PERFORMANCE ANALYSIS OF ENTROPY METHODS ON K MEANS IN CLUSTERING PROCESS

Mhd. Dicky Syahputra Lubis¹, Herman Mawengkang², Saib Suwilo³

¹Departement of Computer Science,
²Department of Mathematics Science,
University Of Sumatera Utara
Medan, Indonesia, 20155

Email: ¹dickylubis91@gmail.com

Abstract. K Means is a non-hierarchical data clustering method that attempts to partition existing data into one or more clusters / groups. This method partitions the data into clusters / groups so that data that have the same characteristics are grouped into the same cluster and data that have different characteristics are grouped into other groups. The purpose of this data clustering is to minimize the objective function set in the clustering process, which generally attempts to minimize variation within a cluster and maximize the variation between clusters. However, the main disadvantage of this method is that the number k is often not known before. Furthermore, a randomly chosen starting point may cause two points to approach the distance to be determined as two centroids. Therefore, for the determination of the starting point in K Means used entropy method where this method is a method that can be used to determine a weight and take a decision from a set of alternatives. Entropy is able to investigate the harmony in discrimination among a multitude of data sets. Using Entropy criteria with the highest value variations will get the highest weight. Given this entropy method can help K Means work process in determining the starting point which is usually determined at random. Thus the process of clustering on K Means can be more quickly known by helping the entropy method where the iteration process is faster than the K Means Standard process. Where the postoperative patient dataset of the UCI Repository Machine Learning used and using only 12 data as an example of its calculations is obtained by entropy method only with 2 times iteration can get the desired end result.

Keywords: K Means, Entropy, Clustering, Data Mining, Weight

1. Introduction

The purpose of this data clustering is to minimize the objective function set in the clustering process, which generally attempts to minimize variation within a cluster and maximize the variation between clusters. (YudiAgusta 2007). Where, the difference between the nearest point and the farthest point in the large amount of data becomes rather weak. In, the method based on the vector angle is suggested to know the angle between the two vectors of the data object. This step seems much more sensitive to high dimension data sets based on distance. (Shu-yin Xia 2015). In two dimensions, the length of the straight line connection becomes two points x and y. In n dimensions, one problem with this method is that the Euclidean distance is sensitive to large values; In other words, sensitive to outliers. In
addition, it will lose negative correlation because they will provide a great distance. (Bernard Chen et al 2005). Entropy method is a method that can be used to determine a weight and make a decision from a set of alternatives. Entropy is able to investigate the harmony in discrimination among a multitude of data sets. By using Entropy criteria with the highest value variations will get the highest weight (Abbas, 2004). Another study mentions that Entropy can take decisions in the selection of subcontracted gloves production (Jamila, 2012). In that study, weighting each criteria and determining which subcontracts are the best choice. Where the selection of the entropy method is to determine the center of the initial centroid to obtain the weight of the value in making Clustering on the K-Means algorithm.

1.2 K means
K-Means is a non-hierarchical data clustering method that attempts to partition existing data into one or more clusters. This method partitions the data into clusters / groups so that data that have the same characteristics are grouped into the same cluster and data that have different characteristics are grouped into other groups. The purpose of this data clustering is to minimize the objective function set in the clustering process, which generally attempts to minimize variation within a cluster and maximize the variation between clusters. (YudiAgusta 2007)

Data clustering using K-Means method is generally done with the following basic algorithm:
1. Determine the number of clusters
2. Allocate data into the cluster randomly
3. Calculate the centroid / average of the data in each cluster
4. Allocate each data to the nearest centroid / average
5. Return to Step 3, if there is still data that moves the cluster or if the centroid value changes, some are above the specified threshold value or if the value changes on the objective function used above the specified threshold value

1.2.1 Distance Space To Calculate Distance Between Data and Centroid
Some distance space has been implemented in calculating the distance between data and centroid. For Euclidean distance space, the distance between two points is calculated using the following formula:

\[ D_{i2} (x_2, x_1) = \|x_2 - x_1\|_2 = \sqrt{\sum_{j=1}^{P} (x_{2j} - x_{1j})^2} \]

Where: P: Dimensions of data

Euclidean is often used because the distance calculation in distance space is the shortest distance that can be obtained between two points taken into account, whereas Manhattan is often used because of its ability to detect special circumstances such as the existence of outliers better. (YudiAgusta 2007)

1.3 Entropy
Entropy method is a method that can be used to determine a weight. Entropy is able to investigate harmony in discrimination between data sets. A set of alternative value data on certain criteria is described in the Decision Matrix (DM). Using the Entropy method, the criteria with the highest value variations will get the highest weight. Thus, the Entropy method can calculate the maximum likelihood (maximum Entropy) for each single data in a set of entities that have different possibilities. Specifically, Entropy is also able to adapt to a set of multiple plural data that have variations varying from one criterion to another. The weighting steps using the Entropy method are as follows (Tiyaswiyoso, 2012)

1.3.1 Normalization of Data Criteria
The normalization formula is as follows:

\[ d_i^k = \frac{x_i^k}{x_{i\text{max}}} \quad d_i = d_i^1, \ldots, d_i^m \]

\[ D_i = \sum_{k=1}^{m} d_i^k \quad i = 1, 2, \ldots, n \]
2. Entropy Calculation

The next step is the measurement of Entropy for each of the i-th criteria.

The formula is:

\[ e_{\text{max}} = \ln m, \]  
then calculate

\[ K = \frac{1}{e_{\text{max}}} \]  

\[ e(d_i) = -K \sum_{k=1}^{m} \frac{d_k^i}{D_i} \ln \left( \frac{d_k^i}{D_i} \right), \quad K > 0 \]

Where:
- \( e_{\text{max}} \) = Maximum entropy
- \( K \) = Entropy constant
- \( e(d_i) \) = Entropy for every attribute / criterion i

After obtaining \( e(d_i) \) for each criterion, then can be determined total Entropy for each criterion, the formula is:

\[ E = \sum_{i=1}^{n} e(d_i), \quad n \text{ is the number of criteria} \]

3. Calculation of Entropy Weight

The next step is to calculate the weights by using the following formula:

\[ \lambda_1 = n - E \left[ 1 - e(d_i) \right], \quad 0 \leq \lambda_1 \leq 1 \]

\[ \sum_{i=1}^{n} \lambda_1 = \pm 1 \]

Where:
- \( \lambda_1 \) = Temporary Entropy weight
- \( n \) = number of attributes / criteria

2. Research methodology

In the process of determining the initial weight value for centroid center with Entropy method is done in two ways:

a. Entropy calculation for each of the I-criteria

\[ e_{\text{max}} = \ln m = \ln 12 = -0.402429 \]

\[ K = \frac{1}{0.4024} = 2.48242 \]

\[ e(d_i) = -K \sum_{k=1}^{m} \frac{d_k^i}{D_i} \ln \left( \frac{d_k^i}{D_i} \right), \quad K > 0 \]

\[ e(d_1) = 0.39516 \]  
\[ e(d_2) = 0.98521 \]

with:

\[ E = \sum_{i=1}^{n} e(d_i), \quad \text{then:} \quad E = 1.38037 \]
b. Calculation of Entropy Weight for each of the i-th criteria

$$\lambda_i = \frac{1}{n-E}[1-e(d_i)], \quad 0 \leq \lambda_i \leq 1$$

Then the result:

$$\lambda_1 = 0.97613, \quad \lambda_2 = 0.02386$$

The value of the data is multiplied by the final entropy weight. Then the value is searched for the highest, middle and lowest values to serve as the center of the initial cluster.

| Table 1. first centre centroid |
|--------------------------------|
| Patient temperature | Patient comfort |
| C1 (Max) | 43.92588614 | 0.358037953 |
| C2 (Median) | 36.11683972 | 0.238691969 |
| C3 (Min) | 35.14070891 | 0.119345984 |

5. Result and Discussion

Following calculations on its iteration:

Where on iteration 1

Input set ke-1 : (36.116, 0.358)
Centre cluster 1: (43.925, 0.358) Centre cluster 2: (36.116, 0.238)
Centre Cluster 3: (35.140, 0.119)

The iteration:

$$C_1=\sqrt{(36.116 - 43.925)^2 + (0.358 - 0.358)^2} = \sqrt{(-7.809)^2 + (0)^2} = 7.809$$

$$C_2=\sqrt{(36.116 - 36.116)^2 + (0.358 - 0.238)^2} = \sqrt{(0)^2 + (0.12)^2} = 0.12$$

$$C_3=\sqrt{(36.116 - 35.140)^2 + (0.358 - 0.119)^2} = \sqrt{(0.976)^2 + (0.239)^2} = 1.005$$

And so on so that the results obtained are shown in the table 2

| Table 2 result of first iteration |
|----------------------------------|
| Patient Number | Temp of Patient | Comfortable of Patient | C1 | C2 | C3 | Shortest Distance |
|----------------|-----------------|------------------------|----|----|----|------------------|
| 1              | 36.11683972     | 0.358037953            | 7.809046425 | 0.119346 | 1.004891 | 0.119345984 |
| 2              | 35.14070891     | 0.238691969            | 8.785987844 | 0.976131 | 0.119346 | 0.119345984 |
| 3              | 37.09297052     | 0.238691969            | 6.833957811 | 0.976131 | 1.955906 | 0.976130803 |
| 4              | 36.11683972     | 0.358037953            | 7.809046425 | 0.119346 | 1.004891 | 0.119345984 |
| 5              | 36.11683972     | 0.238691969            | 7.809958357 | 0 | 0.9834 | 0 |
| 6              | 43.92588614     | 0.358037953            | 8.788419243 | 0.9834 | 0 | 0 |
| 7              | 35.14070891     | 0.119345984            | 8.788419243 | 0.9834 | 0 | 0 |
| 8              | 41.97362453     | 0.238691969            | 1.955906144 | 5.856785 | 6.833958 | 1.955906144 |
| 9              | 35.14070891     | 0.238691969            | 8.785987844 | 0.976131 | 0.119346 | 0.119345984 |
| 10             | 35.14070891     | 0.238691969            | 8.785987844 | 0.976131 | 0.119346 | 0.119345984 |
| 11             | 36.11683972     | 0.358037953            | 7.809046425 | 0.119346 | 1.004891 | 0.119345984 |
| 12             | 36.11683972     | 0.238691969            | 7.809958357 | 0 | 0.9834 | 0 |
Based on the above clustering obtained 2 data on cluster 1, 6 data on cluster 2 and 4 data on cluster 3 then resumed in the iteration process-2 which its way almost the same as the first iteration that only distinguish it is on the determination of its center C Where in this 2nd iteration the center C is determined by searching for the average of each previous cluster. In the 2nd iteration the new cluster center is determined based on the average of each value on each new cluster where in C1 there is 1 member, C2 is 6 members and C3 there are 4 members, then obtained a new cluster center. Where in the 2nd iteration process gets the desired final result because the clusterization result on the iteration to 1 and the 2nd iteration remain or not change as the calculation process stops here.

Table 3 comparison methods K-Means

| Patient Number | K Means With Entropy | K Means Standard |
|----------------|----------------------|-----------------|
|                | Iteration | C1 | C2 | C3 | Iteration | C1 | C2 | C3 |
| 1.             | 2         | 1  | 3  | 1  | 3         | 1  | 1  | 1  |
| 2.             | 2         | 1  | 3  | 1  | 3         | 1  | 1  | 1  |
| 3.             | 2         | 1  | 3  | 1  | 3         | 1  | 1  | 1  |
| 4.             | 2         | 1  | 3  | 1  | 3         | 1  | 1  | 1  |
| 5.             | 2         | 1  | 3  | 1  | 3         | 1  | 1  | 1  |
| 6.             | 2         | 1  | 3  | 1  | 3         | 1  | 1  | 1  |
| 7.             | 2         | 1  | 3  | 1  | 3         | 1  | 1  | 1  |
| 8.             | 2         | 1  | 3  | 1  | 3         | 1  | 1  | 1  |
| 9.             | 2         | 1  | 3  | 1  | 3         | 1  | 1  | 1  |
| 10.            | 2         | 1  | 3  | 1  | 3         | 1  | 1  | 1  |
| 11.            | 2         | 1  | 3  | 1  | 3         | 1  | 1  | 1  |
| 12.            | 2         | 1  | 3  | 1  | 3         | 1  | 1  | 1  |

From the results of research that has been done that Entropy method can be used as a determinant of initial centroid value in K Means. Centroid value obtained from Entropy method shows good performance in studying data during training process. Entropy method yields centroid value that can represent the whole data of rice productivity with a little amount of training data that is only 12 data although with number of iteration 2 times. But the learning process is getting better too. Thus the initial centroid value also yields a better level of effectiveness when compared to standard K Means which takes a random value to determine the initial centroid value. This is due to the nature of the Entropy method of calculation capable of investigating the degree of harmony in a set of data and being able to adapt to a set of plural-plural data that has variations in values that differ from one data to another. So the initial centroid value obtained can learn and represent data from each class quickly. Therefore, the initial centroid value is very influential from the learning process data and the level of effectiveness gained. But overall testing, the initial centroid value generated by the Entropy method has a success rate in clustering data compared to other ways. So it can be concluded that the value of early centroid is very influential from the process of learning data and the level of accuracy obtained. Then the determination of initial centroid value on K Means by using Entropy method can be said better when compared with the determination of the initial centroid value by K Means standard.

6. References

[1] Abbas, A.R. 2004. *Entropy Method for Adaptive Utility Elicitation*. *IEEE Transaction on System, Man, and Cybernetics* 34(2): 169-178.

[2] Agusta, Yudi PhD, “K-Means – Penerapan, PermasalahandanMetodeTerkait”, JurnalSistemdanInformatika Vol. 3 (Pebruari 2007), 47-60

[3] Chen , Bernard, Phang C. Tai, R. Harrison and Yi Pan, “Novel Hybrid Hierarchical-K-means Clustering Method (H-K-means) for Microarray Analysis”
Department of Computer Science, Department of Biology, Georgia State University, Atlanta, USA 2005

[4] Chen, Bernard, Phang C. Tai, R. Harrison and Yi Pan, “Novel Hybrid Hierarchical-K-means Clustering Method (H-K-means) for Microarray Analysis” Department of Computer Science, Department of Biology, Georgia State University, Atlanta, USA 2005

[5] Jamila. 2012. SistemPendukungKeputusanPemilihanSubkontrakProduksiSarungTanganMenggunakanMetode Entropy dan TOPSIS. Seminar Nasional Informatika UPN Yogyakarta. I(2).

[6] Nurul C., Wiharto, & Umi S. (2012). Pengaruh Normalisasi Data pada Jaringan Syaraf Tiruan Backpropagation Gradient Descent Adaptive Gain (BPGDAG) untuk Klasifikasi. JURNAL ITSMART. 2301-7201.

[7] Xia, Shu-yin, Zhong-yang Xiong, Yue-guo Luo, Wei-Xu, “Effectiveness of the Euclidean distance in high dimensional spaces”, College of Computer Science, 2015