Implementation of Data Mining using the Clustering Method (Case: Region of the Actors of Theft Crime by Province)

Frinto Tambunan

Universitas Potensi Utama, Medan, North Sumatra, Indonesia
frintoaja@gmail.com

Abstract

Theft is a behavior that causes harm to victims who are targeted and cause casualties. This study aims to classify areas of theft crimes based on provision by using data mining techniques. Data was obtained from the Indonesian statistical center (Badan Pusat Statistik) consisting of 34 provinces. The grouping technique used is K-Means. Clusters are divided into 3 namely: C1: areas with high crime rates of theft, C2: areas with crime rates of ordinary theft and C3: areas with low theft crime rates. Data processing is done using the help of RapidMiner software. The results of the k-means analysis obtained 17 provinces in Indonesia have the highest theft crime rate (C1), namely: Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Lampung, DKI Jakarta, West Java, Central Java, East Java, Banten, West Nusa Tenggara, East Nusa Tenggara, South Kalimantan, South Sulawesi and Papua. The results of the study concluded that more than 50% of regions in Indonesia still had high rates of crime of theft.

Keywords: *Data mining, clustering, K-Means, Theft.*

1. Introduction

Theft is an unlawful behavior that can cause casualties. Theft often occurs in the city center or shopping centers such as markets because the crowd center can trigger the perpetrators to commit acts of theft. Actors operate individually or in groups. This is triggered by the desire to fulfill economic needs that are increasingly high. In every area theft often occurs even every year the crime rate of theft is increasing. Another factor that causes theft is increased unemployment, economic, environmental and social crisis that is not good.

Based on these problems, researchers want to analyze areas with the highest theft crime rates by province using data mining techniques. There are several settlement techniques that can be done using data mining. Data mining is a method used for processing data, in order to find hidden images of processed data. Data that is processed with data mining methods then produces a new knowledge that comes from old data, the results of processing the data, can be used as information to determine future decisions [1]–[3]. Some of these data mining techniques (1) Classification, (2) Clustering, (3) Estimates and (4) Associations [4]–[7]. From these cases the researchers used the k-means clustering technique to classify data on theft crime cases based on provinces in Indonesia. Some of the advantages of k-means are that the method uses a simple principle, can be explained in non-statistics, the time needed to run it is relatively fast and very flexible and easily adaptable. This has also been proven by several previous researchers who solved the problem using the K-Means method. One of which is [6] with the title Implementation of Data Mining on Rice Imports by the Major Country of Origin Using Algorithm Using K-Means Clustering Method. The results of the study state that k-means can be analyzed and applied to the grouping of rice imports. The result is an assessment based on rice import index with 2 high-imported clusters of countries namely Vietnam and Thailand, 4 medium-level clusters of moderate import countries namely China, India, Pakistan and other 4 low-imported clusters countries namely USA, Taiwan, Singapore and Myanmar. The results of the research can be imported from the main country of origin. Based on this, the results of the research using the k-means method in the case of grouping the regions of theft crimes by province can answer the formulation of the problem that is analyzing and testing the k-means method in cases of theft crimes based on provinces in Indonesia

2. Research Methodology

2.1. K-Means Method

K-Means is a data analysis method or Data Mining method that performs the modeling process without supervision (unsupervised) and is one method of grouping data with system partitions. The purpose of the k-means method is to minimize objective functions that are set in the clustering process by minimizing variations between the data in a cluster and maximizing variations with the data in other clusters [1], [8].

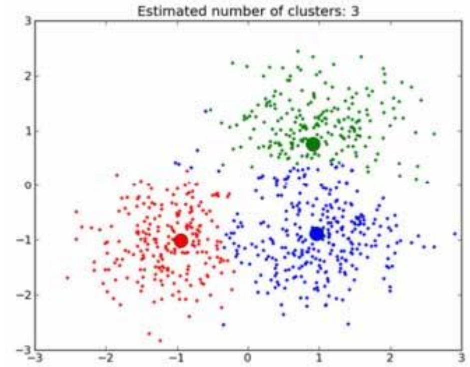

Figure 1. Example of k-means cluster results

2.2. Steps of the K-Means Method

Generally done with the basic algorithm as follows:

a) Determine the number of clusters
b) Allocate data into clusters randomly
c) Calculate the centroid / average of the data in each cluster
d) Allocate each data to the nearest centroid / average
e) Return to Step 3, if there is still data that moves clusters or if changes in the centroid value, there is something above the specified threshold value or if the change in the objective function used is above the specified threshold value [9], [10].

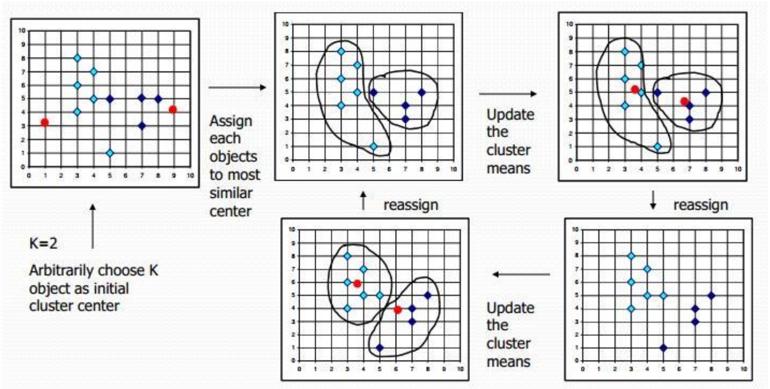
Assign each objects to most similar center
Update the cluster means
reassign
reassign
K=2
Arbitrarily choose K object as initial cluster center
Update the cluster means

Figure 2. Illustration of the K-Means Clustering Process

2.3. Data source

The source of research data was obtained from the Indonesian statistical center (https://www.bps.go.id/) regarding data on theft crimes based on provinces in Indonesia using data from 2008, 2011 and 2018. Then from the criminal journal st site which later managed from data in each province. The data used are data in 2008, 2011, and 2018 consisting of 34 provinces. The data will be taken an average value. The following research data:

Table 1. Research data

No	Province	Years			Average value
		2008	2011	2018	
1	Aceh	140	63	47	83
2	North Sumatra	186	93	141	140
3	West Sumatra	57	55	71	61
4	Riau	149	79	101	110
5	Jambi	66	57	76	66
6	South Sumatra	279	221	430	310
7	Bengkulu	38	33	40	37
8	Lampung	200	186	255	214
9	Kep. Bangka Belitung	24	16	15	18
10	Kep. Riau	21	13	17	17
11	DKI Jakarta	28	35	73	45
12	West Java	297	287	344	309
13	Central Java	132	146	176	151
14	DI Yogyakarta	7	28	20	18
15	East Java	269	290	419	326
16	Banten	78	54	49	60
17	Bali	8	13	23	15
18	West Nusa Tenggara	96	107	146	116
19	East Nusa Tenggara	61	54	67	61
20	West Kalimantan	39	33	37	36
21	Central Kalimantan	54	41	27	41
22	South Kalimantan	117	85	73	92
23	East Kalimantan	55	25	37	39
24	North Kalimantan	-	-	9	9
25	North Sulawesi	22	27	21	23
26	Central Sulawesi	21	24	23	23
27	South Sulawesi	69	54	47	57
28	Southeast Sulawesi	6	23	16	15
29	Gorontalo	3	4	3	3
30	West Sulawesi	23	7	7	12
31	Maluku	-	9	20	15
32	North Maluku	5	2	7	5
33	West Papua	3	13	13	10
34	Papua	64	153	113	110

3. Results and Discussion

3.1. Centroid Data

Determination of the starting point of this cluster is carried out by taking the highest value in the area of the high crime criminals (C1), the average value in the area of normal theft crimes (C2) and the smallest value in the area of low theft crime (C3). Next is the centroid of the data in the first iteration:

Table 2. Early Centroid Data

C1	*C2*	*C3*
289.75	85.2381	18.44792

3.2. Clustering Data

The first cluster iteration process is done by taking the closest distance from each data that is processed. From the average value of the area of the crime of theft in 2008, 2011, 2018 according to the province, grouping was found in the first iteration for the 3 clusters. The regional cluster of perpetrators of high theft crimes (C1), namely 4 provinces: South Sumatra, Lampung, West Java, East Java. Regional clusters of normal theft crimes (C2), namely 14 provinces: Aceh, North Sumatra, West Sumatra, Riau, Jambi, DKI Jakarta, Central Java, Banten, West Nusa Tenggara, East Nusa Tenggara, Central Kalimantan, South Kalimantan, South Sulawesi, Papua and the cluster of low theft (C3) crime areas, namely 16 Provinces: Bengkulu, Kep. Bangka Belitung, Kep. Riau, DI Yogyakarta, Bali, West Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua. Following is the Calculation of the First Center Cluster Iteration and Data Grouping The first iteration can be illustrated in the following table:

Table 3. the Calculation of the First Center Cluster Iteration

No	Province	Average value	C1	C2	C3	Shortest distance
1	Aceh	83	243	5	80	5
2	North Sumatra	140	186	62	137	62
3	West Sumatra	61	265	17	58	17
4	Riau	110	216	32	106	32
5	Jambi	66	260	12	63	12
6	South Sumatra	310	16	232	307	16
7	Bengkulu	37	289	41	34	34
8	Lampung	214	112	136	210	112
9	Kep. Bangka Belitung	18	308	60	15	15
10	Kep. Riau	17	309	61	14	14
11	DKI Jakarta	45	281	33	42	33
12	West Java	309	17	231	306	17
13	Central Java	151	175	73	148	73
14	DI Yogyakarta	18	308	60	15	15
15	East Java	326	0	248	323	0
16	Banten	60	266	18	57	18
17	Bali	15	311	63	11	11
18	West Nusa Tenggara	116	210	38	113	38
19	East Nusa Tenggara	61	265	17	57	17
20	West Kalimantan	36	290	42	33	33
21	Central Kalimantan	41	285	37	37	37
22	South Kalimantan	92	234	14	88	14
23	East Kalimantan	39	287	39	36	36
24	North Kalimantan	9	317	69	6	6
25	North Sulawesi	23	303	55	20	20
26	Central Sulawesi	23	303	55	19	19
27	South Sulawesi	57	269	21	53	21
28	Southeast Sulawesi	15	311	63	12	12
29	Gorontalo	3	323	75	0	0
30	West Sulawesi	12	314	66	9	9
31	Maluku	15	312	63	11	11
32	North Maluku	5	321	73	1	1
33	West Papua	10	316	68	6	6

No	Province	Average value	C1	C2	C3	Shortest distance
34	Papua	110	216	32	107	32

Table 4. Data Grouping The first iteration

No	Province	First Iteration Data Group		
		C1	C2	C3
1	Aceh		1	
2	North Sumatra		1	
3	West Sumatra		1	
4	Riau		1	
5	Jambi		1	
6	South Sumatra	1		
7	Bengkulu			1
8	Lampung	1		
9	Kep. Bangka Belitung			1
10	Kep. Riau			1
11	DKI Jakarta		1	
12	West Java	1		
13	Central Java		1	
14	DI Yogyakarta			1
15	East Java	1		
16	Banten		1	
17	Bali			1
18	West Nusa Tenggara		1	
19	East Nusa Tenggara		1	
20	West Kalimantan			1
21	Central Kalimantan		1	
22	South Kalimantan		1	
23	East Kalimantan			1
24	North Kalimantan			1
25	North Sulawesi			1
26	Central Sulawesi			1
27	South Sulawesi		1	
28	Southeast Sulawesi			1
29	Gorontalo			1
30	West Sulawesi			1
31	Maluku			1
32	North Maluku			1
33	West Papua			1
34	Papua		1	

Based on table 4, the process continues until the last iteration process is the same as the previous iteration. Determination of centroid values will continue to change according to the iteration. The second iteration process until the next will use the help of RapdMiner software. By using RapidMiner software, the iteration process ends in the eighth iteration where the final result of the seventh iteration is the same as the eighth iteration. The following are the last Iteration Calculations and Grouping The latest data on theft crime cases by province as shown in the following table:

Table 5. the Calculation of the Last Cluster Iteration

No	Province	Average value	C1	C2	C3	Shortest distance
1	Aceh	83	10	80	72	10
2	North Sumatra	140	66	137	129	66
3	West Sumatra	61	13	58	50	13
4	Riau	110	36	107	98	36

No	Province	Average value	C1	C2	C3	Shortest distance
5	Jambi	66	7	63	55	7
6	South Sumatra	310	236	307	299	236
7	Bengkulu	37	37	34	26	26
8	Lampung	214	140	211	202	140
9	Kep. Bangka Belitung	18	55	15	7	7
10	Kep. Riau	17	57	14	6	6
11	DKI Jakarta	45	28	42	34	28
12	West Java	309	236	306	298	236
13	Central Java	151	78	148	140	78
14	DI Yogyakarta	18	55	15	7	7
15	East Java	326	252	323	315	252
16	Banten	60	13	57	49	13
17	Bali	15	59	12	3	3
18	West Nusa Tenggara	116	43	113	105	43
19	East Nusa Tenggara	61	13	58	49	13
20	West Kalimantan	36	37	33	25	25
21	Central Kalimantan	41	33	38	29	29
22	South Kalimantan	92	18	89	80	18
23	East Kalimantan	39	35	36	28	28
24	North Kalimantan	9	65	6	2	2
25	North Sulawesi	23	50	20	12	12
26	Central Sulawesi	23	51	20	11	11
27	South Sulawesi	57	17	54	45	17
28	Southeast Sulawesi	15	59	12	4	4
29	Gorontalo	3	70	0	8	0
30	West Sulawesi	12	61	9	1	1
31	Maluku	15	59	11	3	3
32	North Maluku	5	69	2	7	2
33	West Papua	10	64	7	2	2
34	Papua	110	36	107	99	36

Table 6. Data Grouping The last iteration

No	Province	Last Iteration Data Group		
		C1	C2	C3
1	Aceh	1		
2	North Sumatra	1		
3	West Sumatra	1		
4	Riau	1		
5	Jambi	1		
6	South Sumatra	1		
7	Bengkulu			1
8	Lampung	1		
9	Kep. Bangka Belitung			1
10	Kep. Riau			1
11	DKI Jakarta	1		
12	West Java	1		
13	Central Java	1		
14	DI Yogyakarta			1
15	East Java	1		
16	Banten	1		
17	Bali			1
18	West Nusa Tenggara	1		
19	East Nusa Tenggara	1		
20	West Kalimantan	1		

No	Province	Last Iteration Data Group		
		C1	C2	C3
21	Central Kalimantan			1
22	South Kalimantan	1		
23	East Kalimantan			1
24	North Kalimantan			1
25	North Sulawesi			1
26	Central Sulawesi			1
27	South Sulawesi	1		
28	Southeast Sulawesi			1
29	Gorontalo		1	
30	West Sulawesi			1
31	Maluku			1
32	North Maluku		1	
33	West Papua			1
34	Papua	1		

Based on table 6, it can be explained that the results of the final grouping on the area of perpetrators of high theft crimes (C1) are 17 provinces: Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Lampung, DKI Jakarta, West Java, Central Java, Java Timur, Banten, West Nusa Tenggara, East Nusa Tenggara, South Kalimantan, South Sulawesi, Papua. In the area of normal theft crimes (C2) are 2 provinces: Gorontalo, North Maluku and areas of low crime crimes (C3) are 15 Provinces: Bengkulu, Kep. Bangka Belitung, Kep. Riau, DI Yogyakarta, Bali, West Kalimantan, Central Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, Southeast Sulawesi, West Sulawesi, Maluku, West Papua.

4. Conclusion

Based on the results of the study it can be concluded that the K-means method can be analyzed and applied to theft crime cases by province where the results of grouping are obtained more than 50% of provinces in Indonesia still have a high crime theft rate. The provinces are Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Lampung, DKI Jakarta, West Java, Central Java, East Java, Banten, West Nusa Tenggara, East Nusa Tenggara, South Kalimantan, South Sulawesi, Papua . This is an input for the government, especially those who are authorized to become information material in reducing crime rates, especially theft in every province in Indonesia.

References

[1] J. C. Rubio-romero *et al.*, “Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Promosi,” *Ind. Mark. Manag.*, vol. 1, no. 1, pp. 1–9, 2014.

[2] A. P. Windarto, “Penerapan Data Mining Pada Ekspor Buah-Buahan Menurut Negara Tujuan Menggunakan K-Means Clustering,” *Techno.COM*, vol. 16, no. 4, pp. 348–357, 2017.

[3] M. ko. Dicky Nofriansyah, S.Kom., *Konsep Data Mining Vs Sistem Pendukung Keputusan.pdf*, Ed.1, Cet. Yogyakarta: Deepublish, 2014.

[4] E. Elisa, “Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti,” *J. Online Inform.*, vol. 2, no. 1, p. 36, 2017.

[5] K. Singh, D. Malik, and N. Sharma, “Evolving limitations in K-means algorithm in data mining and their removal,” *IJCEM Int. J. Comput. Eng. Manag. ISSN*, vol. 12, no. April, pp. 2230–7893, 2011.

[6] A. P. Windarto, "Implementation of Data Mining on Rice Imports by Major Country of Origin Using Algorithm Using K-Means Clustering Method," *Int. J. Artif. Intell. Res.*, vol. 1, no. 2, pp. 26–33, 2017.

[7] S. Sudirman, A. P. Windarto, and A. Wanto, "Data Mining Tools | RapidMiner : K-Means Method on Clustering of Rice Crops by Province as Efforts to Stabilize Food Crops In Indonesia," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 420, no. 12089, pp. 1–8, 2018.

[8] N. Kaur, J. K. Sahiwal, N. Kaur, and P.- Punjab, "Efficient K-Means Clustering Algorithm Using Ranking Method," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 1, no. 3, pp. 85–91, 2012.

[9] K. Fatmawati and A. P. Windarto, "Data Mining: Penerapan Rapidminer Dengan K-Means Cluster Pada Daerah Terjangkit Demam Berdarah Dengue (Dbd) Berdasarkan Provinsi," *Comput. Eng. Sci. Syst. J.*, vol. 3, no. 2, p. 173, 2018.

[10] B. Supriyadi, A. P. Windarto, T. Soemartono, and Mungad, "Classification of natural disaster prone areas in Indonesia using K-means," *Int. J. Grid Distrib. Comput.*, vol. 11, no. 8, pp. 87–98, 2018.

Authors

1st Author
Frinto Tambunan
Universitas Potensi Utama, Medan, North Sumatra