# A HEURISTIC K-MEANS CLUSTERING ALGORITHM BY KERNEL PCA

*Mantao Xu and Pasi Fränti*

University of Joensuu
P. O. Box 111, 80101 Joensuu, Finland
{xu, franti}@cs.joensuu.fi

**ABSTRACT**

K-Means clustering utilizes an iterative procedure that converges to local minima. This local minimum is highly sensitive to the selected initial partition for the K-Means clustering. To overcome this difficulty, we present a heuristic K-means clustering algorithm based on a scheme for selecting a suboptimal initial partition. The selected initial partition is estimated by applying dynamic programming in a nonlinear principal direction. In other words, an optimal partition of data samples in the kernel principal direction is selected as the initial partition for the K-Means clustering. Experiment results show that the proposed algorithm outperforms the PCA based K-Means clustering algorithm and the kd-tree based K-Means clustering algorithm respectively.

## 1. INTRODUCTION

K-Means is a well-known technique in unsupervised learning and vector quantization. The K-Means clustering is formulated by minimizing a formal objective function, mean-squared-error distortion.

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

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 samples;
$P = \{ p(i) \mid i = 1, \ldots N \}$ is class label of $X$;
$C = \{ c_j \mid j = 1, \ldots k \}$ are $k$ cluster centroids.

Due to its simplicity for implementation, the conventional K-Means can be applied to a given clustering algorithm as a postprocessing stage to improve the final solution [1]. However, the main challenge for the conventional K-Means is that its classification performance highly relies on the selected initial partition. In other words, with most of randomized initial partitions, the conventional K-Means algorithm converges to a locally optimal solution. An extended version of K-Means, the K-Median clustering, serves a solution to overcome this limitation. 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. However, in practice, there were no efficient solutions known to most of the formulated K-Median problems that are NP-Hard [2]. A more advanced technique [3] is to formulate the K-Means clustering as a kernel machine in a highly dimensional feature space. Namely, the kernel machine solves $k$-clustering problem in a highly dimensional Hilbert space instead of its input space.

The optimization of k-clustering problems in $d$-dimensional space has proved to be NP-hard in $k$, however, for one-dimensional feature space, a scheme based on dynamic programming [8] can serve as a tool to drive a globally minimal solution. Hence, a heuristic approach to estimate the initial partition for K-Means clustering is to tackle the clustering optimization problem in some one-dimensional component space. Motivated by Wu's work on color quantization [9], this can be solved by dynamic programming in the principal component subspace. In particular, a nonlinear curve can be selected as this principal direction, i.e. a kernel principal component [5]. Developed by Scholkopf et al. [6], the kernel principal component analysis (KPCA) is a state-of-art technique for feature extraction with an underlying nonlinear spatial structure, which transfers the input data into a higher dimensional feature space. In this sense, a kernel trick is utilized to perform operation in the new feature space, where data samples are more separable. Since the best principal direction can be selected only from $d$-number of principal components in the linear PCA, the estimated initial partition could be far from the global optima in the case of high dimensional data source. However, the kernel PCA can provide the same number of principal components as the number of input data samples. In a larger sense, data samples are more separable in the nonlinear principal curve direction than in the linear one. Hence, an initial partition closer to the global optima can

0-7803-8554-3/04/$20.00 ©2004 IEEE.

be obtained by applying dynamic programming in the nonlinear principal curve subspace.

In this paper, a heuristic K-Means clustering algorithm is investigated based on the kernel PCA and dynamic programming. A biased distance measurement, the Delta-MSE dissimilarity, is incorporated into the proposed clustering algorithm instead of using the Euclidean distance. In next section, we describe the heuristic K-Means algorithm by using kernel PCA and dynamic programming. In section 3, we briefly review the technique of the kernel principal component analysis. Section 4 introduces the Delta-MSE dissimilarity for the K-Means algorithm. In experimental section, the proposed algorithm is compared to the two existing clustering approaches: the PCA based suboptimal K-Means algorithm [9] and the kd-tree based K-Means clustering algorithm [4]. Finally, conclusions are drawn in section 6.

**input**: Datasets $X$
Number of clusters $k$
Number of principal components $m$
**output**: Class membership $P_{\mathrm{OPT}}$

**Function** HeuristicKMeans($X$, $k$, $m$)
$W$ ← solve $m$ number of kernel principal directions of $X$;
$f_{min}$ ← ∞
**for** $j$ = 1 **to** $m$
  $X_{PJ}(j)$ ← project $X$ on each kernel principal direction $w(j)$;
  $P_1(j)$ ← solve $k$ optimal clustering problems on each scalar variable $X_{PJ}(j)$ by dynamic programming;
  $P(j)$ ← solve K-Means clustering problem $d$-dimensional input space with initial partition $P_1(j)$;
  *fratio* ← calculate F-ratio of $P(j)$
  **if** *fratio* < $f_{min}$ **then**
    $P_{\mathrm{OPT}} \leftarrow P(j)$
    $f_{min} \leftarrow$ *fratio*
  **end if**
**end for**

**Figure 1.** Pseudocodes of heuristic K-Means

## 2. HEURISTIC K-MEANS CLUSTERING

As mentioned earlier, the conventional K-Means algorithm typically converges to a local minimum of mean-squared-error (MSE). The K-Means algorithm is often initialized with a randomly chosen partition. However, in this sense, there is no guarantee of convergence to the global optima. The optimization problem of $k$-clustering in $d$-dimensional feature space has been proved to be $NP$-complete in $k$. Encouraged by the success of kernel PCA [5,6], we apply the kernel PCA in estimation of the suboptimal initial partition instead of using only the $d$ number of principal components in Wu's work on color quantization [9]. The nonlinear principal components are constructed by performing PCA in the higher dimensional feature space expanded by Mercy kernel functions. Application of dynamic programming in each nonlinear principal direction leads to an optimal partition of data samples in the projection subspace. Among the output optimal partitions in $m$ number of principal directions, the partition with the minimum F-ratio clustering validity index is selected as the initial partition for K-Means clustering. The selection strategy leads to a smaller distortion between the suboptimal initial partition and the globally optimal solution. We present the pseudocodes of the proposed heuristic clustering algorithm in figure 1.

## 3. KERNEL PCA

Principal component analysis is one of the most popular techniques for feature extraction. The principal components of input data $X$ can be obtained by solving the eigenvalue problem of the covariance matrix of $X$. This conventional PCA can be generalized as a nonlinear one, the kernel PCA, by $\Phi$: $R^d \rightarrow F$, a mapping from input data space to a highly dimensional feature space $F$. The space $F$ and therewith also the mapping $\Phi$ might be very complicated. To avoid this problem, the kernel PCA employs a kernel trick to perform feature space operations by explicitly using the inner product between two points in the feature space:

$$(\Phi(x_i), \Phi(x_j)) \rightarrow K(x_i, x_j) \tag{2}$$

Thus, its covariance matrix can be written as:

$$W_\Phi = \frac{1}{N}\sum_{i=1}^{N} \Phi(x_i) \cdot \Phi(x_i)^{\mathrm{T}} \tag{3}$$

For any eigenvalue of $W_\Phi$, $\lambda \geq 0$, and its corresponding eigenvectors $V \in F \backslash \{0\}$, the equivalent formulation of eigenvalue problem [6] in $F$ can be defined as:

$$N\lambda\alpha = K\alpha \tag{4}$$

where eigenvector $V$ is spanned in space $F$ as:

$$V = \sum_{i=1}^{N} \alpha_i \Phi(x_i) \tag{5}$$

and where $K_{ij} = K(x_i, x_j)$ and $\alpha = (\alpha_1, \alpha_2 \ldots \alpha_N)^{\mathrm{T}}$. For the kernel component extraction, we compute projection of each data sample $x$ onto eigenvector $V$

$$(\Phi(x), V) = \sum_{i=1}^{N} \alpha_i K(x_i, x) \tag{6}$$

The kernel PCA allows us to obtain the features with high-order correlation between the input data samples. In nature, the kernel projection of data samples onto the kernel principal component might undermine the nonlinear spatial structure of input data. Namely, the inherent nonlinear structure inside input data is reflected with most merit in the principal component subspace.

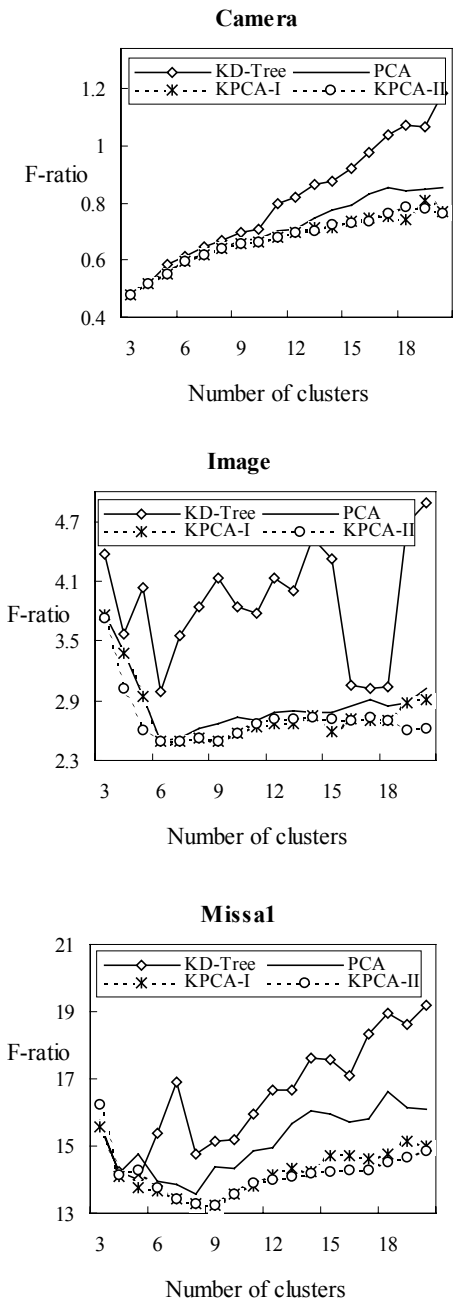


**Figure 2**: Fratio distortions obtained by using the four different K-Means clustering algorithms.

## 4. Delta-MSE dissimilarity

Instead of using the Euclidean distance, we incorporate a heuristic distance measurement, Delta-MSE dissimilarity, into the K-Means clustering as proposed in [10]. This dissimilarity is analytically induced from the clustering MSE function by moving a data sample from one cluster to another, which is calculated as the change of the within-class variance caused by this movement.

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

$$v_{ij}(x) = \frac{n_j}{n_j + 1} \| x - c_j \|^2 - \frac{n_i}{n_i - 1} \| x - c_i \|^2 \qquad (7)$$

The first part in the right hand side, the increased variance of cluster $j$, denotes the biased dissimilarity between $x$ and $c_j$. The second part, representing the decreased variance of cluster $i$, denotes the dissimilarity between $x$ and $c_i$. Thus, the Delta-MSE dissimilarity between data point $x_i$ and the cluster centroid $c_j$ is written as:

$$D_{MSE}(x_i, c_j) = \begin{cases} n_j \| x_i - c_j \|^2 / (n_j + 1), & p(i) \neq j \\ n_j \| x_i - c_j \|^2 / (n_j - 1), & p(i) = j \end{cases} \qquad (8)$$

It is worth noting that the sparser the cluster is, the more different the Delta-MSE dissimilarity can be in comparison to the $L_2$ square distance. In the repartition of data samples driven by this dissimilarity, each sample is inclined to join or leave sparse clusters more frequently than dense clusters. Thus, the heuristic dissimilarity enables the proposed clustering procedure to converge to a solution closer to the global optima.

## 5. EXPERIMENTAL RESULTS

We have conducted experiments on the $k$-clustering problems of 5 real datasets from UCI machine learning repository [7] and the datasets from 6 standard images: *Bridge* and *Camera* are the datasets with 4×4-blocks from image Bridge and Cameraman; *Housec5* and *Housec8* - quantized to 5 bits and 8 bits per color respectively; *Missa1* and *Missa2* are the datasets with 4x4 vectors from the difference image of frame 1 and 2 for Miss American and the difference image of frame 2 and 3 respectively. We studied the proposed K-Means algorithm by two dynamic programming methods. In the first method denoted as KPCA-I, we implemented the dynamic programming by the MSE distortion defined only on the projection subspace. In the second method denoted as KPCA-II, we considered the MSE distortion defined on the whole $d$-dimensional input space in design of dynamic programming. Of course, in practice, one can view this approach as a heuristic algorithm for selecting the initial partition for the K-means clustering. We also compared the two proposed approaches with the two existing clustering algorithms: the PCA-based suboptimal K-Means algorithm (denoted as PCA) and the kd-tree based K-Means clustering algorithm (denoted as KD-Tree). The kd-tree based K-Means algorithm selects the initial cluster centroids from the $k$-bucket centers of a kd-tree developed also by principal component analysis.

The four K-Means clustering approaches, PCA, KPCA-I and KPCA-II and KD-Tree, are tested over the five datasets from UCI repository and the six image datasets. The performances of the 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 approaches over the datasets: *Camera*, *Image* (image segmentation data from UCI) and *Missa1*. The F-ratio validity index is presented as the function of the number of clusters $k$. It can be observed that the two proposed methods in general outperform the other comparative algorithms. In particular, as the number of cluster $k$ is increased, their clustering performances are much improved against the two others. Among the four clustering approaches, the proposed K-Means algorithms by the kernel PCA yield better results than the others. We also compare clustering results from the four algorithms with number of clusters $k$ = 10 in table 1-2. Not surprisingly, the proposed heuristic K-Means algorithms achieve better F-ratio validity indices than the others.

**Table 1**: Performance comparisons of the four K-Means clustering algorithms on the five real datasets from UCI.

| Datasets | KD-Tree | PCA | KPCA-I | KPCA-II |
|---|---|---|---|---|
| *boston* | 3.687 | 3.512 | 3.402 | 3.338 |
| *glass* | 4.838 | 4.185 | 3.699 | 3.644 |
| *heart* | 6.442 | 6.380 | 5.989 | 6.091 |
| *image* | 3.843 | 2.733 | 2.575 | 2.575 |
| *thyroid* | 2.687 | 1.868 | 1.802 | 1.769 |

**Table 2**: Performance comparisons of the four K-Means clustering algorithms on the six image datasets.

| Datasets | KD-Tree | PCA | KPCA-I | KPCA-II |
|---|---|---|---|---|
| *bridge* | 2.213 | 2.225 | 2.117 | 2.087 |
| *camera* | 1.166 | 0.8671 | 0.8268 | 0.7676 |
| *housec5* | 1.224 | 1.223 | 1.112 | 1.111 |
| *housec8* | 0.4733 | 0.4586 | 0.4319 | 0.4338 |
| *missa1* | 19.19 | 16.10 | 15.10 | 14.84 |
| *missa2* | 21.35 | 16.26 | 15.01 | 14.89 |

## 6. CONCLUSION

We have proposed a new approach to the $k$-clustering problem based on the kernel PCA and dynamic programming. Application of dynamic programming in the nonlinear principal direction obtained by the kernel PCA estimates a suboptimal initial partition for the K-means clustering. Since data samples are more separable in the nonlinear principal direction than in the linear one, an initial partition closer to the global optimum is achieved by the proposed selection scheme. A heuristic distance measurement, Delta-MSE function, is also incorporated into the proposed K-Means clustering algorithm instead of the Euclidean distance. Experiment results show that the proposed algorithm in general outperforms the two existing K-Means algorithms compared in this work. In particular, by increasing the number of clusters, its classification performance is improved against the two other algorithms.

## 7. REFERENCES


[1] P. Fränti, J. Kivijärvi and O. Nevalainen: "Tabu search algorithm for codebook generation in VQ," *Pattern Recognition*, 31 (8), pp. 1139-1148, August 1998.

[2] M.R. Garey and D.S. Johnson, *Computers and Intractability*: *A Guide to NP-Completeness*, W. H. Freeman, New York, 1979.

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

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

[5] B.Schölkopf, A. Smola, and K.R. Müller, "Kernel principal component analysis," *Advances in Kernel Methods - Support Vector Learning*, pp.327-352, MIT Press, Cambridge, MA, 1999.

[6] B. Schölkopf, S. Mika, A. Smola, G. Rätsch, and K.R. Müller, "Kernel PCA pattern reconstruction via approximate pre-images," *Proceedings of the 8th International Conference on Artificial Neural Networks, Perspectives in Neural Computing*, pp. 147-152. Springer Verlag, Berlin, 1998.

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

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

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

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