Attainment of K-Means Algorithm using Hellinger distance

Stancu Ana-Maria Ramona
"Dimitrie Cantemir" Christian University
ana_maria_ramona@yahoo.com
Cristescu Marian Pompiliu
"Lucian Blaga" University of Sibiu
marian.cristescu@ulbsibiu.ro
Stoica Liviu Constantin
Academy of Economic Studies, Bucharest
stoica.liviu.constantin@gmail.com

Abstract

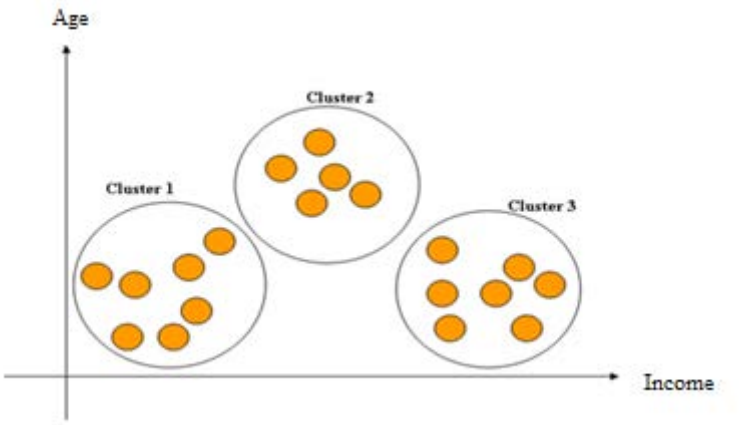
In this article in the first part I will begin with an introduction to unsupervised learning methods, focusing on the K-Means clustering algorithm, which is achieved with the help of the Euclidian distance. In the second part we modified the K-Means algorithm, that is, it was achieved with the help of the Hellinger distance, after which the clustering time was compared and a parallel was made between the two algorithms (the K-Means algorithm achieved with the Euclidean distance and the K-Means algorithm achieved with Hellinger distance). As a result of the two algorithms I found that the number of groups is the same, and the number of iterations is different.

Key words: algorithm, cluster, distance, iteration, group
J.E.L. classification: O31, O32

1. Introduction

Clustering is a statistical method used to group multidimensional data. That is, it is useful to summarize large amounts of information, and each group contains several points with similar features. (*Stancu Ana-Maria Ramona, Mocanu Mihaela, "Metode utilizate în Data Mining", Conferința Internațională 2017, Revista Knowledge Horizons–Economics, pp. 43-47, ISSN: 2066-1061*)

Figure no. 1 Clusters of people by age and income



For example, if we have the attributes: age and revenue, then the segmentation algorithm groups into data sets as follows:

- Cluster 1: includes the young population with low income;
- Cluster 2: includes middle-aged population with income;
- Cluster 3: includes the old-aged population with low-income;

Segmentation is an undirected data mining operation in which there is no attribute to conduct the training process, and all input parameters are treated equally. Most clustering algorithms build their model by iterations that stop when the pattern is fully covered, that is, when the limits of these segments are stabilized.

It is known that clustering of a time series applied to financial-banking data is valid only if the fluctuations in the group are correlated and the fluctuations between the groups are slightly correlated or not correlated at all. Financial-banking data may comply with GnetXP, otherwise they can be reprocessed. Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects, groups called clusters.Unlike classification, clustering does not require the establishment of a variable to be studied as a form of grouping by observation and not by learning. Clustering algorithms verify groups in data sets and try to obtain an optimal boundary between elements based on these clusters.

Clustering algorithms operate with two types of data:

1) the data matrix (or object-variable structure) that represents m objects with n variables (also called measures or attributes). The structure has the form of a relational table or a matrix of size m * n

$$\begin{bmatrix} x_{11} & \cdots & x_{1d} & \cdots & x_{1n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{id} & \cdots & x_{in} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{m1} & \cdots & x_{md} & \cdots & x_{mn} \end{bmatrix}$$

2) the difference matrix (object-object structure) contains a collection of proximities of all pairs of objects and is represented as a n * n matrix:

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ d(2,1) & 0 & 0 & 0 \\ d(3,1) & d(3,2) & 0 & 0 \\ \vdots & \vdots & \vdots & 0 \\ d(n,1) & d(n,2) & \cdots & 0 \end{bmatrix}$$

where d (i, j) represents the difference between objects i and j.

a. Quality of clusters

The quality of clusters is measured by differences and similarities. Generally, the difference *d (i, j)* is a non-negative number close to 0 if it is very similar to j and gets bigger as objects differ.

Differences can be obtained by subjective measurements or correlation coefficients used in statistics: Pearson and Spearman.

b. Classification of clustering methods:

b1. Partition methods

Given a database of n objects or data tuples, a partitioning method will make k partitions in which each partition represents a cluster with k n and satisfies the following conditions:

1) each cluster contains at least one object;
2) each object must belong to a single cluster.

Given k representing the number of partitions to build.Such a method creates an initial partition using an iterative reallocation technique that attempts to improve partition by moving objects from one group to another.

The general criterion of a good partition is that objects in the same cluster must be "close" or similar, while objects belonging to two different clusters must be as "distant" as possible or very different.

Most applications use one of the following two heuristic methods:

i. The k-medoids algorithm (proposed by Kaufman and Rousseeuw in 1987) where each cluster is represented by one of the objects closest to the center of the cluster (medoid);
ii. The average k-algorithm (proposed by McQueen in 1967) where each cluster is the average value of its objects.

These heuristic clustering methods work well for finding spherical clusters in small and medium databases.

Variants of the k algorithm:

- k-Mode algorithm or k prototype algorithm (proposed by Huang, 1998);
- Expectation Maximization algorithm EM (Expectation Maximization, Lauritzen, 1995).

Variants of the kernel algorithm k:

- PAM (Partition around medoids) used in small and medium databases;
- CLARA (Clustering large applications) used in large databases.

2. The K-Means algorithm

The K-Means algorithm was first proposed in 1957 by Stuart Lloyd, developed in 1967 by Mac Queen, following a more efficient version in 1975/1979 proposed and published by Fortran,Hartigan and Wong, an algorithm used in the analysis of clusters using partitioning methods. (*JuanyingXie, Shuai Jiang 2010 - "A simple and fast algorithm for global K-means clustering", 2010 Second International Workshop on Education Technology and Computer Science, 978-0-7695-3987-4/10 $26.00 © 2010 IEEE DOI 10.1109/ETCS.2010.347*)

The K-Means algorithm - cluster initiator, is a method of partitioning formed of spherical-shaped groups. This algorithm uses statistical methods used in the grouping of attributes. The K-Means algorithm is a numerical algorithm designed to find all the cluster positions by minimizing the distance between them and the data points.

The algorithm divides the n objects into k partitions (clusters), and each partition represents a group that is easy to put into practice and can be applied to large data sets.

To accomplish this algorithm, I followed the following steps:

1. We created k clusters by choosing them from a randomly chosen number of data;
2. We calculated the arithmetic mean for each group formed;
3. We assigned each record to the closest cluster using the simulation formula.

Given two forms x and Y with $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ și $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ we define the Euclidean distance as follows:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

1. We assigned each attribute to a cluster, after which we recalculated the arithmetic mean of all groups in the dataset;
2. This process continues with step 3 until no point can be moved and the procedure is complete.

In 2009, Pakira modified the K-Means algorithm so that it moved the center of each group making sure there were no empty clusters.The comparison between the original and the modified algorithm has shown that the original algorithm has the number of iterations higher than the one that was modified. For numerical examples that produce empty clusters, the method proposed by him could not be compared to any other method because the algorithm avoids empty clusters. (*Pakhira, Malay K.: "A Modified k-means Algorithm to Avoid Empty Clusters". International Journal of Recent Trends in Engineering, Vol.1, No. 1, May 2009*). The methods for clustering algorithm were proposed by Bradley and Fayyad in 1998, Wu in 2008, and in 2004 the algorithm was proposed by Khan and Ahmed, an algorithm that creates complex procedures being rather expensive. (*Wu, F. X.: "Genetic weighted k-means algorithm for clustering large-scale gene expression data". BMC Bioinformatics, vol. 9, 2008*)

3. The K-Means algorithm using the Hellinger distance

Based on our research, we found that the algorithm can be implemented in C / C ++ using other distances alongside the Euclidean distance. Among the distances they can implement in the K-Means algorithm are: Baire distance, Bhattacharyya distance, Hamming distance, Hellinger distance, Mahalanobis distance, etc. From the distances outlined above we studied and implemented the K-Means algorithm in C/C ++ using the Hamming and Hellinger distances.

The studied distance is Hellinger. The K-Means algorithm takes the input vectors $\{v_1, v_2, \dots, v_n\}$, $v_i \geq 0$and normalizes them to a unit of length, and the vectors are compared according to their cosine similarity. That is, the distance between v_i and v_j is given by: (*Nathan Stein,VinayKashyap, Xiao-Li Meng, David van Dyk, H-MEANS IMAGE SEGMENTATION TO IDENTIFY SOLAR THERMAL FEATURES, Department of Statistics, Harvard University, Cambridge MA 02138 USA, Harvard-Smithsonian Center for Astrophysics, Statistics Section, Imperial College London, Conference 2012*)

$$d_{cos}(v_i, v_j) = 1 - \frac{v_i^T * v_j}{\|v_i\| * \|v_j\|}$$

Hellinger distance calculates the deviation between two probabilistic distributions that are easy to calculate and are included in the range [0, 1].

Definition: Let M and N discrete probabilistic distributions with $M = (m_1, m_2, \dots, m_k)andN = (n_1, n, \dots, n_k)$. Then Hellinger distance will be:

$$d_H(M, N) = \frac{1}{\sqrt{2}} * \sqrt{\sum_{i=1}^{k} \left(\sqrt{m_i} - \sqrt{n_i}\right)^2}$$

So Hellinger distance is directly related to the Euclidian norm, ie: (*Quang Vu Bui, KarimSayadi, Marc Bui, A Multi-Criteria Document Clustering Method Based on Topic Modeling and Pseudoclosure Function, Proceedings of the Sixth International Symposium on Information and Communication Technology, pages 38-45, Hue City, Viet Nam — December 03 - 04, 2015)*

$$d_H(M, N) = \frac{1}{\sqrt{2}} * \left\|\sqrt{M} - \sqrt{N}\right\|_2$$

Figure no. 2 K-Means algorithm implemented in C / C ++ (Hellinger distance)

```
.....
float radical(float r)
{
return sqrt(r);
}
.....
r =2;
cout<< "Radicalullui " << r << " este: " << radical(r);
  t=1/radical(r);
cout<< "\n Valoarealui t " << " este: " << t;
.....
for(int i = 0; i < k; i++)
 {
for(int j = 0; j < nr; j++)
        {
dist[i][j] = t*(pow(abs(k_val[i] - num[j]),2));
 }
 }
.....
```

where t = 1 / sqrt (2)

4. Comparison of clustering time between the two algorithms

In the first part of the research we compared time K-Means clustering algorithm of Euclidean distance achieved with K-Means algorithm developed by Hellinger distance.

For comparison we have chosen 60 times the initial number of cluster centroids as, the other parameter is the number of iterations required. After testing it cares during the running application to obtain the final solution. Thus, we used a number of different K-Means clustering algorithm and will take one of the values: 1, 5, 10. In terms of time required, we implemented and tested algorithm K-Means using two distances (Euclidean distance, distance Hellinger) and noticed that it took on average 20 seconds to obtain the final solution, and the results we have synthesized Table 1.

Table no. 1 Parallel execution time using three distances

	Euclidean distance	Hellinger distance
k = 1	35,498 seconds	19,054 seconds
k = 5	25,484 seconds	19,017 seconds
k = 10	17,507 seconds	16,831 seconds

5. Comparison of the two algorithms

In the second part of the research, we conducted a comparison between the K-Means algorithm and the Euclidean distance using the K-Means algorithm we used the Hellinger distance.

Table no. 2 Parallel algorithms (using two distances)

Algoritmul K-Means (distanțaEuclidiană)	**Algoritmul K-Means (distanțaHellinger)**
Iteration 1 0 1 4 1 0 1 4 1 0 9 4 1 16 9 4 Group 1: (1 = 1) Group 2: (2 = 2) Gorup 3: (3, 4, 5 = 4) *Iteration 2* 0 1 9 1 0 4 4 1 1 9 4 0 16 9 1 Group 1: (1 = 1) Group 2: (2, 3 = 2) Group 3: (4, 5 = 4)	*Iteration 1* 0 0 2 0 0 0 2 0 0 6 2 0 11 6 2 Group 1: (1,2 = 1) Group 2: (3 = 3) Group 3: (4, 5 = 4) *Iteration 2* 0 2 6 0 0 2 2 0 0 6 0 0 11 2 0 Group 1: (1,2 = 1) Group 2: (3, 4 = 3) Group 3: (5 = 5)
	Iteration 3 0 2 11 0 0 6 2 0 2 6 0 0 11 2 0 Group 1: (1,2 = 1) Group 2: (3, 4 = 3) Group 3: (5 = 5)

After applying the K-Means algorithm using Euclidean distance algorithm and K-Means distance achieved using Hellinger found that the number of groups is the same, and the number of iterations is different.

6. Conclusion

Clustering is the process of splitting a database into groups of similar records so that members of the same groups are as close to each other as possible, and the groups are as far away from each other. In the first part, I compared the clustering time between the K-means algorithm achieved with the help of the Euclidean distance and the algorithm K-means achieved using the Hellinger distance and I noticed that the execution time for the algorithm achieved using the Hellinger distance is smaller than the algorithm achieved using the Euclidean distance. In the second part I realized a parallel between the two algorithms and noticed that I have the same number of groups and the number of iterations is higher for the algorithm achieved using the Hellinger distance

7. References

- JuanyingXie, Shuai Jiang 2010 - "A simple and fast algorithm for global K-means clustering", 2010 Second International Workshop on Education Technology and Computer Science, 978-0-7695-3987-4/10 $26.00 © 2010 IEEE DOI 10.1109/ETCS.2010.347
- Nathan Stein,VinayKashyap, Xiao-Li Meng, David van Dyk, H-MEANS IMAGE SEGMENTATION TO IDENTIFY SOLAR THERMAL FEATURES, Department of Statistics, Harvard University, Cambridge MA 02138 USA, Harvard-Smithsonian Center for Astrophysics, Statistics Section, Imperial College London, Conference 2012
- Pakhira, Malay K.: "A Modified k-means Algorithm to Avoid Empty Clusters". International Journal of Recent Trends in Engineering, Vol.1, No. 1, May 2009
- Quang Vu Bui, KarimSayadi, Marc Bui, A Multi-Criteria Document Clustering Method Based on Topic Modeling and Pseudoclosure Function, Proceedings of the Sixth International Symposium on Information and Communication Technology, pages 38-45, Hue City, Viet Nam — December 03 - 04, 2015
- Stancu Ana-Maria Ramona, Mocanu Mihaela, "*Metode utilizate în Data Mining*", Conferința Internațională 2017, Revista Knowledge Horizons–Economics, pp. 43-47, ISSN: 2066-1061
- Wu, F. X.: "Genetic weighted k-means algorithm for clustering large-scale gene expression data". BMC Bioinformatics, vol. 9, 2008