reddy. Novel techniques to reduce search space in multiple minimum supports-based frequent pattern mining algorithms. In

Proceedings of the 14th International Conference on Extending Database Technology, pages 11–20, 2011. (cited on Page 35)

Yun Sing Koh and Sri Devi Ravana. Unsupervised rare pattern mining: a survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(4):1–29, 2016. (cited on Page 20, 30, and 37)

Yun Sing Koh and Nathan Rountree. Finding sporadic rules using apriori-inverse. In *Advances in Knowledge Discovery and Data Mining: 9th Pacific-Asia Conference, PAKDD 2005, Hanoi, Vietnam, May 18-20, 2005. Proceedings 9*, pages 97–106. Springer, 2005. (cited on Page 38 and 64)

Sunil Kumar and Krishna Kumar Mohbey. A review on big data based parallel and distributed approaches of pattern mining. *Journal of King Saud University-Computer and Information Sciences*, 34(5):1639–1662, 2022. (cited on Page 135)

Chan Man Kuok, Ada Fu, and Man Hon Wong. Mining fuzzy association rules in databases. *ACM Sigmod Record*, 27(1):41–46, 1998. (cited on Page 40)

K Prasanna Lakshmi and CRK Reddy. Fast rule-based heart disease prediction using associative classification mining. In *2015 International Conference on Computer, Communication and Control (IC4)*, pages 1–5. IEEE, 2015. (cited on Page 105)

Yeong-Chyi Lee, Tzung-Pei Hong, and Wen-Yang Lin. Mining association rules with multiple minimum supports using maximum constraints. *International Journal of Approximate Reasoning*, 40(1-2):44–54, 2005. (cited on Page 34)

Brian Lent, Arun Swami, and Jennifer Widom. Clustering association rules. *Proceedings - International Conference on Data Engineering*, pages 220–231, 1997. (cited on Page 78 and 90)

Bing Liu, Wynne Hsu, and Yiming Ma. Mining association rules with multiple minimum supports. In *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 337–341, 1999. (cited on Page 19, 33, 36, and 51)

Ruilin Liu, Kai Yang, Yanjia Sun, Tao Quan, and Jin Yang. Spark-based rare association rule mining for big datasets. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 2734–2739. IEEE, 2016. (cited on Page 39 and 42)

Sainan Liu and Haoan Pan. Rare itemsets mining algorithm based on rp-tree and spark framework. In *AIP Conference Proceedings*, volume 1967. AIP Publishing, 2018. (cited on Page 39 and 43)

Shengjie Liu, Chunjie Zhou, Jialong Li, and Xiaoyun Lu. Rare association rule mining based on reinforcement learning. In *Proceedings of the 2023 12th International Conference on Computing and Pattern Recognition*, pages 79–86, 2023. (cited on Page 7)

Yifeng Lu, Florian Richter, and Thomas Seidl. Efficient infrequent pattern mining using negative itemset tree. *Complex Pattern Mining: New Challenges, Methods and Applications*, pages 1–16, 2020. (cited on Page 13, 14, 19, 20, 30, 36, 51, and 64)

José María Luna, Philippe Fournier-Viger, and Sebastián Ventura. Frequent itemset mining: A 25 years review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1329, 2019. (cited on Page 5, 6, 15, 16, 63, and 102)

M Marimuthu, M Abinaya, KS Hariesh, K Madhankumar, and V Pavithra. A review on heart disease prediction using machine learning and data analytics approach. *International Journal of Computer Applications*, 181(18):20–25, 2018. (cited on Page 104)

Xifeng Mi. The mining algorithm of maximum frequent itemsets based on frequent pattern tree. *Computational Intelligence and Neuroscience*, 2022(1):7022168, 2022. (cited on Page 64)

Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7:81542–81554, 2019. (cited on Page 104)

Pranav Motarwar, Ankita Duraphe, G Suganya, and M Premalatha. Cognitive approach for heart disease prediction using machine learning. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pages 1–5. IEEE, 2020. (cited on Page 104)

Shalini Zanzote Ninoria and SS Thakur. Review on high utility rare itemset mining. In *Social Networking and Computational Intelligence: Proceedings of SCI-2018*, pages 373–388. Springer, 2020. (cited on Page 13 and 14)

Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. Methods and tools for causal discovery and causal inference. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(2):e1449, 2022. (cited on Page 136)

Mert Ozcan and Serhat Peker. A classification and regression tree algorithm for heart disease modeling and prediction. *Healthcare Analytics*, 3:100130, 2023. (cited on Page 129)

Francisco Padillo, José María Luna, and Sebastián Ventura. Mining perfectly rare itemsets on big data: an approach based on apriori-inverse and mapreduce. In *Intelligent Systems Design and Applications: 16th International Conference on Intelligent Systems Design and Applications (ISDA 2016) held in Porto, Portugal, December 16-18, 2016*, pages 508–518. Springer, 2017. (cited on Page 38)

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. (cited on Page 85)

Jian Pei, Jiawei Han, Hongjun Lu, Shojiro Nishio, Shiwei Tang, and Dongqing Yang. H-mine: Hyper-structure mining of frequent patterns in large databases. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 441–448. IEEE, 2001. (cited on Page 30)

Jyothi Pillai, OP Vyas, and Maybin Muyeba. Huri - a novel algorithm for mining high utility rare itemsets. In *Advances in Computing and Information Technology: Proceedings of the Second International Conference on Advances in Computing and Information Technology (ACITY) July 13-15, 2012, Chennai, India-Volume 2*, pages 531–540. Springer, 2013. (cited on Page 40)

Saeed Piri, Dursun Delen, Tieming Liu, and William Paiva. Development of a new metric to identify rare patterns in association analysis: The case of analyzing diabetes complications. *Expert Systems with Applications*, 94:112–125, 2018. (cited on Page 43)

Ansel Y Rodríguez-González, Fernando Lezama, Carlos A Iglesias-Alvarez, José Fco Martínez-Trinidad, Jesús A Carrasco-Ochoa, and Enrique Munoz de Cote. Closed frequent similar pattern mining: Reducing the number of frequent similar patterns without information loss. *Expert Systems with Applications*, 96:271–283, 2018. (cited on Page 16)

Cristóbal Romero, José Raúl Romero, Jose María Luna, and Sebastián Ventura. Mining rare association rules from e-learning data. In *Educational Data Mining 2010*. ERIC, 2010. (cited on Page 44)

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215, 2019. (cited on Page 5)

Heungmo Ryang, Unil Yun, and Keun Ho Ryu. Discovering high utility itemsets with multiple minimum supports. *Intelligent Data Analysis*, 18(6):1027–1047, 2014. (cited on Page 40)

Kanimozhi SC Sadhasivam and Tamilarasi Angamuthu. Mining rare itemset with automated support thresholds. *Journal of Computer Science*, 7(3):394, 2011. (cited on Page 37)

KR Seeja and Masoumeh Zareapoor. Fraudminer: A novel credit card fraud detection model based on frequent itemset mining. *The Scientific World Journal*, 2014, 2014. (cited on Page 42)

S Selvarani and M Jeyakarthic. Rare itemsets selector with association rules for revenue analysis by association rare itemset rule mining approach. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, 14(7):2335–2344, 2021. (cited on Page 36 and 51)

Sunidhi Shrivastava and Punit Kumar Johari. Analysis on high utility infrequent itemsets mining over transactional database. In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 897–902. IEEE, 2016. (cited on Page 13 and 14)

Manu Siddhartha. Heart disease dataset (comprehensive). ieee dataport. https://dx.doi.org/10.21227/dz4t-cm36, 2020. Accessed: 2023. (cited on Page 105)

Archana Singh and Rakesh Kumar. Heart disease prediction using machine learning algorithms. In *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, pages 452–457. IEEE, 2020. (cited on Page 11)

Avadh Kishor Singh, Ajeet Kumar, and Ashish K Maurya. An empirical analysis and comparison of apriori and fp-growth algorithm for frequent pattern mining. In *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, pages 1599–1602. IEEE, 2014. (cited on Page 27 and 90)

KM Mehedi Hasan Sonet, Md Mustafizur Rahman, Pritom Mazumder, Abid Reza, and Rashedur M Rahman. Analyzing patterns of numerously occurring heart diseases using association rule mining. In *2017 twelfth International Conference on Digital Information Management (ICDIM)*, pages 38–45. IEEE, 2017. (cited on Page 105)

Laszlo Szathmary, Amedeo Napoli, and Petko Valtchev. Towards rare itemset mining. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, volume 1, pages 305–312. IEEE, 2007. (cited on Page 19, 30, and 38)

Laszlo Szathmary, Petko Valtchev, Amedeo Napoli, and Robert Godin. Efficient vertical mining of minimal rare itemsets. In *CLA*, pages 269–280. Citeseer, 2012. (cited on Page 37)

Wiem Taktak and Yahya Slimani. Ms-fp-growth: A multi-support vrsion of fp-growth agorithm. *International Journal of Hybrid Information Technology*, 7(3):155–166, 2014. (cited on Page 35)

Akbar Telikani, Amir H Gandomi, and Asadollah Shahbahrami. A survey of evolutionary computation for association rule mining. *Information Sciences*, 524: 318–352, 2020. (cited on Page 33)

Caroline V Tew, Christophe G Giraud-Carrier, K Tanner, and Scott H Burton. Behavior-based clustering and analysis of interestingness measures for association rule mining. *Data Mining and Knowledge Discovery*, 28:1004–1045, 2014. (cited on Page 23)

Sofya S Titarenko, Valeriy N Titarenko, Georgios Aivaliotis, and Jan Palczewski. Fast implementation of pattern mining algorithms with time stamp uncertainties and temporal constraints. *Journal of Big Data*, 6(1):37, 2019. (cited on Page 1 and 33)

Hannu. Toivonen, Mika. Klemettinen, Pirjo. Ronkainen, Kimmo. Hätönen, and Heikki. Mannila. Pruning and grouping discovered association rules. *Workshop Notes of ECML Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases*, pages 47–52, 1995. (cited on Page 78 and 90)

Luigi Troiano and Giacomo Scibelli. A time-efficient breadth-first level-wise lattice-traversal algorithm to discover rare itemsets. *Data Mining and Knowledge Discovery*, 28:773–807, 2014. (cited on Page 30)

Luigi Troiano, Giacomo Scibelli, and Cosimo Birtolo. A fast algorithm for mining rare itemsets. In *2009 ninth International Conference on Intelligent Systems Design and Applications*, pages 1149–1155. IEEE, 2009. (cited on Page 38)

AR Troncoso-García, María Martínez-Ballesteros, Francisco Martínez-Álvarez, and Alicia Troncoso. A new approach based on association rules to add explainability to time series forecasting models. *Information Fusion*, 94:169–180, 2023. (cited on Page 5)

Sidney Tsang, Yun Sing Koh, and Gillian Dobbie. Rp-tree: rare pattern tree mining. In *Data Warehousing and Knowledge Discovery: 13th International Conference, DaWaK 2011, Toulouse, France, August 29-September 2, 2011. Proceedings 13*, pages 277–288. Springer, 2011. (cited on Page 19, 20, 30, 38, 49, 50, and 64)

Kuladeep Tummala, C Oswald, and B Sivaselvan. A frequent and rare itemset mining approach to transaction clustering. In *Data Science Analytics and Applications: First International Conference, DaSAA 2017, Chennai, India, January 4-6, 2017, Revised Selected Papers 1*, pages 8–18. Springer, 2018. (cited on Page 38)

Sunitha Vanamala, L Padma Sree, and S Durga Bhavani. Eclat_rpgrowth: finding rare patterns using vertical mining and rare pattern tree. In *Computer Networks, Big Data and IoT: Proceedings of ICCBI 2020*, pages 161–176. Springer, 2021. (cited on Page 19 and 30)

Chen-Shu Wang and Jui-Yen Chang. Misfp-growth: Hadoop-based frequent pattern mining with multiple item support. *Applied Sciences*, 9(10):2075, 2019. (cited on Page 36)

Ke Wang, Liu Tang, Jiawei Han, and Junqiang Liu. Top down fp-growth for association rule mining. In *Advances in Knowledge Discovery and Data Mining: 6th Pacific-Asia Conference, PAKDD 2002 Taipei, Taiwan, May 6–8, 2002 Proceedings 6*, pages 334–340. Springer, 2002. (cited on Page 9)

Wei Wang, Jiong Yang, and Philip S Yu. Efficient mining of weighted association rules (war). In *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 270–274, 2000. (cited on Page 41)

Gary M Weiss. Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter*, 6(1):7–19, 2004. (cited on Page 49)

Cheng-Hsiung Weng. Mining fuzzy specific rare itemsets for education data. *Knowledge-Based Systems*, 24(5):697–708, 2011. (cited on Page 41 and 44)

World Health Organization. Cardiovascular diseases, 2021. URL https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1. (cited on Page 102)

Cheng-Wei Wu, JianTao Huang, Yun-Wei Lin, Chien-Yu Chuang, and Yu-Chee Tseng. Efficient algorithms for deriving complete frequent itemsets from frequent closed itemsets. *Applied Intelligence*, pages 1–22, 2022. (cited on Page 64)

Xiaoying Wu, Dimitri Theodoratos, and Nikos Mamoulis. Discovering closed and maximal embedded patterns from large tree data. *Data & Knowledge Engineering*, 133:101890, 2021. (cited on Page 16)

Tiantian Xu and Xiangjun Dong. Mining frequent patterns with multiple minimum supports using basic apriori. In *2013 Ninth International Conference on Natural Computation (ICNC)*, pages 957–961. IEEE, 2013. (cited on Page 33 and 34)

Chetna Yadav, Shrikant Lade, and Manish K Suman. Predictive analysis for the diagnosis of coronary artery disease using association rule mining. *International Journal of Computer Applications*, 87(4), 2014. (cited on Page 105)

Guizhen Yang. The complexity of mining maximal frequent itemsets and maximal frequent patterns. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 344–353, 2004. (cited on Page 17)

Huazhong Yang, Zhongju Chen, Huajian Yang, and Maojin Tian. Predicting coronary heart disease using an improved lightgbm model: performance analysis and comparison. *IEEE Access*, 11:23366–23380, 2023. (cited on Page 104)

A Yashudas, Dinesh Gupta, GC Prashant, Amit Dua, Dokhyl AlQahtani, and A Siva Krishna Reddy. Deep-cardio: recommendation system for cardiovascular disease prediction using iot network. *IEEE Sensors Journal*, 2024. (cited on Page 102 and 104)

Mohammed J. Zaki. Generating non-redundant association rules. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, page 34–43, 2000a. (cited on Page 24 and 78)

Mohammed Javeed Zaki. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390, 2000b. (cited on Page 30)

Mohammed Javeed Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, and Wei Li. New algorithms for fast discovery of association rules. *3rd international Conference on Knowledge Discovery and Data Mining*, pages 283–286, 1997. (cited on Page 80)

Shichao Zhang and Xindong Wu. Fundamentals of association rules in data mining and knowledge discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(2):97–116, 2011. (cited on Page 13)

I herewith assure that I wrote the present thesis independently, that the thesis has not been partially or fully submitted as graded academic work and that I have used no other means than the ones indicated. I have indicated all parts of the work in which sources are used according to their wording or to their meaning.

Magdeburg, April 22, 2025