Clustering Methods and Bound Value in Classify Density Traffic Accident Areas

Hsien-Tsung Chang, Hieu Nguyen.

Department of Computer Science and Information Engineering
Chang Gung University, Taoyuan County, 33302, Taiwan
E-mail: smallpig@widelab.org

Abstract

Nowadays, Traffic accidents (TAs) causes much damage regarding human and asset. TAs records have been stored and published in many Open Data sources. In this article, we propose a method of using clustering algorithm Density-based Spatial Clustering of Applications with Noise (DBSCAN) to classify the TAs' records to find the Density Traffic Accident Areas (DTAA). We also discuss the optimal variables in DBSCAN that need to consider when applied in real urban areas. We emphasize the characteristics of DTAA by the Bound Value (BV) and modeling some important traffic characteristics. We then evaluate the performance and efficiency between DBSCAN and K-mean clustering methods. The result clusters and characteristics can be easily adapted to real traffic applications for increase the travel safety.

Keywords: DBSCAN, Traffic Accident, Bound Value, Clustering, Density Traffic Accident Areas.

1. Introduction

In recent years, Traffic Accident is one of the most brutal "disease" happening in the world. The World Health Organization (WHO) reported in their recent reports[*] that every year, there are about 1.2 million deaths because of Traffic Accident; around 50 million others suffer injuries and disability.

Another report, named Safe and Sustainable Roads[**] said that Traffic Accident is now the deadliest killer for kids and young people of age 10 to 24. If there are no urgent actions, the death numbers will raise from 1.3 million to 2 million deaths every year. Every day there are about 3500 deaths caused by Traffic Accident with 70% of them happen in developing countries. Most of the people died while driving bicycle and motorbike. The report also showed that Traffic Accident cause more deaths for kids and young children than malaria, diarrhea, and HIV.

Traffic Accident is not only caused deaths but also cause a heavy burden of economic losses for the victims, their families and the nation. The victims losses money for health treatment, some of them losses their outcome due to their disability. Those who caused Traffic Accident also losses because they need to pay the treatment, also their time involve with laws.

In this new era of Big Data, more and more Open Data sources have been published. The government wants to keep this information on track, for a safer and simpler way of store data. The Open Data sources are the great help for research on analyze and classify historical data. They contain a lot of different type of data; include the records of Traffic Accident. Some of these Open Data sources are New York City Open Data and Great Britain Open Data[***].

In the past decades, there has been an interest in using historical data to analyze or classify them to reduce Traffic Accidents. Based on the results, they developed different methods to archive this. While some of them used tracking vehicle method to detect Traffic Accident (1), others used the analysis results to predict the places where Traffic Accidents likely to happen (2, 3, 4, 5, 6), some studies used clustering methods to classify the Traffic Accident into clusters based on certain parameters (7, 8, 9, 10). In order to understand more about the cause of Traffic Accidents, some authors studied about the Traffic Accidents' characteristics (11, 12, 13, 14) which can help them find the similarity of these accidents.

In this study, we use the common Density-Based Spatial Clustering of Application with Noises (DBSCAN) algorithm (15, 16) to form the Traffic Accidents records into clusters based-on the Traffic Accident's locations and density. After that, we use the Bound Value to identify the similarity or characteristics of each cluster. The result gave us the clusters and their abnormal characteristics that we can later use in suggestion methods to avoid these dangerous areas or suggest the one who has the responsibility to check for the cause why that location have so much Traffic Accident happen based on the abnormal characteristics.

The contribution of our work can be summarized as:

- We used DBSCAN algorithm to form the Traffic Accident Open Data source into clusters while considered the input parameters so we can have most

[*] http://www.fiafoundation.org/our-work/make-roads-safe

[**] http://www.who.int/healthinfo/global_burden_disease

[***] https://nycopendata.socrata.com/
https://data.gov.uk/

DOI: 10.12792/iciae2017.023

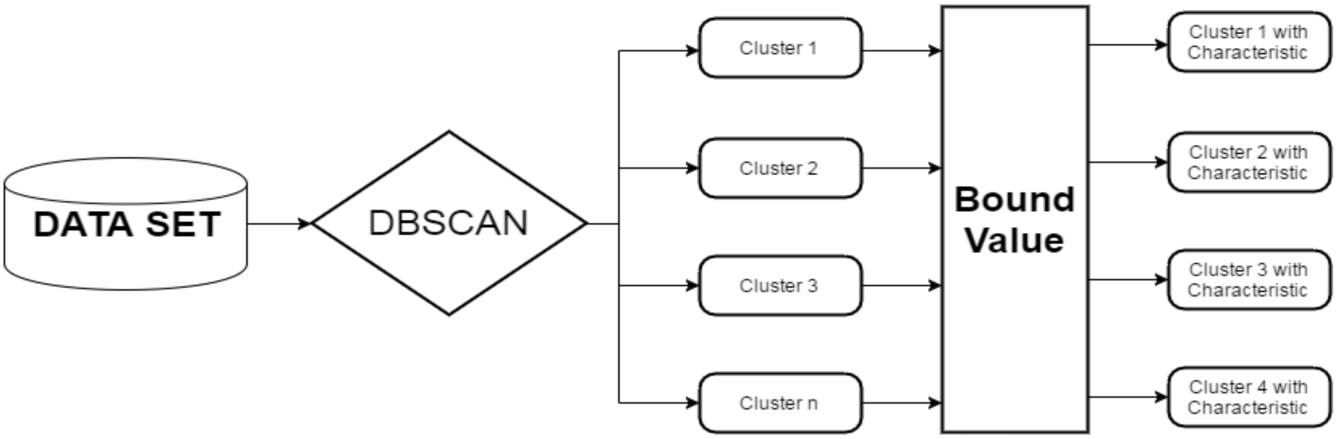
DATA SET
DBSCAN
Cluster 1
Cluster 2
Cluster 3
Cluster n
Bound Value
Cluster 1 with Characteristic
Cluster 2 with Characteristic
Cluster 3 with Characteristic
Cluster 4 with Characteristic

Fig. 1. Overview.

suitable parameters while applying this method to real life problems.

- We listed out some environment characteristics of Traffic Accidents and combined them into cases that can happen.
- We then used Bound Value to classify them to find the similarity of these accidents.
- We evaluated our approach with the Open Data set from NYC and Great Britain. We also compared it with the K-mean method in term of performance and efficiency.

The organized sections of this paper are as follow: In Section 2, we continue to discuss about the related works. In Section 3, we described our proposed method. Section 4 is our experimental results. Finally, we conclude our study and briefly discuss future work in Section 5.

2. Related Works

There are some studies proposed the approaches to classifying TAs using different clustering algorithms. Kumar and Toshniwal et. Al (7) using k-modes, an enhanced version of the K-mean clustering algorithm to form the TAs clusters. Tessa et. al (9) used Geographical Information Systems (GIS) and Kernel Density Estimation to add attribute data to existing TAs areas, then used the K-mean clustering algorithm to form the areas which have similar attribute data into clusters. Sandor and Peter et. al (8) used Density-Based Spatial Clustering of Applications with Noises (DBSCAN) on GPS coordinates of TAs in ordered to form the Black spots where the number of accidents is higher than other areas.

There are many comparisons between K-mean clustering and DBSCAN through studies, not only in TAs but also in many different objects. Jeffrey et. al (17) used K-mean and DBSCAN algorithms to classify the network traffic, the result shows that K-mean has faster performance while DBSCAN has better clusters. Nejdet et. al (10) compared K-mean, DBSCAN and Expectation Maximization (18) while presented a real-time traffic accident detection method; the result showed that DBSCAN gave the best promising result.

In this paper, we use DBSCAN as our main algorithm to cluster the TAs based on their coordinates. Simply because DBSCAN is the Density algorithm, which suite best for finding areas with higher density.

Ai and Xiang et. al (14) used fuzzy clustering to form the TAs then analyzed their characteristic in term of vehicle types to find the most dangerous driving behavior. Kumar and Toshniwal et. Al (7) listed the characteristics of TAs in both subjectivity and environment aspect. They then used Association Rule Mining Algorithm to add the rules to each cluster, the clusters with strong rules will be taken for analysis the cause of TAs. We, however, want to consider more about how to apply the clustering method and TAs' characteristics to specific real-life location, so we came up with the Bound Value.

3. Proposed System Design

In this section, we present how we used DBSCAN algorithm with customize parameters to form the clusters. After that, we used Bound Value to classify the result clusters into clusters with characteristics to understand the cause of these dangerous areas. An overview of our works is described in Figure 1.

3.1 Density-Based Spatial Clustering of Applications with Noises (DBSCAN)

DBSCAN is Density-Based Spatial Clustering of Application with Noise; this method focuses on identifying more density areas over the fewer density areas. This clustering algorithm requires two parameters that are distance e and a minimum number of points minPts. In this algorithm, the points separated into core points, boundary points and noises. A point p is a core point if there are at least minPts points (include p) that are within the distance e of it. All the points within distance e of a core point become part of the cluster and called boundary points, from these boundary points, continue to check if any other points are within the distance e of them and make them part of the cluster as well. The points that are not reachable by neither core points nor boundary points called noises.

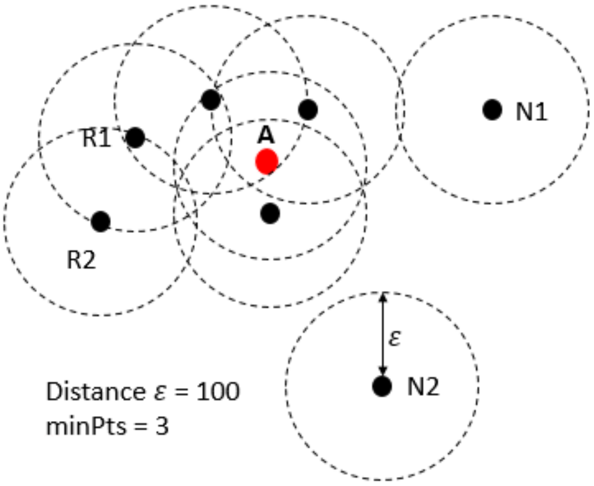

Fig. 2. DBSCAN example.

In Figure 2, A is considered as a core point because it has 4 points (include it also) within its distance e. R1 and R2 are not within A's distance but R1 link with A's boundary point, and R2 link with R1 so they are considered as boundary points and a part of A's cluster as well. N1 and N2 are not reachable by neither core points nor boundary points, so we consider them as noises.

3.2 The Data Set

For the data sources, we used two set of data from New York City open data and Great Britain open data. The NYC open data is more detailed than GB open data regarding traffic vehicles and traffic throughput, but less regarding traffic weather condition and lighting condition. Because of that, for different required fields we used different data set to suite the best. For using the DBSCAN algorithm to cluster the data sets to identify the DTAA, we used the NYC open data with four attributes: The date that TA happen (TDate), the time that TA happen (TTime), the latitude (LAT) and the longitude (LONG).

Our target was to cluster this data set into some specific DTAA, which can help us identify the areas that the TA are most likely happen. Working on the position of the TA, LAT, and LONG, we first considered about the distance e since it is a variable parameter, which need to be careful selected. Because even the TA can occur anytime, anywhere, but they still have some constraints that we can consider. Consider an urban area and a countryside; we can say that the chance for TA happen is much higher in the urban area than countryside since the traffic volume and frequency in urban areas are much higher. Also in urban areas, an area which has a lot of intersections or crowded places usually have the higher chance that TA will happen than an area which doesn't have these.

3.3 DBSCAN Parameters

We first tried to select e = 500 meters and minPts = 5, the data set was 1000 records from NYC open data. After clustering used DBSCAN, we found 2 clusters which have 996 records and only 4 noises left. That mean the distance we chose was too long so most of the records can link to each other within 500 meters. We then looked at the map around the areas and realized that we can divided most of the living areas into blocks with so many small intersections about 100 to 280 meters long. Most of the accidents happen in the middle of the intersection, and they just linked with other intersection due to our large distance selected.

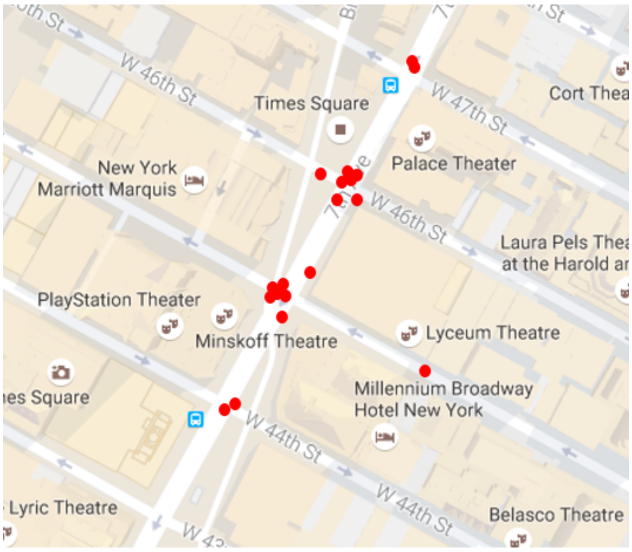

Fig. 3. Different clusters along the street.

We then selected e = 100 meters and minPts = 5 with the same data set. After used DBSCAN we got 25 clusters which have 166 records, the others 834 were noises. We realized that most of these clusters happen in big intersections, some of the TAs are happen in exact same position. Of course, these clusters are good, but then we found out some clusters were close to each other, more than 100 meters of course but they were along on the same street, just different side of the intersection. One example showed in Figure 3.

After that, we chose e = 180 meters and minPts = 7, that gave us the result of 33 clusters which have 326 records, others 674 were noises. The shape of the clusters are intersections and expand a little bit to the connected streets. However, increased the distance led us to an interesting problem, because there are so many intersections in a small area and they are not far to each other, so increase the distance increase a chance that a cluster will expand to others intersections, others streets. That may lead to inaccuracy clusters we want to see since the covered areas are too large, it is meaningless when it comes to DTAA term.

After many tried, we think that the parameter distance e should be between 140 to 220 meters when applying for urban areas, the minimum numbers of points should be between 4 to 7. These parameters may variable based on the area's attributes, the further between the intersections, the greater the distance e. For an area that has the lower frequency that TA may happen, the lower minimum numbers of points should be.

3.4 Bound Value and Traffic Accidents' Characteristics

We considered only the environment characteristics that affect the chance that TA will happen. We showed these attributes in Table 1 bellow.

Table 1. Traffic Accidents Attributes.

TA time	T1 (12:00PM – 6:00AM)
	T2 (6:00AM – 12:00AM)
	T3 (12:00AM – 6:00PM)
	T4 (6:00PM – 12:00PM)
Light condition	L1 (Day time)
	L2 (Dust with street light)
	L3 (Dust without street light)
	L4 (Dark with street light)
	L5 (Dark without street light)
Weather condition	W1 (Normal)
	W2 (Strong wind)
	W3 (Rain)
	W4 (Fog)
	W5 (Others)

From the Table 1, we have in total 100 combinations that shown the characteristics of the TA. Although this may not be enough and need to be vary based on different places, locations we apply it. Nevertheless, take it for example, if an area we considering suite for this table then almost all the TA happen will fall into these 100 combinations. Of course, we cannot expect that these combinations are equal to probability with each other. Since there are very low chance that, the normal days in a year will be equal to the strong wind days or rain days in a year, etc. Therefore, if we consider these combinations as equal, that will lead to inaccuracy analysis. Therefore, we came up with the Bound Value (BV).

This BV is a parameter that variable based on locations. Because each location has different traffic volume in daytime, also different in light condition and weather condition. Therefore, to identify this BV we need to consider manually for each case.

Take the TA time and NYC traffic volume 2012-2013 of Manhattan district for example. We calculated and got the result like shown in Table 2. The detailed shape is in Figure 4.

Table 2. Traffic Volume of Manhattan.

TA time	Number of vehicles
T1 (12:00PM – 6:00AM)	32440
T2 (6:00AM – 12:00AM)	81706
T3 (12:00AM – 6:00PM)	101599
T4 (6:00PM – 12:00PM)	78595

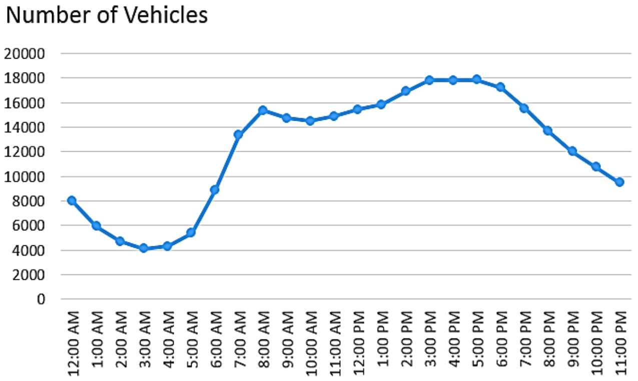

Fig. 4. Manhattan Traffic Volume Pattern.

After checked the records in NYC open data and GB open data, we realized that the pattern of traffic volume is nearly the same. The volume in T1 about 10% total traffic volume of the day, volume in T2 is about 30%, volume in T3 is about 35%, and volume in T4 is about 25%. Theoretically, we can say that the greater the traffic volume is, the greater the chance that TA can happen. So if in one area, the TA happen mostly in time T1 – which has the lowest traffic volume of the day – then that area sure have some problem we need to check carefully.

For each TA attribute in Table 1, we have N cases. BV is the boundary value, BV(x) is the percentage of case x take part in N cases, or different way to say, BV(x) is the probability that x can happen between N cases. If BV(W1) = 0.4 that mean the probability that a TA happen on a normal day is 40%, the rest 60% of the TA will fall into W2, W3, W4. We have:

$$\sum_{x=1}^{N} BV(x) = 1 \quad (1)$$

For each combination of a total of 100 combinations, for example, the combination T1-L1-W1, we used C to denoted the probability that a TA will fall into this combination is BV(T1) x BV(L1) x BV(W1). We will have in total 100 C that is the probability for 100 combinations.

$$\sum_{C=1}^{100} C = 1 \quad (2)$$

Once we had C in theory, we can easily find C, in reality, using the TA open data records, just find there are how many TA happen which have the characteristic as combination x among the total number of TA. After we have C in theory and C in reality, we can compare them together and check if these both values are normal or abnormal. If:

$$\begin{aligned}
&C_{real} < C_{theory} &&: Ignore\\
&C_{real} = C_{theory} &&: Normal - Ignore\\
&C_{real} > C_{theory}\ under\ 15\% &&: Normal - Ignore\\
&C_{real} > C_{theory}\ greater\ than\ 15\% &&: Abnormal
\end{aligned}$$

For example, if we only consider the TA attributes like shown in Table 3 bellow:

Table 3. TA Attributes example.

Attribute	Cases	BV
TA time	T1(0:00AM – 12:00AM)	0.45
	T2(12:00AM – 24:00PM)	0.55
Light condition	L1 (Day time)	0.62
	L2 (Dark)	0.38

So we can see that we have total 4 combinations, and C_{theory} of each combination can be calculated by multiple the BV.

$$C_{theory} of\ T1 - L1 = 0.45\ X\ 0.62 = 0.279$$

Then we will check in the open data source for $C_{real} of\ T1 - L1$. For example, if there are 47 TAs that fall into T1-L1 combination in the total of 123 TAs we are considering, then:

$$C_{real} of\ T1 - L1 = \frac{47}{123} = 0.285$$

We then compare $C_{theory} of\ T1 - L1$ and $C_{real} of\ T1 - L1$, the result is the 3rd case, $C_{real} of\ T1 - L1 > C_{theory} of\ T1 - L1$ under 15%, so this combination is normal. Therefore, if there are any combination that falls into 4th case $C_{real} > C_{theory}\ greater\ than\ 15\%$, then we may consider this combination is the characteristic of the cluster we are considering.

4. Experimental and Evaluation

In this section, we experiment and evaluate the efficiency of the DBSCAN algorithm; we also evaluate the performance of the common K-mean clustering algorithm and compare them together since these both are frequently used algorithms in Traffic problems. We then discuss the strengths and weaknesses of these algorithms when address to TA clustering.

We used the latest NYC open data set that collected from Manhattan district in the first half year 2016. The data set contains around 20500 records of TA all over Manhattan district. This data set is very detailed about TA coordinates, date time and type of vehicle.

We separated this data source into several smaller source to test the performance and compare the efficiency of each algorithm when running on different data set size. We separated into 1000 records, 2000 records, 5000 records, 10000 records, and the original source.

Many studies used K-mean clustering algorithm to classify the events (7, 17, 10). We at first tried to use it to form clusters as well, but since it need the manual number of cluster and only form the clusters based on average distance, it is not really suite our purpose of identifying DTAA, hence we end up used DBSCAN because it perfectly match when it comes to density problems. Bellow parts are the results we compared the performance and efficiency of K-mean and DBSCAN.

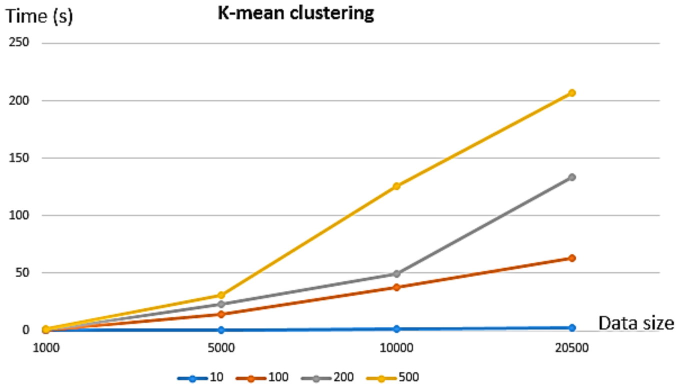

Fig. 5. K-mean clustering performance.

As we can see in Figure 5, the performance of K-mean when handle big records data set is good. However, if the records of data set or the cluster we input increased, the performance of K-mean slow down dramatically. It is because the application need to calculate all of the distances from each point to the centroids (cluster's center) we defined. So the more records and centroids, the more calculating needed.

Compare to K-mean algorithm while handle the same records of data set, the DBSCAN algorithm require a lot more time to execute, this is because every record needs to calculate the distance between it and all the records left. Not like K-mean, only need to calculate the distance between points and centroids we defined.

Figure 6 show the performance of DBSCAN while running 5000 records data set with different value of parameters. We ran 8 different pairs [e=140, MinPts = 4], [e=140, MinPts = 7], [e=180, MinPts = 4], [e=180, MinPts = 7], [e=200, MinPts = 4], [e=200, MinPts = 7], [e=220, MinPts = 4], [e=220, MinPts = 7] as shown in the Figure 5. For the same minimum number of points, when we increased

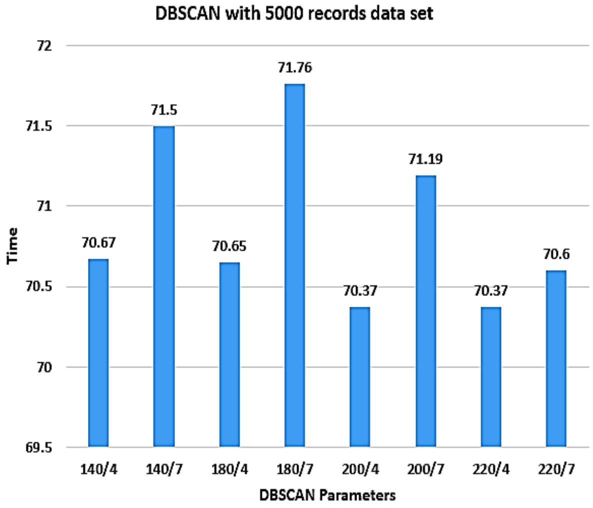

Fig. 6. DBSCAN performance.

the distance, the time required for executed slightly reduced. For the same distance, when we increased the minimum number of points, the time required increased.

Table 4. DBSCAN formed clusters.

Parameters (e/MinPts)	Clusters	Noises
140/4	184	379
140/7	152	1121
180/4	37	74
180/7	31	264
200/4	30	52
200/7	24	165
220/4	18	34
220/7	16	101

Table 4 show the result of each executed. With the same distance e, when we increased the MinPts, we got less number of clusters and more noises; this is because more MinPts required mean fewer core points will be detected. In the other hand, when we increased the distance e, the clusters formed decreased; that is because the coverage of the clusters will be larger.

We also found out that not all the clusters formed are useful, since there will be some clusters with very large coverage, due to the expandability of DBSCAN and the fact that the data set is too large, the probability that some huge clusters formed will be there.

This led us into further considering about the scale of the areas and data sets we should use. If the area is not so big and we input a large data set at once then most likely, the result will be only one or two big clusters that are useless. However, if we use a small data set into a big area, the chance for clusters to form is quite small. So in our case, when we applied the algorithm for Manhattan district, because the TA happen there was so much, so we separated the data set into monthly records data set, which was around 2000 to 5000 records and we think that reasonable enough.

5. Conclusion and Future Work

In this article, we present an approach to classifying historical data set with DBSCAN algorithm to find the Density Traffic Accident Areas. We can use these formed clusters in suggestion or prediction applications to avoid these dangerous areas or give feedback to the one have the responsibility about the existing dangerous.

We also discuss the Bound Value, one of the parameters that we need to consider when working with the clusters' characteristics. If we can identify the characteristics of these dangerous clusters more accuracy, we will be able to identify the cause why traffic accidents happen in these areas so frequently.

Lastly, we developed a clustering application with C++ to cluster the open data set using DBSCAN and K-mean algorithm. We compared their performance and efficiency. We compared the results of DBSCAN while using different parameters.

As future works, we will try to learn more about the traffic accidents' characteristics to find the way to identify the cause more accuracy. We also want to try different methods of clustering to improve our current method. We are also building an application that can suggest the drivers take the safer travel route using these clusters we found above; this application also can request the government to check some dangerous areas that have suspicious accidents based on the clusters' characteristics.

References

(1) Shunsuke Kamijo, Yasuyuki Matsushita, Katsushi Ikeuchi, Fellow, IEEE, and Masao Sakauchi; "Traffic Monitoring and Accident Detection at Intersections." IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 1, NO. 2, JUNE 2000.

(2) Rob Eenink, Martine Reurings (SWOV), Rune Elvik (TOI), João Cardoso, Sofia Wichert (LNEC), Christian Stefan (KfV). "Accident Prediction Models and Road Safety Impact Assessment: recommendations for using these tools."

(3) Poul Greibe, Danish Transport Research Institute, Knuth Winterfeldts Allé, DK-2800 Kgs. Lyngby, Denmark; "Accident prediction models for urban roads." Accident Analysis and Prevention 35, 273–285, 2003.

(4) Yisheng Lv and Shuming Tang; "Real-time Highway Traffic Accident Prediction Based on the k-Nearest Neighbor Method." International Conference on Measuring Technology and Mechatronics Automation, 2009.

(5) Tao Lu and Yan Lixin; "The traffic accident hotspot prediction: Based on the logistic regression method". The 3rd International Conference on Transportation Information and Safety, Wuhan, P. R. China, June 25 – June 28, 2015.

(6) Weiming Hu, Xuejuan Xiao, Dan Xie, Tieniu Tan, Fellow, IEEE, and Steve Maybank, Member, IEEE; "Traffic Accident Prediction Using 3-D Model-Based Vehicle Tracking". IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, VOL. 53, NO. 3, MAY 2004.

(7) Sachin Kumar and Durga Toshniwal; "A data mining framework to analyze road accident data" . Journal of Big Data (2015) 2:26 DOI 10.1186/s40537-015-0035-y, 2015.

(8) S. Szenasi and P. Csibad; "Clustering algorithm in order to find accident black spots identified by GPS coordinates". 14th International Multidisciplinary Scientific GeoConference SGEM 2014, www.sgem.org, SGEM2014 Conference Proceedings, ISBN 978-619-

7105-10-0 / ISSN 1314-2704, Book 2, Vol. 1, 497-504 pp, June 19-25, 2014.

(9) Tessa K. Anderson; "Kernel density estimation and K-means clustering to profile road accident hotspots". Accident Analysis and Prevention 41 (2009) 359–364, 2009.

(10) Nejdet DOGRU and Abdulhamit SUBASI; "Comparison of clustering techniques for traffic accident detection". Turkish Journal of Electrical Engineering & Computer Sciences. Turk J Elec Eng & Comp Sci (2015) 23: 2124 – 2137, 2015.

(11) Tibebe Beshah1, Shawndra Hill; "Mining Road Traffic Accident Data to Improve Safety: Role of Road- elated Factors on Accident Severity in Ethiopia". Artificial Intelligence for Development, Papers from the 2010 AAAI Spring Symposium, Technical Report SS-10-01, Stanford, California, USA, March 22-24, 2010.

(12) Karen Zimmerman, Ali A. Mzige, Pascience L. Kibatala, Lawrence M. Museru and Alejandro Guerrero; "Road traffic injury incidence and crash characteristics in Dar es Salaam: A population based study". Accident Analysis and Prevention (2011).

(13) Phillipo L Chalya1, Joseph B Mabula, Ramesh M Dass, Nkinda Mbelenge, Isdori H Ngayomela, Alphonce B Chandika and Japhet M Gilyoma; "Injury characteristics and outcome of road traffic crash victims at Bugando Medical Centre in Northwestern Tanzania". Journal of Trauma Management & Outcomes, 6:1, 2012.

(14) Ai-Zeng Li and Xiang-Hong Song; "Traffic Accident Characteristics Analysis Based on Fuzzy Clustering". IEEE Symposium on Electrical & Electronics Engineering (EEESYM), 2012.

(15) Martin Ester , Hans-peter Kriegel , Jörg Sander and Xiaowei Xu; "A density-based algorithm for discovering clusters in large spatial databases with noise". KDD-96 Proceedings, AAAI (www.aaai.org), 1996.

(16) Dr. Mohammed Ali Hussain, Dr. R. Satya Rajesh and Md. Abdul Ahad; "A Study of DBSCAN Algorithms for Spatial Data Clustering Techniques". IJCST Vol. 3, ISSue 4, oCT - DeC 2012.

(17) Jeffrey Erman, Martin Arlitt and Anirban Mahanti; "Traffic Classification Using Clustering Algorithms". Proceedings of the 2006 SIGCOMM workshop on Mining network data, Pages 281-286, New York, NY, USA ©2006.

(18) Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B Met; 39: 1- 38, 1977.