University of Leeds
SCHOOL OF COMPUTER STUDIES
RESEARCH REPORT SERIES
Report 95.18

Fast K-MEANS Clustering Algorithms

by

M. B. Al-Daoud, N. B. Venkateswarlu & S. A. Roberts
Division of Operational Research and Information Systems

June 1995

Abstract– K-MEANS is one of the most popular clustering algorithms. The CPU time required by K-MEANS is often unacceptable, particularly for large problems. In this article, some new techniques are presented to reduce CPU time. Experiments on two data sets gave 90% savings.

1 Introduction

Clustering techniques have received attention in many areas such as image processing applications for data compression. For large clustering problems such as Vector Quantization [2], the time required by the K-MEANS algorithm [5] is unacceptable, due to the amount of time required to compute nearest neighbours [6, 7].

Recently, Zaki et al. [9] developed a classifying method called Ensemble Average (EA) method. Venkateswarlu et al. [8] applied the EA method to classifying remote sensing images and reported that the method demands less CPU time than Euclidean classifier method, although both methods gave the same classification.

In this article, first we investigate the efficiency of the EA method to clustering problems and propose using it with K-MEANS. Further, we propose some variants of the EA-K-MEANS. The results are compared with the K-MEANS algorithm using two real data sets.

2 K-MEANS Algorithm

The K-MEANS algorithm [5] is based on minimising the sum of squared distances $d_i(\mathbf{X})$ from all input vectors $\mathbf{X}$ in the cluster domain to their cluster centres.

Let $\mathbf{X}_1$, $\mathbf{X}_2$, ... $\mathbf{X}_N$ be the input vectors to be clustered; let $\mathbf{M}_i$ be cluster centres involved. Thus, each input vector $\mathbf{X}$ is assigned to cluster (class) c_i if $d_i(\mathbf{X}) < d_j(\mathbf{X})$ for all $j \neq i$, $j = 1, \ldots, K$

2.1 Ensemble Average (EA) Algorithm

The EA method is a new non-parametric classification procedure in which the ensemble average (mean) of the input vectors in each cluster is computed. Then, an input vector $\mathbf{X}$ is assigned to cluster c_i if

$$\mathbf{X}^T(\mathbf{M}_i - \mathbf{M}_j) \geq T_{ij}, \quad \forall j \neq i \tag{1}$$

where T_{ij} is a threshold value defined as

$$T_{ij} = \frac{(\mathbf{M}_i^T\mathbf{M}_i)[\mathbf{M}_i^T(\mathbf{M}_i - 2\mathbf{M}_j)] - (\mathbf{M}_j^T\mathbf{M}_j)[\mathbf{M}_j^T(\mathbf{M}_j - 2\mathbf{M}_i)]}{2(\mathbf{M}_i^T\mathbf{M}_i + \mathbf{M}_j^T\mathbf{M}_j - 2\mathbf{M}_i^T\mathbf{M}_j)} \tag{2}$$

Since T_{ij} is symmetric only the upper triangle need be computed.

To assign an input vector to its nearest class, the EA method requires $(K-1)D$ multiplications, D is the dimensionality, while the original K-MEANS requires KD multiplications.

2.2 Modified Ensemble Average (MEA) Algorithm

In this algorithm we propose a new logic (termed PNND) which is based on the Nearest Neighbouring Distance (NND) [3] and Expanded Distance (ED) [8].

The NND of a cluster is one-half of the distance to its nearest cluster in D-space. In the ED method, $d_i(\mathbf{X})$ is expressed as:

$$d_i(\mathbf{X}) = (\mathbf{X}^T\mathbf{X} - \mathbf{W}^T\mathbf{M}_i + \mathbf{M}_i^T\mathbf{M}_i) \tag{3}$$

where $\mathbf{W} = 2\mathbf{X}$.

The PNND logic is as follows: if input vector $\mathbf{X}$ is assigned to cluster q (in a previous iteration) and the distance between $\mathbf{X}$ and q is less than the NND of q then $\mathbf{X}$ is assigned to q; otherwise, apply the EA algorithm.

Thus, an input vector $\mathbf{X}$ is assigned to cluster c_i if

$$(\mathbf{X}^T\mathbf{X} - \mathbf{W}^T\mathbf{M}_q < NND(q) - \mathbf{M}_q^T\mathbf{M}_q) \tag{4}$$

Otherwise, apply the EA algorithm.

2.3 PCA-MEA Algorithm

In this algorithm we use Principal Component Analysis (PCA) with MEA. Given a data set with D variables , it is possible to construct a new set of p variables, $p \leq D$ which are a linear transformation of the original dimensions [1].

The PCA-MEA algorithm is as follows:

1. Conduct a linear transformation to reduce the original dimension to a smaller one with 95% preserved information.
2. Run the MEA algorithm, with the reduced dimension.
3. Reconstruct the generated cluster centres by conducting an inverse transformation.

3 Experimental Results and Discussion

To evaluate the proposed algorithms two data sets have been used. The first set represents the widely used Baboon image, the second contains a data extracted from one minute of speech.

All algorithms were implemented in C++ programming language and executed on a Sun work station. The CPU time is measured in seconds. The number of dimensions, D, varies between 4 and 32 and the number of clusters, K, varies between 2^2 and 2^{10}. The approach of Katsavounidis et al. [4] was used to initialise the clusters for all methods.

Figure 1 shows the performance of each algorithm, with varying number of clusters (K). The MEA and PCA-MEA algorithms performed better than K-MEANS in all cases. Also the EA algorithm performs better than the original K-MEANS when $K < 512$ and $D < 8$, although this improvement is marginal.

We have also tested the algorithms by varying dimensionality, while the number of clusters and samples (input vectors) are fixed. The results (Table 1) show that the performance of the MEA and PCA-MEA algorithms continue to perform the best.

The clustering obtained from the PCA-MEA algorithm are very close to those of the K-MEANS algorithm. This small difference is likely to be acceptable as long as we are seeing 90% savings in CPU time. It is expected that one or two further iterations may be needed to achieve the same results as those of the K-MEANS. This process will be cheap if the MEA algorithm is used.

4 Conclusion

In this article, new strategies have been incorporated into the K-MEANS clustering algorithm. These strategies were tested on two data sets. The results show that the percentage of CPU time savings varies between 60 to 90%. The new strategies represent efficient tools to clustering problems.

References

[1] C. Chatfield and A. Collins, Introduction to Multivariate Analysis. London: Chapman and Hall, 1980.

[2] A. Gersho and R. Gray, Vector Quantization and Signal Compression. Boston: Kluwer, 1993.

[3] M. Hodgson, “Reducing computational requirements of the minimum distance classifier,” Remote Sensing of Environments, vol. 25, 1988, pp. 117–128.

[4] I. Katsavounidis, C. Kuo and Z. Zhang, “A new initialization technique for generalized Lioyd iteration,” IEEE Signal Processing Letters, vol. 1, no. 10, 1994, pp. 144–146.

[5] J, MacQueen, “Some methods for classification and analysis of multivariate observations,” Proc. 5th Berkeley Symposium on Mathematics, Statistics and Probability, 1967, pp. 281–297.

[6] M. Soleymani and S. Morgera, “An efficient nearest neighbor search method,” IEEE Transactions on Communications, vol. COM-35, 1987, pp. 677–679.

[7] N. B. Venkateswarlu and P. S. V. S. K. Raju, “Fast isodata clustering algorithms,” Pattern Recognition, vol. 25, no. 3, 1992, pp. 335–342.

[8] N. B. Venkateswarlu and P. S. V. S. K. Raju, “A new fast classifier for remotely sensed imaginary,” International Journal of Remote Sensing, vol. 14, no. 2, 1993, pp. 383–389.

[9] F. Zaki, A. Abd El-Fattah, Y. Enab and S. El-Konyaly, “An ensemble average classifier for pattern recognition machines,” Pattern Recognition, vol. 21, no. 4, 1988, pp. 372–332.

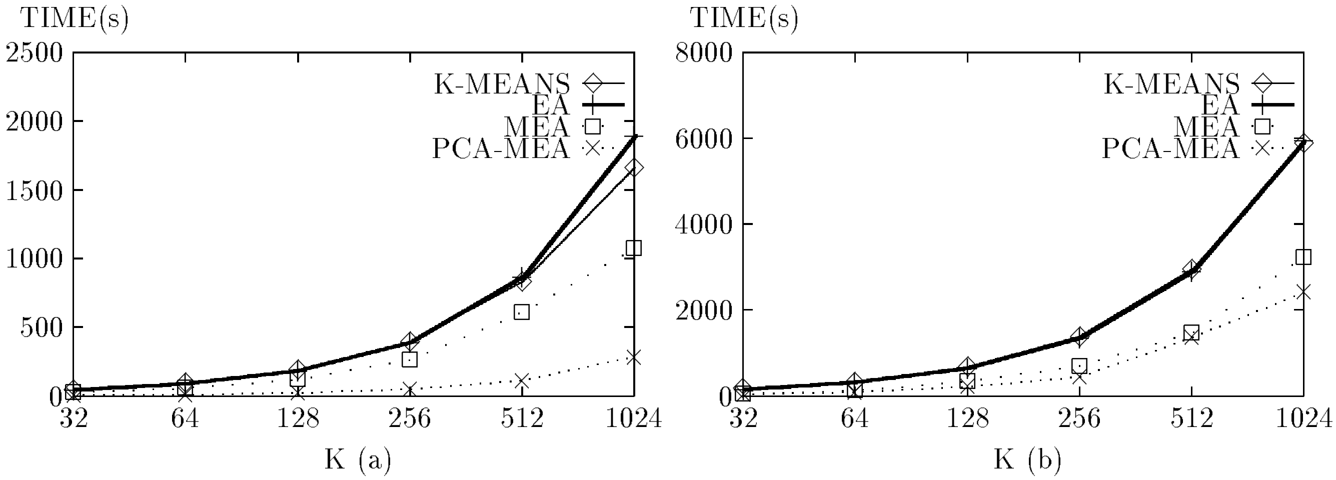

Figure 1: (a) Image: $D = 8$, $N = 8192$, No. of PCs=2. (b) Speech, $D = 8$, $N = 29000$, No. of PCs=6, first 20 iterations.

	IMAGE				SPEECH			
D	4	8	16	32	4	8	16	32
K-MEANS	45	98	183	343	163	348	650	1214
EA	45	98	186	357	153	345	643	1225
MEA	33	76	125	242	48	114	276	448
PCA–MEA	5	15	33	55	42	91	220	358

Table 1: CPU time for both data sets (N=2048,K=256) with different dimensions, first 20 iterations.