Clustering Analysis of Traffic Accident in Semarang City

Wiwik Budiawan[1,*], *Bambang* Purwanggono[1]

[1]Department of Industrial Engineering, Diponegoro University, Semarang - Indonesia

Abstract. Traffic accidents are one of the global issues that require serious handling. Accidents occur in different places with different incidents, which makes it difficult to determine which areas have a high degree of traffic accidents. Information about areas prone to accidents is needed by the community and law enforcement. Such information can be taken into consideration for the supervision and anticipation action especially for the police. In this study made a cluster to analyze the areas prone to accidents in the city of Semarang. The method used is cluster analysis where the grouping to determine the vulnerability of an area. The result of the research stated that the level of traffic accident vulnerability is mostly happened in Semarang - Semarang regency passing through Semarang regency. In addition, the level of vulnerability in the city of Semarang occurred on weekdays. From the validation results that have been made, the suitability of the hazardous modeling area that has been formed is: Occurs more likely on weekdays (Monday, Thursday, Friday and Sunday); At an average Kilometer of 19.75-Direction B; During Afternoon and Evening; Small and Large Vehicle Types; Cloudy, Drizzle and Rain.

Keywords: **clustering analysis; traffic accident; data mining.**

1 Introduction

The increasing number of residents in the city of Semarang every year causes the need for transportation is also increasing (1,602,717 people in 2016). Population growth in Semarang City in 2016 reached 0.47% per year [1]. An increase in population will indirectly increase the risk of growing transportation problems. Transportation problems according to Tamin [2] are not only limited to the limited number of modes of transportation. The rapid development of transportation will indirectly increase the risk of growing traffic problems (traffic accidents). Traffic accidents according to the Law of the Republic of Indonesia No. 22 of 2009 is an incident on an unexpected and unintentional highway involving a vehicle with or without other road users that results in human casualties and / or property losses. Data from the Police of the City of Semarang Resort (Polrestabes-Semarang) states that in 2017 there were 936 traffic accidents in Semarang [3].

This clearly needs attention and effective handling because it relates to losses suffered by the community during an accident. Polrestabes-Semarang recapitulates the number of accidents, number of victims, and total material losses in an area to be presented in the form of descriptive statistical analysis.

Generally, traffic accident data analysis is presented as an information on the results of description statistics ([4][5][6]), which includes: (a) frequency distribution, (b) periodic data (data arranged in a time sequence), (c) weighting (the value used to calculate the accident index based on the characteristics of each accident), (d) z-score technique (analysis based on raw standards), (e) cumulative summary technique (procedure used to identify accident locations, or (f) stick diagram analysis (used to classify similar types of accidents).

Data processing with new descriptive statistical techniques reveals a small portion of information hidden in the database. Important information that supports decision making to reduce and prevent traffic accidents, such as patterns of causes of accidents and trends that develop due to accidents have not yet been presented [7]. Another limitation of descriptive statistics is that it does not show causality between parameters or to recognize similarities in phenomena that may be hidden in the data [8].

Data mining is known as a technique for summarizing data by finding unexpected relationships, finding patterns that can be understood and useful for data owners (Larose, 2005). Several studies related to the use of data mining in processing traffic accident datasets in Indonesia, such as the use of Apriori [9] and Naïve Bayes techniques [10] to predict traffic accidents. Use of Association techniques [11] relating to accident predictions based on the event rules contained in the dataset. However, research related to data clustering as a basis for identifying areas prone to traffic accidents is still low, so research is needed on data clustering.

[*] Corresponding author: wiwikbudiawan@ft.undip.ac.id

2 Proposed Method

Traffic accidents are defined as unexpected and unintentional road events. Traffic accidents involve vehicles that cause human casualties or property losses. In Semarang, the party authorized to record traffic accidents is the Polrestasbes-Semarang Traffic Unit. There are seven main information archived as digital documents as follows:

1. Case file number
2. Day, time and place of incident
3. Type, brand and license plate number of the vehicle involved in the accident (Antara)
4. Brief description of events
5. Condition of accident victims divided into 3 types, namely: death (MB), severe injury (LB), or minor injury (LR).
6. Material losses (meticulous)
7. Actions taken by the Police (Giat / CB)

From these data, the Polrestasbes Semarang every month recapitulates the number of accidents, the number of victims who died, serious injuries, minor injuries, and total material losses. The results of this recapitulation are one of the main information to identify areas prone to traffic accidents. According to Wedasana [4], the determination of accident-prone areas ideally considers historical data, so that the Polrestabes-Semarang usually also refer to the number of accidents in recent years.

Accident-prone areas are a location where the number of accidents is high with the incidence of recurrent accidents in a space and a relatively the same time span, caused by a particular cause. To identify accident-prone areas there are two stages that must be passed [4], namely:

a. Study the history of accidents (accident history) from all study areas and then choose locations that are considered to be accident-prone.
b. Study in detail the selected location to find the treatment that can be done.

The search for knowledge in data, also known as Knowledge discovery in Databases (KDD) [12], is defined as data extraction that has the potential for valuable information that is implicit and not previously recognized.

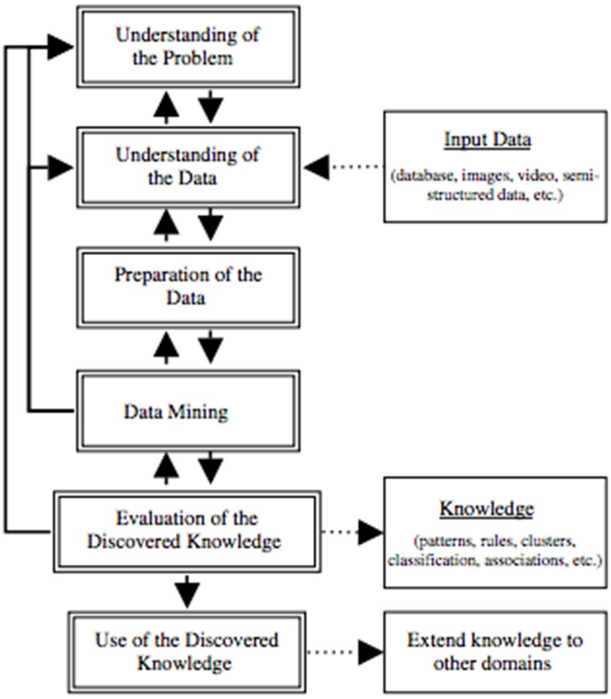
Understanding of the Problem
Understanding of the Data
Input Data
(database, images, video, semi-structured data, etc.)
Preparation of the Data
Data Mining
Evaluation of the Discovered Knowledge
Knowledge
(patterns, rules, clusters, classification, associations, etc.)
Use of the Discovered Knowledge
Extend knowledge to other domains

Fig.1. Steps in Knowledge Data Process (Pal and Jain, 2005)

There are a number of stages in the KDD process, however, there are basically three main stages [13] as shown in Figure 1, namely:

a. Pre-processing, related to data collection and retrieval (data collection), data cleaning (data cleaning), and data selection and transformation (data selection and transformation).

   From traffic accident data, there are a number of key information selected to be used as parameters or criteria for clustering, which are: time, location of the incident, type of vehicle involved, and condition of the victim due to an accident. This information is generally stored in the form of descriptions / descriptions.

   For computing needs, descriptive information (nominal) needs to be changed in the form of ordinal, interval, or ratio data types. In this case the simplification and generalization process of data applies to obtain information that can be further processed.

   - day
   - Location (km)
   - Directions (A or B)
   - Vehicle Type (Small, Medium, Large)
   - Weather (Bright, Cloudy, Drizzle, Rain, Heavy Rain)
   - Victims (Light Injuries, Weight, Death)

b. Data mining, related to the process of data exploration to find patterns or rules that have not been identified before, can be interpreted, and have the potential to be exploited [14]. There are a number of models or techniques that can be used to find these patterns, such as: anomaly detection, rule learning, clustering, classification, regression, and summarization.

In this study two algorithms (K-means and Hierarchical Clustering) are used. The K-means algorithm is used because there are several categorical data including the day of the event, direction, type of vehicle, and weather. Then the categorical data is converted to numeric.

Clustering technique with Hierarchical Clustering algorithm aims to classify data based on the similarity of characteristics possessed by the data. Hierarchical Clustering is a group analysis method that attempts to build a hierarchy of data groups. The stages in Hierarchical Clustering are:

1. Calculate Matrix Distance between data (Manhattan Distance or Euclidian Distance)
2. Combine the two closest groups based on the specified proximity parameters.
3. Update Matrix The distance between data to represent the closeness between the new group and the remaining group.
4. Repeat steps 2 and 3 until only one group remains.

c. Post-processing, related to the evaluation of results and visualizing them in a form that is easily understood by the user. After the clustering process is complete, then it is carried out then entering the third stage of post-processing data in the form of analysis and visualization of results.

Linoff and Berry [15] assert that the results of data mining cannot be submitted entirely to computer machines, human intervention is still needed to analyze the results. In other words, it takes an expert who knows well the mining datasets, and understands the statistics and structure of the mathematical model that underlies the work of the software [15].

In this study, it takes the role of an expert who understands well the characteristics of traffic accident data to compare the results of clustering and determine the status of road objects in the city of Semarang. Can be interpreted the results of this study are: Cluster 1 has characteristics: Occurs more likely on Monday, Thursday, Friday, and Sunday (Working Days), At an average Kilometer 19.75, Direction B, At Evening and Night, Small and Large Vehicle Types , Weather Cloudy, Drizzle, and Rain, with Severe Wounds. For Cluster 2 has Characteristics: Occurs more likely on Tuesday, Wednesday, and Saturday, at an average Kilometer of 34,0273, Direction A, Morning and Afternoon, Medium Vehicle, In Bright Weather and Heavy Rain, with Light Injuries and dies.

3 Conclusion

Based on the mapping of traffic accident data clustering results that have been carried out, there are several things that can be concluded, namely: The traffic accident data clustering system with the Hierarchical Clustering method can be used to classify road objects based on similar characteristics to the number of victims, the type of vehicle involved, and the number of accidents that occur within a certain time span. The traffic accident data clustering system requires the role of a traffic expert to analyze the results of clustering and determine the classification of the status of the level of highway vulnerability.

Moreover, by reducing the number of accidents, a further traffic jam caused by occurring accident can be greatly reduced. Reduced number of traffic accident will also prevent carbon emission from the engine combustion so that a low carbon society can be achieved.

References

1. BPS, “Indicator of Population Growth in Semarang Municipality, 2010 - 2016,” 2017.
2. O. Z. Tamin, *Transportation Planning and Modelling (in Bahasa)*. Bandung: Institut Teknologi Bandung, 1997.
3. S. Polrestabes-Semarang, “Traffic Accident Report (in Bahasa),” 2018.
4. A. S. Wedasana, “Analisis Daerah Rawan Kecelakaan dan Penyusunan Database Berbasis Sistem Informasi Geografis (Studi Kasus Kota Denpasar),” Udayana University, 2011.
5. Depkimpraswil, “Controling of Prone Traffic Accident Area (in Bahasa),” Jakarta, 2004.
6. M. E. Bolla, Y. Messah, and M. B. Koreh, “Analysis of Prone Traffic Accident Area (Case Study on Timor, Kupang)(in Bahasa),” *J. Civ. Eng.*, vol. 2, no. 2, pp. 147–156, 2013.
7. A. Z. Li and X. H. Song, “Traffic Accident Characteristics Analysis Based on Fuzzy Clustering,” in *IEEE Symposium on Electrical & Electronics Engineering*, 2012, pp. 468–470.
8. Y. Chen, L. Chun, H. Wu, and W. Sun, “Identification of Black Spot on Traffic Accidents and its Spatial Association Analysis Based on Geographic Information System,” in *Seventh International Conference on Natural Computation*, 2011, pp. 143–150.
9. C. S. Harahap, *Design of Accident Prediction Application using Apriori Algorithm (in Bahasa).* Medan: Pelita Informatika Budi Darma, 2013.
10. W. Yunanto, M. Hariadi, and M. H. Purnomo, “Geospatial Information Smart Visualisation and Timeline Hybrid,” Institut Sepuluh Nopember, 2012.
11. N. Ransi, “Application of Classification Algorithm Based on Predictive Association Rules (in Bahasa),”

Gadjah Mada University, 2014.

12. M. C. Thomas, W. Zhu, and J. A. Romagnoli, Data mining and clustering in chemical process databases for monitoring and knowledge discovery," *J. Process Control*, vol. 67, pp. 160–175, 2018.
13. U. Fayyad, *Advances in Knowledge Discovery and Data Mining*. MIT Press, 1996.
14. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2005.
15. G. S. Linoff and M. J. Berry, *Data Mining Technique: For Marketing, Sales, and Customer Relationship Management*, 3rd ed. Wiley and Sons, 2011.