The Implementation of K-Means Algorithm for Clustering Traffic Accident Rates on the Highway

Anita Ahmad Kasim [a,1], Siti Uyun Mubarak [b,1]

[a,] Information Technology Department, Faculty of Engineering Universitas Tadulako, Palu Indonesia
[b,] Study Program of Informatics, Faculty of Engineering, UniversitasTadulako, Palu, Indonesia

[1]nita.kasim@gmail.com*; [2]uyunmubarak@gmail.com

ARTICLE INFO

Article history
Received
Revised
Accepted

Keywords
Data mining
Clustering
K-Means

ABSTRACT

Introduction: The increase of population in Palu city has result in increased vehicle ownership and increased the risk of traffic problems such as traffic accidents. So far, the accident data in the Palu Resort Police Station has not been fully utilized by the interests of related parties. Therefore, the accumulation of data will be processed by data mining techniques. This study aims to cluster the level of accidents in Palu city based on the age of the perpetrators, where the results of the clustering will be used as consideration for the more targeted socialization of traffic accidents. Based on the results of testing with 2 different centroid initialization methods, the results obtained indicate that centroid initialization using the ranking method has an SSE value of 233.0690397 while centroid initialization using a random method has an SSE value of 356.42304. It proves that centroid initialization using ranking method has better clustering results compared to centroid initialization experiments using random methods.

1. Introduction

The increasing land use in Palu city such as tourism areas, shopping areas, recreation areas, and housing areas is one of the impacts that occur due to an increase in the population of Palu city. The emergence of these areas makes the higher mobility of people from one location to another. It causes an increase in the number of vehicle owners which then increase the risk of traffic problems, namely traffic accidents.

So far, the accident data in the Palu city Police Station has not been fully utilized by the interests of related parties. Therefore, the accumulation of the data will be processed with data mining techniques to cluster the accidents that occur based on the age of the accident perpetrators. The results of the accident rate clustering will be used as a material for socializing more targeted traffic accidents and as a preventive measure to assist the police in reducing the number of accidents.

Data mining is defined as the process of discovering new patterns from very large data sets, including methods that are slices of artificial intelligence, machine learning, statistics, and database system. Data mining is known as a big data, which has four characteristics: high-volume, high-variety, high-velocity, and high-veracity [1].Clustering is the process by which objects are classified into groups called clusters. In terms of grouping, the problem is grouping unlabeled collections into meaningful groups without prior information. Each label associated with the object is obtained solely from data [2].

K-means is included in the cluster partition i.e. every data must be included into a particular cluster and it is possible for each data to be included in the certain cluster at one stage of the process. The next stage moves to another cluster. K-means separates data into separate k regions, where k is a

positive integer number. The K-means algorithm is very well known for its ease and ability to cluster large data and outliers very quickly [3].

Research related to traffic accidents has been carried out previously with the title “Analysis of Traffic Accidents Rate with the Method of Association Rule Using Apriori Algorithm”. This study aimed to look for the relationship between the factors that caused accidents with accident categories of material loss, mild, moderate, and severe. The variables in this study included the type of accident, light conditions, time of occurrence and geometric shapes [4].

Another research that discusses traffic accidents has also been carried out with the title "Road Traffic Accidents Classification System in Boyolali City Using Naïve Bayes Method". The purpose of this study was to determine the classification of accidents whether they occurred frequently or not. The naïve bayes method applied in this system was to calculate the greatest probability on predetermined variables such as cause, day, age, place of occurrence, and time [5].

Another research was also conducted with the title “Analysis of Highway Traffic Accidents in Semarang City Using Method K-means Clustering”. The purpose of this study was to determine the mapping of accident events based on age, vehicle type, causative factors, and types of days where the accident occurred and the compilation of an accident mapping database system based on the accident data analysis information system in Semarang that is able to increase the level of alertness of road users [6].

Based on the previous studies, this study aims to cluster the accident rates which will be grouped into 3 clusters based on the age of perpetrators. The three clusters will be labeled in the form of a few, medium, and many accident rate using 4 parameters: the gender of the perpetrators, the vehicle transmission type, the road models, and the number of vehicles involved.

2. Method

In this study, several stages of the KDD (Knowledge Discovery in Database) process can be seen in Fig 1.

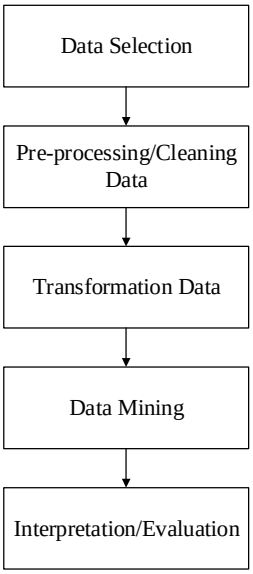

Fig 1. KDD Process

A. Data Selection

The accident data used in this study amounted to 857 data consisting of the accidents from 2017 to 2018. From the accident data, there are only a few parameters or information that can be used to determine the age of frequent accidents and their causes.

Not all parameters in the accident data can be included to do the clustering process. This is due to incomplete data issued on the existing accident report. Only parameters that affect the occurrence of an accident will be included. Therefore, data selection is performed to select data from several factors that contain the main information to be used as cluster attributes/parameters.

After the data selection stage, the data for the study object consists of 4 parameters:

a. The gender of the perpetrators : gender perpetrators of accident consist of male and female.
b. Vehicle transmission type : the vehicle transmission type used by the accident perpetrators is divided into 2 namely Automatic Transmission (AT) and Manual Transmission (MT).
c. Road models : road models at the scene experienced by the accident perpetrators include straight roads, bends, and intersections.
d. Number of vehicles involved : the number of vehicles involved in accidents is single, multiple, and consecutive.

B. Pre-processing/Data Cleaning

Pre-processing/data cleaning is used to eliminate inconsistent data. Data cleaning is performed on the accident data included in a hit-and-run, where the identity of the accidents perpetrators and vehicle transmission type used by the accident perpetrators are unknown. After pre-processing/data cleaning, there are 689 accident data that will be included in the clustering process.

C. Transformation Data

At this stage, transformation data is carried out by changing the accident data which is descriptive in the form of nominal data types and does not change the information contained there in. The example of accident data before transformation data can be seen in Table 1 while the accident data after transformation data can be seen in Table 2.

Table 1. Accident Data before *Transformation Data*

Identity of Perpetrators/Actors	Vehicle Identity	Number of Vehicles Involved	Road Models
Lk, Fahri, 16 years, Private Work, Western Palu, (Heavy Injury)	SM, Yamaha Mio	Multiple	Intersection

Table 2. Accident Data after *Transformation Data*

Actors Age	Gender		Transmission Type		Road Models			Number of Vehicles Involved		
	Male	Female	AT	MT	Straight	Bend	Intersections	Single	Multiple	Consecutive
16	1	0	1	0	0	0	1	0	1	0

D. Data Mining

After transformation data has been successfully carried out, the accident data can be processed to the clusters using the K-means algorithm. In this study, the number of cluster in the accident data is divided into three clusters including cluster 1 which is a cluster with a few accident rates, cluster 2 is a cluster with a medium accident rate, and cluster 3 is a cluster with many accident rates.

E. Interpretation/Evaluation

The results of the clustering process will providea pattern or information that is a member of the age of accident perpetrators in Palu city, which included in a few, moderate, and many accidents. The age of the accident perpetrators who are members of cluster 3 (many accident rates) will be the focus of the police to conduct socialization.

The process of K-means algorithm contained in the system can be seen in Figure 2. The process begins when the user inputs the accident data, then the system will save the data to the database. Furthermore, the system will perform a K-means algorithm process where each data will be

Tadulako Science and Technology Journal
Vol 1, No. 1, December 2020, pp. 7-12

measured by the initial centroid that has been determined in each cluster using the Euclidean distance theory. The Euclidean distance is represented in Equation (1).

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \qquad (1)$$

For d_{ij} is the distance of the object between the data value and the cluster center value; p is the number of data dimensions; x_{ik} is the data value of the dimension *k*; x_{jk} is the cluster center value of the dimension *k*.

The distance from the results of the calculation of the Euclidean distance will be compared and selected the closest distance between data with cluster center. This distance indicates that the data is in a group with the closest cluster center.

After the members of each cluster are known, the new centroid will be calculated based on the average of the data in each cluster. If there is a change in centroid value in the current iteration and the previous iteration, the distance data will be recalculated. If there is no centroid change, the cluster results will be saved to the database and then displayed to the user.

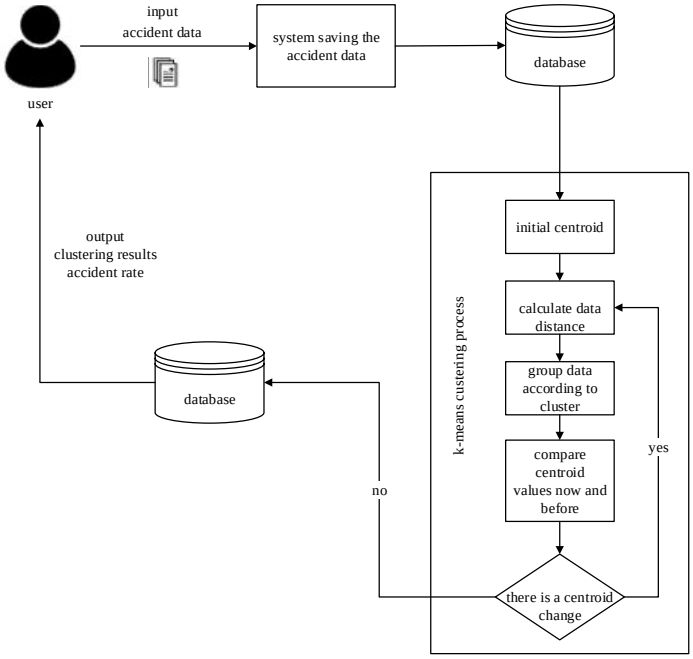

Fig 2. Algorithm Process *K-means*

3. Results and Discussion

The K-means algorithm is very sensitive to initial centroids. The initial centroid difference will give a different clustering result and if the initial centroid given is a bad centroid, it can be ascertained that the clustering results are also not good [7]. Therefore, the initial centroid value in traffic accident data is done by comparing 2 different centroid initialization methods i.e. centroid initialization using random methods and centroid initialization using ranking methods. The results of observing 2 centroid initialization methods are different, so the calculation of the value of *SSE* (Sum of Square Error) will be performed to find better clustering results. The *SSE* formula can be seen in Equation (2).

$$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2 \qquad (2)$$

SSE is the sum squares of all clusters; k is the number of clusters; n_j is the sum of squares of each cluster; x_{ij} is the closest distance value to the data; $\bar{x}_j$ is the average value of each other.

A. Centroid Initialization Using Random Methods

Centroid initialization using random methods is done by taking the initial centroid value in each cluster randomly. This value is derived from the value owned by a certain age that represents each cluster. After getting the results of clustering with centroid initialization using a random method, the *SSE* (Sum of Square Error) values are calculated as follows:

SSE = 238.1097346 + 64.245994 + 54.0673

= 356.42304

B. Centroid Initialization Using Ranking Methods

Centroid initialization using ranking methods is done by taking the initial centroid value in each cluster by finding the lowest value, middle value and the highest value. After getting the results of clustering with centroid initialization using a ranking method, the *SSE* (Sum of Square Error) values are calculated as follows:

SSE = 83.07735568 + 35.33367327 + 114.658011

= 233.0690397

From the *SSE* calculation, the results are obtained that the centroid initialization using the ranking method has the smallest *SSE* value compared to the centroid initialization using the random method. Based on this finding, it can be concluded that the centroid initialization using the ranking method has better clustering results.

In the initial centroid initialization using the ranking method, the level of traffic accidents in cluster 1 (few accident rate) is a cluster that has 31 members, while cluster 2 (medium accident rate) has 15 members, and cluster 3 (many accident rates) has total of 8 members.

The age of the accident perpetrators included in cluster 1 members are age 13, 26, 31, 33, 35, 36, 37, 38, 39, 40, 43, 44, 45, 46, 47, 48, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, and 65 years. In this cluster, most of the age of the perpetrators is included in the category of late adulthood, early elderly and late elderly. Furthermore, the age of the accident perpetrators included in cluster 2 members are ages 12, 14, 15, 24, 25, 27, 28, 29, 30, 32, 34, 41, 42, 49, and 50 years. This cluster is dominated by the age of the accident perpetrators who are in the early adulthood category. In addition, the age of the accident perpetrators included in cluster 3 members are aged 16, 17, 18, 19, 20, 21, 22, and 23 years. The perpetrators in this cluster arein the category of late adolescence.

The gender of the accident perpetrators in cluster 1, cluster 2, and cluster 3 placesmale as the gender that frequently becomes the perpetrator of the accident than women. As for the type of transmission in the form of manual transmission in cluster 1 and cluster 2, it is the type of transmission that is most widely used by the accident perpetrators. For cluster 3, the type of transmission that is frequently used by the accident perpetrators is automatic transmission.

Furthermore, the road model where accidents often occur in cluster 1, cluster 2, and cluster 3 places the straight road as the most common road model experienced by the perpetrators of the accident, followed by the crossing road model and the bend road model. Furthermore, the number of vehicles involved in cluster 1, cluster 2, and cluster 3 put the number of dual vehicles in the first position followed by the number of single vehicles and the number of consecutive vehicles.

4. Conclusion

Based on the testing and analysis of the design of a cluster system of traffic accident rates on the highway, the following conclusions can be drawn:

1. Centroid initialization experiment using the ranking method has SSE value 233.0690397 while centroid initialization using random method has SSE value of 356.42304. These findings prove that the centroid initialization using ranking methods has better clustering results than the centroid initialization experiments using random methods.
2. The age of accident perpetrators included in cluster 3 (many accident rates) is in the age range of 16 to 23 years. This age will be the focus of the police in conducting socialization.

References

[1] Suyanto, *Data Mining For Classification and Data Clustering*, Revised. Informatics, 2018.

[2] O. K. I. A. Saputra, "Application Of K-Means Clustering Algorithm For Grouping Of Markisa Fruits Based On The Application Of K-Means Clustering Algorithm," 2017.

[3] W. S. Azis and D. Atmajaya, "Grouping Student Interest Reading Using K-Means Method," *Ilk. J. Ilm.*, vol. 8, no. 2, p. 89, 2016.

[4] W. Alimuddin, E. Tungadi, and Z. Saharuna, "Traffic Accident Rate Analysis with Association Rule Method Using Apriori Algorithm Analysis of Traffic Accident Rate with Association Rule Method," no. January, 2018.

[5] S. B. Utami and Y. Al Irsyadi, "Road Traffic Accident Classification System in Boyolali City Using the Naïve Bayes Method," *J. Mitra Manaj.*, vol. 2, no. 4, pp. 273–285, 2018.

[6] M. S. Fajar, "In Semarang City Using the K-Means Clustering Method," 2015.

[7] A. Maududie and W. C. Wibowo, "Improvement of k-means initialization using minimum forest graphs," *Pros. Semin. Ilm. Nas. Komput. and Sist. Intelijen (KOMMIT 2014)*, vol. 8, no. Kommit, pp. 8–15, 2014.