"Dimitrie Cantemir" Christian University
Knowledge Horizons - Economics
Volume 9 , No. 4, pp. 25 - 27
P-ISSN: 2069-0932, E-ISSN: 2066-1061

www.orizonturi.ucdc.ro

MAKING THE K-MEANS ALGORITHM USING THE HAMMING DISTANCE

Liviu Constantin STOICA[1] , Marian Pompiliu CRISTESCU[2], Ana-Maria Ramona STANCU[3]

[1] PhD, Academy of Economic Studies, Bucharest, E-mail: stoica.liviu.constantin@gmail.com
[2] Associate Professor, "Lucian Blaga" University, Sibiu, E-mail: marian.cristescu@ulbsibiu.ro
[3] Assistant Lecturer, "Dimitrie Cantemir" Christian University, Bucharest, E-mail: ana_maria_ramona@yahoo.com

Abstract

In this article I will start in the first part with an introduction to unsupervised learning methods, specifically the K-Means clustering algorithm, the algorithm achieved using the Euclidean distance.

In the second part, we modified the K-Means algorithm, which is to say, I realized it with the help of Hamming, and then we compared the clustering time, and we made a parallel between the two algorithms (the K-means algorithm conducted with the Euclidean distance and the algorithm K-Means achieved using the Hamming distance).

Key words:
algorithm, cluster, distance, iteration, group
JEL Codes:
C 55
C 61

1. K-MEANS ALGORITHM

The K-Means algorithm is the most used part algorithm (Macqueen, 1967) that has become one of the most used algorithms. Its popularity is given by the simplicity of implementation, scalability, efficiency and speed of convergence. [1]

The K-Means algorithm is an unsupervised classification method that involves grouping objects in a data set in multiple clusters, and each cluster contains a set of objects in a specific category.

The most significant advantages of the K-Means algorithm:

- The algorithm is simple to understand and easy to use;
- It is used in large data sets with thousands of recordings;
- Minimizing the objective of the quadratic error function has increased efficiency, etc.

The algorithm also presents some disadvantages from which they can recall:

- How the initialisation of centroids is not specified;
- The number of clusters must be known beforehand;
- The final result depends on the initial positions of the centroids and the Metrica used to measure the distance. [2]

2. K-MEANS ALGORITHM USING THE HAMMING DISTANCE

The Hamming distance bears the name of Richard Hamming who introduced the concept of error codes in his work in the year 1950, and the bit analysis is used in various disciplines such as: information theory and coding, cryptography, etc.

Definition: either X and Y vectors, then the Hamming distance will be:

$$d(X,Y) = \sum |X_i - Y_i|$$

Fig. 1 – Algoritm K-Means implemented in C/C++ (Hamming distance)

```
.....
for(int i = 0; i < k; i++)
  {
for(int j = 0; j < nr; j++)
        {
dist[i][j] = abs(k_val[i] - num[j]);
}
  }
.....
```

In order to assess the performance of K-Means algorithms using the three distances (Euclidean distance, Hellinger distance and Hamming distance) I implemented in C/C++ and studied them using a different number of clusters and values.

In order to compare the performance of the K-Means algorithm, we have chosen the Euclidean distance as the basis for the achievement of the test.

3. COMPARING THE CLUSTERING TIME BETWEEN THE TWO ALGORITHMS

Comparing the clustering time between the K-means algorithm achieved with the Euclidean distance and the K-means algorithm achieved with the two distances (Hellinger distance and Hamming distance).

For the comparison of time we chose 60 clusters as the initial number of centroids, and the other required parameter is the number of iterations. As a result of the test I am interested in the time when the application runs until the final solution is obtained. Thus, we used a varied number of clusters for the K-Means algorithm and I will take one of the values: 1, 5, 10.
As for the time required we have implemented and tested the K-Means algorithm using the three distances (Euclidean distance, Hellinger distance and Hamming distance) and I noticed that we needed an average of 20 seconds to obtain the final solution, and The results obtained we synthesized in *table* 3.4

Table 1 – Parallel running time using the three distances

	Euclidean distance	Hamming distance
k = 1	35,498 sec	17,377 sec
k = 5	25,484 sec	17,368 sec
k = 10	17,507 sec	17,507 sec

4. PARALLEL BETWEEN THE TWO ALGORITHMS

In the second part of the research, we developed a parallel between the K-means algorithm using the Euclidean distance and the K-means algorithms in which we used the Hamming distance.

K-Means algorithm (Euclidian distance)	**K-Means algorithm (Hamming distance)**
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) Group 3: (3, 4, 5 = 4)	*Iteration 1* 0 1 2 1 0 1 2 1 0 3 2 1 4 3 2 Group 1: (1 = 1) Group 2: (2 = 2) Group 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 2* 0 1 3 1 0 2 2 1 1 3 2 0 4 3 1 Group 1: (1 = 1) Group 2: (2, 3 = 2) Group 3: (4, 5 = 4)

At the same time, we have altered the number of values entered into the two algorithms and the value of K.

In the K-Means algorithms with the help of the Euclidean dimension and the Hamming I noticed that if the value of K (cluster number) is 3 I have a single iteration, three groups and the execution time is greater than 40 sec, and if its value increases the time of Execution is lower.

6. CONCLUSIONS

Following the application of the K-means algorithm using the Euclidean distance and the K-means algorithm achieved with the Hamming distance I found that the K-means algorithm achieved with the help of the Hamming distance has the same number of iterations as the initial algorithm (The K-means algorithm achieved using the Euclidean distance), the groups in the K-means algorithm made using the Hamming distance correspond to the initial algorithm groups (the K-means algorithm achieved using the Euclidean distance) and The common characteristic of the three algorithms is that in every iteration I have three groups.

7. BIBLIOGRAPHY:

[1] Drd. Ing. Ioan Agavriloaei, "Modele și algoritmi de Web Mining", Teză de doctorat, Universitatea Tehnică "Gheorghe Asachi", Facultatea de Automatică și Calculatoare, Iași 2012

[2] Stancu Ana-Maria Ramona, "Soluții de extragere a cunoștințelor din volume mari de date"-Teza de doctorat, ASE 2017