Unbalanced Data Clustering with K-Means and Euclidean Distance Algorithm Approach Case Study Population and Refugee Data

NM Faizah[1], Surohman [2], L Fabrianto [2], and Hendra[2] R Prasetyo[2]

[1]Information Systems Department, Tama Jagakarsa University, Jakarta, Indonesia.
[2]Master of Computer Science-Postgraduate Program, STMIK Nusa Mandiri, Jakarta, Indonesia

*novianti@jagakarsa.ac.id

Abstract. There is a lot of data that does not have a pattern and unbalanced that is difficult to classify, such as the total population of each country in the world is very varied especially when compared with the number of refugees from each of these countries, China and India numbered more than 2 billion people but the number of refugees is only 0, 01%, while Syria around 70% of the 18 million more residents are refugees. By using the K-Means algorithm, we can group countries that have similar characteristics of population and number of refugees, the average percentage of refugees to the population in each cluster is the character of the cluster. The methodology used in this study are: measures the distance of the data using the Euclidean distance formula, runs the K-Means algorithm, calculate the percentage value of each cluster and find conclusions from the characteristics of the clusters formed. We found how machine learning made a pattern of data without political and social issues, the result is K-Means describing machine learning can grouped every country in cluster that make sense.

I. Introduction

The total population of the world is approximately 7.5 billion that spread unevenly in every country in the world. There are 2 countries with very large populations, China and India, under these two countries there are several countries with average populations are only 15% of India population.

The number of refugees in worldwide not evenly distributed in quantity and distribution, data used is dataset from Humanitarian Data Exchange (HDX) UNHCR Global Trends: Forced Displacement in 2017 Data[6]. And population data for each country throughout the world is obtained from Wikipedia, which contains data for each country in an update, even real-time [7].

By using the K-Means algorithm, we create four groups countries that have similar characteristics of population and number of refugees, the average percentage of refugees to the population in each cluster is

the character of the cluster, the cluster to be formed is four classes, the first class is the lowest percentage, the second class the middle percentage, the third class the biggest percentage and the fourth class formed because the population of China and India is very large which will form their own cluster so it does not affect the grouping of the other three clusters.

The purpose of this study is to look for the results of clustering on unbalanced data using the K-means and Euclidean Distance algorithms with examples of population data and refugee data in each country and averaged to see the percentage results as a characteristic of each cluster. Function of clustering is the processes of making an imbalanced data into group that have similarities. Clustering can interpreted a collection of unsupervised data mining methods, which aim to sort out an entire data set into several smaller sizes [1].

2. Method

2.1 K-Means clustering.

This method is one of the oldest and very popular among practitioners because of the ease of implementation and the speed of the process [2]. This algorithm makes groups or clusters based on their attribute values into a number of K clusters, algorithm flowchart can be seen in Figure 1 [3], and the expected results of clustering as illustrated in Figure 2.

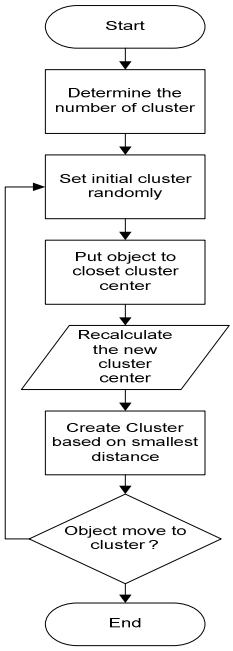

Figure 1. Flowchart of K-Means clustering algorithm

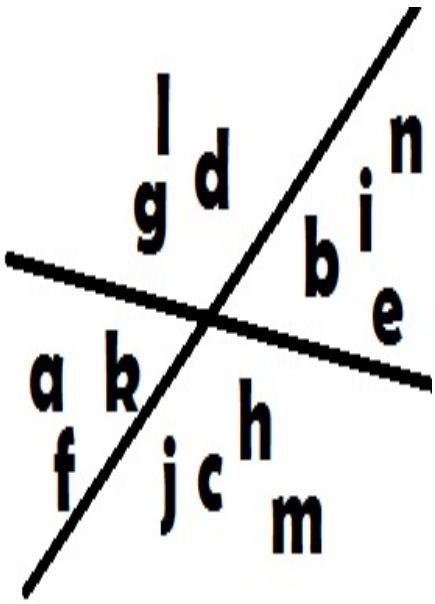

Figure 2. Clustering Result

The steps in running clustering use K-Means, as follows [3]:

1. It starts with determining the number k as a cluster that you want to form.
2. Determine a number of data as the center of the cluster, can be taken randomly
3. Calculate the distance of each input data to each cluster center (centroid) using the Euclidean distance formula (Euclidean Distance) to find the closest distance from each data to the centroid.
4. Grouping each data based on its proximity to the centroid (the shortest distance).

5. Update the value of the centroid by averaging the cluster in question.
6. Perform the 2nd step iteration until step 5 if the centroid value keeps changing.

Step 6 is finished when the centroid value in the last iteration does not change and is used as a parameter for data classification.

2.2 *Euclidean Distance*.

Formula to measure the distance between one object and another there are many ways, such as: Manhattan Distance, Chebyshev Distance, Canberra Distance, Hamming Distance and Euclidean Distance. Each form has its own formula [4]. Euclidean Distance is the most popular for calculating the distance between two objects, calculating the square root of the difference in coordinates between two objects, the formula is as follows [5]:

$$d_{ij} = \sqrt{\sum_{k-1}^{n} \quad (x_{ik} - x_{jk})^2} \tag{1}$$

The Euclidean Distance formula is available on the Rapid miner program is one algorithm that is quite commonly used in various applications.

3. Result and Discussion

The attributes used in this paper are the population in each country (Population) and the number of refugee in each country (Refugee), refugee data obtained from the HDX dataset repository in 2017[6] and population data obtained from Wikipedia[7]. The steps used in this research are combine population and refugee data as seen in Table 1., run the K-means algorithm and draw conclusions from the clustering formed. The process of merging population and refugee data is done by matching the names of each country, in this case there are 207 countries.

Table 1. Sample of combine dataset from HDX repository in 2017 and Wikipedia

No	Country	Populations	Refugees
1	Afghanistan	31,575,018	5,336,582
2	Albania	2,870,324	30,941
3	Algeria	42,545,964	10,482
205	Yemen	28,915,284	2,187,305
206	Zambia	16,887,720	448
207	Zimbabwe	14,848,905	39,923

Populations and refugees data will be through data processing with a clustering process that uses the K-means algorithm to get the results of the desired grouping.

The steps of the K-Means algorithm in this case are as follows:

1. Determine the number of cluster, in this case the author's uses four clusters.
2. Determine the initial center of the cluster (centroid) randomly, Table 2, is initial centroid.

Table 2. Initial centroid

No	Country	Populations	Refugees
4	American Samoa	56,700	3
54	Denmark	5,806,015	24
153	Russian Federation	146,877,088	100,286

207	Zimbabwe	14,848,905	39,923

3. Calculate the distance from the centroid to the point of each object using the Euclidian Distance formula

[7] $D_{(i,f)} = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \ ... + (x_{ki} - x_{kj})^2}$

The calculation of the first iteration produces a new centroid and clustering
4. Grouping each data based on proximity to the centroid.
5. Update the centroid value and recalculate if the centroid and cluster values keep changing. The iteration is stopped if the number of cluster and centroid value is stable. Table 3, shows all iterations.

Table 3. Recapitulation of all iterations

Iterations	Centroid		C1	Centroid		C2	Centroid		C3	Centroid		C4
	Population	Refugees		Population	Refugees		Population	Refugees		Population	Refugees	
1st Iteration	42,550,943	1,108,075	23	246,894,281	550,036	22	24,109,090	973,594	17	4,852,469	98,516	145
2nd Iteration	64,622,137	896,445	31	471,811,656	437,133	9	22,783,974	787,659	33	3,898,105	76,611	134
3rd Iteration	101,381,638	757,563	28	1,020,986,333	120,569	3	26,186,130	860,554	42	3,898,105	76,611	134
4th Iteration	134,844,709	587,081	20	1,367,390,000	179,590	2	31,098,784	911,521	50	3,979,222	79,312	135
5th Iteration	160,025,882	817,695	14	1,367,390,000	179,590	2	39,003,410	881,984	48	4,672,044	99,881	143
6th Iteration	179,291,024	517,723	11	1,367,390,000	179,590	2	47,014,003	814,580	42	5,541,046	186,393	152
7th Iteration	195,322,598	407,992	9	1,367,390,000	179,590	2	52,910,842	919,673	39	6,163,071	182,755	157
8th Iteration	204,745,613	446,896	8	1,367,390,000	179,590	2	57,422,547	935,743	36	6,731,976	192,480	161
9th Iteration	215,932,129	510,716	7	1,367,390,000	179,590	2	61,783,455	818,027	34	7,176,010	225,175	164
10th Iteration	215,932,129	510,716	7	1,367,390,000	179,590	2	65,662,419	926,244	30	7,783,517	219,903	168
11th Iteration	215,932,129	510,716	7	1,367,390,000	179,590	2	66,725,089	957,839	29	7,943,644	218,646	169
12th Iteration	215,932,129	510,716	7	1,367,390,000	179,590	2	67,776,531	992,041	28	8,116,239	217,345	170
13th Iteration	215,932,129	510,716	7	1,367,390,000	179,590	2	67,776,531	992,041	28	8,116,239	217,345	170

6. The iteration process is stopped and the clustering result of data is available. Table 4, is clustering obtained.

Table 4. Clustering results

Cluster	Number of		
	Countries	Population	Refugees
Class 1	7	1,511,524,905	3,575,013
Class 2	2	2,734,780,000	359,180
Class 3	28	1,879,742,878	27,777,161

Class 4	170	1,379,760,579	36,514,030

To see the characteristics of each cluster formed, we make a percentage ratio of refugees to population in each cluster, so the description of the characteristics of each cluster is easy to draw conclusions, as seen at Table 5.

Table 5. Percentage of refugees

Cluster	Number of			Refugees
	Countries	Population	Refugees	%
Class 1	7	1,511,524,905	3,575,013	0.24
Class 2	2	2,734,780,000	359,180	0.01
Class 3	28	1,879,742,878	27,777,161	1.48
Class 4	170	1,379,760,579	36,514,030	2.65

The cluster position of each country during the K-Means algorithm process is not fixed and when the process stops the cluster position of each country is fixed. Table 6. Sample of fixed cluster position.

Table 6. Sample of cluster position

No	Country	Populations	Refugees	Cluster member
1	Afghanistan	31,575,018	5,336,582	4
25	Brazil	209,850,000	7,658	1
38	China	1,395,190,000	310,616	2
41	Colombia	50,038,400	7,901,909	3
42	Comoros	850,688	933	3
43	Congo, Republic of	5,399,895	134,327	3
87	Indonesia	265,015,300	17,885	1
171	Somalia	15,181,925	3,203,155	4
181	Syrian Arab Rep.	18,284,407	13,288,372	4

The following is description of the characteristics of each cluster:

- Class 1: Describe a collection of countries with a large population (150 million - 300 million people) and around 0.24% of the populations are refugees, countries included in the Class 1 include: Indonesia, Brazil, Bangladesh, United States, Russia and Pakistan.
- Class 2: A group of countries that has a very large population, China and India have the percentage of refugees is very little 0.01%.
- Class 3: Consists of 28 countries with a large number of average populations and the percentage of refugees almost 1.5%, several countries in this class such as: Colombia, Congo, Ethiopia, Iraq, Sudan, Myanmar and Ukraine, become countries as contributors quite a lot of refugees.

- Class 4: This is a group of countries with the least number of populations but the highest percentage of refugees namely 2.65%, countries like Afghanistan, Syria, Somalia and South Sudan are countries with a very dominant number of refugees.

4. Conclusions

The K-Means algorithm is very helpful to form several clusters based on characteristics of unbalanced data, with the K-Means algorithm we can create labels from a data set that we find difficult to make a reference to separate them. Result of this study is prove that machine learning using K-means algorithm and Euclidean Distance formula can make a pattern of data ignore political and social issues, result each member of cluster are make sense according real situation.

There are several formulas for calculating distances, such as: Manhattan Distance, Chebyshev Distance, Canberra Distance, and Hamming Distance, available on Rapid miner is Euclidean Distance. We can use a formula that suits the needs of the data to be clustered.

The next study authors or other researcher can approaching political and social issues with more detailing to compare to machine learning clustering with or without K-Means algorithm.

References

[1] Lindawati 2008 Data Mining Dengan Teknik Clustering Dalam Pengklasifikasian Data Mahasiswa Studi Kasus Prediksi Lama Studi Mahasiswa Universitas Bina Nusantara, *Seminar Nasional Informatika (semnasIF 2008), ISSN :1979-2328*.

[2] Suyanto 2018 *Machine Learning, Tingkat Dasar Dan Lanjut* Bandung : Informatika

[3] Zeyad Safaa Younus, Dzulkifli Mohamad, Tanzila Saba, Mohammed Hazim Alkawaz 2014 Content-based image retrieval using PSO and k-means clustering algorithm in *Arab J Geosci*

[4] Murti, Darlis Heru, Nanik Suciati, and Daru Jani Nanjaya. "Clustering data non-numerik dengan pendekatan algoritma k-means dan hamming distance studi kasus biro jodoh." *JUTI: Jurnal Ilmiah Teknologi Informasi* 4.1 (2005): 46-53.

[5] Advanced Projects R&D 2005 *Euclidean Distance raw, normalized, and double-scaled coefficients.* Available: https://www.pbarrett.net/techpapers/euclid.pdf [accessed on 4 December 2018]

[6] Humanitarian Data Exchange (HDX) UNHCR Global Trends: Forced Displacement in 2017 Data . Available: https://data.humdata.org/dataset/unhcr-global-trends-forced-displacement-in-2017 [accessed on 5 January 2019]

[7] Real time Populations Data. Available : https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population [accessed on 5 February 2019]