EXCELLENCIA
INTERNATIONAL MULTI-DISCIPLINARY JOURNAL OF EDUCATION
VOLUME 1, ISSUE NO. 5, (2023) | ISSN: 2994-9521
http://multijournals.org/index.php/excellencia-imje/index

Clustering Method Of Distributed Technologies In Data Flow Management

Nazarov F, Rashidov A, Pardayev M, Sunnatova S.

(Samarkand state university)

Abstract:

Grouping data into groups based on certain rules is known to increase data flow, not only to extract meaning from data, but also to increase the efficiency of large data processing. A uniform distribution of data is particularly effective in approaches such as distributed computing or parallel processing. The main reason for this is that dividing the data into as many equal clusters as possible allows for the highest performance results in these approaches. But the human-factor-based uniform distribution of data is a complex process due to the impossibility of pre-planning the data in the data stream and the size of the data. Therefore, in the study, the application of the clustering method of distributed technologies in data flow management was considered.

Introduction

Nowadays, data analysis, which is, identifying hidden contents from a data set, and separating them into useful, reliable, and interrelated data groups is one of the topics of modern research [1]. In the modern digital age, the increase in the volume of data flow increases the relevance of this research topic. Because as the flow of information increases, people's natural data analysis capabilities are limited, and the process of extracting meaningful information from big data

becomes more complicated [2-4]. In such a situation, the most effective solution for finding hidden patterns in data is unsupervised learning based on data clustering approaches according to certain criteria. Unsupervised clustering is an approach to grouping data by grouping or separating those with similar patterns, and those with different patterns from the data without human intervention [5].

In these clustering approaches, the grouping of data mainly focuses on characteristics such as similarity and uniqueness of the data, and usually ignores or does not consider the number of elements in the groups to be very important. But in real life, there are such problems that not only the similarity and difference of data are important in solving these problems, but also the equality of the number of elements in clusters is an important factor. For example, when storing and processing data using parallel and distributed computing mechanisms. In these mechanisms, the more evenly the data is distributed to the parts of the system, the higher the efficiency. Because in this case, all distributed functions within the system perform the same number of tasks, and all parts of the system complete their tasks at the same time. In other words, they do not wait for each other's tasks to be completed. This is the most important indicator of achieving the highest efficiency in the work process. Therefore, achieving equal size distribution based on unsupervised clustering algorithms was taken as the major research objective.

In order to solve the problem presented during the research, the capabilities of hierarchical, K-means, Bisecting K-means, and DBSCAN existing unsupervised clustering algorithms to divide data into clusters of equal size are analyzed. However, due to the limited capabilities of existing algorithms, a new unsupervised clustering algorithm is proposed to achieve uniform distribution of data.

K-means clustering

The K-means clustering algorithm is a center-based grouping method, and the objects in the set are clustered according to which of the central points known as centroids are close. The number of centroids is the same as the number of clusters requested by the user. Centroids are initially chosen arbitrarily. Therefore, clustering will not be perfect in the initial state. Minimization of the sum of squares of the errors of the objects in the clusters relative to the centroids is used to make the clusters perfect (1).

$$S = \sum_{i=1}^{k} \sum_{x_j \in C_i} dist(c_i, x_j)^2 \quad (1)$$

where k - is the number of clusters, c_i - is the cluster, c_i is the centroid coordinate of the i cluster, x_j is the j element coordinate in the c_i cluster. It is known that S reaches its minimum value when the derivative of S with respect to c_i is equal to 0 (2).

$$\frac{\partial S}{\partial c_i} = \frac{\partial S}{\partial c_i} \sum_{i=1}^{k} \sum_{x_j \in C_i} dist(c_i, x_j)^2 = 0 \quad (2)$$

Based on this formula, the optimal value of the centroid of the n^{th} cluster can be found as follows.

$$\frac{\partial S}{\partial c_n} = \frac{\partial S}{\partial c_n} \sum_{i=1}^{k} \sum_{x_j \in C_n} (c_n - x_j)^2 = 0 \tag{3}$$

If the number of objects in the n^{th} cluster is m, then the following solution comes from equation 3

$$2 \cdot \sum_{j=1}^{m} c_n = 2 \cdot \sum_{j=1}^{m} x_j \quad \Rightarrow \quad m \cdot c_n = \sum_{j=1}^{m} x_j \quad \Rightarrow \quad c_n = \frac{\sum_{j=1}^{m} x_j}{m} \tag{4}$$

The conclusion from formula 4 is that the centroids reach optimal coordinates when they are equal to the average value of the coordinates of the objects in the cluster.

In the K-means algorithm, objects are divided into clusters of equal size compared to hierarchical clustering. This is especially evident when objects are located at the same density in the coordinate system.

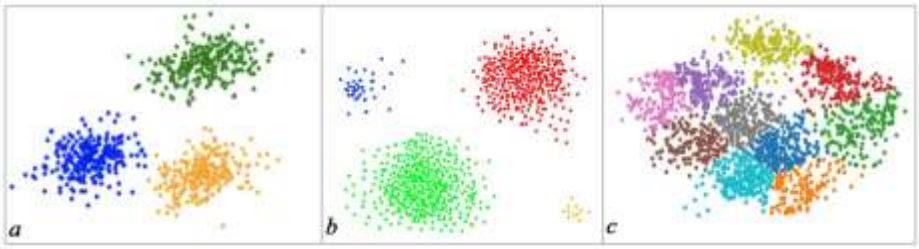

Fig. 1. The result of object clustering using K-means

In Fig.1.a, the objects are divided into 3 clusters, and the size of each cluster is almost the same. But if these objects are divided into 2 or 4 clusters, the sizes of the clusters will be very different from each other. In Fig. 1.b, it can be seen that two clusters are large in size, and the other two are small in size. Fig. 1.c shows a perfect set of objects, i.e., a set of objects that are relatively evenly distributed over space. When this set is divided into an arbitrary number of clusters, clusters of almost the same size are formed.

References

1. Anil K. Maheshwari Business Intelligence and Data Mining [B]. Business Expert Press, LLC, 222 East 46th Street, New York, NY 10017, 2015, 162 p
2. Akhatov A., Nazarov F., & Rashidov A. Mechanisms of information reliability in big data and blockchain technologies [C]. ICISCT 2021: Applications, Trends and Opportunities, 3-5.11.2021, doi: 10.1109/ICISCT52966.2021.9670052
3. Akhatov A., Nazarov F., & Rashidov A. [C]. Increasing data reliability by using bigdata parallelization mechanisms. ICISCT 2021: Applications, Trends and Opportunities, 3-5.11.2021, doi: 10.1109/ICISCT52966.2021.9670387
4. Jumanov I. Djumanov O., & Xolmonov S. Mechanisms of image recovery optimization in the system for recognition and classification of micro-objects [C]. AIP Conference Proceeding 2686, 020009 (2022), doi.org/10.1063/5.0113052
5. Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar. Introduction to data mining [B]. Second edition. New York, NY : Pearson Education, 2019

6. Akhatov A., Sabharwal M., Nazarov F. & Rashidov A. Application of cryptographic methods to blockchain technology to increase data reliability [C]. 2nd ICACITE 2022 doi: 10.1109/ICACITE53722.2022.9823674
7. Jumanov, I.I., Xolmonov, S.M. Optimization of identification of non-stationary objects due to information properties and features of models [C]. IOP Conference Series: Materials Science and Engineering, 2021, doi:10.1088/1757-899X/1047/1/012064