International Journal of
Science and Engineering Investigations
vol. 7, issue 81, October 2018
ISSN: 2251-8843

Implementation K-Means Clustering Analysis of Traffic Accident in Semarang City Using Weka Interface

Wiwik Budiawan[1], Singgih Saptadi[2], Ary Arvianto[3], Pertiwi Andarani[4]
[1,2,3]Department of Industrial Engineering, Diponegoro University
[4]Department of Environmental Engineering, Diponegoro University
([1]wiwikbudiawan@ft.undip.ac.id)

Abstract- Traffic accidents are a major problem in society because they can cause loss in terms of costs and human safety. Data Mining has proven to be a reliable technique for analyzing traffic accident data and providing productive results. Analysis of traffic accident data currently only focuses on identifying the cause of the accident. Accidents occur more frequently in certain locations. Analysis on these locations can help identify the causes of accidents at these locations. Of the 501 accidents data recorded in the Semarang toll road database, the data was selected to be 491 data. From 491 data then grouped based on the location of the accident into 93 data. The data is then analyzed using the K-Means Clustering algorithm with the help of the Weka Interface application. The results of the analysis show the frequency of accidents at each location has the potential to cause accidents. The results of the analysis show that there are three clusters formed from the Weka Interface application (based on K-Means Algorithm). The first cluster instances are 20 data (22%) with cluster centroids at location 10 (2 Km), the second cluster instances are 10 data (11%) with cluster centroids at location 9 (1,8 Km) , and the third cluster instances are 63 data (68%) with cluster centroids at location 38 (7,6 Km).

Keywords- *Accidents, Data Mining, Clustering , K-Means, WEKA*

I. Introduction

Semarang Toll Road is a toll road that connects Semarang City (West, East and South Semarang). The Semarang Toll Road Network consists of three sections, namely for Section A, Section B, and Section C. Section A is the road between Krapyak - Jatingaleh (Length of 8 kilometers, width of each lane is 3.6 meters, and climbing lane with a width of 3 m) . Section A is operated since 1987. Section B is the road between Jatingaleh - Srondol (Length 6 kilometers and width of each lane is 3.6 meters). Section B is operated since 1983. Section C is the road between Jatingaleh - Kaligawe (10 kilometers long, width of each lane is 3.6 meters, and climbing lane with a width of 3 m). Section C is operated since 1997. General Conditions of Semarang Toll Road are roads that are mostly in hilly areas [1].

Based on data from the Semarang toll road manager, there were 501 accidents in 2007 to 2017 (an average of 50 accidents per year). Accident classification is divided into four categories (Material Damage Accident, Minor Accident, Major Accident, and Fatal Accident). "Damaged material" category is a category of accidents that have no human casualties. "Minor" category is a category of accidents that cause minor injuries to humans. The "Major" category is a category of accidents that cause serious injuries to humans. The "Fatal" category is a category of accidents that cause death in humans. The percentage of accident distribution can be seen in the following graph:

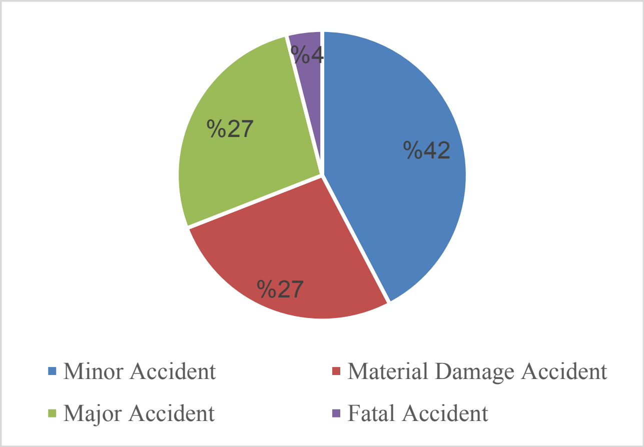

Figure 1. Presentage of Accodent in Semarang Toll Road

From the accident data, there are various factors that cause accidents, among others: driver factors (lack of anticipation, carelessness, sleepiness, drunkenness, and disorderly); vehicle factors (broken tires, slips, loose brakes and engine damage); road factors (road damage, lack of road equipment, and road maintenance work); and environmental factors (smoke, fog and weather).

Traffic accidents are a major problem in life because they can cause losses in terms of costs and human safety. Beshah and Hill [2] states that every year around the world, more than 1.2 million people die, and 50 people are injured in accidents. A study conducted by [3] states that the main cause of death after kardiovascular is road traffic accidents.

Traffic accidents are influenced by several factors due to the condition of the driver, road characteristics, environment and weather [4]. Until now, accident-prone areas have increased which has resulted in many casualties. In the world of computer science, data mining is widely known as a technique for summarizing data in different ways than the usual ones, finding unexpected relationships, finding patterns that are understandable and useful for data owners [5]. Various types of data mining techniques [2] [6] such as association techniques, classification, and clustering are widely used in analyzing areas prone to traffic accidents.

Clustering of accident-prone areas can be used as information for drivers. KMeans clustering method is used to group data that has the same characteristics and group them into a cluster. Therefore, by using one of the data mining methods, this study can classify accident-prone areas based on the K-Means Clustering method.

At present, data has become a very important requirement. Rapid development in information technology makes all information can be stored and accessed. This also encourages the emergence of a very large database system, the data warehouse. The problem that then arises is how to find out the information contained in a very large data warehouse. Knowledge discovery in Database (KDD) is defined as potential, implicit and unknown information extraction from a set of data. The KDD process involves the results of the data mining process (the process of extracting a pattern of data) and then changing the results accurately into easily understood information.

There are several kinds of information / knowledge search techniques in KDD. There are quantitively approaches, such as probabilistic and statistical approaches. Some approaches utilize visualization techniques, classification approaches such as inductive logic, pattern searching, and decision tree analysis. Other approaches include deviation, trend analysis, genetic algorithms, artificial neural networks and a mixed approach of two or more of the existing approaches.

Wright [7] divides the six most essential elements in the information / knowledge search technique in KDD, namely: working on large amounts of data, required efficiency with regard to data volume, prioritizing accuracy / accuracy, requiring high-level language usage, using several forms of learning automatically, and produce interesting results.

II. Methods

The purpose of this study is to get an overview of the number of clusters and cluster centers from existing accident data. This study uses the K-Means algorithm to determine cluster members. This cluster is then used to classify any vulnerable areas. K-Means algorithm is a clustering algorithm that groups data based on the center of the cluster (centroid) closest to the data. Besides, K-Means is also included as a supervised clustering technique so that at the beginning of computing the number of clusters that have been desired in advance [8].

The purpose of the K-Means Algorithm is to group data by maximizing the similarity of data in one cluster and minimizing the similarity of data between clusters. A measure of the similarity used in a cluster is a function of distance. So maximizing the similarity of data is obtained based on the shortest distance between the data to the centroid point.

The K-Means algorithm has the following stages[9]:

1. The initial stages carried out in the process of data clustering using the K-Means algorithm is the formation of the cj centroid starting point. In general, the formation of centroid starting points is randomly generated. The number of cj centroids generated is in accordance with the number of clusters specified at the beginning.
2. After k centroid is formed, then calculated the distance of each data xi with the centroid j to k, denoted by d (xi, cj). There are several measures of distance used as a measure of the similarity of a data instance, one of which is Euclidean distance [8].Euclidean distance calculation as in the following equation:
3. A data will be a member of a cluster provided that it has the minimum distance between all existing clusters.
4. Group data that is a member of each cluster.
5. Renew the center value of the cluster that can be calculated by searching for the average value according to the number of members of each cluster according to the formula
6. Repeat steps 2-5 until no more data moves.

The software used in this study is Weka. The purpose of using software this is comparing the results with theoretical calculations with the results obtained with the process in the Weka Interface. The Weka Interface research tool, as shown in Figure 2, is a Java-based open source data mining application. This application was first developed by the University of Waikato in New Zealand. Weka has many machine learning algorithms that can be used to generalize or formulate a collection of sampling data. One of them is clustering using the K-Means algorithm.

Waikato Environment for Knowledge Analysis
Version 3.8.3
(c) 1999 - 2018
The University of Waikato
Hamilton, New Zealand

Figure 2. Weka Interface Research Tool

Sharma & Paul [10] states that clustering techniques have wide use and currently have a tendency to increase along with the amount of data that continues to grow. K-means is a simple technique for cluster analysis. The aim is to find the best division of entity n into group k (called a cluster), so that the total distance between group members and entroids is appropriate, regardless of the minimized group. Each entity belongs to a cluster with the closest mean. This results in partitioning the data space into Voronoi Cell.

Research Location:

The study was conducted on the Semarang (Indonesia) Toll Road in sections A, B, and C. Currently the Semarang Toll Road manager only collects operational data regarding the number of vehicles passing and accident report data.

III. Results and Discussion

A. Data

The data used in this study are data obtained from the Semarang (Indonesia) Toll Road Manager. The data used in this study are accident data for the period of 2007 to 2017 (number of accidents are 501). The initial steps that need to be done before processing are data selection, data cleaning, and data transformation.

B. K-Means algorithm

K-Means algorithm is arranged based on the distribution of objects. In this algorithm, the first element in the cluster can be selected to be used as the cluster's centroid point. The KMeans method algorithm will repeat the following steps until stability (no object can be moved):

1. Determine the coordinates of the midpoint of each cluster. In this data the initial cluster center is made into two pieces. The application of the K-Means algorithm on clustering will produce two clusters based on the number of accidents occurring.
2. Determination of the cluster center is used as a reference to perform calculations on each row of the test data table (distance to the cluster center).
3. Grouping of objects based on the minimum distance of the results in stage 2.

C. Testing with the Weka Interface Application

The stages in the Weka interface application are the same as the stages in the K-Means algorithm. The convenience obtained when using this application is graphic info in representing the output of the algorithm.

1. The coordinates of the initial midpoint are equated with the stages in the k-means algorithm, namely three cluster centers. Three cluster centers are calculated from the cluster analysis hierarchy.

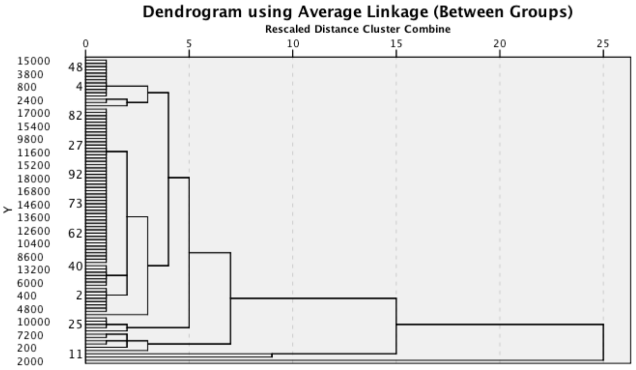

Figure 3. Dendrogram Using Average Linkage (Between Groups)

2. The three clusters are then used as references to calculate each row of data.

```
Initial starting points (random):
Cluster 0: Lok10,0.6,1,0.4,0
Cluster 1: Lok9,0,0.571429,0.2,0
Cluster 2: Lok38,0,0.142857,0,0
```

3. Grouping of objects based on the minimum distance of the results in stage 2.

```
kMeans
======

Number of iterations: 8
Within cluster sum of squared errors:
97.92685600907035

Initial starting points (random):

Cluster 0: Lok10,0.6,1,0.4,0
Cluster 1: Lok9,0,0.571429,0.2,0
Cluster 2: Lok38,0,0.142857,0,0

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute      Full Data      0       1       2
               (93.0)   (20.0)  (10.0)   (63.0)
==============================================
KM       Lok1       Lok3       Lok1       Lok2
Minor     0.1656       0.51      0.14 0.0603
Major     0.0952       0.0786    0.4286    0.0476
Material 0.0882        0.15      0.34      0.0286
Fatal     0.1183     0.15          0      0.127

Time taken to build model (full training data) : 0
seconds

=== Model and evaluation on training set ===

Clustered Instances

0      20 ( 22%)
1      10 ( 11%)
2      63 ( 68%)
```

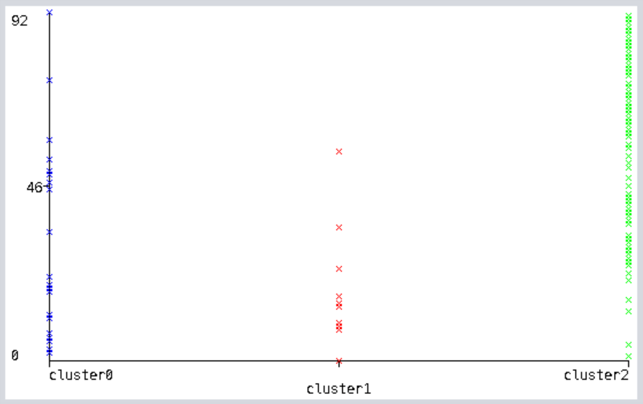

Figure 4. Plot of Cluster Instances

IV. CONSLUSION

Based on the results of data processing using the Hierarchy and Weka Interface Cluster Analysis (KMeans algorithm) there are three clusters formed. The three clusters are sequentially starting from very vulnerable areas (location cluster centers 10/2 Km, cluster instances 20), vulnerable areas (location cluster centers 9 / 1.8 Km, cluster instances 10), and quite vulnerable areas (location cluster centers 38 / 7.6 Km, cluster instances 63).

REFERENCES

[1] Jasa Marga, "Semarang Toll Road Section A-B-C," Semarang, 2018.

[2] T. Beshah and S. Hill, "Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia," in *The 2010 AAAI Spring Symposium*, 2010.

[3] D. Bener, A; Crundall, "Road traffic accidents in the United Arab Emirates compared to Western countries," *Adv. Transp. Stud. an Int. J.*, vol. A, no. 6, 2005.

[4] A. Pakgohar, R. S. Tabrizi, M. Khalili, and A. Esmaeli, "The role of human factor in incidence and severity of road crashes based on the CART and LR regression: A data mining approach," *Procedia Comput. Sci.*, vol. 3, pp. 764–769, 2011.

[5] E. Prasetyo, Data Mining Concept and Application with Matlab (in Bahasa). Andy Offset, 2012.

[6] M. Sowmya and P. Ponmuthuramalingam, "Analyzing the Road Traffic and Accidents with Classification Techniques," *Int. J. Comput. Trends Technol.*, vol. 5, no. 4, 2013.

[7] P. Wright, "Knowledge Discovery in Databases: Tools and Techniques," Magazine XRDS: Crossroads, The ACM Magazine for Students - Special issue on networks and distributed systems, pp. 23–26, 1998.

[8] D. T. Larose, "Hierarchical and k- Means Clustering," in *Discovering Knowledge in Data: An Introduction to Data Mining*, Wiley, 2005.

[9] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2012.

[10] B. R. Sharma and A. Paul, "Clustering Algorithms: Study and Performance Evaluation Using Weka Tool," *Int. J. Curr. Eng. Technol.*, vol. 3, no. 3, 2013.

Wiwik Budiawan was graduated from Diponegoro University majoring in Industria Engineering, she continued her study at Bandung Institute of Technology and joined Ergonomic and Work System Engineering Laboratory in 2010. He has received some awards and grant concerning on safety transportation, accident prevention and academic excellence during her graduate degree. He did researches about accident analysis and prevention in Mass Transportation. Currently He is working at Department of Industrial Engineering, Faculty of Engineering, Diponegoro University,Indonesia.