International Journal of Applied Sciences and Smart Technologies, Vol. 1, No.1 (2019) 38 - 48
http://e-journal.usd.ac.id/index.php/IJASST/index
ISSN: 2655-8564

Implementation of k-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta

Eduardus Hardika Sandy Atmaja [1*]

[1] *Informatics Study Program, Faculty of Science and Technology, Sanata Dharma University, Yogyakarta, Indonesia*
[*]*Corresponding Author: edo@usd.ac.id*

(Received 02-05-2019; Revised 21-05-2019; Accepted 21-05-2019)

Abstract

The increase in crime from day to day needs to be a concern for the police, as the party responsible for security in the community. Crime prevention effort must be done seriously with all knowledge that they have. To increase police performance of crime prevention effort, it is necessary to analyze crime data so that relevant information can be obtained. This study tried to analyze crime data to obtain relevant information using clustering in data mining. Clustering is a data mining method that can be used to extract valuable information by grouping data into groups that have similar characters. The data used in this study were crime patterns which were then grouped using K-medoids clustering algorithm. The obtained results in this study were three crime groups, namely high crime level with 4 members, medium crime level with 6 members and low crime level with 8 members. It is expected that this information can be used as material for consideration in crime prevention effort..

Keywords: important, keyword, reflect, research

1 Introduction

Crime is any act that is prohibited by public law to protect the public and given punishment by the state. These acts are punished because they are violating the social norms such as act that conflict with legal norms, social norms and religious norms that applied in the society [1]. The existence of punishment applied by law enforcement does not make the criminals undermine their intentions, and in fact criminal in Yogyakarta are increasing widespread.

The increase of criminal cases in the society can result in losses both materially and immaterially. For this reason, efforts are needed from law enforcement to reduce crime in the society. Such efforts can be done by finding relevant information related to crime. Such information can be obtained by processing and analyzing crime data owned by the police.

The crime data owned by the Yogyakarta Police is still stored in the manual form such as register books and excel. The data is only stored and is not used to produce any information. Where the data can be processed and analyzed to produce valuable information in efforts to prevent crime. Data mining is a proper technique to extract important information from a data set.

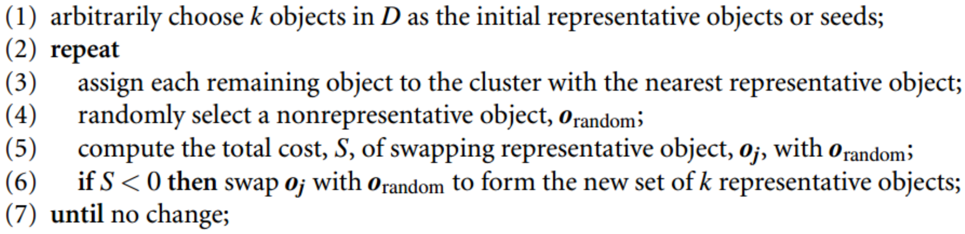
Crime data owned by the police can be processed using data mining to become crime patterns that represent relationship between crimes. The research was successfully done by Atmaja [2], the result was crime patterns presented in graph form. The weakness in that study is that there is no clear grouping on crime level form generated crime patterns. This study tried to refine previous research by groupings crime patterns into three categories, namely high crime level, medium crime level and low crime level.

Clustering is one of the data mining techniques that aims to group data based on information found in the data [3]. The grouping is based on the similarity between data so the data in the same cluster is homogeneous. Thus clustering is a very appropriate method for classifying crime patterns into high, medium and low crime level.

Researches on implementation of clustering method have been done, as done by Singh et. al. [4]. They tried to implement K-means clustering algorithm by using three different distance measurements namely Euclidean, Manhattan and Chebychev. The result is that the implementation of K-means algorithm using Euclidean distance measurements can produce the best group from the other distance measurements. So it can be concluded that the best pair for K-means algorithm is the Euclidean distance measurement.

Research on the use of Euclidean distance in K-means algorithm has been successfully done by Atmaja [5]. The aim of his study was to cluster crime data into three categories, namely high, medium and low crime level. Although the objective of the research was achieved, K-means algorithm is classified as an ineffective algorithm because it involves too much noise and outliers caused by the average selection of clusters [6].

This study tried to improve previous study by replacing K-means algorithm with K-medoids algorithm. K-medoids algorithm is one of the clustering algorithms that are not influenced by outliers or other extreme variables [6]. K-medoids work by determining the center point of existing data without performing an average calculation as in K-means. The following is the K-medoids algorithm [6]:



```
arbitrarily choose $k$ objects in $D$ as the initial representative objects or seeds;
repeat
    assign each remaining object to the cluster with the nearest representative object;
    randomly select a nonrepresentative object, $o_{random}$;
    compute the total cost, $S$, of swapping representative object, $o_j$, with $o_{random}$;
    if $S < 0$ then swap $o_j$ with $o_{random}$ to form the new set of $k$ representative objects;
until no change;
```

Figure 1. K-medoids algorithm

The result of this study is crime patterns that have been divided into three groups, namely high, medium and low crime level. It is expected that the police can use this information to improve crime prevention efforts in the society.

2 Research Methodology

Research methodology done by this research is activity steps to implement K-means algorithm to cluster crime patterns from Yogyakarta Police data which are presented in Figure 2. Figure 2 shows research methodology which began with literature study to study relevant theories related to solve problems. The next step was data collecting related to research, in this case the processed data was crime data from Yogyakarta Police. The crime data that has been collected then processed using association techniques in data mining to produce association rules that described crime patterns. Generated rules was used as input to K-medoids algorithm to produce crime patterns accompanied by grouping based on low, medium and high crime level. The next step was result analyzing that has been obtained to find out whether the objective achieved or not. Finally, the result analysis will draw conclusions from the research that has been done. Suggestions were also given to correct existing disadvantages to be applied in the future research.

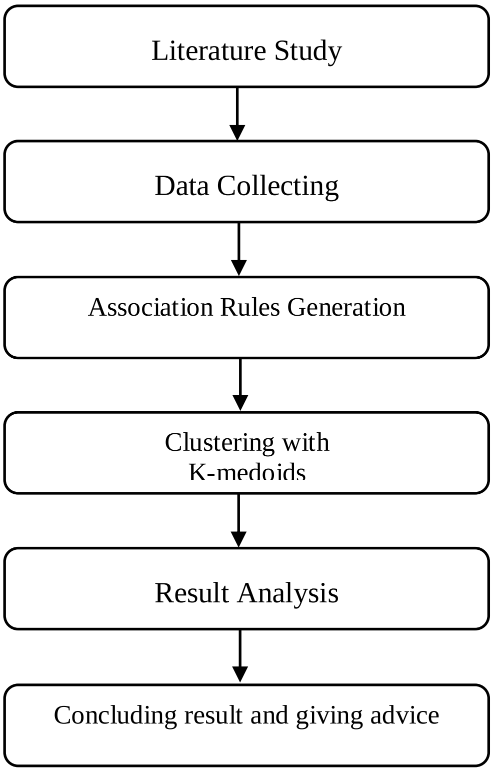


Figure 2. Research methodology

3 Results and Discussions

This research was completed through several stages described in the following steps.

3.1 Crime Patterns

There are 18 samples of crime patterns as results of association technique processing accompanied by support and confidence. The data will be grouped using the K-medoids algorithm based on variable support and confidence. These data are presented in Table 1.

Table 1. Crime patterns

No.	Rules	Support	Confidence
1	IF Embezzlement THEN Theft	0.02	0.03
2	IF Theft THEN Embezzlement	0.02	0.29
3	IF Embezzlement THEN Deception	0.54	0.81
4	IF Deception THEN Embezzlement	0.54	0.82
5	IF Embezzlement THEN Document Forgery	0.02	0.03
…	…	…	…
18	IF Unpleasant Act THEN Defamation	0.02	0.38

3.2 Determining Initial Medoids

In the first stage, three medoids were randomly selected from data sample in Table as shown in Table 2.

Table 2. Three initial medoids

	Medoid		
	C1	C2	C3
Support	0.54	0.08	0.03
Confidence	0.81	0.12	0.30

3.3 Calculating Euclidean Distance Iteration 1

The next step is euclidean distance calculation from each data to the three selected medoids. Euclidean distance is calculated based on the following formula [6]:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2} \tag{1}$$

Here, $d(i,j)$ represents distance between data and medoid, x_{i1} denotes support value in each data, x_{j1} is medoid (c) for support, x_{i2} denotes confidence value in each data and x_{j2} is medoid (c) for confidence. Table 3 presents results from euclidean distance calculation on each data along with medoid information which has the shortest distance to the data.

Table 3. Rules with euclidean distance

Rules	Support	Confidence	Distance to Medoid			Shortest Distance
			C1	C2	C3	
1	0.02	0.03	0.937	0.108	0.270	0.108
2	0.02	0.29	0.735	0.180	0.014	0.014
3	0.54	0.81	0.000	0.829	0.721	0.000
4	0.54	0.82	0.010	0.838	0.728	0.010
5	0.02	0.03	0.937	0.108	0.270	0.108
6	0.02	0.41	0.656	0.296	0.110	0.110
7	0.09	0.13	0.815	0.014	0.180	0.014
8	0.09	0.97	0.478	0.850	0.673	0.478
9	0.02	0.03	0.937	0.108	0.270	0.108
10	0.02	0.41	0.656	0.296	0.110	0.110
11	0.08	0.12	0.829	0.000	0.187	0.000
12	0.08	0.94	0.478	0.820	0.642	0.478
13	0.03	0.30	0.721	0.187	0.000	0.000
14	0.03	0.86	0.512	0.742	0.560	0.512
15	0.02	0.23	0.779	0.125	0.071	0.071
16	0.02	0.69	0.534	0.573	0.390	0.390
17	0.02	0.44	0.638	0.326	0.140	0.140
18	0.02	0.38	0.675	0.267	0.081	0.081

From Table 3, it can be seen that medoid C1 has 5 members rules {3,4,8,12,14}, medoid C2 has 5 members rules {1,5,7,9,11} and medoid C3 has 8 members rules {2,6,10,13,15,16,17,18}.

3.4 Calculating Total Cost Iteration 1

Calculating total cost is the final step from iteration 1, by summing the shortest distance from data in Table 3. So the total cost is 2.734.

3.5 Determining Random Medoids Iteration 2

The process continues to iteration 2 by selecting a new random medoid from the data to replace the medoid C3 temporarily. The selection of a new medoid should not be the same as one of the medoids that has been selected. Table 4 shows three medoids for iteration 2.

Table 4. Three medoids iteration 2

	Medoid		
	C1	C2	C Random
Support	0.54	0.08	0.03
Confidence	0.81	0.12	0.86

3.6 Calculating Euclidean Distance Iteration 2

After a new medoid has been determined, the next step is to recalculate the euclidean distance for each data based on three medoids from Table 4. The results is shown in Table 5.

Table 5. Rules with euclidean distance iteration 2

Rules	Support	Confidence	Distance to Medoid			Shortest Distance
			C1	C2	C3	
1	0.02	0.03	0.937	0.108	0.830	0.108
2	0.02	0.29	0.735	0.180	0.570	0.180
3	0.54	0.81	0.000	0.829	0.512	0.000
4	0.54	0.82	0.010	0.838	0.512	0.010
5	0.02	0.03	0.937	0.108	0.830	0.108
6	0.02	0.41	0.656	0.296	0.450	0.296
7	0.09	0.13	0.815	0.014	0.732	0.014
8	0.09	0.97	0.478	0.850	0.125	0.125
9	0.02	0.03	0.937	0.108	0.830	0.108
10	0.02	0.41	0.656	0.296	0.450	0.296
11	0.08	0.12	0.829	0.000	0.742	0.000
12	0.08	0.94	0.478	0.820	0.094	0.094
13	0.03	0.30	0.721	0.187	0.560	0.187
14	0.03	0.86	0.512	0.742	0.000	0.000
15	0.02	0.23	0.779	0.125	0.630	0.125
16	0.02	0.69	0.534	0.573	0.170	0.170
17	0.02	0.44	0.638	0.326	0.420	0.326
18	0.02	0.38	0.675	0.267	0.480	0.267

From Table 5, it can be seen that medoid C1 has 2 members rules {3,4}, medoid C2 has 12 members rules {1,2,5,6,7,9,10,11,13,15,17,18} and medoid C3 has 4 members rules {8,12,14,16}.

3.7 Calculating Total Cost Iteration 2

Calculating total cost is the final step from iteration 2, by summing the shortest distance from data in Table 5. So the total cost is 2.416. To determine the next iteration, total cost from iteration 2 is compared with iteration 1, which is 2,416 > 2,734. Because the total cost of iteration 2 is not greater than iteration 1, the iteration is continued to iteration 3 and the medoid C Random replaces medoid C3.

3.8 Determining Random Medoids Iteration 3

The process continues to iteration 3 by selecting a new random medoid from the data to replace the medoid C3 temporarily (C Random from iteration 2). The selection of a new medoid should not be the same as one of the medoids that has been selected. Table 6 shows three medoids for iteration 3.

Table 6. Three medoid iteration 3

	Medoid		
	C1	C2	C Random
Support	0.54	0.08	0.02
Confidence	0.81	0.12	0.44

3.9 Calculating Euclidean Distance Iteration 3

After a new medoid has been determined, the next step is to recalculate the Euclidean distance for each data based on three medoids from Table 6. The results is shown in Table 7.

Table 7. Rules with euclidean distance iteration 3

Rules	**Support**	**Confidence**	**Distance to Medoid**			**Shortest Distance**
			C1	**C2**	**C3**	
1	0.02	0.03	0.937	0.108	0.410	0.108
2	0.02	0.29	0.735	0.180	0.150	0.150
3	0.54	0.81	0.000	0.829	0.638	0.000
4	0.54	0.82	0.010	0.838	0.644	0.010
5	0.02	0.03	0.937	0.108	0.410	0.108
6	0.02	0.41	0.656	0.296	0.030	0.030
7	0.09	0.13	0.815	0.014	0.318	0.014
8	0.09	0.97	0.478	0.850	0.535	0.478

Rules	Support	Confidence	Distance to Medoid			Shortest Distance
			C1	C2	C3	
9	0.02	0.03	0.937	0.108	0.410	0.108
10	0.02	0.41	0.656	0.296	0.030	0.030
11	0.08	0.12	0.829	0.000	0.326	0.000
12	0.08	0.94	0.478	0.820	0.504	0.478
13	0.03	0.30	0.721	0.187	0.140	0.140
14	0.03	0.86	0.512	0.742	0.420	0.420
15	0.02	0.23	0.779	0.125	0.210	0.125
16	0.02	0.69	0.534	0.573	0.250	0.250
17	0.02	0.44	0.638	0.326	0.000	0.000
18	0.02	0.38	0.675	0.267	0.060	0.060

From Table 7, it can be seen that medoid C1 has 4 members rules {3,4,8,12}, medoid C2 has 6 members rules {1,5,7,9,11,15} and medoid C3 has 8 members rules {2,6,10,13,14,16}.

3.10 Calculating Euclidean Distance Iteration 3

Calculating total cost is the final step from iteration 3, by summing the shortest distance from data in Table 7. So the total cost is 2.510. To determine the next iteration, total cost from iteration 3 is compared with iteration 2, which is 2.510 > 2.416. Because the total cost of iteration 3 is greater than iteration 2, the iteration stops.

3.11 Results

Each medoid represents 1 group of crime level based on support and confidence. C1 represents high crime level, C2 represents medium crime level and C3 represents low crime level. The results of crime patterns grouping are shown in Tables 8, 9 and 10.

Table 8. High level crime patterns

No.	Rules	Support	Confidence
1	IF Embezzlement THEN Deception	0.54	0.81
2	IF Deception THEN Embezzlement	0.54	0.82
3	IF Fiduciary THEN Embezzlement	0.09	0.97
4	IF Information violation and electronic transaction THEN Deception	0.08	0.94

Table 9. Medium level crime patterns

No.	Rules	Support	Confidence
1	IF Embezzlement THEN Theft	0.02	0.03
2	IF Embezzlement THEN Document Forgery	0.02	0.03
3	IF Embezzlement THEN Fiduciary	0.09	0.13
4	IF Deception THEN Document Forgery	0.02	0.03
5	IF Deception THEN Information violation and electronic transaction	0.08	0.12
6	IF Persecution THEN Beating	0.02	0.23

Table 10. Low level crime patterns

No.	Rules	Support	Confidence
1	IF Theft THEN Embezzlement	0.02	0.29
2	IF Document Forgery THEN Embezzlement	0.02	0.41
3	IF Document Forgery THEN Deception	0.02	0.41
4	IF Persecution THEN Domestic Violence	0.03	0.3
5	IF Domestic Violence THEN Persecution	0.03	0.86
6	IF Beating THEN Persecution	0.02	0.69

Tables 8, 9 and 10, show that some crime patterns are classified as high and some others are classified as low. Information about high level crime can be used by the police to prevent potential crime in the society.

4 Conclusions

It can be concluded that K-medoids algorithm can be used to cluster crime patterns into three crime levels namely, 4 rules classified as high level crime, 6 rules classified as medium level crime and 8 rules classified as low level crime. Suggestions that can be given based on the results of this study are:

1) There is a need to compare some distance method for K-medoid algorithm. Thus, it can be known the most appropriate distance calculation method for K-medoid algorithm.
2) There is a need to apply weighting mechanism for each variable because not all variables have the same interests and priorities.

Acknowledgements

The author thanks *Polisi Resor Kota Yogyakarta* (POLRESTA Yogyakarta) who provided crime data in this research by hiding some sensitive variables regarding victims and criminals.

References

[1] J. E. Sahetapy and B. M. Reksodiputro, *Paradoks dalam Kriminologi*, Rajawali, Jakarta, (1982).

[2] E. H. S. Atmaja, *Visualisasi Aturan Asosiasi Berbasis Graph untuk Data Tindak Kejahatan*. Media Teknika, Vol. 12 No. 1 (2017) 46-57.

[3] P. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, Addison-Wesley, Boston, (2006).

[4] A. Singh, A. Yadav and A. Rana, *K-means with Three different Distance Metrics*, International Journal of Computer Applications, Vol. 67 No. 10 (2013) 13-17.

[5] E. H. S. Atmaja, *Pengelompokan Tingkat Kriminalitas di Kota Yogyakarta dengan Menggunakan Metode K-Means Clustering*, Seminar Nasional Riset dan Teknologi Terapan 2018 (RITEKTRA 2018), (2018).

[6] J. Han, *Data Mining: Concepts and Techniques*, *Second Edition*, Morgan Kaufmann, San Francisco, (2006).