ISSN:0975-9646

An Effective Determination of Initial Centroids in K-Means Clustering Using Kernel PCA

M.Sakthi and Dr. Antony Selvadoss Thanamani

Department of Computer Science, NGM College, Pollachi, Tamilnadu.

Abstract--- **Clustering is considered as the task of dividing a data set such that elements within each subset that is similar between themselves and are dissimilar to elements belonging to other subsets. Clustering techniques usually belong to the group of undirected data mining tools; these techniques are also sometimes referred to as "unsupervised learning" because there is no particular dependent or outcome variable to predict. Cluster analysis is most common in any discipline that involves analysis of multivariate data. K-Means is one of the most widely used algorithms in clustering techniques because of its simplicity and performance. The initial centriod for K-Means clustering is generated randomly. The performance of K-Means clustering is highly affected when the dataset used is of high dimension. The accuracy and time complexity is highly affected because of the high dimension data. Hence, the initial centroid provided must be appropriate. For this purpose, the dimensionality reduction technique called Principal Component Analysis (PCA) is used. For better performance, this paper uses the Kernel Principal Component Analysis (KPCA) for deciding the initial centroid. The experimental result shows that the proposed clustering technique results in better accuracy and the time complexity is also reduced.**

Keywords--- **K-Means, Principal Component Analysis (PCA), Kernel Principal Component Analysis (KPCA), Centroid**

I. INTRODUCTION

ONE of the most essential modes of understanding and learning is categorizing data into reasonable groups. For instance, a general scheme of scientific categorization puts organisms into a system of ranked taxa includes domain, kingdom, phylum, class, etc. Cluster method is defined as the formal study of techniques and approaches for grouping objects based on the evaluated intrinsic relationship or characteristics. Category tags are not utilized in cluster analysis that label objects with previous identifiers, that is class labels. The data clustering (unsupervised learning) is differentiated from categorization or discriminant analysis (supervised learning) due to the unavailability of category information. The main goal of this clustering technique is to generate structure in data and thus it is exploratory in nature.

Due to the improvement in detecting and storage technology and remarkable development in applications like digital imaging, internet search and video surveillance numerous huge-volume and high-dimensional data sets have been generated. It is calculated that the digital universe roughly used 281 exabytes in 2007, and it is estimated to be 10 times that size by 2011. (One exabyte is ~1018 bytes or 1,000,000 terabytes). The majority of the data is stored digitally in electronic media that offers high potential for the development of automatic information investigation, classification, and retrieval techniques. In addition to the increase in the quantity of information, on the other hand, the variety of available data has also increased drastically. The popularity of RFID tags or transponders due to their low cost and small size has led to the deployment of millions of sensors that transmit data regularly. E-mails, blogs, transaction data, and billions of Web pages create terabytes of new data every day. Many of these data streams are unstructured, adding to the difficulty in analyzing them.

Thus, due to increase in both the volume and the variety of data, it is necessary that there should be advancements in methodology to automatically understand, process, and summarize the data. Clustering technique is used widely to deal with this problem. The most commonly and widely used clustering is K-Means [12, 13] because of its simplicity and accuracy. For K-Means clustering, initial parameters like number of clusters and initial centriods are need to be provided [9, 10]. When large dataset [14] is used in clustering, K-Means will misclassify some data and also the time complexity will be more. To overcome this problem, the initial centroid mentioned should be effective. For this purpose Principal Component Analysis (PCA) [11] is employed in [16]. In this paper PCA is replaced by Kernel Principal Component Analysis (KPCA) [7] for effective determination of initial centroids.

II. RELATED WORKS

Zhang Zhe *et al.,* [1] proposed an improved K-Means clustering algorithm. K-means algorithm [8] is extensively utilized in spatial clustering. The mean value of each cluster centroid in this approach is taken as the Heuristic information, so it has some limitations such as sensitive to the initial centroid and instability. The enhanced clustering algorithm referred to the best clustering centroid which is searched during the optimization of clustering centroid. This increases the searching probability around the best centroid and enhanced the strength of the approach. The experiment is performed on two groups of representative dataset and from the experimental observation, it is clearly noted that the improved K-means algorithm performs better in global searching and is less sensitive to the initial centroid.

Hai-xiang Guo *et al.,* [2] put forth an Improved Genetic k-means Algorithm for Optimal Clustering. The value of k must be known in advance in the traditional k-means approach. It is very tough to confirm the value of k accurately in advance. The author proposed an enhanced genetic k-means clustering

(IGKM) and builds a fitness function defined as a product of three factors, maximization of which guarantees the formation of a small number of compact clusters with large separation between at least two clusters. Finally, the experiments are conducted on two artificial and three real-life data sets that compare IGKM with other traditional methods like k-means algorithm, GA-based technique and genetic k-means algorithm (GKM) by inter-cluster distance (ITD), inner-cluster distance (IND) and rate of separation exactness. From the experimental observation, it is clear that IGKM reach the optimal value of k with high accuracy.

Yanfeng Zhang *et al.,* [3] proposed an Agglomerative Fuzzy K-means clustering method with automatic selection of cluster number (NSS-AKmeans) approach for learning optimal number of clusters and for providing significant clustering results. High density areas can be detected by the NSS-AKmeans and from these centers the initial cluster centers with a neighbor sharing selection approach can also be determined. Agglomeration Energy (AE) factor is proposed in order to choose a initial cluster for representing global density relationship of objects. Moreover, in order to calculate local neighbor sharing relationship of objects, Neighbors Sharing Factor (NSF) is used. Agglomerative Fuzzy k-means clustering algorithm is then utilized to further merge these initial centers to get the preferred number of clusters and create better clustering results. Experimental observations on several data sets have proved that the proposed clustering approach was very significant in automatically identifying the true cluster number and also providing correct clustering results.

Xiaoyun Chen *et al.,* [4] described a GK-means: an efficient K-means clustering algorithm based on grid. Clustering analysis is extensively used in several applications such as pattern recognition, data mining, statistics etc. K-means approach, based on reducing a formal objective function, is most broadly used in research. But, user specification is needed for the k number of clusters and it is difficult to choose the effective initial centers. It is also very susceptible to noise data points. In this paper, the author mainly focuses on option the better initial centers to enhance the quality of k-means and to minimize the computational complexity of k-means approach. The proposed GK-means integrates grid structure and spatial index with k-means clustering approach. Theoretical analysis and experimental observation show that the proposed approach performs significantly with higher efficiency.

Trujillo *et al.,* [5] proposed a combining K-means and semivariogram-based grid clustering approach. Clustering is widely used in various applications which include data mining, information retrieval, image segmentation, and data classification. A clustering technique for grouping data sets that are indexed in the space is proposed in this paper. This approach mainly depends on the k-means clustering technique and grid clustering. K-means clustering is the simplest and most widely used approach. The main disadvantage of this approach is that it is sensitive to the selection of the initial partition. Grid clustering is extensively used for grouping data that are indexed in the space. The main aim of the proposed clustering approach is to eliminate the high sensitivity of the k-means clustering approach to the starting conditions by using the available spatial information. A semivariogram-based grid clustering technique is used in this approach. It utilizes the spatial correlation for obtaining the bin size. The author combines this approach with a conventional k-means clustering technique as the bins are constrained to regular blocks while the spatial distribution of objects is irregular. An effective initialization of the k-means is provided by semivariogram. From the experimental results, it is clearly observed that the final partition protects the spatial distribution of the objects.

Huang *et al.,* [6] put forth the automated variable weighting in k-means type clustering that can automatically estimate variable weights. A novel approach is introduced to the k-means algorithm to iteratively update variable weights depending on the present partition of data and a formula for weight calculation is also proposed in this paper. The convergency theorem of the new clustering algorithm is given in this paper. The variable weights created by the approach estimates the significance of variables in clustering and can be deployed in variable selection in various data mining applications where large and complex real data are often used. Experiments are conducted on both synthetic and real data and it is found from the experimental observation that the proposed approach provides higher performance when compared the traditional k-means type algorithms in recovering clusters in data.

III. METHODOLOGY

K-Means Clustering

Given a data set of data samples, a desired number of clusters, k, and a set of k initial starting points, the k-means clustering algorithm finds the desired number of distinct clusters and their centroids. A centroid is defined as the point whose coordinates are obtained by computing the average of each of the coordinates (i.e., feature values) of the points of the jobs assigned to the cluster. Formally, the k-means clustering algorithm follows the following steps.

Step 1: Choose a number of desired clusters, k.

Step 2: Choose k starting points to be used as initial estimates of the cluster centroids. These are the initial starting values.

Step 3: Examine each point in the data set and assign it to the cluster whose centroid is nearest to it.

Step 4: When each point is assigned to a cluster, recalculate the new k centroids.

Step 5: Repeat steps 3 and 4 until no point changes its cluster assignment, or until a maximum number of passes through the data set is performed.

The time complexity is higher for using the K-Means algorithm if large dataset is used for clustering. Therefore, dealing with high dimensional data using clustering techniques, evidently a complicated task in terms of higher number of data included. For the purpose of improving the efficiency, the noisy and outlier data may be rejected and reduce the execution time and it is necessary to decrease the no. of variables in the original data set. Principle Component Analysis is a common method for identifying patterns in high dimensional data. This paper uses Kernel Principle Component Analysis (KPCA) for dimensionality reduction. This will help in better initialization of centroids for clustering.

The proposed technique performs data clustering with the help of Principal component obtained from KPCA. It clusters the provided data set into k sets. The median of every set can be used as better initial cluster centers and then assign every data points to its nearest cluster centroid. The Proposed method is illustrated in Figure 1.

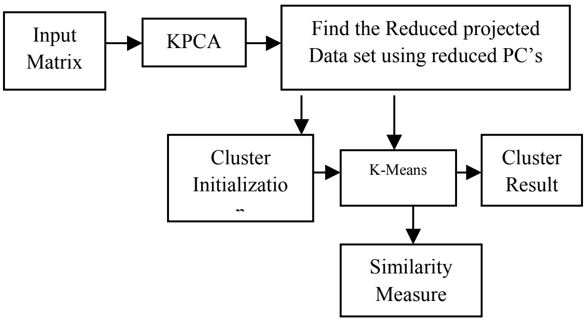

Figure 1: Proposed Method

The proposed technique involves the following algorithm for clustering.

Algorithm 1: The proposed method

Step 1: Reduce the dimension of the data into d dimension and determine the initial centroid of the clusters by using Algorithm 2.

Step 2: Assign each data point to the appropriate clusters by K-Means algorithm.

In the above provided algorithm the data dimensions are decreased and the initial centroids are identified systematically so as to construct clusters with better accuracy.

Algorithm 2: Dimension reduction and finding the initial centroid using KPCA.

Step 1: Reduce the D dimension of the N data using Kernal Principal Component Analysis (KPCA) and prepare another N data with d dimensions (d<D).

Step 2: The Principal components are ordered by the amount of variance.

Step 3: Choose the first principal component as the principal axis for partitioning and sort it in ascending order.

Step 4: Divide the Set into k subsets where k is the number of clusters.

Step 5: Find the median of each subset.

Step 6: Use the corresponding data points for each median to initialize the cluster centers.

The initial centroids of the clusters are supplied as input to K-Means. It initiates by forming the initial clusters according to the relative distance of every data-point from the initial centroids. The Euclidean distance is helpful in identifying the closeness of every data point to the cluster centroids. For every data-point, the cluster to which it is assigned and its distance from the centroid of the nearest cluster are noted. For every cluster, the centroids are recomputed by taking the mean of the values of its data-points. This computation of initial centroids will help in better clustering result and also reduces the time complexity.

Kernel Principal Component Analysis

Principal Component Analysis (PCA) is a basis transformation to diagonalize an estimate of the covariance matrix of the data $x_k, k = 1, \dots . l, x_k \in \mathrm{R}^N$, $\sum_{k=1}^{l} x_k = 0$, defined as

$$C = \frac{1}{l}\sum_{j=1}^{l} x_j x_j^T \qquad (1)$$

The new coordinates in the Eigenvector basis, i.e. the orthogonal projections onto the Eigenvectors, are called principal components.

This paper generalizes this setting to a nonlinear one of the following kind. Suppose initially the data nonlinearly is mapped into a feature space F by

$$\Phi: \mathrm{R}^N \rightarrow F, x \rightarrow \mathrm{X} \qquad (2)$$

PCA can be still performed in arbitrarily large dimensionality for certain choices of Φ by using the kernel functions.

Assume for the moment that the data mapped into feature space, $\Phi(x_1), \dots, \Phi(x_l)$ is centered, i.e. $\sum_{k=1}^{l} \Phi(x_k) = 0$. To do PCA for the covariance matrix

$$C = \frac{1}{l}\sum_{j=1}^{l} \Phi(x_j)\,\Phi(x_j{}^{\mathrm{T}}) \qquad (3)$$

It is necessary to find Eigenvalues $\lambda \geq 0$ and Eigenvectors $V \in F \backslash \{0\}$ satisfying XV=CV. Substituting (3), it is noted that all solutions V lie in the span of $\Phi(x_1), \dots, \Phi(x_l)$. This implies that the equivalent system may consider is

$$\lambda(\Phi(x_k).V) = (\Phi(x_k).CV) \text{ for all } k = 1, \dots, l \qquad (4)$$

and that there exist coefficients $\alpha_1, \dots, \alpha_l$ such that

$$V = \sum_{i=1}^{l} \alpha_i \Phi(\mathrm{X_i}) \qquad (5)$$

Substituting (3) and (5) into (4), and defining an l X l matrix K by

$$K_{ij} := \left(\Phi(x_i).\Phi(x_j)\right) \qquad (6)$$

Obtained

$$l\lambda K\alpha = K^2\alpha \qquad (7)$$

where α denotes the column vector with entries $\alpha_1, \dots, \alpha_l$. To identify solutions of (7), the Eigenvalue problem is solved as

$$l\lambda\alpha = K\alpha \qquad (8)$$

for nonzero Eigenvalues. Clearly, all solutions of (8) do satisfy (7). Moreover, it can be shown that any additional solutions of (8) do not make a difference in the expansion (5) and thus are not interesting.

The solutions α^k belonging to nonzero Eigenvalues are normalized by requiring that the corresponding vectors in F be normalized, i.e. is $(V^k.V^k)=1$. By virtue of (5), (6) and (8), this translates into

$$\begin{aligned} l &= \sum_{i,j=1}^{l} \alpha_i^k \alpha_j^k \left(\Phi(x_i).\Phi(x_j)\right) = (\alpha^k.K\alpha^k) \\ &= \lambda_k(\alpha^k.\alpha^k) \end{aligned} \qquad (9)$$

For principal component extraction, we compute projections of the image of a test point Φ(x) onto the Eigenvector V^k in F according to

$$\left(V^k.\Phi(X)\right) = \sum_{i=1}^{l} \alpha_i^k(\Phi(\mathrm{X_i}).\Phi(\mathrm{X})) \qquad (10)$$

Note that neither (6) nor (10) requires the $\Phi(x_i)$ in explicit form - they are only needed in dot products. Therefore, we are able to use kernel functions for computing these dot products without actually performing the map Φ. For some choices of a kernel k(x, y), it can be shown by methods of functional analysis that there exists a map Φ into some dot product space F (possibly of infinite dimension) such that k computes the dot product in F. Kernels which have successfully been used in Support Vector Machines include polynomial kernels

$$k(x,y) = (x.y)^d \qquad (11)$$

Radial basis function $k(x,y) = exp(-\|x-y\|^2/(2\sigma^2))$, and sigmoid kernels $k(x,y) = tanh(k(x,y) + \Theta)$ It can be shown that polynomial kernels of degree d correspond to a map Φ into a feature space which is spanned by all products of d entries of an input pattern, e.g., for the case of N = 2; d = 2,

$$(x.y)^2 = (x_1^2, x_1x_2, x_2x_1, x_2^2)(y, y_1y_2, y_2y_1, y_2^2)^T \qquad (12)$$

Substituting kernel functions for all occurrences of (Φ (x). Φ (y)), the following algorithm is obtained for kernel PCA (Fig. 2): the dot product matrix is computed (cf. Eq. (6)) K_{ij} = (k(x_i, x_j))$_{ij}$, solve (8) by diagonalizing K, normalize the Eigenvector expansion coefficients α^n by requiring Eq. (9), and extract principal components (corresponding to the kernel k) of a test point x by computing projections onto Eigenvectors (Eq. (10), Fig. 3).

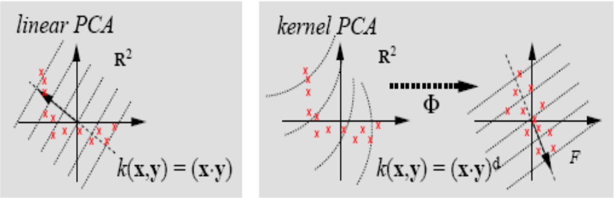

Figure 2: Basic idea of kernel PCA: by using a nonlinear kernel function k instead of the standard dot product, we implicitly perform PCA in a possibly high-dimensional space F which is nonlinearly related to input space. The dotted lines are contour lines of constant feature value.

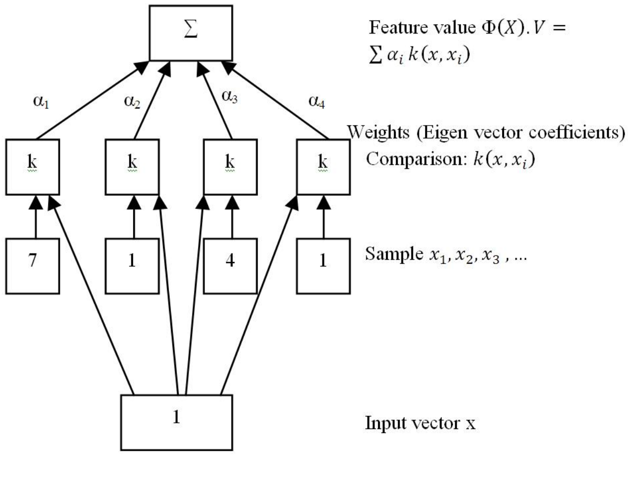

Figure 3: Kernel PCA feature extraction for an OCR task (test point x, Eigenvector V).

IV. Experimental Results

The experiment is conducted on iris data sets from UCI machine learning repository [15] for evaluating the proposed clustering algorithm. The Iris flower data set (Fisher's Iris data set) is a multivariate data set. The dataset comprises of 50 samples from each of three species of Iris flowers (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from every sample; they are the length and the width of sepal and petal, in centimeters. Based on the combination of the four features, Fisher has developed a linear discriminant model to distinguish the species from each other. It is used as a typical test for many classification techniques. This database has four continuous features consisting of 150 instances: 50 for each class. The initial centroid for standard k-means algorithm is selected randomly. The experiment is conducted 8 times for different sets of values of the initial centroids, which are selected randomly. In each experiment, the accuracy and time was computed and taken the average accuracy and time of all experiments.

Table 1 Resulted Accuracy for Iris Dataset

No. of Cluster	Algorithm	Run	Accuracy (%)
K=3	K-Means	8	81.25
	K-Means + KPCA	1	90.95

Table 2 Execution Time for Iris Dataset

No. of Cluster	Algorithm	Run	Time (sec)
K=3	K-Means	8	0.31
	K-Means + KPCA	1	0.19

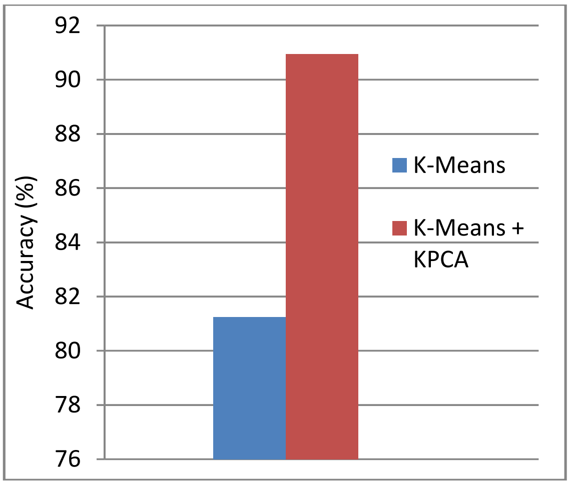

Figure 4: Resulted Accuracy for Iris Dataset

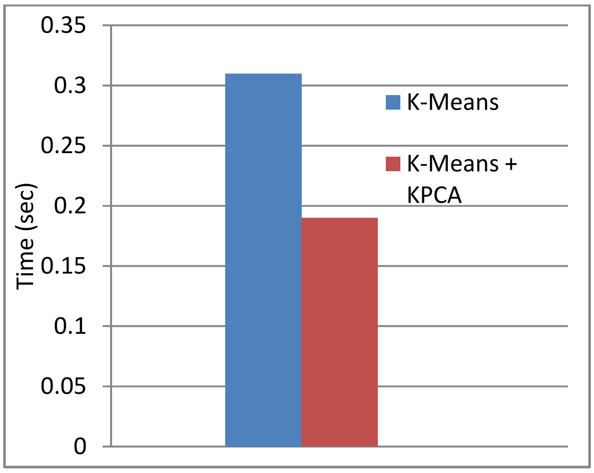

Figure 5: Resulted Accuracy for Iris Dataset

The accuracy obtained by using K-Means and proposed method is shown in table 1 and figure 4. From the table, it can be seen that the average accuracy for 8 run using the K-Means algorithm is 81.25 %, whereas the proposed clustering technique results in better accuracy of 90.95 %. The execution time for using K-Means and proposed method is shown in table 2 and figure 5. From the table, it can be seen that the average execution time for 8 run using the K-Means algorithm is 0.31 seconds, whereas the proposed clustering technique take only lesser time for i.e., 0.19. These results show that the proposed technique results in better accuracy in less time when compared to conventional K-Means clustering technique.

V. CONCLUSION

The need of analyzing and grouping of data is required for better understanding and examination. This can be solved by using the clustering technique which groups the similar kind data into a particular cluster. One of the most commonly and widely used clustering is K-Means clustering because of its simplicity and performance. The initial centroid for clustering is generated randomly before clustering. If the dataset used is large, then the performance of K-Means will be reduced and also the time complexity is increased. To overcome this problem, this paper focuses on altering the initial cluster centroid effectively. For this purpose, Kernel Principal Component Analysis (KPCA) is used in this paper. The experimental result shows that the proposed technique results in better accuracy and also the time complexity is reduced

REFERENCES

[1] Zhang Zhe, Zhang Junxi and Xue Huifeng, "Improved K-Means Clustering Algorithm", Congress on Image and Signal Processing, Vol. 5, Pp. 169-172, 2008.

[2] Hai-xiang Guo, Ke-jun Zhu, Si-wei Gao and Ting Liu, "An Improved Genetic k-means Algorithm for Optimal Clustering", Sixth IEEE International Conference on Data Mining Workshops, Pp. 793-797, 2006.

[3] Yanfeng Zhang, Xiaofei Xu and Yunming Ye, "NSS-AKmeans: An Agglomerative Fuzzy K-means clustering method with automatic selection of cluster number", 2nd International Conference on Advanced Computer Control, Vol. 2, Pp. 32-38, 2010.

[4] Xiaoyun Chen, Youli Su, Yi Chen and Guohua Liu, "GK-means: an Efficient K-means Clustering Algorithm Based on Grid", International Symposium on Computer Network and Multimedia Technology, Pp. 1-4, 2009.

[5] Trujillo, M., Izquierdo, E., "Combining K-means and semivariogram-based grid clustering", 47th International Symposium, Pp. 9-12, 2005.

[6] Huang, J.Z., Ng, M.K., Hongqiang Rong and Zichen Li, "Automated variable weighting in k-means type clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 5, Pp. 657-668, 2005.

[7] Mantao Xu nd Franti, P., "A heuristic K-means clustering algorithm by kernel PCA", 2004 International Conference on Image Processing, Vol. 5, Pp. 3503-3506, 2004.

[8] Muhr, M. and Granitzer, M., "Automatic Cluster Number Selection Using a Split and Merge K-Means Approach", 20th International Workshop on Database and Expert Systems Application, Pp. 363-367, 2009.

[9] Donghai Guan, Weiwei Yuan, Young-Koo Lee, Andrey Gavrilov and Sungyoung Lee, "Combining Multi-layer Perceptron and K-Means for Data Clustering with Background Knowledge",Advanced Intelligent Computing Theories and Applications, Springer-Verlag, Vol. 2, Pp. 1220-1226, 2007.

[10] Chen Zhang and Shixiong Xia, "K-means Clustering Algorithm with Improved Initial Center", Second International Workshop on Knowledge Discovery and Data Mining, Pp. 790-792, 2009.

[11] Chris Ding and Xiaofeng He, “k-means Clustering via Principal component Analysis”, In Proceedings of the 21st international conference on Machine Learning, Banff, Canada, 2004.

[12] Fahim A.M,Salem A.M, Torkey A and Ramadan M.A: An Efficient enchanced k-means clustering algorithm,Journal of Zhejiang University,10(7): 1626-1633,2006.

[13] Fahim A.M,Salem A.M, Torkey F. A., Saake G and Ramadan M.A: An Efficient k-means with good initial starting points, Georgian Electronic Scientific Journal: Computer Science and Telecommunications, Vol.2, No. 19,pp. 47-57, 2009.

[14] Ismail M. and Kamal M.: Multidimensional data clustering utilization hybrid search strategies,Pattern Recognition Vol. 22(1),PP. 75-89, 1989.

[15] Merz C and Murphy P, UCI Repository of Machine Learning Databases, Available: ftp://ftp.ics.uci.edu/pub/machine-Learning-databases.

[16] Bernhard Scholkopf, Alexander Smola, Klaus-Robert Muller, "Kernel Principal Component Analysis", 1999.