Comparative Analysis Of Data Mining Using The Rought Set Method With K-Means Method

Marnis Nasution, Deci Irmayani, Ronal Watrianthos, Sudi Suryadi, Ibnu Rasyid Munthe

Abstract: The purpose of this article is to compare between two data mining methods. Namely rought sets and k-means which both types of data are mining for clustering. Data mining itself is a method used to explore knowledge from a pile of data which so far has only been archived. While the clustering method itself is one method used to classify tendency, either the rought set method or k-means itself is used to find tendency or classify data. Both the method of rought set and k-means have the advantages of each according to needs. It is important to know what the advantages of each method are before deciding to use which method to use.

Index Terms: Data Mining, Rought Set, K-Means

1. INTRODUCTION

Data mining itself is a series of processes to explore new knowledge from information or data that hasn't been known manually. Through data mining, new information that has been unknown is unknown. One of them is classifying tendencies. One example is looking for logic abilities of prospective students[1]. The logical ability of a computer student is very important to have because computer student needs that ability in the development of science, well when learning the base or when going deep into computer science on the higher level. However, unfortunately, the ability of logic can't be directly measured to prospective students who will enroll the college. The college only gives the basic skills. After the students attend the lecture then the ability of logic can be known from the students by observing the students score from logic and algorithm subjects. Using data from students which already have logic and algorithm grades then using data mining technique will uncover the knowledge from students data[2]. So that for the next prospective students can be known the logic ability before they become students by using the grade from their basic skill test and which school they're from. Some data mining methods which can be used are rought set and ke-means. both of the methods are equally can uncover the data knowledge, set of data by looking at tendency.

2 RESEARCH METHOD

2.1 Data Mining

Data mining is the process of posing various queries and extracting useful information, patterns, and trends often previously unknown from large quantities of data possibly stored in a database. Essentially, for many organizations, the goals of data mining include improving marketing capabilities, detecting abnormal patterns, and predicting the future based on past experiences and current trends. By increasing the size of a database, its supporting decision-making becomes more difficult. Moreover, the data might be from multiple sources and multiple domains. Thus, the integrity of a data also should be considered in a database approach[1]. Some of the data mining techniques include those based on rough sets, inductive logic programming, machine learning, and neural networks, among others. The main tasks of data mining are classification (finding rules to partition data into groups), association (finding rule to make associations between data), and sequencing (finding rules to order data). With the increased use of computers, there is an ever increasing volume of data being generated and stored. The sheer volume held in corporate databases is already too large for manual analysis and, as they grow, the problem is compounded. Furthermore, in many companies data is held only as a record or archive. Data mining encompasses a range of techniques, which aim to extract value from volume and form the foundation of decision making[1]

2.2 Knowledge Discovery in Database (KDD)

Along with the data's increasing growth, some large-scale databases already have gone far beyond the degree which artificial could analyze, but KDD is an effective way to solve the problems above. KDD is a newly-involved research area which is based on the combination of artificial intelligence and machine learning technology. As a decision-making support process, KDD, this is mainly based on artificial intelligence, machine learning, pattern recognition, and statistics, highly-automated analysis massive data, data mining, and thus makes correct decisions. The most important step of the KDD processes is the data mining and the goal of data mining which is to mine the concealed and significant knowledge from the massive data [2].The data mining processes are shown as follows Figure.1.

RESULT

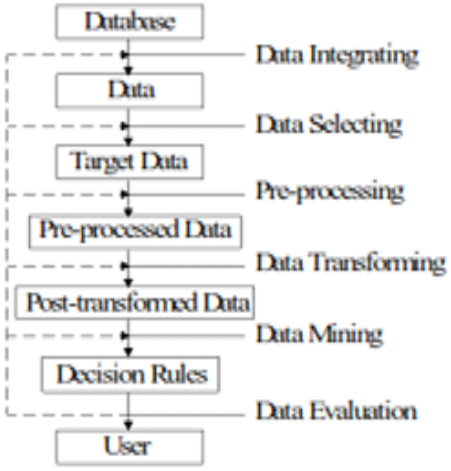

Figure 1. *The Procedure of data mining*

From Figure.1, we know that the procedure of data mining contains 3 stages, data preparation, data mining, and data evaluation, and the stage of data preparation includes 4 steps, such as data integrating, data selecting, pre-processing and

data transforming. The decision rules are abstracted by data mining. All the steps have connection with data evaluation.

2.3 Data Mining Rought Set

The Rough Set Theory has had a significant impact in the field of data analysis and as a result has attracted the attention of researchers worldwide. Owing to this research, various extensions to the original theory have been proposed and areas of application continue to widen. As Jensen observes, many rough set based clinical decision models are available to assist physicians, in particular the inexperienced ones, to recognize patterns in symptoms and allow for quick and efficient diagnosis. Results available support the premise that systems based on RST give accuracy and reliability that is comparable to Physicians though accurate input data is required. Such systems, in conjunction with other ICT facilities, can be particularly helpful in remote areas of developing countries where healthcare infrastructure is patchy. Detailed discussion on Rough Sets has been presented by the authors in[3].

2.3 Data Mining K-Means

From a practical point of view, clustering analysis is one of the main tasks of data mining. It is now used in many areas like knowledge discovery, pattern recognition and so on. Many clustering analysis algorithm are available of which the most wellknownis the K-means algorithm which is based on division. Clustering can enable users to find the relevant documents more easily. This paper aimed to investigate the websites that are in top in one cluster and other sites in second cluster and for top ranking we need URL, back-links, in-links, length of title are required. “Clustering based on k-means” that it is closely related to a number of other clustering and location problems which include the Euclidean k-medians which minimize the sum of distances to the nearest center, and the geometric k-center problem, which aimed to minimize the maximum distance from every point to its closest center[4]. K-Means clustering is a very popular algorithm to find the clusters in a dataset by iterative computations. It has the advantage of simple implementation and finding at least local optimal clustering. K-Means algorithm is employed to find the clustering in dataset. The algorithm is composed of the following steps:

1. Initialize k cluster centers to be seed points. (These centers can be randomly produced or use other ways to generate).
2. For each sample, find the nearest cluster center, put the sample in this cluster and recomputed centers of the altered cluster (Repeat n times).
3. Exam all samples again and put each one in the cluster identified with the nearest center (don’t recomputed any cluster centers). If members of each cluster haven’t been changed, stop. If changed, go to step 2[4].

DISCUSSION

3.1 K-Means Algorithm Process

K-Means Algorithm process uses the rapid miner can be seen at the following:

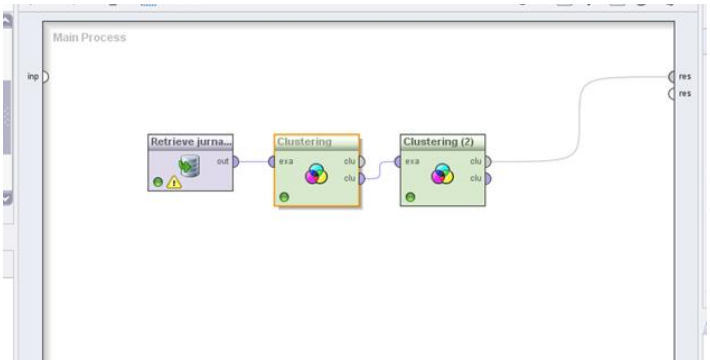

***Figure 2**. The K-Means Algorithm Process Using Rapid Miner*

By using K-means clustering modeling as shown above, with initialization of 2 clusters, then the number of cluster_0 there are 13 items, cluster_1 there are 14 items with a total number of data as many as 27.

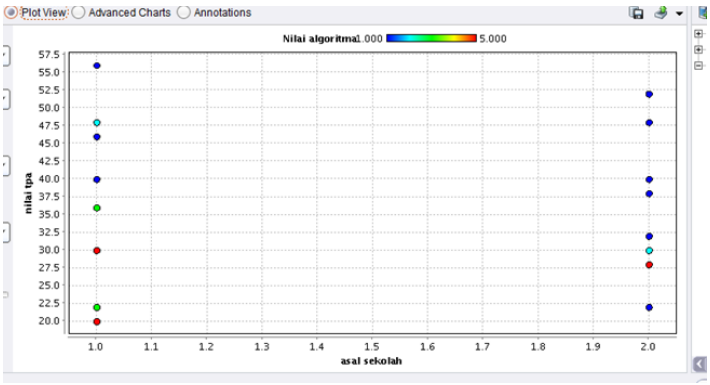

***Figure 3**. The Result of K-Means Algorithm Using Rapid Miner*

From the results of clustering from the display with scatter diagram, it can be seen clearly that the clustering of students from vocational school is far superior to students from high school, even the TPA score of students who come from vocational schools does not significantly affect their logic ability[5].

3.2 Rought Set Algorithm Process

The rule generated by rought set:

School origin(SMA) AND department (IPS) AND tpa score(2) => Algorithm score(B) OR Algorithm Score(A) OR Algorithm Score(E) OR Algorithm Score(C)

School origin(SMK) AND departmen(Otomotif) AND tpa score(4) =>Algoritma score(A)

School origin(SMK) AND departmen(Akuntansi AND tpa score(3) =>Algoritma score(A)

School origin(SMK) AND departmen(Otomotif) AND tpa score(3) =>Algoritma score(A)

School origin(SMA) AND departmen(IPA) AND tpa score(4) =>Algoritma score(B) OR Algoritma score(A)

School origin(SMA) AND departmen(IPS) AND tpa score(3)

=>Algoritma score(B) OR Algoritma score(A)

School origin(SMK) AND departmen(Listrik) AND tpa score(3) =>Algoritma score(A)

School origin(SMK) AND departmen(Lainnya) AND tpa score(2) =>Algoritma score(A)

School origin(SMA) AND departmen(IPS) AND tpa score(1) =>Algoritma score(E)

School origin(SMK) AND departmen(Perkantoran) AND tpa score(1) =>Algoritma score(B)

School origin(SMA) AND departmen(IPA) AND tpa score(5) =>Algoritma score(A)

School origin(SMA) AND departmen(IPA) AND tpa score(2) =>Algoritma score(A)

School origin(SMA) AND departmen(IPA) AND tpa score(3) =>Algoritma score(C)

School origin(SMK) AND departmen(Lainnya) AND tpa score(3) =>Algoritma score(B)

School origin(SMK) AND departmen(Penjualan) AND tpa score(2) =>Algoritma score(C)

School origin(SMK) AND departmen(TKJ) AND tpa score(4) =>Algoritma score(A)

School origin(SMK) AND departmen(TKJ) AND tpa score(2) =>Algoritma score(B)

School origin(SMK) AND departmen(Perkantoran) AND tpa score(3) =>Algoritma score(A)

School origin(SMK) AND departmen(TKJ) AND tpa score(1) =>Algoritma score(E)

School origin(SMK) AND departmen(Akuntansi) AND tpa score(4) =>Algoritma score(A)

School origin(SMK) AND departmen(Penjualan) AND tpa score(1) =>Algoritma score(A)

School origin(MA) AND departmen(IPS) AND tpa score(1) =>Algoritma score(C)

From the rules generated by rosetta, it can be seen clearly the difference between students from high school and vocational school. Students from the vocational School whether majoring in technology or not having a larger degree to obtain high grades in logic and algorithm subjects[5].

4 CONCLUSION

From the two previous kinds of research to measure the logic ability of students with AMIK Labuhan Batu Case Study, each of the previous studies can show that there is a relation between the origin of student school, the results of the student basic skills test with the student's logic abilities. Where if the rought set method is used then the results to be obtained are in the form of if-then rules, from that rule it can be seen that the students with origin from vocational high school have a diverse logic ability and that is also affected from the academic skill test, the better the academic ability, the better the logic ability. While the results of the research using data mining k-means can be seen that diversity for k-means results cannot be seen in a more detailed tendency. So that from the two research measuring students' logic ability using both the rought set and using the k-means it can be seen that the rought set can provide more detailed result according to the condition attributes. The more attribute conditions the more possibilities will be found while for k-means the results only show a greater tendency. Both methods can be used properly in accordingly with needs, if the process of looking to determine the tendency is greater then it is better to use k-means while if needed a detailed explanation it'll be better to use the rought set method

REFERENCES

[1]. M. Hossein and F. Zarandi, “Application of Rough Set Theory in Data Mining for Decision Support Systems (DSSs),” J. Ind. Eng. 1, vol. 1, pp. 25–34, 2008.

[2]. L. I. Wanqing, M. A. Lihua, and W. E. I. Dong, “Data Mining Based on Rough Sets in Risk Decision-making : Foundation and Application,” WSEAS Trans. Comput, vol. 9, no. 2, pp. 113–123, 2010.

[3]. M. Sudha, “Comparative Analysis between Rough set theory and Data mining algorithms on their prediction,” vol. 13, no. 7, pp. 3249–3260, 2017.

[4]. M. Jindal and N. Kharb, “K-means Clustering Technique on Search Engine Dataset using Data Mining Tool,” vol. 3, no. 6, pp. 505–510, 2013.

[5]. M. Nasution, “IMPLEMENTASI DATA MINING K-MEANS UNTUK MENGUKUR KEMAMPUAN LOGIKA MAHASISWA (STUDI KASUS : AMIK LABUHAN BATU),” Informatika, vol. 5, no. 1, 2017.