Iterative K-Means Algorithm Based on Fisher Discriminant

Mantao Xu
University of Joensuu
P. O. Box 111
FIN-80101 Joensuu
Finland
xu@cs.joensuu.fi

Pasi Fränti
University of Joensuu
P. O. Box 111
FIN-80101 Joensuu
Finland
franti@cs.joensuu.fi

Abstract – *K-Means clustering is a well-known tool in unsupervised learning. The performance of K-Means clustering, measured by the F-ratio validity index, highly depends on selection of its initial partition. This problematic dependency always leads to a local optimal solution for k-center clustering. To overcome this difficulty, we present an intuitive approach that iteratively incorporates Fisher discriminant analysis into the conventional K-Means clustering algorithm. In other words, at each time, a suboptimal initial partition for K-Means clustering is estimated by using dynamic programming in the discriminant subspace of input data. Experimental results show that the proposed algorithm outperforms the two comparative clustering algorithms, the PCA-based suboptimal K-Means clustering algorithm and the kd-tree based K-Means clustering algorithm.*

Keywords: K-Means clustering, discriminant analysis, dynamic programming.

1 Introduction

K-Means clustering is a famous unsupervised learning technique in the context of pattern recognition and machine learning. The objective of the convectional K-Means clustering is to dig out the inherent partition inside data objects, namely, to search an optimal partition of data objects with the minimum value of the mean distortion function. Thus, the K-Means clustering is an optimization problem described by the minimization of the MSE function:

$$\text{minimum} MSE(P) = \frac{1}{N}\sum_{i=1}^{N} \| x_i - c_{p(i)} \|^2 \qquad (1)$$

where

N is the number of data samples;
k is the number of clusters;
d is the dimension of data vector;
$X = \{ x_1, x_2, \ldots\ldots x_N \}$ is a set of N data vectors;
$P = \{ p(i) \mid i = 1,\ldots N \}$ is class label of X;
$C = \{ c_j \mid j = 1,\ldots k \}$ are k cluster centroids.

The main challenge for the conventional K-Means clustering is that its classification performance highly depends on its initially selected partition. In other words, with most of the randomized initial partitions, the conventional K-Means algorithm converges to a locally optimal solution. The extended versions of K-Means such as K-Median [1], adaptive K-Means [2] and kernel K-Means [4] were recently developed to overcome this local optimality problematic. The K-Median algorithm searches each cluster centroid from data samples such that the centroid minimizes the summation of the distances from all data points in the cluster to it. The optimal adaptive K-Means provides the conventional K-Means algorithm with an enhancement of fast convergence by approximating an optimal clustering solution with an adaptive learning rate. The improvement made the by this adaptive K-Means algorithm is based on the optimality criterion that clusters in the underlying partition of data source have the same variances when the number of clusters is large enough. The optimality criterion also provides a biased distance measurement [2] by weighting the square distance with the within-class variance. A state-of-art technique to attack the *k*-center clustering problem is the kernel version of K-Means clustering, which expresses its distance function in a form of kernel product of two data samples in a higher dimensional space, where data samples are more separable. Namely, the kernel machine solves the *k*-center clustering problem in a highly dimensional Hilbert space instead of it original feature space.

The optimization problem of *k*-center clustering in *d*-dimensional feature space has been proved to be *NP*-complete in *k*. The solution for *k*-center clustering in one dimensional space, however, can be solved by dynamic programming in O(*kN*) time [7]. An intuitive approach to estimate an initial partition closer to the global optimum is to apply the dynamic programming technique over some one-dimensional subspace of input data. In particular for Wu's work on color quantization [8], this subspace was estimated by using principal component analysis (PCA) on input data. In other words, the dynamic programming technique can be performed over each principal component subspace obtained by PCA. Since the best principal direction can be selected only from *d* number of principal components, this estimated initial partition might be still far from the global optimum in the case of high dimensional data source. A departure from this limitation is to iteratively incorporate both the linear Fisher discriminant and the dynamic programming technique into the K-Means clustering. The initial partition of the K-

Means clustering at each iteration is estimated by using dynamic programming in the discriminant subspace of input data. The input class partition for the Fisher discriminant analysis at each iteration is selected by the output partition of the K-Means clustering at the former iteration.

In this work, a suboptimal K-Means clustering algorithm is investigated based on the multi-class Fisher discriminant and dynamic programming. In particular, a biased non-symmetric distance measurement, the Delta-MSE dissimilarity, is incorporated into the proposed clustering algorithm. In second section, we describe the suboptimal K-Means clustering algorithm. In section 3, we briefly review the multi-class Fisher discriminant. In section 4, a heuristic biased dissimilarity, the Delta-MSE function, is introduced for K-Means clustering. In the experimental section, the proposed approach is compared to the other two K-Means clustering algorithms: the PCA-based suboptimal K-Means algorithm [8] and the *kd-tree* based K-Means clustering algorithm [5]. Finally, conclusions are drawn in section 6.

Function SubOptimalKMeans(X, k, m)
input: Dataset X
Number of clusters k
Number of iterations m
output: Class labels P_{OPT}

$C \leftarrow$ Randomly choose cluster centroids from X;
$P \leftarrow$ K-Means(X, C, k);
$f_{min} \leftarrow \infty$
for $j = 1$ **to** m
$w \leftarrow$ solve Fisher discriminant based on class label P;
$X_w \leftarrow$ project input data X into discriminant direction w;
$P_w \leftarrow$ optimally solve k-center clustering problems on X_w using dynamic programming;
$C, P \leftarrow$ K-Means(X, P_w, k);
$fratio \leftarrow$ calculate F-ratio of P
if $fratio < f_{min}$ **then**
$P_{\mathrm{OPT}} \leftarrow P$
$f_{min} \leftarrow fratio$
end if
end for

Fig. 1. Pseudo-code for the proposed algorithm.

2 Suboptimal K-Means Clustering

As mentioned earlier, the conventional K-Means algorithm typically converges to a local minimum of mean square error (MSE). The algorithm is often initialized by a randomly chosen initial partition. However, in this sense, there is no guarantee of convergence to the global minimum. Motivated by Wu's optimal solution for scalar quantization [7] and solution for color quantization [8], we iteratively apply the multi-class Fisher discriminant in estimation of the suboptimal initial partition instead of using only the d number of principal components. The Fisher discriminant at each iteration can be constructed from the output class assignments obtained by the K-Means clustering at previous iteration. The application of dynamic programming in the discriminant direction leads to a suboptimal partition of data source in the discriminant subspace. Thus, the output suboptimal partition can be selected as the initial partition of K-Means clustering at next iteration. Namely, K-Means clustering and Fisher discriminant are performed once at each iteration. We have presented the pseudocodes of the proposed suboptimal K-Means clustering algorithm in figure 1.

3 Multi-class Fisher discriminant

Discriminant analysis is a powerful tool in finding a direction that best reveals the classification structure. The goal of its application in this work is to apply the discriminant classifier to form a convex partition in the projection subspace that best matches the partition obtained by K-Means clustering in original feature space. After the discriminant direction is determined, one can apply dynamic programming to all the projected data samples to improvingly optimize the partition in the projection subspace. The multi-class Fisher discriminant [8] lends us a tool to design a classifier that approximates the partition achieved by the convectional K-Means clustering algorithm. The separation of input classes in the projection direction w can be measured by the so-called F-ratio validity index, $F(w)$, defined as the ratio of between class variance and within class variance:

$$F(w) = k \frac{\sum_{j=1}^{M} n_j (w^T (c_j - \bar{x}))^2}{\sum_{i=1}^{N} (w^T (x_i - c_{p(i)}))^2} \tag{2}$$

where n_j is the sample size of class j. The multi-class linear Fisher discriminant is derived by the minimization of the F-ratio validity index in equation (2), i.e.,

$$w = \arg\min_{w} k \cdot \frac{w^T \mathbf{S}_W w}{w^T \mathbf{S}_B w} \tag{3}$$

where $\mathbf{S}_B$ is between class covariance matrix and $\mathbf{S}_W$ is within class covariance matrix respectively:

$$\begin{aligned} \mathbf{S}_B &= \sum_{j=1}^{M} n_j (c_j - \bar{x})(c_j - \bar{x})^{\mathrm{T}} \\ \mathbf{S}_W &= \sum_{i=1}^{N} (x_i - c_{p(i)})(x_i - c_{p(i)})^{\mathrm{T}} \end{aligned} \tag{4}$$

The discriminant direction w can be estimated by computing the leading eigenvector of matrix $\mathbf{S}_W^{-1}\mathbf{S}_B$.

4 Delta-MSE dissimilarity

In this work, Instead of L_2 square distance, a heuristic distance measurement, the Delta-MSE dissimilarity, was taken into account for the K-Means algorithm as proposed in [9]. The Delta-MSE dissimilarity is analytically induced from the clustering MSE distortion by moving a given data sample from one cluster to another cluster. The dissimilarity is calculated as the change of the within-class variance caused by this movement. The design approach

of Delta-MSE always takes into account the dynamic nature of the K-Means partition process, in which cluster parameters (cluster size) are subject to change all the time in the clustering algorithm.

Assuming that a data sample x is moved from cluster i to cluster j, the change of the MSE function caused by this move is:

$$v_{ij}(x) = \frac{n_j}{n_j + 1} \| x - c_j \|^2 - \frac{n_i}{n_i - 1} \| x - c_i \|^2 \qquad (5)$$

The first part in the right hand side, representing the increased variance of cluster j caused by this move, denotes the biased dissimilarity between x and c_j. as D_{MSE} (x, c_i). The second part, representing the decreased variance of cluster i, denotes the dissimilarity between x and c_i as D_{MSE} (x, c_i). Thus, the Delta-MSE dissimilarity between data point x_i and the cluster centroid c_j can be defined as:

$$D_{MSE}(x_i, c_j) = w_{ij} \cdot \| x_i - c_j \|^2 \qquad (6)$$

and can be weighted by:

$$w_{ij} = \begin{cases} n_j / (n_j + 1) & p(i) \neq j \\ n_j / (n_j - 1) & p(i) = j \end{cases} \qquad (7)$$

It is worth noting that the sparser the cluster is, the more different the Delta-MSE dissimilarity can be in comparison to the L_2 distance. The weight w_{ij} makes the biased dissimilarity bigger than L_2 square distance if x_i is allocated in the cluster and smaller than L_2 square distance otherwise. In the repartition of data samples driven by the biased dissimilarity, each sample is inclined to join or leave the sparser clusters more frequently than the denser clusters. Accordingly, the reassignments of data samples into their closest clusters are driven with the Delta-MSE dissimilarity more frequently than with the L_2 square distance. Thus, the biased dissimilarity enables the suboptimal clustering algorithm with a faster convergence to the global optimum.

5 Experimental results

We have conducted experiments on the k-clustering problems of 5 real datasets from UCI machine learning repository [4]. In the experiments, we studied the proposed suboptimal K-Means clustering by two dynamic programming methods. In the first method denoted as LFD-I, we implemented dynamic programming by the MSE distortion function defined on the projection subspace. Although the first method converges to a global minimum [7] in one-dimensional projection subspace, in the second method denoted as LFD-II, we consider the MSE distortion function defined on the d-dimensional feature space in design of dynamic programming. Of course, in practice, one can view this approach not only as an approximation algorithm but also as a heuristic algorithm. We also compared the two proposed methods with the two alternative clustering algorithms: the PCA-based suboptimal K-Means algorithm and the *kd-tree* based K-Means clustering algorithm. The *kd-tree* based K-Means algorithm selects its initial cluster centroids from the k- bucket centers of a so-called *nested PCA kd-tree* structure. This *kd-tree* structure can be constructed by reccursively using principal component analysis. The two comparative K-Means clustering algorithms are here denoted as PCA and KD-Tree respectively.

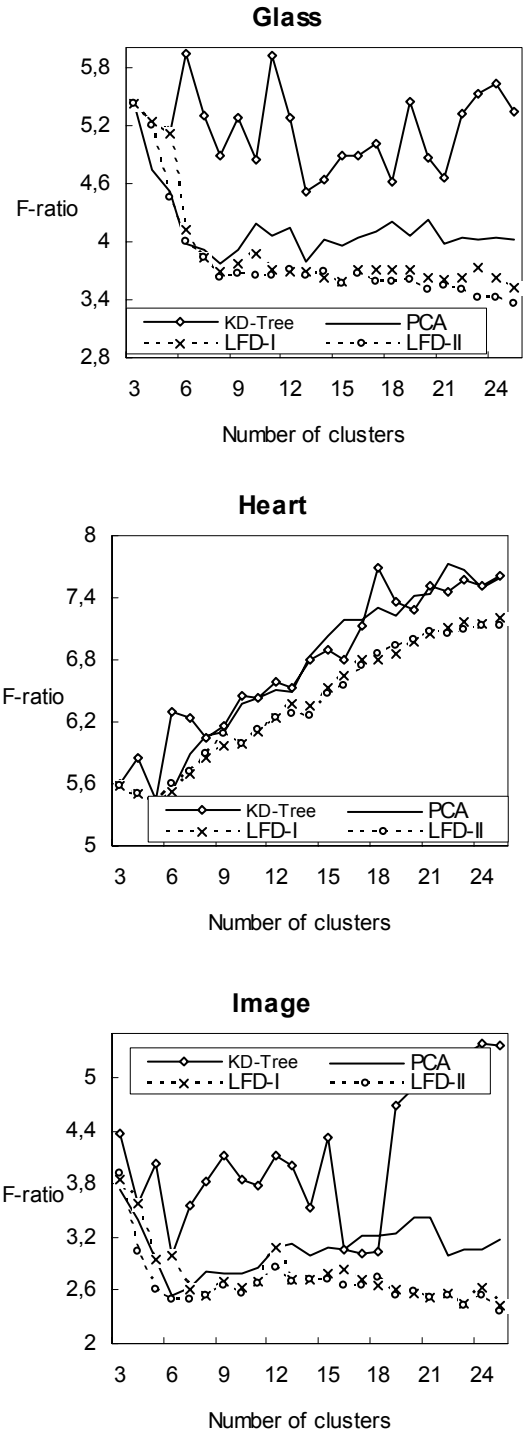

Fig. 2. Fratio distortions obtained by using the four different K-Means clustering algorithms.

The four K-Means clustering approaches are tested: PCA, LFD-I and LFD-II and KD-Tree over the five famous datasets from UCI machine learning repository [6] as *boston*, *heart*, *glass*, *image* and *thyroid*. The clustering performances of the four K-Means clustering algorithms are measured by the F-ratio clustering validity index. Figure 2 plots the F-ratio validity index obtained by the four K-Means algorithms over the datasets: *glass*, *heart* and *image*. The F-ratio validity index is presented as the function of the number of clusters *k*. It can be observed that the two proposed algorithms in general outperform the other two comparative algorithms. In particular, with the number of cluster *k* increased, their clustering performance gains are much improved against the other two comparative algorithms.

Among the four clustering algorithms, the proposed suboptimal K-Means algorithms based on the multi-class Fisher Discriminant analysis yield better results than the other tow algorithms. We also compared clustering results from the four algorithms on the practical number of clusters for each dataset. Table 1 displays the F-ratio validity indices on the practical number of clusters for each dataset. Not surprisingly, the suboptimal K-Means algorithms based on the Fisher discriminant analysis achieve better F-ratio validity indices than the other two algorithms.

Table 1: Performance comparisons of the four K-Means algorithms on the practical numbers of clusters

Datasets	*k*	KD-Tree	PCA	LFD-I	LFD-II
boston	9	4.083	3.526	3.515	3.515
glass	6	5.931	3.974	3.966	3.984
heart	5	5.436	5.420	5.410	5.410
image	7	3.556	2.622	2.615	2.499
thyroid	3	2.265	2.265	2.264	2.264

6 Conclusion

We have proposed a new approach to the k-center clustering problem based on the linear Fisher discriminant analysis and the dynamic programming technique. The linear Fisher discriminant analysis serves as a tool of finding the subspace that best matches the classification structure obtained by the conventional K-Means clustering algorithm. Application of dynamic programming in the linear discriminant subspace improves the clustering partition of the K-means algorithms. The improved partition is considered as the initial partition of K-Means clustering in next iteration. Thus, a design technique for the iterative K-Means clusterings can be constructed by iteratively by incorporating the Fisher discriminant analysis and the dynamic programming technique. Experiment results show that the proposed approach in general outperforms the other two comparative K-Means algorithms: the PCA based suboptimal K-Means clustering algorithm and the *kd-tree* based K-Means clustering algorithm. In particular, by increasing the number of clusters, its classification performance gains are improved against the two comparative K-Means algorithms.

References

[1] P. S. Bradley, O. L. Mangasarian and W. N. Street, "Clustering via Concave Minimization", *Advances in Neural Information Processing Systems 9* (*NIPS9*), 368-374, MIT Press, Cambridge, MA 1997.

[2] C. Chinrungrueng and C. H. Sequin. Optimal adaptive k-means algorithm with dynamic adjustment of learning rate. *IEEE Transactions on Neural Network*, Int. J. Tracking in Aerospace, 1(6):157–169, 1995.

[3] D. H. Foley and J. W. Sammon. A optimal set of discriminant vectors. *IEEE Transactions on Computers*, vol. 3, no. 24, pp. 281-289, 1975.

[4] M. Girolami, "Mercer Kernel Based Clustering in Feature Space", *IEEE Trans. on Neural Networks*, 13(4); 780 – 784, 2002.

[5] A. Likas, N. Vlassis and J. J. Verbeek, "The Global *K*-means Clustering Algorithm", *Pattern Recognition*, 36 (2): 451-461, 2003.

[6] UCI Repository of Machine Learning Databases and Domain Theories. http: //www.ics.uci.edu/ ~mlearn/MLRepository.html, 2003.

[7] X. Wu and K. Zhang, "Quantizer Monotonicities and Globally Optimal Quantizer Design Algorithms", *IEEE Trans. on Information Theory*, vol. 39, no. 3, p. 1049-1053, May 1993.

[8] X. Wu, "Color Quantization by Dynamic Programming and Principal Analysis", ACM Trans. on Graphics, vol. 11, no. 4 (TOG special issue on color), p. 348-372, Oct. 1992.

[9] M. Xu, "Delta-MSE Dissimilarity in GLA-based Vector Quantization", *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, (ICASSP'04)*, Montreal, Canada, 2004. (to appear)