The Clustering and Fuzzy Logic Methods Complex for Big Data Processing

Larysa Globa, Rina Novogrudska and Andrii Liashenko
Igor Sikorsky Kyiv Polytechnic Institute, 37 Peremohy avenue, Kyiv, Ukraine
lgloba@its.kpi.ua, rinan@ukr.net

Keywords: Fuzzy Logic, Clustering Algorithms, Smart System, Statistical Numerical Data, Fuzzy Knowledge Bases, Fuzzy Logical Rules.

Abstract: Currently, telecom operators are facing a problem that is conditionally called "Big Data". The telecom industry is growing rapidly and dynamically, new technologies are emerging (IoT, M2M, D2D, P2P), new companies are using them, new information and communication services are being introduced to automate production processes, and so on. Methods of statistical analysis, A\B testing, data fusion and integration, Data Mining, machine learning, data visualization are used in the Big Data processing and analysis, but due to the fact that large amounts of Big Data are not structured, come in real-time with various delays related to bandwidth and network congestion, in each case the processes of processing and analysis of Big Data are extremely costly in terms of time and resources. As a result, telecom operators need not only to process large amounts of data but also to extract knowledge from them. However, the analytical processing of large data is characterized by blurred boundaries, which determine certain logical relationships between data. This study proposes the flexible complex of clustering and fuzzy logic methods for big data processing, which increases the speed and reliability of their processing in network nodes, as well as an architectural solution for analysis and processing Big Data realization using micro-services, which increases system scalability and reduces the load on the servers that process them. Experimental studies have confirmed the effectiveness of the proposed modifications. Studies of the K-means algorithm when processing 1500 rows in 3 columns showed decreasing in execution time by 2 seconds. Studies of the Fuzzy C-means algorithm have shown a reduction in execution time by almost 2 times. The validity of the developed fuzzy knowledge base for the K-means and fuzzy C-means algorithms increased by 9% and 4%, respectively.

1 INTRODUCTION

The rate estimation of in data volume increasing in communication networks is determined by the following trends: population growth [1], increasing of the number of mobile users, the number of Internet users and the number of social network users. These trends are driving an ever-increasing amount of content and data in the digital space. According to the source [2], the amount of data generated every second is more than 30,000 gigabytes. At the same time, improving the network infrastructure of information platforms for the provision of modern digital services is quite a complex and time-consuming and costly task. The state of the current infrastructure of telecom services requires the knowledge extraction from statistical data sets to effectively support and process significant amounts of information from both users and from various services. Currently, analytics [3] are used to obtain some knowledge from statistical data sets, but due to the fact that large amounts of big data are not structured, they come in real time with various delays associated with bandwidth and network congestion, simple statistical analysis of data is not enough. It is necessary to apply a set of methods that would allow to process, analyze data and form knowledge from them, using different modifications of clustering algorithms, to form logically connected groups of data that define logical dependencies. Choosing the right clustering algorithm is an important step in this process. K-means clustering algorithm is one of the most popular and simplest clustering technologies used in practice [4,5,6]. But the usual K-means algorithm has problems with the accuracy of cluster center selection. This algorithm requires solving the

problem of initialization of cluster centers and finding the right number of clusters [7].

At the same time, Big Data is characterized by fuzzy and requires the participation of experts during their analysis. Based on this, it is proposed to analyze them using fuzzy logic, forming fuzzy logical statements (rules) such as *IF… AND… THAN*, which are best for human perception. However, it is not possible to use a single method or algorithm to develop fuzzy logic rules. This process requires the creation of reliable sets of clusters from which fuzzy logical rules are formed, and to achieve this it is necessary to use clustering algorithms, construction of membership functions, formation of fuzzy logical rules for the transition from numerical data to logical statements.

The stages of developing the system for the formation of sets of fuzzy logical rules (fuzzy knowledge base) are determined by the following stages: expert estimation - loading of pre-cleared and structured statistics, the transition from data clusters to membership and union functions, and hence the transformation of numerical values into terms of linguistic variables (formation of fuzzy rules), checking the correctness of the model. In addition, the system for forming a fuzzy knowledge base should be flexible, take into account the peculiarities of data flows and increase the efficiency (speed and reliability) of processing large amounts of data. In this study, this is achieved by using a set of methods for constructing fuzzy logical rules based on clustering algorithms that focus on the features of statistical data sets processed in telecommunication systems, as well as architectural solutions for development Big Data analysis and processing using microservices.

The paper is structured as follows: Section 2 contains state of art analysis of Big Data processing methods problem. Section 3 explains the approach to be solved by proposed flexible complex of clustering and fuzzy logic methods for Big Data processing. Section 4 introduces the approach for the fuzzy knowledge base development. Section 5 presents the approach for the architecture development of the system based on fuzzy knowledge base. Section 6 shows efficiency of the proposed flexible complex of clustering and fuzzy logic methods for Big Data processing usage. Section 7 includes the summary and outlook on future work.

2 STATE OF THE ART AND BACKGROUND

Features of Big Data such as the speed of data generation, the complexity of their analysis has necessitated the use of machine learning and artificial intelligence. Along with the evolution of data analysis computer methods, their analysis is also based on traditional statistical methods. Processing and analysis of Big Data is performed in the conditions of streaming data as they appear, and then apply various methods of data analysis as they are created, to find behavioral patterns and trends. As the amount of data increases, so are developed the methods used to process it.

Methods of Big Data processing and appropriate tools analytics are classified [8]:

1) A / B testing - these data analysis tools involve comparing the control group with different test groups to determine which ways to influence or change will improve a given objective variable.

2) Data merging and integration is a common tool used in Big Data analytics. Different data from different sources are integrated together and data analysis is performed by combining methods of statistics and machine learning in database management. An example of such an analysis in the telecom industry is the collection of customer data to determine which customers are most likely to a proposal respond.

3) Data Mining - a method of data analysis designed to search for previously unknown patterns in large arrays of information. These patterns make it possible to make effective management decisions and optimize business processes. Data Mining methods include teaching associative rules, classification, cluster analysis, regression analysis, detection and analysis of deviations, and more.

4) Machine learning - methods of artificial intelligence, which are a method of data analysis that allows the machine, robot or analytical system to conduct independent learning by performing a group of similar tasks.

5) Artificial neural networks are mathematical models, as well as their software or hardware implementation, built on the principle of organization and functioning of biological neural networks - networks of nerve cells of a living organism.

6) Visualization of analytical data - a means of presenting statistical information in a form that is better perceived by humans, in the form of diagrams, drawings using animation to determine the results of the expert or for further analysis.

All these methods of processing Big Data and extracting patterns from them are characterized by the fact that they depend on the structure and type of data, for example, if signals can be analyzed visually, then such methods are not effective for other data. Since searching for previously unknown patterns in large amounts of information provides some knowledge about the behavior of a system from sets of statistical data, these methods are promising for the telecommunications industry.

In recent years, cluster analysis is widely used in Data Mining as one of the main methods [9, 4]. The purpose of cluster analysis is to assign to objects to homogeneous data sets (clusters). Assigning objects is done so that the objects are similar to each other in one cluster and different in others. The method of summarizing the observed data in clusters is determined on the basis of statistical information that can be stored in database tables or files, because at the beginning of the study there is no prior knowledge.

There are many clustering algorithms, the paper [10] proposes the structure of algorithm categorization. Different clustering algorithms can be broadly classified as follows:

Separate algorithms: in such algorithms all clusters are defined quickly. The initial groups are specified and redistributed into associations. In other words, distribution algorithms divide data objects into several partitions, where each partition is a cluster. These clusters must meet the following requirements:

- each group must contain at least one object,
- each object must belong to exactly one group.

For example, in the K-means algorithm, the center is the mean of all points and coordinates, which is the arithmetic mean. There are many other partitioning algorithms such as K-modes, PAM, CLARA, CLARANS and FCM.

Hierarchical algorithms: data is organized in a hierarchical way depending on proximity. Proximity defines intermediate nodes. The initial cluster is gradually divided into several clusters as the hierarchy continues. The process continues until the stop criterion is reached (often determined by the number of k clusters). However, the hierarchical method has a serious drawback, which is that once a step (merger or division) is performed, it cannot be undone. BIRCH, CURE, ROCK and Chameleon are some of the well-known algorithms in this category.

Density Based: Here, data objects are divided based on their density, connectivity, and boundary areas. A cluster is an associated dense component that grows in any direction that determines density. Density-based algorithms are able to detect clusters of arbitrary shape. The total data point density is analyzed to determine the functions of the data sets that affect its assignment to a particular cluster. DBSCAN, OPTICS, DBCLASD and DENCLUE are algorithms that use this method to filter noise and detect arbitrary clusters.

Grid-based: The space of data objects is divided into grids. The main advantage of this approach is its fast-processing time, because the data set is passed once to calculate statistical values for grids. Collected grid data makes grid-based clustering methods independent of the data objects number that use a single grid to collect specific statistics and then perform clustering in the grid rather than directly in the database. The performance of a grid-based method depends on the size of the grid, which is usually much smaller than the size of the database. However, for very irregular data distributions, the use of a single homogeneous grid may not be sufficient to obtain the required clustering quality or to meet processing time requirements. Wave-Cluster and STING are typical examples of this category.

When defining clustering methods, it is necessary to use specific criteria to assess the relative strengths and weaknesses of each algorithm in relation to the multidimensional properties of Big Data, the most important of which are volume, speed and diversity.

Volume: refers to the ability of a clustering algorithm to work with large amounts of data. To control the selection of the appropriate clustering algorithm for the Volume property, the following criteria are considered:

- data set size;
- high dimensional processing;
- processing of emissions / noisy data.

Variety: refers to the ability of a clustering algorithm to process different types of data (numerical, categorical, and hierarchical). To control the selection of the appropriate clustering algorithm for the Variety property, a criterion such as data set type is considered.

Velocity: refers to the clustering algorithm speed during the big data processing. To control the selection of the appropriate clustering algorithm for the Velocity property, the following criteria are considered:

- complexity of the algorithm;
- performance at runtime.

Clustering algorithms perform effectively with either numerical data or categorical data; most of them do not process well mixed categorical and

numerical data types, with large size and data set, and in case of errors. As the number of measurements increases, the data becomes sparser, so measuring the distance between pairs of points becomes impractical, and the average density of points anywhere in the data is likely to be low. The process of clustering large amounts of data takes too much time, which can be impractical. The K-Means and FCM (Fuzzy C-Means) algorithms are among the most efficient algorithms that meet the requirements for processing large amounts of numerical data in telecom systems. FCM clustering algorithm is more flexible than K-Means algorithm, it allows to form a base of fuzzy logical rules for business processes of telecom operators.

This research focuses on creating a set of clustering methods and fuzzy logic for Big Data processing, which take into account the peculiarities of statistical data flows, increase the speed and reliability of their processing in network nodes, and on developing the architecture of Big Data analysis and processing using micro-services. This increases the scalability of the system and reduces the load on the servers that perform their processing.

3 THE APPROACH TO DEVELOP THE FLEXIBLE COMPLEX OF CLUSTERING AND FUZZY LOGIC METHODS FOR BIG DATA PROCESSING

The Fuzzy C-means algorithm (FCM) is a separate algorithm, like the K-means algorithm. The main difference is that a point can belong to all centers of clusters, but with its degree of affiliation, which can range from 0 to 1. The higher the degree, the more probably it is that the object belongs to that cluster. The FCM clustering algorithm is more flexible than the K-Means algorithm because it uses some ambiguity as the value of the membership function. This allows to determine whether a sample object belongs to clusters with different degrees of membership without losing existing logical connections on cluster boundaries. This gives the possibility to realize the transition to fuzzy logical statements (fuzzy knowledge) [4, 11].

FCM algorithm convergence criteria:

For each element of the measurements sample, the sum of the its belonging degrees to the clusters should be equal to:

$$\sum_{i=1}^{c} \mu_{ij} = 1, \qquad \forall j = 1, \ldots . N,$$

The value of the affiliation degree is limited by the interval [0,1]:

$$\mu_{ij} \in [0,1], \forall i = 1, \ldots ., c, \;\; \text{i} \;\; \forall j = 1, \ldots . N$$

FCM clustering is performed by minimizing the objective function (1):

$$J = \sum_{i=1}^{n} \sum_{k=1}^{c} \mu_{ik}^{q} \, |x_i - v_k|^2, \qquad (1)$$

where

J – the objective function,
n – the number of objects in the data sample,
c – number of clusters,
μ – fuzzy membership value from the table,
q – fuzzy coefficient (value > 1),
x_i– the value of the *i-th* object in the sample,
v_k– the cluster center,
$||x_i - v_k|$-Euclidean distance determined by (2):

$$||x_i - v_k| = \sqrt{\sum_{i=1}^{n} (x_i - v_k)^2}. \qquad (2)$$

The calculation of the cluster center is determined by (3):

$$v_k = \frac{\sum_{i=1}^{n} \mu_{ik}^{q} x_i}{\sum_{i=1}^{n} \mu_{ik}^{q}}. \qquad (3)$$

The fuzzy membership table is calculated using (4):

$$\mu_{ik} = \frac{1}{\sum_{l=1}^{c} \left(\frac{|x_i - v_k|}{|x_i - v_l|}\right)^{\frac{2}{q-1}}}. \qquad (4)$$

Implementation steps:

Step 1: Setting the number of clusters, fuzzy parameter (*constant value> 1*), and stop parameter.

Step 2: Initializing the matrix of affiliation degrees.

Step 3: Setting the cycle counter *k = 0.*

Step 4: Calculating the centroids of the cluster, calculating the value of the objective function *J.*

Step 5: For each object and for each cluster calculating the value of membership in the matrix.

Step 6: If the value of *J* between successive iterations is less than the stop condition, then stop; otherwise set *k = k + 1* and go to step 4.

Step 7: Obtaining the membership matrix and the end of the algorithm.

However, this algorithm has significant shortcomings, namely the initial initialization of cluster centers and the correct number of clusters can affect the accuracy of the fuzzy logic rules development and subsequently on the construction of fuzzy knowledge bases [10].

Since the initial centers of clusters have a strong influence on obtaining the final sets of clusters, the

process of their formation depends on the choice of starting points as the initial centers of clusters [12]. As a rule, the initial centers of clusters are selected randomly, which directly affects the accuracy of cluster construction. If a cluster center is initialized as a "remote" point, it may simply not have associated points, and more than one cluster may be associated with a single cluster center. Similarly, more than one centroid can be initialized in one cluster, which will lead to poor clustering. Correct calculation of the initial centers of clusters allows to obtain more accurate and efficient groups of clusters, as well as to reduce the complexity of the clustering process over time. Studies conducted in [13] have identified as an effective way to initially initialize clusters methods K-means ++ with the simultaneous use of the method of "elbow", which avoids the initial centroids, which are located close to each other.

K-means ++ is similar to initializing random points because it randomly selects data points to use as initial centroids. However, instead of selecting these points uniformly at random, K-means ++ selects them sequentially so as to induce the initial centroids to be distributed. In particular, the probability that a point will be chosen as the starting center is proportional to the square of its distance to the existing starting centroids [14].

The "elbow" method is based on determining the sum of squares in the middle of clusters (WCSS - Within Cluster Sum of Squares).

$$WCSS = K_n \sum_{P_i \text{ in Cluster } k}^{n} distance(P_i, C_i)^2$$

A cluster that has a small sum of squares is more compact than a cluster that has a large sum of squares. The "elbow" method considers the nature of the change in the WCSS scatters with the increasing number of groups *k*. Combining all n observations into one group, we obtain the largest intracluster dispersion, which will decrease to *0* for $k \rightarrow n$. At some point, the decrease in this dispersion slows down - this happens at a point called the "elbow".

The disadvantage of the elbow method is that it measures only the general characteristics of clustering, but the algorithm of the "elbow" method is effective if the time to find the number of clusters is important to obtain the result [13]. Some clustering methods allow building fuzzy logical rules after the clustering process.

In the theory of fuzzy logic, true values of statements can take any value of truth from the interval of real numbers [*0*; *1*]. This provision allows building a logical system in which you can make approvals with uncertainty and assess the degree of truth of statements. One of the concepts of fuzzy logic is the concept of elementary fuzzy expression. In the set theory, an element either belongs to the set or not. The theory of fuzzy sets is based on the concept of partial belonging to the set: each element belongs to the fuzzy set partially. [15]. The fuzzy set is defined by the "membership function", which corresponds to the concept of "characteristic function" in classical logic. The membership function is an important element in fuzzy logic. On the one hand, it provides a convenient tool for analytically presenting the degree of membership of a given term, on the ordinate axis - *f(x)* is always delayed range from *0* (clearly does not belong) to *1* (clearly belongs) - the degree of membership, and the axis abscissa - quantitative indicators of the corresponding term. On the other hand, the membership function makes it possible to perform various operations on fuzzy sets [4]. The fuzzification procedure is to determine the degree to which a variable (for an example, measurement) belongs to a fuzzy set. The defuzzification procedure is to determine the numerical value of a variable based on the degree of its belonging to a fuzzy set. Fuzzy logic rule bases are the most commonly used tools in analytical software systems with fuzzy logic. They apply rules in the form such as: *IF "condition" THEN "result"*.

The main difficulty of the "Fuzzy inference" block is that it is not possible to form a rules base in advance due to unstructured data and their large volume. To develop a fuzzy knowledge base, the fuzzy inference mechanism is often used, which is called the Mamdani mechanism [4, 15,16]. In order to specify the number of fuzzy rules, the number of linguistic terms into which the input variables of statistical data are divided without an expert, it is necessary to identify the structure of the fuzzy system. Such identification is performed using fuzzy cluster analysis.

4 THE FUZZY KNOWLEDGE BASE DEVELOPMENT

The proposed approach to designing a fuzzy knowledge base works only with numerical data, which are presented in tabular form. This situation is typical for the technical infrastructure of telecom operators and is acceptable for fuzzy logic procedures in the fuzzification phase, which operates

with numerical data to determine the fuzzy value of the linguistic variable term.

The input data comes in the form of a table in CSV format. In this data structure, each column is an object property. And each line is an instance of the object state. The efficiency study of a set of methods for building a fuzzy knowledge base was conducted on the example of data on the processor's power consumption at different loads. These data were obtained from the Technical University of Dresden [17], where:

FR - the frequency with which the data is processed;

TR - number of threads;

EN - the energy that will be consumed by the processor (Figure 1).

	FR	*TR*	*EN*
1	*1300*	*2*	*637,727*
2	*2400*	*4*	*3448,939*
…	….	…….	……

Figure 1: A fragment of the table with input data.

From the point of view of the fuzzy model, the columns in the table are arranged so that the first *N*-columns (*FR* and *TR*) are conditions, antecedents, which are the left part of the fuzzy logical rule, and the last column (*EN*) is a consequent or fuzzy logical the conclusion on the right side of the fuzzy inference. The rule has the form *IF… AND… .TO*. The result of the fuzzy rule is a combination of propositions combined by "*TA*" operators. The "*OR*" instruction is not used to generate the result.

If tabular data are visualized, we get an *N*-dimensional space in which each instance of the state of the object, so the string will be displayed as a point, and each property of the object will be an axis (Figure 2) [18].

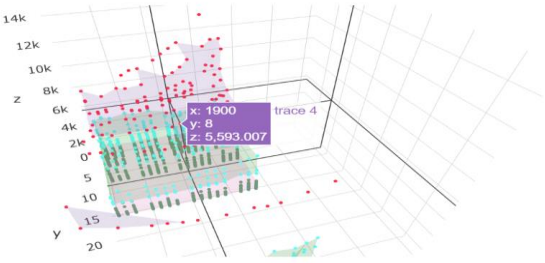

Figure 2: Representation of a point in space.

The next step is to determine that this structure is responsible for the linguistic variable rules and what is terms set of a certain linguistic variable. A rule contains a set of object characteristics, that is columns in a tabular structure. The set of these columns also determines the structure of the resulting linguistic rule. An object characteristic is a linguistic variable rule. A linguistic variable is an axis in *N*-dimensional space that represents term sets. *Term* - a description of the value of the column (for example, for data on the energy efficiency of computer servers - these will be the values for frequency, number of flows, and energy consumption in form as *low*, *high*, *medium*). The meaning of terms in the form of linguistic variables is determined by an expert.

Consider an example of the converting from numerical statistics, which have a tabular data structure, to a fuzzy set. In this case, the measuring space is a two-dimensional plane, in which the abscissa will be the value of the condition (antecedent), and the y-axis will be the final value (consequent) [18]. To move to ambiguity, it is necessary to find membership functions for each linguistic variable. The clustering procedure is performed that will divide all the data from the input space into cluster groups. After clustering, each cluster will be a term set (Figure 3).

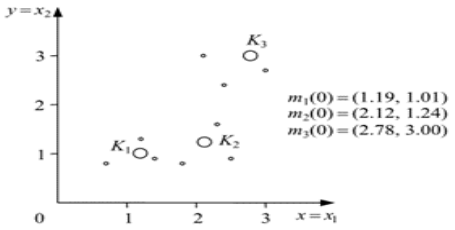

Figure 3: Clustering for two dimensional spaces.

The membership function is a function that has values from *0* to *1* on the ordinate axis, and on the abscissa axis the numerical values of this term. Membership functions can be of different types: triangular, trapezoidal or Gaussian. The proposed method uses the Gaussian function.

The Gaussian function has the form: $\mu(x) = e^{(-\frac{(m_{xi}-x)^2}{2\sigma^2})}$, where x – the center of a cluster, σ – standard deviation. The mathematical expectation for this function is the center of the cluster, and the standard deviation is the measure of the scatter of points near the center of the cluster. To find the width σ it's possible to use the $\frac{|m_{xi} - m_{x(i+1)}|}{N_x}$, where m_{xi} – the center of a cluster, $m_{x(i+1)}$ – the center of the next cluster, N_x – number of clusters.

After determining the membership functions, it is necessary to design them for each axis of the *N*-dimensional space of each cluster (terms) of the

Gaussian membership function of the form (Figure 4) [18,4].

Formation of fuzzy logical rules. After the step of building membership functions, each object in the input sample will create a separate new rule in the form *If.. AND… AND*. The mechanism of rule formation works so as to process all received statistical data [19, 4]. The system cannot have two identical rules, as this will use more computing resources or there will be conflicting inconsistencies, which is not correct. In Figure 5 shows the value of the Gaussian function (membership function) at a given point. Then the largest value of all values is determined, which indicates that the point refers to a fuzzy set (to the corresponding term).

Thus, for each numerical value of the string (object characteristics) there is a corresponding term set, which will be a fuzzy logical rule. The example is shown in Table 1.

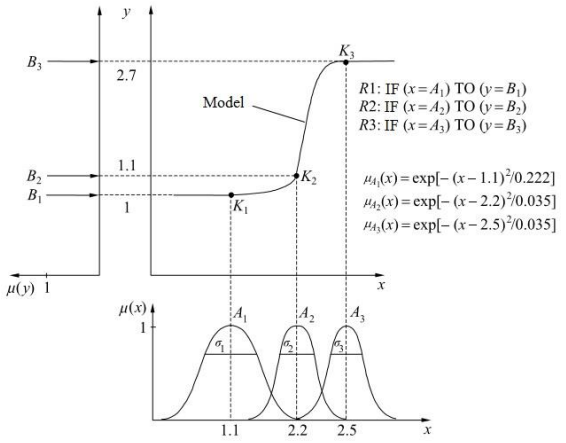

Figure 4: Designing of membership functions for each linguistic variable.

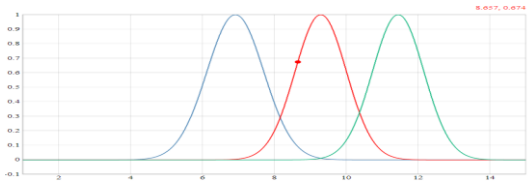

Figure 5: The value of the function at the appropriate point.

Table 1: The fragment of data converted into corresponding term sets.

FR	*TR*	*EN*
low	*low*	*low*
high	*low*	*high*
high	*low*	*low*
middle	*low*	*high*

Based on table1 the rules set in the form ***If… AND .. Then***:

IF ***FR –> low*** *AND* ***TR –> low*** *THEN* ***EN –> low***

IF ***FR –> high*** *AND* ***TR –> low*** *THEN* ***EN –> high***

IF ***FR –> high*** *AND* ***TR –> low*** *THEN* ***EN –> low***

IF ***FR –> middle*** *AND* ***TR –> low*** *THEN* ***EN –> high.***

At this stage, converting the data into terms of linguistic variables, conflicting and duplicate rules can form due to the different nature of the data or because the data during clustering is not properly distributed between clusters at the boundaries of the cluster distribution. In this case, you need to check the set of rules for correctness that can be done by analyzing the metagraphs.

Thus, the modified method for developing fuzzy logic rules for Big Data consists of the following steps:

1) Cleaning up the input data and presentation of them in the correct tabular structures;

2) Fuzzy clustering of the input data based on the FCM algorithm with the first primary initialization of the cluster's centers used the K-means++ algorithm and getting the correct number of the cluster's centers by the algorithm "elbow";

3) Formation of fuzzy logical rules for numerical data converting into a appropriate term-set.

4) The quality analysis of fuzzy logic rules development based on their visual analysis by the metagraphs theory.

Fuzzy logical rules should be as precise as possible, so clustering is an important element of methods complex. If at the final stage of clustering the centers of clusters are found incorrectly, the data will be incorrectly divided, because of it there will be an error in construction of membership functions. To prevent this, training procedures should be performed on different numerical data sets. As a result, the algorithmic and time complexity of the method will increase, but more important is the correctness and accuracy of developing fuzzy logical rules.

The proposed approach to Big Data processing has shown the possibility of designing fuzzy logical rules using numerical clustering methods from numerical statistical data sets to create a fuzzy knowledge base, which greatly simplifies the process of analyzing the effectiveness of telecom services infrastructure.

5 THE ARCHITECTURE OF THE SYSTEM BASED ON FUZZY KNOWLEDGE BASE

When designing real data analysis systems, there is a problem of high load of computing resources due to the large amount of data processed in the system. Microservice architecture is used in Big Data analysis systems to prevent congestion (Figure 6). Microservice architecture is an approach to creating a software package that involves the use of several small application services, each of which corresponds to a limited context. These software services run on different servers and interact with each other over the network, for example via HTTP [20]. The essence of microservice architecture is that each logical part of the system is allocated as a separate micro-service that can be easily connected and integrated into the system, regardless of what technology it uses in the implementation. In Figure 6 shows the architecture of the proposed system for processing Big Data based on the microservice approach.

Each part of the system is allocated as a separate project, which was hosted on a separate server [21].

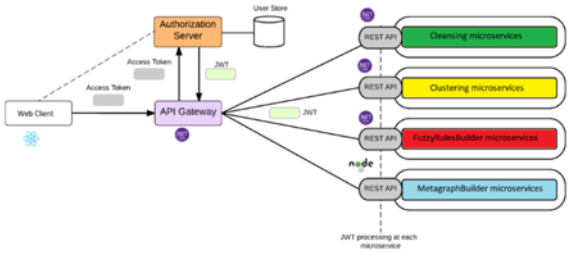

Figure 6: System architecture based on microservice approach.

The system is a website, the client part of which is written in Java-Script using the React library. The proposed architecture uses the OpenId Connect protocol. The Gateway API serves as a gateway between client services there, aggregates data from different services, and is responsible for the logic of executing service requests. The Gateway API performs load balancing, which distributes tasks across multiple network devices to optimize resource depletion, reduce query service time, scale cluster horizontally, and provide resiliency.

Cleansing microservice - a service for cleaning the data that the user uploads to the system. The service uses various algorithms to clean data from incorrect data. For example, the downloaded file may have different data emissions or incorrect format (there may be words instead of numbers). This service provides for data cleaning clustering algorithms. The service uses .Net core technology and is written in the C # programming language.

Clustering microservice - a service for clustering data received after data cleaning. This microservice uses K-means ++ and FCM clustering algorithms, as well as algorithms for initializing initial clusters and finding the number of clusters to analyze data and select similar objects into a homogeneous group. This step is required to construct membership functions for each feature of the object, and then to develop fuzzy logical rules using previously found membership functions. The service uses .Net core technology and is written in the C # programming language.

FuzzyRulesBuilder microservice - a service for designing fuzzy logical rules from statistics. It uses a method that finds the value of the Gaussian function at a point and relates this point to a fuzzy term of a linguistic variable, the service is realized in .Net Core technology.

MetagraphBuilder microservice is a service for displaying and visualizing a metagraph based on fuzzy logical rules. This service filters duplicate rules and builds a metagraph to verify the fuzzy rules of the knowledge base. The service uses Node.js technology.

All services use REST - a protocol for the interaction of components of a distributed system in the network. Steps of the business process of developing fuzzy logical rules from statistical data:

1) Uploading numerical statistics is a step in which the user uploads data to the system. This data must already have the appropriate tabular structure in the form of a CVS file.

2) Initial cluster initialization and finding the number of clusters is a system step in which the system finds the optimal input parameters for clustering before performing clustering: finding the correct number of clusters and performing initial initialization of cluster centers for input sampling.

3) Clustering of uploaded data is a system step in which the system clusters data with pre-selected FCM and K-means ++ clustering algorithms.

4) Initialization of terms for each linguistic variable is a step in which the user can initialize the terms of each linguistic variable or column from the input set. This stage runs after finding the correct number of clusters and performing the clustering procedure because the number of terms for each linguistic variable determines the number of clusters into which the input sample will be divided.

5) Designing of function graphs for each linguistic variable is a system step in which Gaussian membership functions are built. For each function, the mathematical expectation is the center of the cluster, and the standard deviation is the measure of the scatter of the data points around the cluster

6) Converting of statistical numerical data into appropriate term sets is a system step, at the stage of which the system determines for each number in the line the corresponding set of terms. This is done by finding the maximum value of the function at the point for this number.

7) Generating a JSON file with fuzzy logical rules is a system step, in which the system builds a fuzzy logical rule for each line of input statistics, and then passes it to the stage of removing conflicting and duplicate rules.

8) Checking the quality of development of fuzzy logical rules through their visual analysis in the form of metagraphs, retraining as needed.

6 THE EFFICIENCY OF PROPOSED CLUSTERING AND FUZZY LOGIC METHODS

To test the efficiency of constructing fuzzy logic rules in the proposed system, the efficiency is considered as the reliability of the construction of fuzzy logic rules and the time complexity of different clustering algorithms.

Reliability of results was performed by the method presented in [4] and the proposed modified method.

Algorithmic complexity depends on the amount of data received at the input of the clustering procedure. The time complexity of the algorithm K-means *O(ncdi)*, and FCM algorithm $O(ndc^2 i)$, where *n* is the number of points, input data, *d* is the dimension of space, *c* is the number of clusters, *i* - number of iterations for which clustering will be performed.

The experiment was performed with data in the two-dimensional plane. 1500 random points were pre-generated. The input data sequence is divided into 3 clusters by clustering algorithms K-means++ and FCM with different primary initializations of the starting points of the cluster centers. The number of clusters was chosen by the algorithm for finding the number of clusters, namely the "elbow" algorithm.

The number of iterations and the time spent running each of the algorithms for the same data in the same environment (on the same computer) were measured.

The results of the evaluation of efficiency are given in Table 2.

Table 2: Analysis of algorithmic complexity of K-Means and FCM algorithms.

	Primary initialization	Algorithmic complexity	Time spent (seconds)	Number of iterations
K-Means	Random	$O(ncdi)$	1.540	37
FCM	Random	$O(ndc^2 i)$	9.440	115
K-Means	kmeans++	$O(ndc^2 i + + ncdi + nd)$	0.033	8
FCM	kmeans++	$O(2ndc^2 i + nd)$	5.680	74

From the results we can conclude that the time complexity of the K-means algorithm is better than FCM, but the initial initialization greatly affects the final result of clustering, which reduces the time spent and the number of iterations. With the correct initialization, the K-means and FCM algorithms converge in much less time.

To verify the correctness of fuzzy logic rules development, it's needed to conduct an experiment in which there will be two samples of data: training and test. The test sample of data will contain ready-made and already formed fuzzy logical rules, and the training sample will contain only ordinary statistical data, from which fuzzy logical rules will be built by a modified method.

The test sample used data on energy efficiency of servers [17]. The expert used 3 terms for each linguistic variable. The terms had the following meanings: *low, middle, high*. Two experiments were performed using K-means and FCM clustering algorithms with initial K-means ++ initialization.

The training sample had 1,500 rows of columns, which contained the values of data processing frequency, number of streams and energy consumed by the server. The algorithm for finding the number of clusters showed that the number of clusters will be 3. The results of the experiment are given in Table 3.

Table 3: Results of the tests for reliability.

Algorithm	Number of clusters	Number of samples in the input sample	The number of correctly formed samples	Spent time, ms (milliseconds)	Rreliability, C (%)
K-Means (kmeans++)	3	1500	1257	5790	86, 4
FCM (kmeans++)	3	1500	1473	14540	98, 2
K-Means (random)	3	1500	1159	7950	77, 2
FCM (random)	3	1500	1421	35770	94, 7

Thus, it's possible to conclude that the modified method of constructing fuzzy logic rules has reduced the execution time for the two algorithms K-means and Fuzzy C-means, which are used in the proposed method. For K-means the time decreased by about 2 seconds, and for FCM the time decreased by about 2 times.

The reliability of built fuzzy logic rules increased for K-means and FCM algorithms by 10% and 4%, respectively. The reliability of developing fuzzy logic rules using fuzzy FCM is quite high for both methods (normal and modified), but the time complexity of this algorithm is greater than K-means. The choice of algorithm is determined by the characteristics of the data sets, based on the needs of analysis and subject area.

7 CONCLUSIONS

The analysis of characteristics, features and methods of Big Data processing allows to define:

1) Big Data are characterized by different sizes, are structured or unstructured, have different speed of their receipt, a significant amount is simultaneously obtained from different sources, belong to such information that is difficult to process using traditional processes and tools.

2) Some Big Data processing methods are not suitable due to poor structure (numerical data can be in multidimensional space) when managing computing processes in telecommunications nodes, so such data is recommended to analyze in the form of fuzzy logical rules that are close to human understanding.

3) The analysis of clustering algorithms allowed to determine the feasibility of using algorithms FCM, K-Means++ and the algorithm "elbow" to process numerical statistics, extract knowledge from them, so converting to a fuzzy knowledge base, which will be used at the stage of fuzzy inference.

4) A modified method of development fuzzy logic rules for Big Data processing is proposed, the feature of which is the use of algorithms for initialization of cluster centers and finding the number of clusters, using criteria such as reliability of fuzzy logic rules and computational complexity of algorithms.

5) The fuzzy knowledge base was trained on statistical numerical data, which allowed to increase the reliability of the designing of fuzzy logical rules and reduce the operating time of the proposed set of methods.

6) The software of the system and architectural solution for its development with the use of microservices is created, such solutions allow to increase the flexibility of Big Data processing processes and their productivity during data clustering, designing of fuzzy logical rules based on the proposed modified method.

7) The efficiency of the modified method is experimentally proved, which is confirmed by the fact that for K-means algorithm when processing 1500 rows in 3 columns the execution time decreased by 2 seconds, and for FCM execution time was reduced by almost 2 times. The reliability of the designed fuzzy knowledge base for K-means and FCM algorithms increased by 9% and 4%, respectively.

Future researches will focus on further study of the Big Data characteristics and approaches for their processing in information and communication systems, especially such as 5/6 G, and peculiarities of their processing based on data streams specifications.

REFERENCES

[1] Digital 2019: Global Internet use accelerates, [Online]. Available: https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates.

[2] M. Nathaz, W. James, Big data. Principles and best practices of scalable real-time data systems 1st Edition, 2015, pp. 185-192, [Online]. Available: https://www.amazon.com/Big-Data-Principles-practices-scalable/dp/1617290.

[3] E. Nada, Advances in Data Mining. Applications and Theoretical Aspects / E. Nada, E. Ahmed. // Big Data Analytics: A Literature Review Paper. – 14th Industrial Conference, ICDM 2014, St. Petersburg, Russia, July 16-20, 2014. Proceedings, Lecture Notes in Computer Science, Springer, pp. 214-227.

[4] Y. Buhaienko, L. S. Globa, A. Liashenko, and M. Grebinechenko, “Analysis of clustering algorithms for use in the universal data processing system”, in Proc. International scientific and technical conf. Open Semantic Technologies for Intelligent Systems (OSTIS-2020), Minsk, 2020, pp. 101-104.

[5] K. Hribernik, Z. Ghrairi, C. Hans, and D. Thoben, “Co-creating the Internet of Things - First experiences in the participatory design of Intelligent Products with Arduino”, in Proc. 17th International Conference on Concurrent Enterprising, Aachen, Germany, 2011, pp. 1-9.

[6] X. Lei, Z. Yan, and Y. ChunLi, “The application and implementation research of smart city in China”, in Proc. 2012 International Conference on System Science and Engineering(ICSSE), China, 2012, pp. 288-292.

[7] P. Fränti and S. Sieranoja, “How much can k-means be improved by using better initialization and

repeats?", Pattern Recognition, vol. 93, no. 2, pp. 95-112, 2019. doi:10.1016/j.patcog.2019.04.014.

[8] Concepts and Characteristics of Big Data Analytics, [Online]. Available: https://www.iunera.com/kraken/fabric/big-data/.

[9] M. Zgurovsky and Y. Zaychenko, "Big Data: Conceptual Analysis and Applications", Springer Nature Switzerlend, 2020, pp. 1-42.

[10] A. Fahad, N. Alshatri, Z. Tari et al., "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis", IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 3, pp. 267-279, Sept. 2014, doi: 10.1109 / TETC.2014.2330519.

[11] S. Ghosh and S. Kumar Dubey, "Comparative Analysis of K-Means and Fuzzy CMeans Algorithms", International Journal of Advanced Computer Science and Applications (IJACSA), vol. 4, no. 4, pp. 35-39, 2013.

[12] K. A. Abdul Nazeer and M.P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", in Proc. of the World Congress on Engineering 2009, London, 2009, ISBN: 978-988-17012-5-1.

[13] L. S. Globa, Y. M. Buhaienko, I. O. Ishchenko, and A. V. Liashenko, "Approach to determining the number of clusters in a data set", in Proc. International scientific and technical conf. Open Semantic Technologies for Intelligent Systems (OSTIS-2019), Minsk, 2019, pp. 151-154.

[14] Initialization: Where do you start?, [Online]. Available: http://www.salientiastuff.com/k-means-clustering-part-2.html.

[15] A. Pegat, "Sushchnost teorii nechetkikh mnozhestv", Nechetkoye modelirovaniye i upravleniye, 2015, (3th ed.), pp. 13-19, ISBN 3-7908-1385-0.

[16] F. Shevri, "Fuzzy logic", Schneider Electric, 2009, vol. 31, pp. 1-30, [Online]. Available: https://profsector.com/media/catalogs/566dd6af08f6c.pdf

[17] A. V. Lyashenko, "The approach to building fuzzy logical rules for big data", International scientific and technical conference: Modern challenges in telecommunications, 2020, [Online]. Available: http://conferenc.its.kpi.ua/proc/article/view/201702.

[18] A. Pegat, "Samoorganizuyushchiyesya i samonastraivayushchiyesya nechetkiye modeli", Nechetkoye modelirovaniye i upravleniye, 2015, (3th ed.), pp. 506-520, ISBN 3-7908-1385-0.

[19] V. V. Kurdecha, I. O. Ishchenko, and A. H. Zakharchuk, "Data processing method in distributed network Internet of Things", Young Scientist, 2017, vol. 10, no. 50, pp. 75-81.

[20] Microservice architecture, [Online]. Available: https://itnan.ru/post.php?c=1&p=320962.

[21] .NET Microservices: Architecture for Containerized .NET Applications, [Online]. Available: https://docs.microsoft.com/ru-ru/dotnet/architecture/microservices/.