JOSEIT JOURNAL OF SYSTEMS ENGINEERING AND INFORMATION TECHNOLOGY
http://jurnal.iaii.or.id/index.php/JOSEIT

Acceleration and Clustering of Liver Disorder Using K-Means Clustering Method with Mahout's Library

Tariq Mohammed Abdullah Bin Samer[1] and Cahyo Darujati[2]
[12]Faculty of Computer Science, Narotama University, Surabaya, Indonesia

*[1]*tariqbinsameer15@gmail.com*, [2]*cahyo.darujati@narotama.ac.id*

Abstract

Evaluation of liver disorders was performed to observed and clustered in Big Data environment applications. However, since liver disorder is a common illness, global awareness of such cases can be life threatening, therefore the urge to avoid and study must be essential. The idea of parallel computing is established on the basis of the K-means method. The MapReduce framework is used to complete multi-node data processing, and a solution to the MapReduce K-Means method is given. The ultimate goal is to establish clusters that allow each entity to be examined and assigned to a certain cluster. These algorithms are designed to accelerate computations, reduce the volume of enormous data that must be computed, and improve the efficiency of arithmetic operations. The combination of theoretical analysis and experimental evaluation is very significant.

Keywords: k-means, clustering, big data, mahout.

1. Introduction

In all technological fields, parallel processing is utilized to tackle practical problems, such as improving the efficiency of program execution and making better use of computing resources. Strong hardware and architectural parallel programming skills are required from programmers for the initial stage of parallel processing, and there are still higher standards in the field of complicated and challenging parallel control logic. In the age of big data, a cluster environment based on a system architecture has been developed. Parallel processing's constraints are changing, and today's technology centers on the development and application of technological models. As a result, it is feasible to use distributed parallel processing technologies in the Big Data computing environment.

Since liver disorders are one of the most common diseases and their causes are based on daily routine lifestyles, prevention and treatment of liver disorders are crucial, such as food poisoning caused by fast food, drinking excessive alcohol, and drugs. Eventually, such diseases may lead to diseases such as fatty liver disease and cirrhosis, which is one of several diseases caused by several viruses, including hepatitis A, B, and C [1]. The University of California, Irvine (UCI) machine learning repository is a very useful resource to obtain open source and free datasets for machine learning. It is also a collection of databases, domain theories, and data generators that the machine learning community uses to simulate the effectiveness of machine learning algorithms. [2]

One of the ways to examine patients is by taking a blood sample to be tested. Medical professionals are expected to be able to identify the multiple types of liver disorders. Ultimately, the overarching aim is to establish clusters with settings that allow each entity to be analyzed to be placed into a unique cluster. [1] This can offer suggestions to Surabaya province authorities in the prevention of liver disease and local health awareness campaigns. To carefully examine the enormous amount of data produced by recent applications, in more detail, the primary goal is to arrange the data into clusters so that the items are grouped in the same cluster when they are similar in their depiction of specific metrics and different from objects of other groups.

Clustering can be considered as an unsupervised learning concept from the perspective of automated learning. Hadoop is a distributed file system and an open-source MapReduce application for handling large amounts of data. On top of Hadoop, the MapReduce paradigm is used to implement the Apache Mahout clustering algorithms. employing Apache Mahout and offering a comparison. In addition, we highlight the clustering methods that are the best for handling large amounts of data. [3].

Nevertheless, the problem of organizing a large volume of data frequently arises. The data is typically compressed using a generalization process to make it easier to analyze. The "clustering" technique, which includes assembling related data into a cluster to enable considerable generalization, is a fundamental stage in this process. [4]

Accepted by editor: 04-08-2023 | Final revision: 08-09-2023 | Online publication: 08-09-2023

The standard K-Means partition clustering method is simple and effective; nevertheless, it uses a lot of memory and is unable to handle very large data sets. K-Means is one of the most widely used clustering strategies [5]. It is clear that K-Means perform best when there are fewer clusters and iterations than data points, but it lacks the scalability and effectiveness of clustering when processing large data sets, which adds to the complexity of the processing time. Furthermore, when the clusters are no spherical or have different sizes or densities, the K-Means algorithm struggles to locate the pertinent clusters. Document data with topic and subtopic structures is the main source of hierarchically organized data for hierarchical clustering. Complex kinds, strong dynamics, and time series features make it impossible to tackle the challenge of big data clustering. To address the problem of a large data volume and high complexity, it is possible to convert the single-node computing approach to parallel computing [6] [7] [8].

To categorize data or objects into various groups without any prior cluster labels, clustering uses unsupervised learning techniques. The most effective application of this method is data collection without a prior label. Technique that can be applied to labeled data to compare the clustering results with the real labels. Consequently, the accuracy of the clustering algorithm is known. That term "supervised learning", which refers to methods that demand data labels. The shortest distance between two objects is calculated to determine how similar they are. The Euclidean distance is one way to figure out how far something is.

The Pythagorean formula can be used to calculate the distance between two points, which is known as the Euclidean distance in mathematics. A geometric vector with a length (magnitude) and direction is a geometric vector that is frequently described as the Euclidean distance. whereas a set of vectors forms the mathematical structure known as a vector space. The vector may be multiplied by real numbers, in addition to other factors. [9]

The distance between vectors or vector lengths can be defined as follows.

$$\|A\| = \sqrt{x_1^2 + y_1^2} \text{ and } \|B\| = \sqrt{x_2^2 + y_2^2} \tag{1}$$

As for calculating the two distances between the two vectors, they are as follows.

$$d\,(\bar{A}, \bar{B}) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{2}$$

K-Means is one of the data mining algorithms that applies the clustering method. If given a set of data X = {x1, x2, ..., xn} where xi = (xi1, xi2, ..., xin) is vector, then the K-Means algorithm divides x into k groups [9]. The first step in determining the number of clusters, k, is to use this algorithm. assembling things into groups depending on their size and the centroid, or center, of the cluster. [10]

The following pseudocode explains the K-Means algorithm: [11]

a. The first stage is to choose k, or the number of clusters.
b. A random process is used to choose the starting value of the centroid or to initialize the center of the group k.
c. By calculating the closest rarity between things, it is possible to determine how similar two objects are. For the purpose of identifying an object with a specific centroid. This step determines the separation between an object and the centroid. If an object is closest to centroid A, it will be grouped with other objects in centroid group A.
d. Calculating the average value of every object in a particular group is once again used to determine the new centroid.
e. A new centroid is used to group the items in each group.

Repeat steps 3 and 4 as necessary until the centroid value remains the same.

Each object is classified according to how closely it resembles a centroid in Figure 1. Purple data items are a feature of centroids. The same color indicates how related the two objects are.

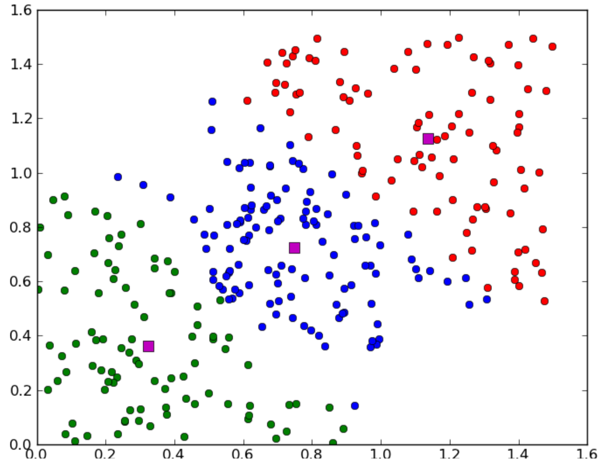


Figure 1. Visualization of K-Means

2. Methods

Data mining pre-processing has been performed and extracted from the website https://archive.ics.uci.edu/ml/datasets/liver + Diseases, in.CSV format, Comma-Separated Values file, which allows data to be saved in tabular format. This involves sequence formatting, as well as most of the metadata to be used in this research, must be readable for the Mahout library.

Figure 2 shows a flow chart diagram that illustrates the research process in general.

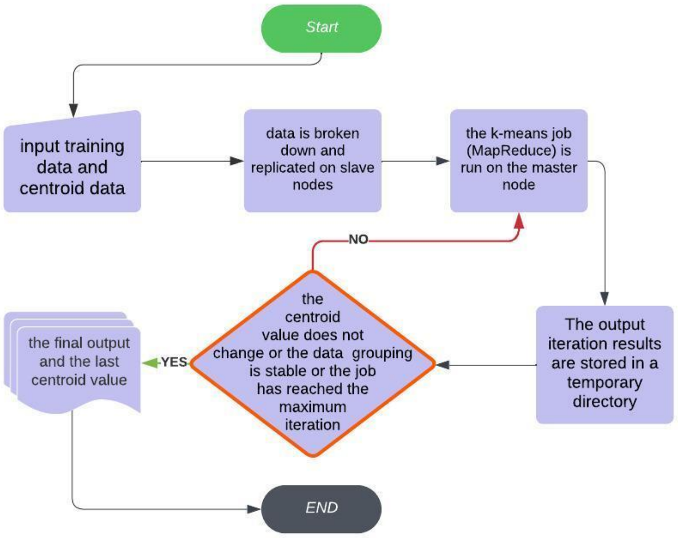

Start
input training data and centroid data
data is broken down and replicated on slave nodes
the k-means job (MapReduce) is run on the master node
NO
the centroid value does not change or the data grouping is stable or the job has reached the maximum iteration
The output iteration results are stored in a temporary directory
YES
the final output and the last centroid value
END

Figure 2. Research Flowchart

2.1 Data

In the liver disorders dataset, the researchers used data sources or data sets from the UCI Machine Learning databank. The following data pertain to liver disorders:

1. Liver disorders is the name of the data set.
2. The information in the data set includes the following.
   a. BUPA Medical Research Ltd. produced the data.
   b. Richard S. Forsyth, 8 Grosvenor Avenue, Mapperley Park, Nottingham NG3 5DX, 0602621676, contributed the data.
   c. On 15 May 1990, the data were first created.
3. The first five factors are the results of blood tests that are believed to be sensitive to liver disorders, of which alcohol consumption is one of the causes. Each line in the bupa.data file represents the result of a man-based test.
4. There are 345 rows of data.
5. Data attributes add up to 7.
6. Liver disorder data do not have missing value or loss of data items.
7. Some information from data attributes, as shown in Table 1:

Table 1. Data Information of Liver Disorder

No.	Attribute	Information
1.	mcv	Mean corpuscular volume, i.e. average corpuscular volume of blood.
2.	alkphos	Alkaline phosphatase, i.e., levels of alkaline phosphate in the blood.
3.	sgpt	Alamine aminotransferase, i.e. natural levels of aminotransferase in the blood.
4.	sgot	Aspartate aminotransferase, i.e. the level of aspartate aminotraspherease in the blood.
5.	Gammagt	Gamma glutamyl transpeptidase, that is, the level of gamma glutamyl transpeptidase in the blood.
6.	drink	The amount of alcoholic beverages consumed per day in half pint units (257 ml per unit).
7.	selector.	This attribute is a class attribute that will later become the source of precision calculations. Worth 1 or 2. If 1 then the instance suffers from liver disease. If it is 2 then the instance is normal.

DOI: https://doi.org/10.29207/joseit.v2i2.5334

2.3 K-Means

Because it is effective in grouping data based on the similarity of qualities, the K-Means clustering approach is used in the analysis of data related to liver disorders. Therefore, it can be determined that a data item in a group or cluster has the same similarity to other data items in the same group. A data item belonging to one group may differ in nature from a data item belonging to another group.

In K-Means Mahout, several factors that must be taken into account include:

1. The input training data must be in sequence format. Sequence files are a type of file that the Apache Mahout library can read. There will be a key-value format structure in the sequence file itself.
2. The format of the centroid data must be a sequence.
3. A directory that retains the K-Means iteration process will be the output or result of the K-Means process.
4. The shortest distance or similarity measure is used to calculate how similar two qualities are. The Euclidean distance is the similarity metric used in this study.
5. The convergence threshold is a predefined value that determines when the K-Means process stops iterating.
6. The greatest number of iterations the system is capable of handling.
7. The K-means system's k values or number of groupings. The number of k to be used for this study is 2.

2.4 Big Data System Schema

Big Data systems were initially established within each node based on the implementation process. A single node cluster is how a Big Data system is implemented on a computer. The next stage is to combine all the nodes into one when each node has the Big Data system implemented on it. A multi node cluster is the name given to this process.

2.4.1 Single-Node Cluster Schema

Figure 3 shows a Hadoop system that is used exclusively on a PC, with a single-node cluster structure. A master node and a slave node are functions of a single computer. Therefore, NameNode, DataNode, SecondaryNameNode, ResourceManager, and NodeManager are the services that are active on this node. Local network connections are made through the router. To assign each computer (or "node" as it is often known) an IP address. In a single network, every computer has an IP address. Each computer can be configured into a multinode cluster after they all have an IP address that is connected to one another in the computer network.

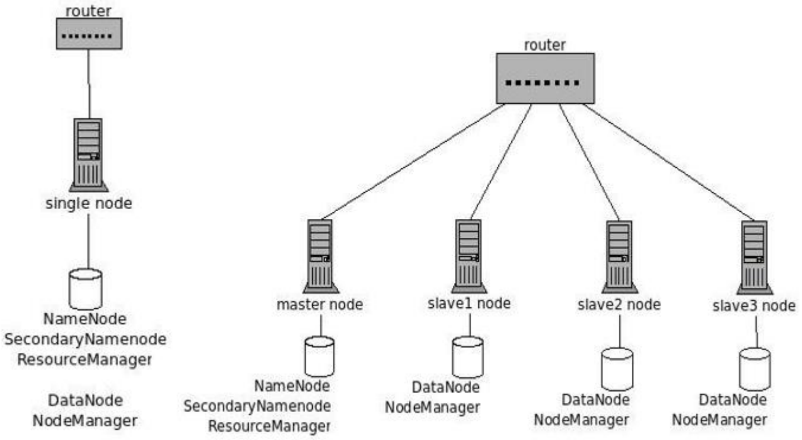


Figure 3. Schematic single node cluster and Schematic multi node cluster

2.4.2 Multi-node cluster schema

Figure 4. illustrates the cluster multi node schema with the Hadoop system set up on various PCs. 4 computers are used in a multi node cluster configuration with 1 master node and 3 slave nodes. A local network connection is made through the router. In the network system, the router performs the function of a gateway, connecting one node to another. Each node is referred to by an IP address. The NameNode, SecondaryNameNode, and ResourceManager services will run by the master node. The NodeManager and DataNode services will run on the slave node.

DOI: https://doi.org/10.29207/joseit.v2i2.5334

Figure 4. Schematic multi node cluster

2.5 System requirements

1. Hardware
   a. Cluster computer specification as shown in Table 2.

Table 2. Cluster Computer Specifications

Node Type	Processor	Memory	hard drive	Network
Master node	Intel® CoreTM i5-5200U-CPU @ 2.20GHz	12GB	500GB	10.1.505.2015 Realtek PCIe GBE family controller
Slave1 node	Intel® CoreTM i5-5200U-CPU @ 2.20GHz	12GB	500GB	10.1.505.2015 Realtek PCIe GBE family controller
Slave2 node	Intel® CoreTM i5-5200U-CPU @ 2.20GHz	12GB	500GB	10.1.505.2015 Realtek PCIe GBE family controller
Slave3 Node	Intel® CoreTM i5-5200U-CPU @ 2.20GHz	12GB	500GB	10.1.505.2015 Realtek PCIe GBE family controller

   b. 4 RJ45 cables
   c. D-Link DES1024D Router.
2. Software
   a. Ubuntu version 20.04.2.0
   Ubuntu is an operating system that uses Linux as its kernel.
   b. Sun™ java 7
   Java is necessary for Hadoop to work on Ubuntu-based systems. Hadoop requires a Java version higher than 5, which is version 5. Java 7 was used in this study.
   c. SSH (Secure Shell)
   To control the entire node, this big data system makes use of SSH access. It is setup by remotely controlling a slave node on a master node.
   d. Apache Hadoop 2.6.0
   Hadoop 2.6.0 is the version used. Yarn functionality is already supported by this version. In addition to Hadoop Distributed File System (HDFS) and MapReduce, Yarn is a crucial component. Yarn controls how the cluster computer uses its resources.
   e. Apache Mahout 0.10.1
   The version of Mahout used is 0.10.1.
   f. Apache Maven 3.3.9
   A library called Apache Maven is used to develop and compile the Mahout library.
   g. Eclipse Kepler
   Eclipse is an integrated development environment (IDE) for writing and managing Apache Mahout source code.
   h. LibreOffice Calc
   Spreadsheets containing numerical data are frequently processed using the LibreOffice Calc office program from the LibreOffice suite.
   i. Nano and Pluma Editors
   Data can be shown using the nano and pluma editors. However, the nanoeditor is only used with a terminal or command line.
   j. MATE terminal
   Applications or software packages, such as Hadoop and Mahout instructions, can run on Linux computers using the MATE Terminal application as a command line.

3. Results and Discussion

Two separate analyses of the results are conducted: the implementation of K-Means using the Mahout library and the other of the performance of the Hadoop system.

3.1 Evaluation of K-Means implementation in a Big Data environment using the Mahout library.

The directories /user/hducer/output on HDFS contain the output results of the K-Means algorithm. Mahout provides a method for evaluating the results of the K-Means computation. The cluster dump command is used as the method. Data items can be grouped using the centroid or group center using this method to build or generate

DOI: https://doi.org/10.29207/joseit.v2i2.5334

an analysis file. The command used to run is $mahout clusterdump, i the last iteration directory, p clusteredPoints and o output_file_analysis. The command to enter the input directory, which is the last iteration directory, is represented by the i or -input parameter. The command to enter the clusteredPoints directory, the final result of data that has undergone a K-Means computing process, uses the parameter p or -pointsDir. The parameter, either o or output, creates a file for output analysis. According to Table 3, the centroid data items used in the manual calculation results and the calculation results produced by the Mahout library are identical. Thus, it can be said that the Mahout library is capable of appropriately performing K-Means computations.

3.2. Performance Analysis of K-Means Implementation in a Big Data Environment Using the Mahout Library. Mahout library-based K-Means computations were performed ten times on various numbers of slave nodes to analyze performance. The performance results are evaluated using the average value. Table 4 shows that the more slave nodes there are, the faster the execution time K means.

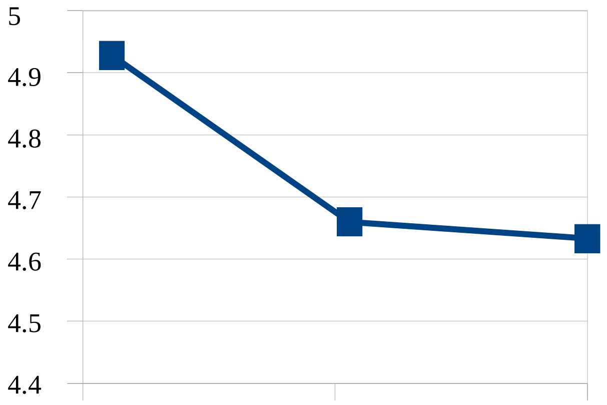
Table 4. Performance of implementing K-Means using the Mahout library in a Big Data environment

	K-Means execution time on slave node (minutes)		
Repetition No.	1	2	3
1	4.523	4.646	2.8
2	5.295	5.029	4.553
3	4.525	4.94	4.946
4	4.96	4.95	5.461
5	4.93	4.058	2.759
6	5.397	5.384	4.967
7	4.001	5.368	4.544
8	4.923	3.631	5.449
	5.297	4.045	5.434
Figure 5. Graph of Hadoop system performance results.			
10	5.428	4.548	5.416
Average	4.9279	4.6599	4.6329

Table 3. Comparison of the results of manual calculations with those from the Mahout library.

Centroid labels	Centroid Data Items						Number of data items
C1	89.95751634	69.2124183	26.81699346	22.87581699	27.05882353	3.14869281	306
VL-27	89.958	69.212	26.817	22.876	27.059	3.149	306
C2	91.68421053	75.47368421	59.55263158	39.05263158	129	5.960526316	38
VL-49	91.684	75.474	59.553	39.053	129	5.961	38

Figure 5. shows that the execution time for running K-Means using the Mahout library decreases as the number of slave nodes increases.



4. Conclusions

The following can be drawn from the research findings of using the MapReduce programming methodology to construct K-Means in a big data environment:

1. The performance findings demonstrate that the Mahout library's K-Means computations execute more quickly the more slave nodes are employed.
2. Data on liver disorders can be used to create K-Means using the Mahout library. Performing a manual K-Means calculation demonstrates this. The output centroid data items from K-Means computations using the Mahout library are identical to those from calculations done manually.

The concept of partition, communication, combination and mapping is built on the basis of the conventional K-means algorithm, and the stages of a parallel K-means algorithm based on MapReduce are designed. The K-means parallel method built on the MapReduce framework provides additional advantages in dealing with time complexity, according to both actual and theoretical results. Calculating the standard K-means algorithm and data load space can be reduced, clustering efficiency can be improved, and the time, space complexity, and loss rate of data points can be reduced by choosing the right cluster number, data block size, iteration number, and number of nodes.

References

[1] Sadhasivam, "Liver disease prediction using machine learning classification," *Webology 18, Special Issue on Information Retrieval and Web Search,* pp. 441-452, 2021. doi: 10.14704/WEB/V18SI02/WEB18293

[2] Markelle Kelly, "University of California Irvine," 2023. [Online]. Available: https://archive.ics.uci.edu.

[3] V. R. Eluri, A comparative study of various clustering techniques on big data sets using Apache Mahout, Muscat: IEEE, 2016. doi: 10.1109/ICBDSC.2016.7460397

[4] Rokach. L., Data Mining and Knowledge Discovery Handbook, Berlin: Germany: Springer, 2010, pp. 22-32. [Online] Available: https://link.springer.com/chapter/10.1007/978-0-387-09823-4_1

[5] Na, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," in *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, Jian, China, 2010. doi: 10.1109/IITSI.2010.74

[6] Dayong, Research on Supply Chain Management Strategy of Longtang Electric Engineering Co. Ltd, Kuala Lumpur: Acta Electronica Malaysia, 2019. doi: 10.26480/aem.01.2019.10.13

[7] Meisam D., Webometrics Analysis of Iranian Universities about Medical Sciences' Websites between September 2016 AND March 2017, Kuala Lumpur: Acta Informatica Malaysia, 2019. doi: 10.26480/aim.01.2019.07.12

[8] Ou Z., A Look at Millennial Attitudes Toward AI Utility in The Class., Kuala Lumpur: Information Management and Computer Science, 2019. doi: 10.26480/imcs.01.2019.07.09

[9] Prasetyo, "Comparison of distance and dissimilarity measures for clustering data with mix attribute types," in *2014 The 1st International Conference on Information Technology, Computer, and Electrical Engineering*, jakarta, 2014. doi: 10.1109/ICITACEE.2014.7065756

[10] Ahmed, "The k-means algorithm," *A comprehensive survey and performance evaluation.*, p. Electronics 9.8: 1295, 2020. [Online] Available: https://doi.org/10.3390/electronics9081295

[11] Sinaga, "Unsupervised K-means clustering algorithm.," Yogyakarta, IEEE access 8, 2020, pp. 80716-

DOI: https://doi.org/10.29207/joseit.v2i2.5334

80727. doi: 10.1109/ACCESS.2020.2988796