International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 20 (2017) pp. 9815-9819

The Clustering of Black Spot using Province Public Data

Ki-Young Lee[1], Myung-Jae Lim[1,*], Jeong-Jin Kang[2],Sung-Jai Choi[3], Eun-Young Kang[4], Sung-Ho Hwang[5]

[1]Department of Medical IT, Eulji University, Seongnam 13135, Korea.
[2]Department of Information and Communication, Dong Seoul University, Seongnam 13117, Korea.
[3]Departmentof Electronic Engineering, Gachon University, Seongnam 13120, Korea.
[4]Department of Computer Software Engineering, DongyangMirae University, Seoul 08221, Korea.
[5]Division of Electronics, Communication & Information, Engineering, Kangwon National University, Samcheok 25913, Korea.

[1,]Corresponding Author: Myung-Jae Lim*

Abstract

Software is an essential element in ensuring competitiveness throughout the country. Advanced foreign and domestic software companies are leading innovation not only in the traditional IT field but also in various fields based on sophisticated software. The development of artificial intelligence machine learning opens up possibilities for the automation of intellectual activities, and its impact is expected to be very large and broad. The high accuracy of artificial intelligence and machine learning requires a large amount of data. In Korea as well as overseas, public data can be accessed and utilized through public data portals. This study clustered the public data of Gyeonggi provincial accidents in order to identify the area of accidents. In this paper, we propose a solution to the problem by specifying the range (Epsilon) and the minimum population (MinPts) in order to characterize the multi-accident area by using DBSCAN (Density Based Spatial Clustering of Applications with Noise).

Keywords: Machine Learning, Clustering, DBSCAN, Public Data, Black Spot

INTRODUCTION

Software is now an integral part of ensuring the competitiveness of the nation as a whole. Overseas and domestic software companies are leading innovation in virtually every field, including automotive, medical, economics, education and culture, as well as traditional IT based on sophisticated software. The development of artificial intelligence and machine learning opens up possibilities for the automation of intellectual activities, and its impact is expected to be very large and broad [1][2][3]. Artificial intelligence machine learning is a key software technology that is essential for the big data and object internet age in which a lot of data is generated [4].

In addition, Big Data is called '21st Century Crude Oil' and is emerging as a new paradigm and new growth engine for various industries such as IT, finance, and distribution. The Big Data industry is expected to continue its high growth. Major governments such as the US, UK and EU are carrying out various projects such as public information disclosure for big data market activation and big data based public service provision. In addition, big data use cases are gradually increasing not only in the governments of major countries but also in domestic and overseas markets [5].

In Korea, public data portals are used to provide data in open data formats such as open API, CSV, HWP, etc., and more and more new data are being created by utilizing domestic public data [6]. In this paper, we investigate the concept of clustering, K-Means, and DBSCAN after investigating the previous studies on clustering and clustering in order to determine the area of accidents. DBSCAN, which is a clustering algorithm, was applied to the data of the current situation of the accident area in Gyeonggi-Do province in 2016 provided by the public data portal.

RELATED RESEARCH

Black Spot

In the past, The "Evaluation of the High-Accident Location Spot-Improvement Program" proposed by Agent, K. R. (1973), to determine the area of accidents involving the Kentucky highway and lowered the probability of accidents at the Kentucky accident site through the site improvement program [7]. The "Identification of Hazardous Rural Highway Locations," proposed by Deacon (1974), requires the involvement of personal judgment in the scope of the accident site proposed by the agent described above, and the accidental nature of the accident. About 35% of the locations surveyed in the field are not appropriate, so management costs are high and there is room for improvement. In order to solve this problem, the highway was divided into Short and Large, and the section length was selected to distinguish the

characteristics of the road. However, since the accident analysis becomes complicated, a method of evenly dividing the accident analysis section by a certain distance has been proposed [8]. The "The Relationship between Truck Accidents and Geometric Design Road Sections: Poisson Versus Negative Binomial Regressions." proposed by Miaou (1994), proposed a slider length method considering the adjacency of the accidental bundle area , because it is difficult to obtain proper results if the interval is too short or the interval is not constant [9].

Clustering of Black Spot

The "Selecting Technique of Accident Sections using K-Mean Method" proposed by Ki-Young Lee, K-Means cluster analysis method was used to classify clusters over a certain length. However, the K-Means cluster analysis method has the limitation that the number of clusters, K, must be determined by the user. In addition, not only the individuals with no meaning in the cluster are included in the cluster [10]. The "A Study on the Identification of Accident Hot Spots Using DBSCAN : Focused on Gyeong-Bu Expressway" proposed by Tae-Kyoung Lee, selected the time - spatial range based on the 2005 Gyeongbu Expressway and the downward line. The causes of highway accidents were analyzed only for drowsiness, over speed, attention neglect, steering wheel operation, which account for more than 85% of all driver accidents. As a result of comparing DBSCAN experiment with K-Means experiment, it is suggested that DBSCAN results in better experimental results [11].

The "A Study on Near-miss Incidents from Maritime Traffic Flow by Clustering Vessel Positions" proposed by Kwang-Il Kim, defined Ship bumper area model for close proximity accident calculation among the eastern ships. Using the ship bumper area model, we propose a proximity accident calculation module for the marine area through proximity accident discrimination equation and vessel location clustering. The proposed proximity accident calculation module was applied to the navigation data of wandering vessel navigation vessel to evaluate ship navigation risk factors such as vessel type, navigation speed and encounter direction. K-Means, a hierarchical clustering algorithm, has been applied to "Study on the calculation of marine traffic proximity accident using vessel location clustering" for clustering [12].

A STUDY ON THE CLUSTERING OF BLACK SPOT

Clustering

Clustering is a kind of classification, and clustering analysis is a process of judging and classifying as meaningful groups in order to solve a specific problem. Partitioning Methods, Hierarchical Methods, Density-based Methods, and Grid-based Methods are also available. Representative clustering algorithms include the K-Means algorithm for partitioning and the DBSCAN algorithm for density-based techniques [13].

K-Means Clustering

The K-Means algorithm quickly and efficiently clusters the model based on the parameter K, which is the number of clusters determined by the user in advance. Each cluster center is selected by the initial data, and the cluster order is determined according to the distance of the data. The K-Means algorithm is often used because it is easy to implement and time complexity is $O(n^2)$ and its execution speed is relatively high. However, this method is a sensitive clustering method in the initial position of the cluster center. In addition, if the initial location is not appropriate, the center of the cluster may remain at the local minimum [14]. The process in which the k-means algorithm is performed is as follows.

1. Determine the number K of clusters.
2. Select the center of the initial K clusters.
3. Assign each object to a nearby cluster based on a given center point. At this time, the distance between the center point and the object is calculated as the Euclidian distance.
4. Calculate the new center point of each cluster around the newly assigned object.
5. If there is no difference between the existing center point and the new center point, stop. Otherwise, go back to 2 and perform again.

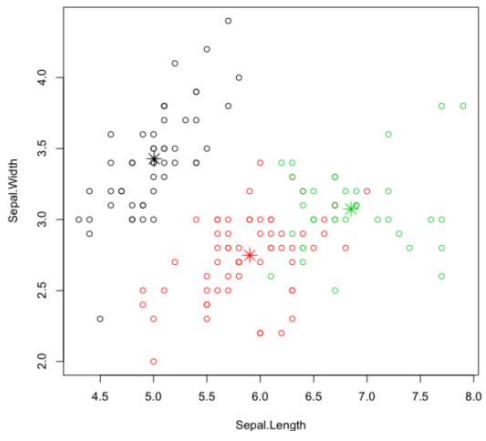

Figure 1. K-Means Clustering

When the above process is performed, it is confirmed that the cluster is generated as shown in the Figure 1. The parameter K, which is the number of clusters, was set to 3, and the clusters were 3 clusters.

DBSCAN Clustering

DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that is suitable for handling spatial data including noise and can distinguish clusters of various shapes and sizes. At this time, the cluster and noise are objectively classified based on the density of the points. DBSCAN's time complexity is $O(n * \log n)$. Compared with the K-Means algorithm, it takes a long time, but it can be practically used because there is no cluster, which is a precedent constraint. The process in which the k-means algorithm is performed is as follows [15].

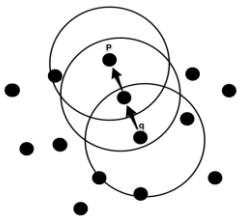

Figure 2. Density-Reachable

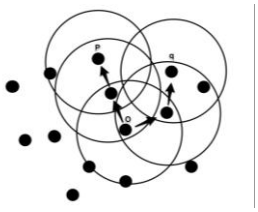

Figure 3. Density-Connected

The Eps-neighborhood of a point p is a set of neighbors within a range Eps from p. In other words, it can be defined as $N_{eps}(p) = \{q \in D | dist(p,q) \leq Eps\}$. The fact that one point p is directly density-reachable from point q means that p must be at the neighbor of q and q is the core point. That is, it has enough neighbor. As a formula, it is $p \in N_{eps}(p)$: , $|N_{eps}(q) \geq MinPts|$ (core point condition). The fact that one point p is densitiy-reachable from point q means that there is a chain that is directly density-reachable from p to q.

As shown in Figure 2, p is density-reachable from q / q is not density-reachable from p As shown in Figure 3, p and q density connected to each other by O. The fact that a point p is density-connected from point q means that there is a density-reachable point O from p and q. Let D be the data set of the points, let the cluster be C, and if Eps and MinPts are sub-sets of non-empty D, then For all points p and q $p \in C$, If p is densitiy-reachable from q $q \in C$, All p, q $p \in C$ P is density-connected to q. When C_i is a cluster in database D, the noise is that it does not belong to any cluster in D. That is, noise = $\{p \in D | \forall i : p \notin C_i\}$. The process in which the DBSCAN algorithm is performed is as follows.

1. Select an arbitrary point that satisfies the core point condition from the database and seed it.
2. Retrieve all density-reachable points from the seed and include them as clusters.

EXPERIMENT AND RESULT

Experiment Environment

Experimental data is position data caused by accidents in Gyeonggi Province, which was updated on September 7, 2016 [16]. For this experiment, 4966 pieces of positional data, which is the product of 1130 latitude and longitude heat of each position data multiplied by the number of accidents, were used as experimental data. This experiment was performed using the statistical program R and the experiment using the R language [17].

Experiment Result

Figure 4 is the result of DBSCAN analysis using the fps library of R 3.1.0 statistical program [18]. The Epsilon value, which is a parameter value for applying the DBSCAN, is set to 0.02, which is a distance of about 2.2 km, and the MinPts value, which is a population number within a range, to 200. Therefore, it is necessary to have 200 objects within a radius of 0.02 to define the same cluster.

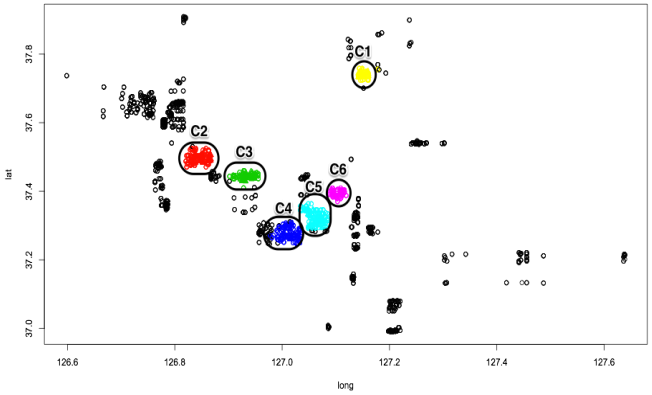

Figure 4. Black Spot Clustering Experiment Result

Table 1. Information of Cluster

Cluster Name	Within the area Number of Jaywalking	Within the area Number of Accidents	Accident Area
C1	122 (46.7%)	261	Uijeongbu
C2	138 (25%)	552	Bucheon
C3	185 (37.4%)	484	Seongnam
C4	309 (47.8%)	646	Suwon
C5	160 (34.9%)	458	Ansan
C6	82 (33.5%)	245	Anyang

Table 1 Shown in, Among the incidents classified as jaywalking, bicycle accidents, elderly walking accidents, and children 's walking accidents, among them, the rate of jaywalking to bicycle accidents with a large number of accidents, The death toll from jaywalking deaths is high, with 17 deaths occurring among a total of 1081 accidents. Therefore, Table 1 was prepared based on the jaywalking. The area of cluster C1 is Uijeongbu city. The total number of occurrences is 261 times, the number of endless jaywalking is 122, and the rate is 46.7%. The area of cluster C2 is Bucheon city. The total number of occurrences is 552 times, the number of jaywalking is 138, and the rate is 25%. Cluster C3 is located in Seongnam city. The total number of occurrences is 484 times, and the number of jaywalking is 185. The rate is 37.4%. Cluster C4 is located in Suwon city. The total number of occurrences is 646 times, the number of jaywalking is 309, and the rate is 47.8%. Cluster C5 is located in Ansan city. The total number of occurrences is 458 times, and the number of jaywalking is 160. The rate is 34.9%. The area of Cluster C6 is Anyang city. The total number of occurrences is 245 times, the number of jaywalking is 82, and the rate is 33.5%.

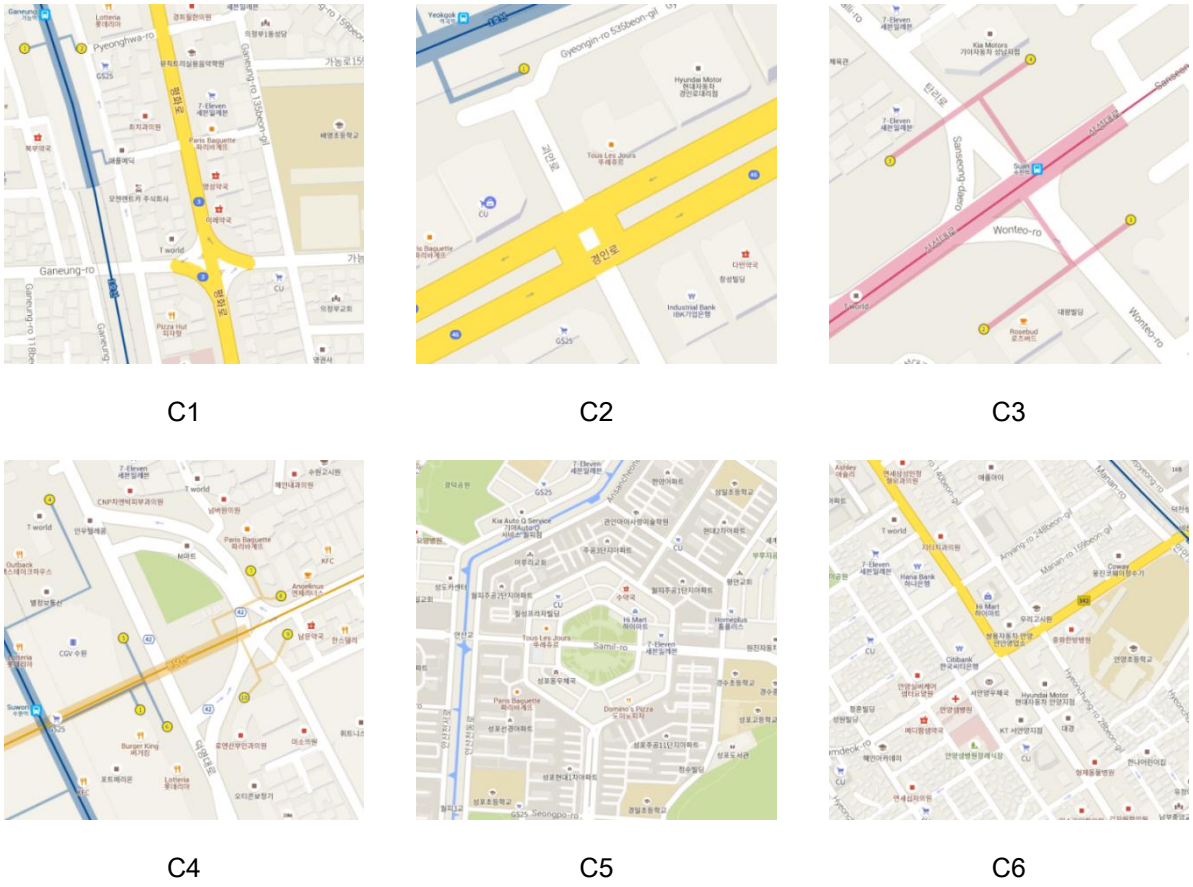

Figure 5: Black Spot of Jaywalking

Figure 5 shows the roads Black Spot where the number of Jaywalking Accidents was the highest among the six clusters. In the case of C1, C2, C3, C4 and C6, it is known that it has station, intersection, crossroad, In the case of C5, the structure of the road is confusing and it is a large apartment complex. In order to reduce the number of Jaywalking accidents, it is important to prevent jaywalking through police officer and fence where was a large number of people in the floating population, crossroad, convenience facilities. If it is difficult to prevent access to the road, it is important to install overpass for don't using the crosswalk indirectly.

CONCLUSIONS

This study clusters data from the Gyeonggi-Do province accident data provided by the public data portal. The DBSCAN algorithm was applied to clustering using fpc library, a statistical program R package. As a result of the experiment, six clusters were confirmed. A table was prepared based on the area of jaywalking accidents with high death rate among accident data, It was found that the structure of a busy street or road with a lot of floating population such as station, intersection, crossroad, convenient facilities was confusing and large-scale residential complex through road status. In order to prevent jaywalking of the road without the crosswalk,

the crosswalk should not be made through the fence or the police officer. If there is a crosswalk but the crosswalk is made and the accident is frequent, an overpass or an additional crosswalk may be added, It is important to make sure that the accident does not occur by going to another route without doing so.

Also, there is a problem that the results of research are different according to the assumption of K like K-Means algorithm. In this study, we overcome the existing constraints by using DBSCAN algorithm to set Epsilon and MinPts, which are cluster conditions, without using K, which is a conventional constraint.

REFERENCES

[1] Park, H.B., Joung, J.O., An Energy Efficient Re-clustering Algorithm in Wireless Sensor Networks. *The Journal of the Institute of Internet, Broadcasting and Communication (JIIBC)*, 15, 3 (2015), 155-161.

[2] Kwon, Y.M. Kwon, Lee, I.R., Kim, M.G., A Study on Clustering of SNS SPAM using Heuristic Method. *The Journal of the Institute of Internet, Broadcasting and Communication (JIIBC)*, 14, 6 (2014), 7-12.

[3] Jun, W.C., A Study on Correlation between Age and Information Ethics Using Information Culture Index. *International Journal of Internet, Broadcasting and Communication (IJIBC)*, 8, 2 (2016), 81-85.

[4] Kim, I. J., Machine learning development trend, industrialization case and activation policy direction. *SPRi Issue Report*, 17 (2016).

[5] Korea Chamber of Commerce and Industry, Big Data Utilization Status and Policy Task Research. (2014).

[6] Kim, Y. J., Yu, B. E., Future society change that artificial intelligence technology development will bring. *KISTEP*, 12 (2016), 52-65.

[7] AGENT, Kenneth R., Evaluation of the High-Accident Location Spot-Improvement Program in Kentucky. (1973).

[8] Deacon, John A., Charles V. Zegeer, and Robert C. Deen., Identification of hazardous rural highway locations. (1974).

[9] Miaou, Shaw-Pin., The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention*, 26, 4 (1994), 471-482.

[10] Lee, K. Y., Jang, M. S., Selecting Technique of Accident Sections using K-mean Method. *International journal of highway engineering*, (2005), 211-219.

[11] Lee, T. K., Chung, J. H., A Study on the Identification of Accident Hot Spots Using DBSCAN : Focused on Gyeong-Bu Expressway. *Journal of Transport Research*, 21, 3 (2014), 55-63.

[12] Kim, K. I., Jeong, J. S., Park, G. K., A Study on Near-miss Incidents from Maritime Traffic Flow by Clustering Vessel Positions. *Journal of Korean Institute of Intelligent Systems*, 24, 6 (2014), 603-608.

[13] Jain, A. K., & Dubes, R. C., Algorithms for clustering data. Prentice-Hall, Inc. (1988).

[14] Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). Machine learning, neural and statistical classification.

[15] Ester, M., Kriegel, H. P., Sander, J., & Xu, X., A density-based algorithm for discovering clusters in large spatial databases with noise. *In Kdd*, 96, 34 (1996), 226-231.

[16] http://data.gg.go.kr/

[17] Matloff, N., The art of R programming. No Starch Pess. 3, (2011).

[18] Christian Hennig, fpc: R Package. https://cran.r-project.org/web/packages/fpc/fpc.pdf, (2015).