A Novel Outlier Detection Applied to an Adaptive K-Means

Sarunya Kanjanawattana

***Abstract*—Data clustering is an important task for data management because it groups similar data into clusters and acquires significant knowledge. K-means is one of the popular clustering algorithms; however, there are several weaknesses such as cluster quality often depended on initial centers and too sensitive to an outlier. To address the problems, this study proposed a new method of initial centers selection based on data density and a novel approach of outlier detection based on data distance. I conducted some experiments to evaluate the methods. For the new method of initial centers selection, I compared the number of iterations and the Silhouette scores from this method and the traditional K-means. For the outlier detection system, I measured the system performance by using a confusion matrix. As the results, the system of the study outperformed the traditional K-means because of higher speed and great accuracy acquired.**

***Index Terms*—K-means, outlier detection, initial centers, a clustering algorithm, local outliers.**

I. Introduction

Despite the huge volume of data, data mining is a useful technique to support people to manage the data because it can extract some knowledge and patterns existing in the data. For example, many previous studies [1] proposed ideas of data mining techniques, e.g., decision tree [2] and association rule [3], to deal with Heart Disease data. The findings from the previous studies may be applied to doctors by helping them to make a treatment decision. A concept of clustering can be also used to find factors and group people who have a risk of suicide [4]. Clustering is a technique of grouping data containing similar characteristics. Fig. 1 demonstrates the concept of a clustering process. It groups similar data characteristics; moreover, the point O in Fig. 2 represents an outlier. It appears in many fields such as machine learning, pattern recognition, image analysis and more. Besides finding the clusters, identifying outliers distinguished from inliers is a challenging task in clustering. An inlier lies within the general population of the observed value or points in the clusters. The inlier is in contrast to an outlier. The outlier is defined as an observation that deviates far from other observations or clusters.

There are several existing clustering algorithms such as K-means [5], DBSCAN [6], [7], K-medoids [8], [9], and BIRCH [10]. K-means is a popular clustering technique nowadays; although there are some obvious weaknesses. First, it is sensitive to the outliers [11], [12]; since a performance of clustering is depended on data purity. Second, it is difficult to realize the exact number of clusters [13]. Third, a number of clustering iteration is based on the size of the dataset and initial centers. Fourth, it deals with the pure numeric dataset only [14]. To improve the K-means, some extended processes have been necessary. For example, a study of MixK-means++ [15] attempted to mitigate the problem of mixed dataset operated to the extension of K-means.

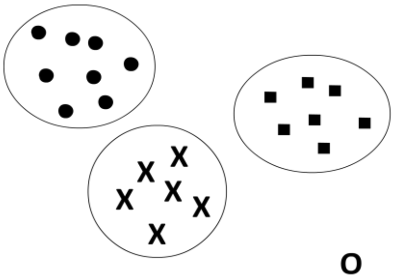


Fig. 1. A concept of the clustering process.

For the task of clustering, a few outliers easily destroy the clusters; then, the problem of outliers should be addressed. However, as a viewpoint of abnormal observation, people focus on detecting the outliers, rather than finding clusters of inliers [16]. The outlier detection has important application in the data mining field such as fraud detection [17], customer behavior analysis [18], and intrusion detection [12]. Based on these perspectives, it is certainly useful for both normal and abnormal observations if the outliers can be identified and separated somehow. For instance, the study from [19] presented a DOPHIN approach that was an algorithm to detect the outliers by distance-based approaches on a large-scale dataset. A study by [20] introduced a Local outlier factor (LOF) algorithm approach on Graphics processing units for Intrusion detection systems. Its main idea was to observe a method that showed how to apply a CUDA based GPU implementation of the k-nearest neighbor algorithm to increase the speed of LOF classification. LOF used a concept of Density-based methods to detect the outliers and capture outliers' degree depending on the density of their local neighborhoods. The algorithm assigned LOF values to instances; then the probability representing LOF value was identified to data instances. A high LOF value represented the data object as potential outliers; while others indicated by a low LOF value were defined as inlier instances [21].

Outlier detection is applicable to industries such as the banking sector and financial, government agencies and insurance. In the financial or business area, outlier detection

Manuscript received September 17, 2018; revised June 1, 2019.

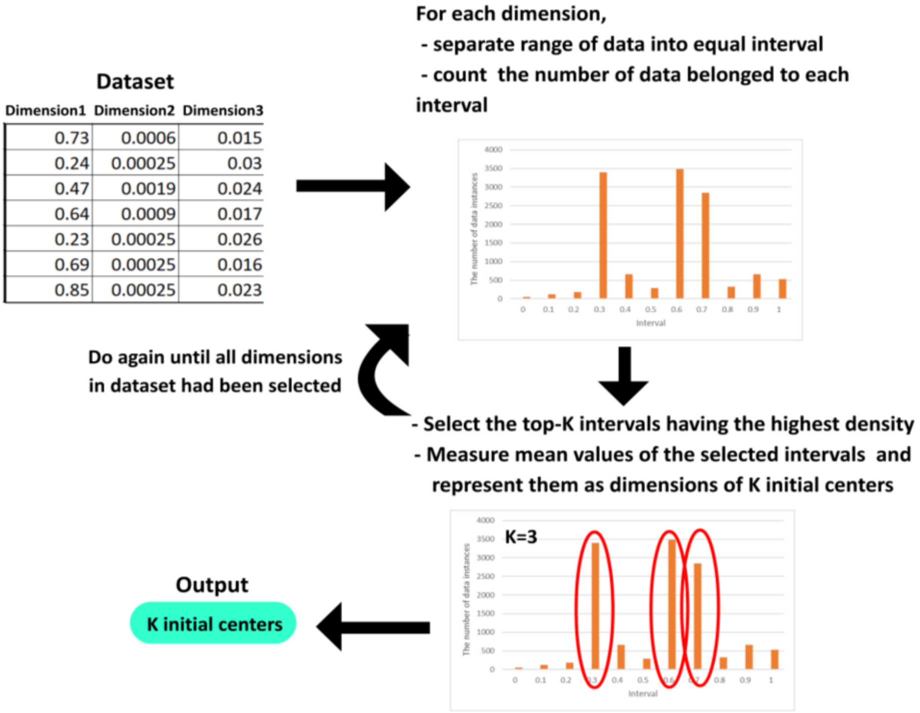
Sarunya Kanjanawattana is with the Compiter Engineering, Institute of Engineering, Suranaree University of Technology, Nakhonratchasima 30000, Thailand (e-mail: Sarunya.k@sut.ac.th).

doi: 10.18178/ijmlc.2019.9.5.841

is usually called fraud detection. Fraud crimes have seemed increasing in recent years; therefore, the outlier detection becomes more significant than ever. Notwithstanding many efforts of the affected institutions, hundreds of millions of dollars are still lost to fraud and tend to increase every year. An electronic statistical book [22] showed that the fraud was detected on the actions of stolen credit cards, misleading accounting practices and forging checks in banking. Frauds in claim insurance had been discovered that reached 25%. A basic solution to detect rare instances is to utilize data mining techniques to identify outliers.

Today, outlier detection has been a challenging problem in the real world. Therefore, many studies attempted to overcome the problem. In this study, I propose a novel method of initial centers selection for K-means including introducing a density-based outlier detection. To select the center, a traditional K-means randomizes K initial centers to be seeds of clusters. Note that K represents the number of clusters, which is a predefined parameter for K-means. However, in this new method, I use a density-based idea to locate the most possible positions in data space to be the centers. In another word, high density or data points massive areas are candidates of the initial centers. Furthermore, I introduce a new method of outlier detection by using a basic idea of the distance-based method. Objectives of this study are to improve a method to select candidates of the initial centers for K-means as well as to present a new approach to outlier detection.

The remainder of the paper is organized into six sections as follows: Section II introduces related works. Section III describes the methodology. Section IV presents the process of experiments and results. Section V discusses the findings. Finally, Section VI presents a conclusion and suggests future works.


Dataset

Dimension1	Dimension2	Dimension3
0.73	0.0006	0.015
0.24	0.00025	0.03
0.47	0.0019	0.024
0.64	0.0009	0.017
0.23	0.00025	0.026
0.69	0.00025	0.016
0.85	0.00025	0.023

For each dimension,
- separate range of data into equal interval
- count the number of data belonged to each interval

Do again until all dimensions in dataset had been selected

- Select the top-K intervals having the highest density
- Measure mean values of the selected intervals and represent them as dimensions of K initial centers

Output

K initial centers

Fig. 2. An illustration of a process of selecting K initial centers for K-means (assuming that K is three).

II. Related Works

During several past decades, there are many works interested in this domain and proposed dominant methods. Similar to this study, the novel K-means clustering and outlier detection had been introduced. I presented the idea of the K-means clustering method based on data density and the outlier detection based on data distance.

K-means clustering is a famous algorithm of clustering introduced by Lloyd [23] since 1982. The method is simple and efficient. Traditional K-means algorithm consists of four steps as following:

1. Choose k initial centers
2. Assign data points X to the closest center
3. Recompute and find new centers
4. Repeat step 2 and 3 until the centers no longer change

However, under several drawbacks, traditional K-means needs to be improved. The K-means clustering is traditionally applicable to numeric dataset only. Kanjanawattana [15] proposed a new K-means method that attempted to omit K-means limitations. It could apply to different types of data attributes, i.e., nominal and numeric data attributes. Moreover, the study also developed an idea based on the K-means++ algorithm [24] that presented a way to choose the initial values or called seeds. It could avoid poor clusters that yield considerable improvement in the final error of K-means. K-means++ was developed based on K-means; thus, its steps was surely similar to K-means, except the first step of algorithm which was a way to choose initial centers. It selected frothe m the data point probability. However, K-means++ algorithm was inapplicable to the mixed-type data set, which combines categorical and the numerical attribute, because the standard K-means is suitable to numeric data set merely.

Most existing research about outlier detection aimed at numerical data sets and cannot straightway handle categorical sets. Moreover, the outlier detection method required quadric time based on the size of the data set. A study [25] introduced a method called Attribute Value Frequency or AVF that was a rapid and flexible outlier detection strategy for categorical data. It linearly scaled with a number of attributes and data points as well as relied on a single data scanned. An experiment with AVF presented that the algorithm's speed increased and provided effective performance.

There were some appropriated methods to attempt to discover outliers in categorical datasets. Most studies presented clustering technique to detect outliers because this is a common idea to detect anomaly among data. However, some studies proposed a method using a concept of frequent itemsets for the categorical attribute. However, there were a few existing studies related to outlier detection using a rule generated from frequent item set. A study [26] explored the effect of applying a condensed representation of the frequent itemsets on the correctness of the outlier detection approach. It called Non-Derivable Item sets or NDI. There were three steps of the NDI. First was to use Apriori algorithm to extract frequent itemsets. Second, the NDI stored a representative subset of all frequent itemsets instead of storing each frequent item. The third was to detect the outliers based on the frequent itemsets. Finally, the outlier detection based on the NDI frequent itemsets algorithm was proposed. This algorithm assigned an outlier score to a point based on the item sets in NDIRep (Data set, min_sup). The results indicated that the NDI-based outlier detection offered an important contribution in terms of rapidity and scalability over frequent itemsets based on outlier detection.

To mitigate the problem of choosing the initial centers, there are numerous related studies such as [27] and [28]. Also, Yedla *et al.* [29] proposed a method to find the better initial centers and to provide an effective solution for assigning the data points along to suitable clusters with low time complexity. This method had not required any additional input; however, the value of K was still necessary.

Outlier detection has been applied for several years in order to monitoring and removing anomalous observations from data. Outliers occurred because of changes in system behavior, mechanical faults, human error, fraudulent behavior, simply through natural deviations in populations, or instrument error. The main idea of outlier detection is to remove the contaminated data and purify them to be ready for mining [30].

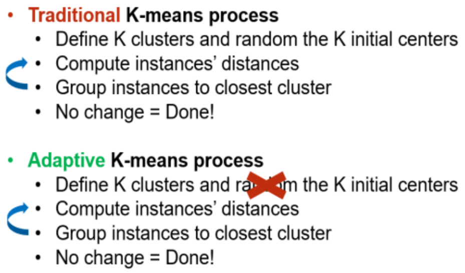
Outlier detection also plays an important role in finance. Visa and Master card exhibited a sales volume of over $190.6 billion at the end of 2005 [31]. The statistics on credit card fraud indicated that approximately $2.8 million was lost from Master card and Visa due to fraud. In conclusion, at least $500 million was lost in a year by credit card fraud. With the significance of fraud detection, many studies attempted to present possible techniques to detect fraudulent behaviors and generate predictable models for credit card fraud detection. Bhattacharyya *et al*. [32] evaluated two data mining approaches, i.e., random forests and support vector machines, altogether with a well-known logistic regression.

Glancy *et al.* [33] proposed a quantitative model to detect abusive the financial report. The model detected the intention of concealing information and exhibited improper information with the US Securities and Exchange Commission (SEC) in annual filings. The input of the model was an entire text document for fraud detection. An accurate and coherent screening tool offered decision support for the beginning of fraud detection. Furthermore, Sanchez *et al.* [34] proposed an approach to extract knowledge from normal behavior patterns. They received lawless invoices from credit card databases for the purpose of prevention and fraud detection; moreover, the method had been practiced in retail companies of Chile.

Regarding the outlier detection, after reviewing several existing studies, K-means usually used to detect abnormal points separated from groups or outliers because clustering results by K-means are sensitive to outliers. Thus, this algorithm is suitable to use for evaluating the outlier detection system [35]. Chawla *et al.* [36] improved their previous method to guaranteed to converge to a local optimum and proposed a new approach to measure data distance presenting in the form of a Bregman divergence. Their method resulted in an improvement in the precision of the outlier detection task by nearly 100%.

III. Methodology

In this study, I introduced two novel methods, i.e., a method of initial centers selection in a process of K-means and a method of outlier detection.



- **Traditional K-means process**
  - Define K clusters and random the K initial centers
  - Compute instances' distances
  - Group instances to closest cluster
  - No change = Done!
- **Adaptive K-means process**
  - Define K clusters and ~~random~~ the K initial centers
  - Compute instances' distances
  - Group instances to closest cluster
  - No change = Done!

Fig. 3. A difference between the adaptive K-means and the traditional K-means.

A. Adaptive K-means Process

A traditional K-means contains four steps. First, the number of desired clusters is defined beforehand, and the initial centers are randomly selected. Second, all instances are computed their distances and grouped into the closest cluster. Then, the algorithm iteratively computes new centers again. Finally, all instances are regrouped to the new centers until the instances do not change the clusters. Fig. 3 presents a difference between the adaptive K-means proposed in this study and the traditional K-means.

In this study, I presented a new simple idea of selecting initial centers based on a density of each data dimensionality explaining within three steps. First, I separated a range of data into intervals. Second, for each data dimension, the number of instances matched to the interval were counted. Third, I selected the intervals with the top-K highest densities and measured the mean of the intervals. The process continues until all dimensions of data had been computed. These mean values represent the K initial centers. Fig. 2 demonstrates my new method of selecting initial centers. However, in this study, only the method of the initial centers had been introduced, other steps of K-means conserved.

B. A Novel Outlier Detection Approach

The initial centers had been selected by the method presented in the previous subsection. After clustering the data by K-means, some proper clusters should be provided. However, the results seemed to be sensitively affected by outliers concealing in the clusters. Therefore, a process of outlier detection is important for improving clustering performance.

I proposed an effective method of outlier detection based on data distance. There are four steps described below.

First, suspected instances are remarked. The system detects them by defining the value of cluster bound. The cluster bound is average distances between member instances in a cluster and its center added to a value of the standard deviation of the distances. If a member instance is further than the cluster bound, it will be observed as a suspected instance.

Second, a value of MinPts should be defined beforehand. MinPts is the number of the nearest neighborhoods whose distances close to an observed instance. For example, if the MinPts is five, five neighborhoods closely away from the observed instance should be selected. Furthermore, the member instances selected as neighborhoods of the suspected instance also have their own neighborhoods.

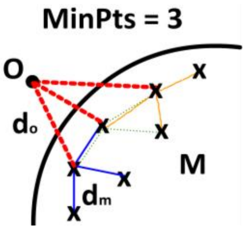


Fig. 4. Detection of a suspected instance by computing the average distances from its closest neighbors (d_o) compared to the average distances within the neighbors (d_m).

Third, this system computes the distances between the suspected instance and its neighborhoods. Fig. 4 displays an example of a suspected instance (Instance O) and its neighborhoods (Instance M). The distances between Instance O and Instance M are represented as d_o, and the distances among neighborhoods are call d_m. Following this step, the system obtains the two average distances, i.e., average d_o and average d_m. Note that, for average d_o, all d_o among neighborhoods are retrieved. Thus, average $\mathrm{d_o}$ come from a summation of all d_o divide by the number of MinPts. Moreover, to acquire average d_m, all neighborhoods of the suspected instance should have their own neighborhoods in the cluster; then, their distances are measured, and the system obtains average distances of among neighborhoods (average d_m).

Fourth, after obtaining the average d_o and d_m, the system compares these obtained values. If the average d_o is greater than $\mathrm{d_m}$, the suspected instance will be judged as an outlier. Otherwise, the average d_o is lower than or equal d_m, the suspected instance will be judged as an inlier. Finally, the process moves back to the previous step by selecting the next suspected instances. The process continues until all suspected instances were judged.

IV. Experiments and Results

Here, I conducted two experiments to evaluate the new method of the initial center's selection and the new method of outlier detection. The experimental data is about Thyroid disease. The dataset contained six numeric attributes and 7200 instances included outliers. In this dataset, it contained the outliers around 7.4% of total or 534 instances. To prepare the dataset for experiments, the dataset had been duplicated into two sets, i.e., the dataset included outliers (Dataset1) and the dataset excluded outliers (Dataset2).

K	Novel K-means		Traditional K-means	
	Iteration	Silhouette	Iteration	Silhouette
2	5	2.07E-04	8	2.07E-04
3	8	3.33E-04	16	3.32E-04
5	27	3.82E-04	37	5.70E-04
7	54	6.38E-04	58	6.20E-04
9	39	8.16E-04	106	8.71E-04
11	30	7.97E-04	109	8.57E-04

Fig. 5. Experimental results of the clustering process.

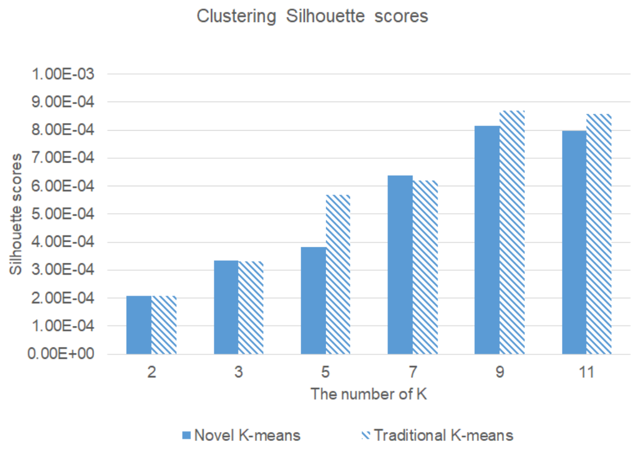


Fig. 6. Silhouette scores comparing the adaptive K-means to the traditional K-means.

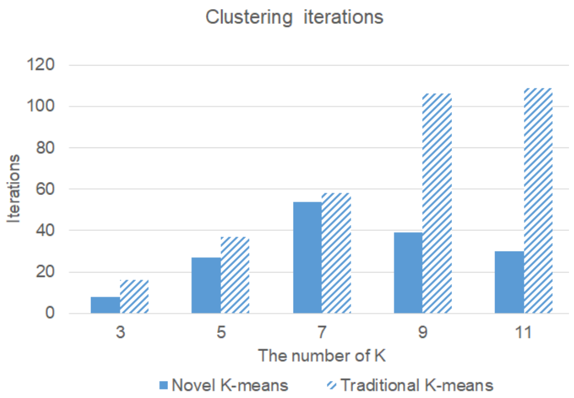
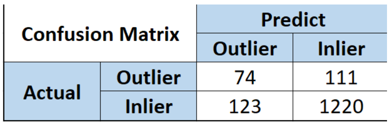


Fig. 7. The number of iterations comparing the adaptive K-means to the traditional K-means.

To evaluate the method of the adaptive K-means as regards to the initial center selection, I used Dataset2 because experimental results presented a true performance of the clustering process. I compared my adaptive K-means to the traditional K-means with different the number of clusters (K) and provided the results such as Silhouette scores and the number of iterations. Note that Silhouette score represents data consistency within individual clusters that indicates how well the data lies within the cluster. A higher value of Silhouette score means an instance properly matched to its cluster. The number of iterations represents the number of clustering process repeated. A higher value of iteration means a slower process.

After processing the dataset to both K-means, I obtained the results as presented in Fig. 5, 6 and 7. Fig. 5 displays statistical results of clustering performance. Figure 6 presents a diagram of Silhouette scores comparing between adaptive and traditional K-means. Fig. 7 illustrates a diagram showing the number of iterations counting during the clustering processes comparing to both K-means.



Confusion Matrix		Predict	
		Outlier	Inlier
Actual	Outlier	74	111
	Inlier	123	1220

Fig. 8. Confusion matrix using to evaluate the outlier detection system.

To validate the outlier detection, Dataset 1 conducted the experiments. The results were presented in a confusion matrix (Fig. 8). From the matrix, precision, recall, accuracy, and F-measures were 0.4, 0.38, 0.84, and 0.38 respectively.

V. Discussions

The study presented a new idea to improve the efficiency of the K-means algorithm. There are two methods produced in this study, i.e., a method of initial center selection and a method of outlier detection. The objectives of this study were to introduce a method to select initial centers and to present a new approach to outlier detection. For evaluations, I separated the experiments into two parts because the experiments should coverage all part of the study.

For the first part of the experiments, I needed to validate the method of initial centers selection by comparing to the traditional K-means. I set up two measurements that are Silhouette scores and the number of iterations. Based on the results, I found that this system processed faster than the traditional K-means. As seen in Fig. 7, if K increased, the traditional K-means provided high iterations; meanwhile, iterations from the adaptive K-means were lower. This happened because the method of this study selected the suitable initial center candidates depended on data density, as different from the traditional K-means, which selected the initial centers by randomizing. Therefore, this system can obtain proper clusters quicker than the traditional one that helps to reduce the number of iterations. Moreover, as shown in Fig. 6, Silhouette scores of the traditional K-means and my adaptive K-means were similar. Thus, it clarified that the adaptive K-means outperformed to the traditional K-means because it worked much faster and the Silhouette scores were not much different.

For the second part of the experiments, the approach of outlier detection should be verified by measuring precision, recall, F-measures, and accuracy. The confusion matrix had been presented as shown in Fig. 8. As observed the precision, recall, the system can identify outliers correctly about 40% of the total. The values of these performance measurements may improve if I reduce the MinPts because this system selected neighborhoods by ranking the top-MinPts instances that have the closest distance; then, the lower MinPts means the small size of neighborhoods and the small average distances within the cluster. The idea of identifying suspected instances is that the distance from the suspected instance to the normal instances in the cluster should be higher than the distance among the normal instances and their neighborhoods within the cluster. This showed that the suspected instance had separated away from those instances in the cluster. However, I observed the accuracy that was up to 84% of the total. The system primarily located cluster instances by defining the bound value, which computed from the average distances within the cluster plus its standard deviation. The instances whose distances are lower than a bound value should be identified as inliers. Due to this idea, this system can identify inlier correctly reached 90%.

However, this system has some limitations found during the study. Originally, K-means needs a predefined parameter K that is a drawback of this algorithm. Unfortunately, another parameter called MinPts had been required for this system as well. Moreover, it is difficult to define the value of MinPts that is suitable for the observed data, as similar to the value of K. Further, the performance is highly depended on the data characteristics. For example, if the data are exceedingly concentrated or hardly separable, this new idea to obtain initial center candidates may not appropriate.

VI. Conclusions

This study improved the K-means algorithm by introducing the method of initial center candidates. The tradition K-means randomizes the K initial center; thus, the cluster quality is uncertainty. Moreover, I attempted to achieve a drawback of K-means that is too sensitive to an outlier. I proposed the novel method of outlier detection based on data distance. The system identified an instance, which was further than a bound value as a suspected instance and judged it as an outlier by comparing the data distance between suspected instances and normal instances.

For the evaluation, I conducted some experiments and obtained significant results. The finding showed that my method provided better performance than the traditional K-means. My method provided the small number of iterations even the number of K was increased. In the meantime, the Silhouette scores were not much different from the result of the traditional K-means. To validate the outlier detection, I obtained satisfied F-measure and accuracy. In the future, I will continue to develop the method to deal with high dimensional data.

References

[1] G. Purusothaman and P. Krishnakumari, “A survey of data mining techniques on risk prediction: Heart disease,” *Indian Journal of Science and Technology*, vol. 8, no. 12, 2015.

[2] C. Jin, L. De-Lin, and M. Fen-Xiang, “An improved ID3 decision tree algorithm,” in *Proc. ICCSE’09. 4th International Conference on Computer Science & Education*, 2009, pp. 127–130.

[3] R. Agrawal, R. Srikant *et al*., “Fast algorithms for mining association rules,” in *Proc. 20th Int. Conf. very Large Data Bases*, 1994, vol. 1215, pp. 487–499.

[4] M. Sinyor, L. P. L. Tan, A. Schaffer, D. Gallagher, and K. Shulman, “Suicide in the oldest old: an observational study and cluster analysis,” *International Journal of Geriatric Psychiatry*, vol. 31, no. 1, pp. 33–40, 2016.

[5] K. Alsabti, S. Ranka, and V. Singh, *An Efficient K-means Clustering Algorithm*, 1997.

[6] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al*., “A density-based algorithm for discovering clusters in large spatial databases with noise,” *Kdd*, vol. 96, no. 34, pp. 226–231, 1996.

[7] T. N. Tran, K. Drab, and M. Daszykowski, “Revised DBSCAN algorithm to cluster data with dense adjacent clusters,” *Chemometrics and Intelligent Laboratory Systems*, vol. 120, pp. 92–96, 2013.

[8] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[9] W. Sheng and X. Liu, "A genetic k-medoids clustering algorithm," *Journal of Heuristics*, vol. 12, no. 6, pp. 447–466, 2006.

[10] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," *ACM Sigmod Record*, 1996, vol. 25, no. 2, pp. 103–114.

[11] V. Hautamäki, S. Cherednichenko, I. Kärkkäinen, T. Kinnunen, and P. Fränti, "Improving k-means by outlier removal," in *Proc. Scandinavian Conference on Image Analysis*, 2005, pp. 978–987.

[12] K.-A. Yoon, O.-S. Kwon, and D.-H. Bae, "An approach to outlier detection of software measurement data using the k-means clustering method," in *Proc. First International Symposium on Empirical Software Engineering and Measurement*, 2007, pp. 443–445.

[13] D. Pelleg, A. W. Moore *et al*., "X-means: Extending k-means with efficient estimation of the number of clusters.," in *Proc. ICML*, 2000, vol. 1, pp. 727–734.

[14] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[15] S. Kanjanawattana, "An extended K-means++ with mixed attributes," in *Proc. The 12th WSEAS International Conference on Applied Computer Science (ACS12)*, 2012, pp. 131–135.

[16] S. R. Gaddam, V. V Phoha, and K. S. Balagani, "K-Means+ ID3: A novel method for supervised anomaly detection by cascading K-Means clustering and ID3 decision tree learning methods," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 345–354, 2007.

[17] C. Phua, V. Lee, K. Smith, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," *arXiv preprint arXiv:1009.6119*, 2010.

[18] M. P. Yadav, M. Feeroz, and V. K. Yadav, "Mining the customer behavior using web usage mining in e-commerce," in *Proc. 2012 Third International Conference on Computing Communication & Networking Technologies*, 2012, pp. 1–5.

[19] F. Angiulli and F. Fassetti, "Dolphin: An efficient algorithm for mining distance-based outliers in very large datasets," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, no. 1, p. 4, 2009.

[20] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," *ACM Sigmod Record*, 2001, vol. 30, no. 2, pp. 37–46.

[21] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *ACM Sigmod Record*, vol. 29, no. 2, pp. 93–104, 2000.

[22] E. S. Textbook. (2010). Tulsa, OK: StatSoft. [Online]. Available: http://www. statsoft. com/textbook/stathome.html

[23] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[24] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proc. the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007, pp. 1027–1035.

[25] A. Koufakou, E. G. Ortiz, M. Georgiopoulos, G. C. Anagnostopoulos, and K. M. Reynolds, "A scalable and efficient outlier detection strategy for categorical data," in *Proc. 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, 2007, vol. 2, pp. 210–217.

[26] M. Fox, G. Gramajo, A. Koufakou, and M. Georgiopoulos, *Detecting Outliers in Categorical Data Sets Using non-Derivable Itemsets*, 2008.

[27] C. Zhang and S. Xia, "K-means clustering algorithm with improved initial center," in *Proc. Second International Workshop on Knowledge Discovery and Data Mining*, 2009, pp. 790–792.

[28] S. S. Khan and A. Ahmad, "Cluster center initialization algorithm for K-means clustering," *Pattern Recognition Letters*, vol. 25, no. 11, pp. 1293–1302, 2004.

[29] M. Yedla, S. R. Pathakota, and T. M. Srinivasa, "Enhancing K-means clustering algorithm with improved initial center," *International Journal of Computer Science and Information Technologies*, vol. 1, no. 2, pp. 121–125, 2010.

[30] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

[31] Spamlaws, "Credit card fraud statistics and facts," *Credit Card Fraud Statistics and Facts*.

[32] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," *Decision Support Systems*, vol. 50, no. 3, pp. 602–613, 2011.

[33] F. H. Glancy and S. B. Yadav, "A computational model for financial reporting fraud detection," *Decision Support Systems*, vol. 50, no. 3, pp. 595–601, 2011.

[34] D. Sánchez, M. A. Vila, L. Cerda, and J.-M. Serrano, "Association rules applied to credit card fraud detection," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3630–3640, 2009.

[35] G. Münz, S. Li, and G. Carle, "Traffic anomaly detection using k-means clustering," in *Proc. GI/ITG Workshop MMBnet*, 2007, pp. 13–14.

[36] S. Chawla and A. Gionis, "k-means: A unified approach to clustering and outlier detection," in *Proc. the 2013 SIAM International Conference on Data Mining*, 2013, pp. 189–197.

Sarunya Kanjanawattana was born in Naknonratchasima, Thailand, in 1986. She received the B.E. degree in computer engineering from Suranaree University of Technology, Nakhonratchasima, Thailand, in 2008, and M. Eng from Asian Institute of Technology, Pathum Thani, Thailand in 2011. In 2017, She graduated from her doctor course with a major of functional control systems from Shibaura Institute of Technology, Tokyo, Japan. In 2011, she joined National Electronics and Computer Technology Center, Thailand, as a research assistance. Her project related to finding an optimal solution to traffic congestion. At the present, she works at Suranaree University of Technology as a lecturer in the department of computer engineering. Her research interests included data mining, machine learning, natural language processing, ontology and computer vision.