Analysis of clustering algorithms for use in the universal data processing system

Buhaienko Y., Globa L.S., Liashenko A., Grebinechenko M.
Information and Telecommunication Department
Institute of Telecommunication Systems
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»
Kyiv, Ukraine
yura.buhaenko@gmail.com
lgloba@its.kpi.ua
andrey.lyashenko44@gmail.com

***Abstract*—Today, humanity is moving to work in the digital space. While processing large amounts of information in digital space, data processing systems can analyze data, create logical inference systems and building some conditions in which data must be homogeneous, similar, and grouped into clusters. Such systems can analyze the data and create logical output systems according to the rules. Some conditions are required to build such systems: the data must be homogeneous, similar to each other and grouped into clusters. One way to help achieve this is through clustered data analysis. There are a large number of cluster analysis methods available today, but not all of them provide positive results. The article offers criteria that allow you to select algorithms that are appropriate for application in a specific subject area, namely the algorithmic complexity of the algorithm, the accuracy of the object's belonging to the cluster, the attribution of the object to one or another cluster at the boundary of two clusters, the ability to build from clusters fuzzy logic rules. Also described is the architecture of the created system, which clusters the data with different algorithms to use from the desired area. The algorithms were tested to solve the problem of data clustering to predict the degree of server energy efficiency at certain values involved in the processor frequency and number of cores calculations. The use of algorithms allowed us to analyze and look at the features of each, to determine the fuzzy C-average as the most appropriate for this task.**

***Keywords*—clustering algorithms, clustering, fuzzy logic**

I. Introduction

Data volumes that require some analysis and processing are growing steadily today. According to a source [1], the average annual amount of data in the world that needs processing is increasing exponentially. As a result, information management takes more time and resources than ever before. In order to minimize the computational complexity of these processes, a number of methods are being actively used and explored today to improve efficiency and minimize the cost of processing large amounts of data.

According to a research made by We Are Social and Hootsuite [2], the number of people using the Internet has

Figure 1. Increasing of average amount of data per year that needs processing

increased over the past year, with more than 1 million daily users appearing from 2018.

- There are now 5.1 billion unique mobile users in the world, an increase of 100 million over the last year.
- The number of internet users reaches almost 4.4 billion, an increase of 366 million over the last year.
- The number of active users of social media reaches 3.4 billion, an increase of 288 million over the last year.
- The number of mobile users on social media reaches 3.2 billion, an increase of 297 million.

Looking at this data, it is clear that the amount of content and data that users produce will also increase steadily. According to a source [3], the amount of data generated every second is more than 30,000 Gigabytes. You can use clustering to analyze this data, which can group similar computing objects into clusters and show how many and how busy the servers are during computing. Each of these groups should be processed independently, which allows to manage the direct process of processing, to determine the type of analysis and work on information, to effectively allocate resources. It is better for analysts to select groups of similar objects and study their features separately than to study the entire sample of data. When designing and developing real

smart systems, there is a problem - the formation of a complex mathematical model that will include various technical features of the equipment and the physical characteristics of the environment (complex calculations may require additional computational resources). In the course of further development, it may be necessary to take into account new conditions, which may lead to a complete revision of the existing model. In this case, different conditions can be defined as some rules (IF ... THAN) using fuzzy logic. To get these rules, you can use experts who set the terms themselves. And you can teach the system to build these rules on its own, using different clustering algorithms. Non-hierarchical algorithms according to [4] are not suitable for large volumes of data. Since the K-Means and Fuzzy C-Means algorithms are universal, hierarchical clustering algorithms are proposed in [5]. In order to have an idea of how to choose the optimal algorithm for a particular subject area, we need to solve the problem of which algorithm is more appropriate to choose. Each algorithm fits to its subject area, according to certain criteria. In [6] it is proposed to determine the algorithmic complexity of the algorithms, and in [7] it is proposed to determine which of the algorithms is best suited for the construction of fuzzy knowledge bases.

Consider solving the problem of predicting the degree of energy efficiency of servers at certain values involved in the process of calculating the processor frequency and number of cores.

Clustering plays an important role in dealing with large amounts of data, both in the modeling technique and as a pre-processing step in many implementations of the data mining process.

II. Algorithms

- K-means algorithm Advantages of the method: ease of use; speed of use; clarity and transparency of the algorithm. Disadvantages of the method: the algorithm is too sensitive to emissions that can distort the mean; slow work on large databases; the number of clusters must be set. The algorithm of clear clustering has certain disadvantages: in the process of work it is very difficult to determine the degree of fuzzy (blurring) of cluster boundaries, and the actual number of clusters cannot be calculated mathematically, but is given by an expert at the beginning of processing.
- Fuzzy C-means fuzzy clustering algorithm. Advantages: The fuzzy definition of an object in a cluster allows you to identify objects on the border into clusters. Disadvantages: Computational complexity, cluster number assignment. [8]

K-means Clustering K-means Clustering is one of the machine learning algorithms. This algorithm is a non-hierarchical, iterative method of clustering, it has become very popular due to its simplicity, clarity of implementation and high quality of work. The basic idea of the k-means algorithm is that the data should be broken down into clusters, after which the center of mass for each cluster obtained in the previous step is iteratively recalculated, then the vectors are again broken down into clusters with new centers of mass. The purpose of the algorithm is to divide n observations into k clusters so that each observation belongs to only one cluster. [9]

Fuzzy C-Means Clustering Fuzzy C-means algorithm is similar to the human thinking style, which in the future will allow to create rules of fuzzy inference of type Mamdani or Sugeno, allows to control degree of blur of borders of clusters and to take into account fuzzy belonging of certain data to a certain group, provides more flexibility, than clear methods, allowing each point in space belongs to several clusters. This algorithm allows each data sample object to be located in each cluster with a different degree of belonging and is less sensitive to emissions. This method should solve the problem of defining the identity of an object in a cluster if it is on the boundary of several clusters. The fuzzy C-mean algorithm is suitable if the clustering objects have the following requirements:

- High dimensionality of data space - objects are described by a large number of attributes, therefore, the algorithm should be adapted to work in high dimensional data spaces.
- Large amount of data.
- Clustering is done to obtain fuzzy rules based on built clusters. [1]

III. Comparison of clustering algorithms

Algorithmic complexity and performance To use clustering algorithms, you need to understand their algorithmic complexity, because typically large datasets need to be clustered. The larger the input sequence enters the clustering input, the longer the algorithm will run. According to [6], the time complexity of the K-means $O(ncdi)$ algorithm and the FCM of the $O(ncd^2i)$ algorithm, where n is the number of points, the input sequence, d is the amount of space dimension, c is the number of clusters, and the number of iterations. That is, as the number of clusters increases, the performance of the FCM algorithm will decrease. The time spent calculating the K-means algorithm will be 2 times greater than that of the FCM algorithm. We can conclude that K-means is the best algorithm for performance.

IV. Building fuzzy knowledge bases

You also need to understand the need for fuzzy systems. If clustering is performed to construct fuzzy rules, then it is better to choose the FCM algorithm. This algorithm makes it possible to see the affiliation of each object to the cluster and to determine which

cluster it belongs to, especially relevant for data that are at the boundary of two clusters, that is, there is some uncertainty about the choice of a point. [7]

Figure 2. The data set

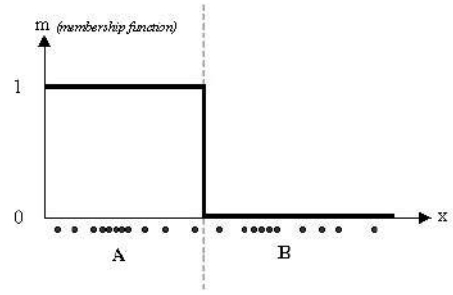

Figure 3. Clustering with clear allocation of clusters

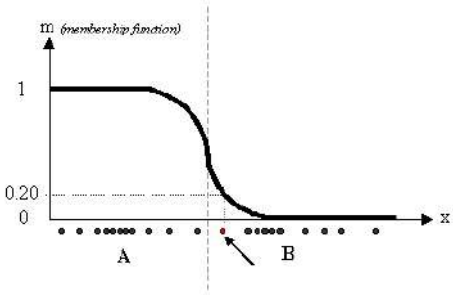

Figure 4. Fuzzy clustering. Each point has its own degree of belonging

In Fuzzy C Means clustering, more accuracy can be achieved to determine the degree of belonging to each point. For example, by adjusting the constant ε, this can be achieved to minimize the criterion J. From the article [7] we can conclude that by reducing ε, in order to minimize the objective function, better data distribution can be achieved. But for such improvement, it should be noted that the execution time is significantly increased, as the number of iterations increases. For K-clustering, it is also suitable for fuzzy systems. However, looking at Figure 3, it can be seen that this clustering "roughly" separates points into clusters that are not accurate enough for points at the cluster boundary.

Practical implementation of the clustering algorithm by example The algorithm was written in C # programming language using the ASP.NET Core platform. In terms of architecture, the application is completely client-server. There is an explicit client level in the application, such as a browser. And an application layer that uses an algorithm to execute a clustering algorithm. The application also uses REST technology. The project uses an ASP.NET Core MVC Web-based API.

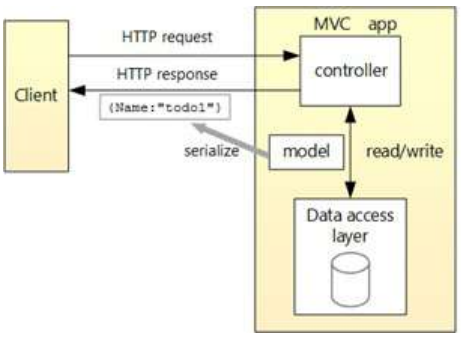

Figure 5. Architecture of the created application

The server side receives the data transmitted in the file through the interface and performs fuzzy clustering with the initial parameters specified by the expert (number of clusters, precision, minimization). The client side of the web application includes a graphical interface for data input for processing, and a page for outputting data results. The clustered data is displayed on the page as a graph written using the open-source Javascript library Plotly.js. Written on the popular libraries d3.js and slack.gl, it is a high-level declarative graph library. For a clearer and more convenient output, we have used the division into three clusters, each responsible for red, green or blue. Each point on the graph represents one dimension in the dataset: its color is represented in RGB (red, green, blue) format, where each parameter is a cluster belonging. With this display method, you can show to which cluster this dimension is most relevant, which data has not been clustered accurately (these points are marked in indeterminate gray). To obtain the result of the experiment, a clustering of server energy efficiency data (number of threads, data processing frequency, energy consumed by the machine) was collected at the Dresden University of Technology. In addition, processing was performed with different clarity of cluster boundaries, which made it possible to correct processing uncertainty.

As a result, an application was developed in which the expert can process his data on different algorithms and also to perform analytic on the basis of visualization, adjusting the parameters - to set up clustering for the conditions of his task.

V. Conclusions

This article provides a comparative analysis of K-means Clustering algorithms and Fuzzy C-means fuzzy clustering algorithm and tested each for data taken from Article [10] on the performance of computing servers. Comparative analysis is carried out on the criterion of algorithmic complexity of algorithms and on the criterion of building fuzzy knowledge bases. And showed that the K-means algorithm has less algorithmic complexity,

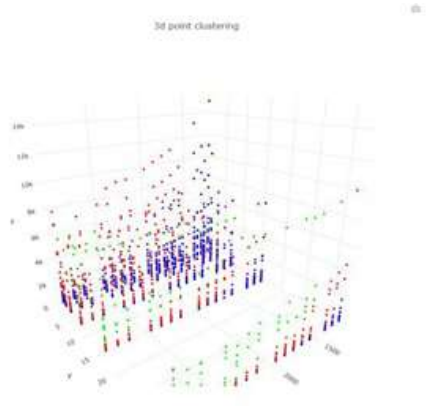

Figure 6. Visualization of clustered data with minimization of criterion J for $\varepsilon < 0.05$

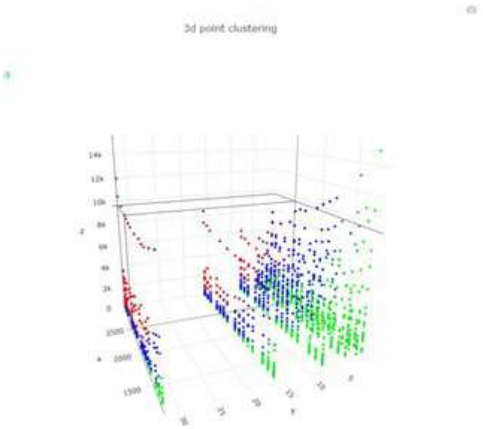

Figure 7. Visualization of clustered data with minimization of criterion J for $\varepsilon < 0.0001$

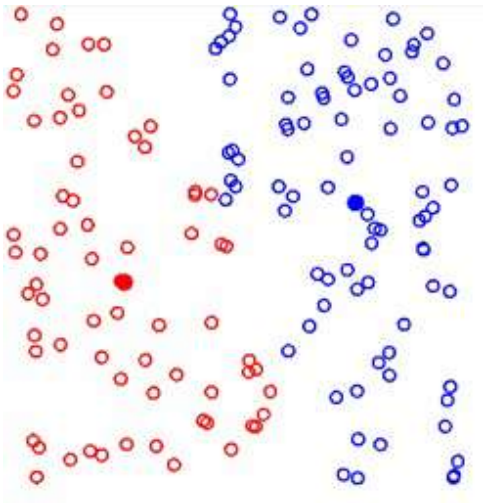

Figure 8. Visualization of the K-means algorithm for 2 clusters in 2-dimensional space

which will allow to spend less resources to execute clustering algorithms. But fuzzy C-means is better suited to building fuzzy knowledge bases, which allows you to more gently divide objects into clusters and show more precisely which cluster the object belongs to. This will give a better opportunity to build fuzzy knowledge bases. In the future, it is planned to use already given clustering algorithms - to build fuzzy knowledge bases, using different types of fuzzy inference rules.

References

[1] Liashenko A., Buhaienko Y. "A clustering method for processing large volumes of data". MODERN CHALLENGES IN TELECOMMUNICATIONS NTUU "KPI", Kiev 2019.

[2] DIGITAL 2019: GLOBAL INTERNET USE ACCELERATES [Electronic resource]. – Resource Access Mode: https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates – 2019

[3] Nathaz Marz, James Warren. Big data. Principles and best practices of scalable real-time data systems. p.[185-192] – 2015 – ISBN-13: 978-1617290343.

[4] Methods of cluster analysis. Iterative methods. [Electronic resource]. – Resource Access Mode: https://www.intuit.ru/studies/courses/6/6/lecture/184?page=4 – 2016 – ISBN: 978-5-9556-0064-2

[5] Pegat A. Fuzzy modeling and control. p. [520-530]. – Moscow, 2009.

[6] Soumi Ghosh, Sanjay Kumar Dubey «Comparative Analysis of K-Means and Fuzzy CMeans Algorithms». International Journal of Advanced Computer Science and Applications(IJACSA), Volume 4 Issue 4. – 2013.

[7] A Tutorial on Clustering Algorithms. Fuzzy C-Means Clustering. [Electronic resource]. – Resource Access Mode: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html – 2014

[8] Magerramov Z., Abdullayev V., Magerramova A. «BIG DATA: PROBLEMS, ANALYSIS METHODS, ALGORITHMS» – 2017

[9] K-means algorithm. [Electronic resource]. – Resource Access Mode: http://algowiki-project.org/en/K-means_clustering – 2018

[10] Dmytro Pukhkaiev, Sebastian Götz «BRISE: Energy-efficient Benchmark Reduction». The 6th International Workshop. Dresden – 2018

Анализ алгоритмов кластеризации для использования в универсальной системе обработки данных

Бугаенко Ю., Глоба Л.С.,
Ляшенко А., Гребинеченко М.

Аннотация – В статье предложены критерии, которые позволяют выбрать алгоритмы, целесообразны для применения в конкретной предметной области, а именно алгоритмический сложность алгоритма, точность принадлежности объекта к кластеру, отнесение объекта к одному или другому кластера на границе двух кластеров, возможность построить из кластеров нечеткие логические правила. Также описана архитектура созданной системы, кластеризирует данные различными алгоритмами в зависимости от потребностей предметной области.

Received 15.12.2019