A Modified Projected K-Means Clustering Algorithm with Effective Distance Measure

B. Shanmugapriya
Lecturer,
Department of Computer Science,
Sri Ramakrishna College of Arts
and Science for Women,
Coimbatore. India.

M. Punithavalli
Director,
Department of Computer Applications
Sri Ramakrishna Engineering College,
Coimbatore. India

ABSTRACT
Clustering high dimensional data has been a big issue for clustering algorithms because of the intrinsic sparsity of the data points. Several recent research results signifies that in case of high dimensional data, even the notion of proximity or clustering possibly will not be significant. K-Means is one of the basic clustering algorithm which is commonly used in several applications, but it is not possible to discover subspace clusters. The subspaces are explicit to the clusters themselves. In this paper, an algorithm called Modified Projected K-Means Clustering Algorithm with Effective Distance Measure is designed to generalize K-Means algorithm with the objective of managing the high dimensional data. The experimental results confirm that the proposed algorithm is an efficient algorithm with better clustering accuracy and very less execution time than the Standard K-Means and General K-Means algorithms.

KEYWORDS
Data Mining, Projected Clustering, K-Means, High Dimensional Data, General K-Means, Efficient Projected Clustering (EPC)

1. INTRODUCTION
One of the most popular data mining approaches which are adequate for numerous applications is clustering. The major reason for its wide range of application is the capability of clustering technique to work on datasets with least or no previous knowledge. This enables clustering as a convenient tool for many real world applications. Clustering is a technique of grouping comparable objects that are similar to each other and dissimilar to the data objects belonging to other clusters based on certain features [1]. Clustering is exploited to assemble items that appear to come naturally together [2]. Several kinds of clustering techniques are available, namely hierarchical vs. partitioned, exclusive vs. overlapping vs. fuzzy and complete vs. partial [3]. Clustering is a kind of unsupervised learning technique that separates data in such a way that comparable data items are assembled together in a set which are referred to as clusters. This technique is essential for condensing and recognizing patterns in data [4].

In recent times, high dimensional data has stimulated the attention of database researchers because of its significant challenges brought to the research community. In huge dimensional space, the distance between a record and its nearest neighbor can approach its distance to the outermost record [5]. In the framework of clustering, the difficulty causes the distance among two records of the same cluster to move toward the distance among two records of various clusters. Conventional clustering approaches possibly will be unsuccessful to recognize the accurate clusters.

Clustering is obviously in need of several techniques to the categorization and association learning approach. Subspace clustering and projected clustering are current research topics in the field of high dimensional space clustering. On the other hand, in high dimensional datasets, conventional clustering approaches are likely to fail based on both accuracy and efficiency [6].

As discussed above, projected clustering has turned out to be a hot research topic because of its ability to cluster huge-dimensional data [7]. Even though, several projected clustering approaches are available based on certain significant user parameters, when incorrect parameter values are taken into consideration, it leads to severe degradation in the performance of that particular clustering technique. In fact, accurate parameter values are not often recognized in real datasets.

In this paper, a modified projected K-Means clustering algorithm with effective distance measure has been proposed that continuously optimizes a comprehensive objective function. In the objective function of this proposed algorithm, an effetive distance measure is used for providing better clustering results in high dimensional data. In order to avoid the value of the objective function from decreasing as a consequence of the exclusion of dimensions, virtual dimensions are incorporated with the objective function. The two necessities for the data values on these virtual dimensions guarantee that the objective function attains its minimum when the real subspace clusters or the clusters in original space are found [8-10]. The standard K-Means clustering approach can be considered as a special case of GKM.

The performance of the proposed algorithm is evaluated using the two standard datasets of UCI machine learning repository, namely Iris and Wine dataset based on the clustering accuracy and execution time. Its confusion matrices are also provided at the end of experimental results section.

2. LITERATURE SURVEY
Projected clustering has come out as a potential solution to the disputes related to clustering in huge dimensional data. A projected cluster is a subset of points in cooperation with a subset of features, in a manner that the cluster points project onto a least range of values in all the features, and are equally distributed in other features. Existing approaches for projected clustering based on the constraints whose correct values are hard to fix by the user, or are incapable to recognize projected clusters with a small number of appropriate features. Moise et al., [11] developed a strong algorithm for projected clustering that can efficiently determine projected clusters in the data at the same time as reducing the number of parameters needed as input. In contradiction of all existing techniques, this approach can determine, under very common circumstances, the accurate number of projected clusters.

In huge-dimensional data, clusters can be present in subspaces that cover themselves from conventional clustering approaches.

The clustering accuracy can be considerably reduced if inaccurate values are used. In real circumstances, it is not often possible for users to provide the parameter values precisely, which causes sensible difficulties in implementing these approaches to real data. Yip et al., [12] examined the chief confronts of projected clustering and recommended that these approaches are required to be based deeply on user parameters. According to this investigation, a new approach has been developed that utilizes the clustering status to change the internal thresholds dynamically not including the user parameters.

Ada Wai-chee Fu et al., [13] recommended the EPC (Efficient Projected Clustering) algorithm to determine the sets of interrelated dimensions and the position of the clusters. This approach is comparatively different from existing techniques and has the following merits: (a) there is no condition on the input concerning the amount of natural clusters and the average cardinality of the subspaces; (b) it can manage clusters of uneven shapes; (c) it generates better clustering outputs compared to the best existing method; (d) it has better scalability.

Irrelevant attributes accumulates noise to high-dimensional clusters and makes conventional clustering techniques unsuitable. In recent times, several approaches that determine projected clusters and their related subspaces have been developed. The similarity among mining frequent itemsets and determining dense projected clusters around random points are considered. According to this, Man Lung Yiu et al., [14] developed a technique that enhances the effectiveness of a projected clustering algorithm (DOC). This method is an optimized adaptation of the frequent pattern tree growth technique exploited for mining frequent itemsets. Numerous methods that make use of the branch and bound model to effectively determine the projected clusters have also been developed.

3. METHODOLOGY

3.1 The General Objective Function for Projected Clustering

For projected clustering consider $\boldsymbol{X} = \{\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_n}\}$ as a collection of $\boldsymbol{n}$ data points in a d-dimensional space $\mathbb{R}^{\boldsymbol{d}}$ with dimensions $\boldsymbol{D} = \{\boldsymbol{D_1}, \boldsymbol{D_2}, \dots, \boldsymbol{D_d}\}$, in which data point $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_d}\}$. A subspace of $\mathbb{R}^d$ is the space spanned by the dimensions in set D_i, where $D_i \subseteq D$. When the subsets, $X_i \subseteq X, 1 \le i \le k$ satisfy the following condition $X_1 \cup X_2 \cup \dots X_k = X$ and $X_i \cap X_j = \Phi$, $i \ne j$ for each i and j, $P = \{X_1, X_2, \dots, X_k\}$ is called a k–partition of X. The major objective of projected clustering is to obtain a k-partition of X in such a way that the data points in all data subsets X_i form a cluster in the subspace spanned by the dimensions in D_i that is related to X_i. This X_i is considered as a cluster in X.

The objective function of the conventional K-Means clustering approach is a squared-error function,

$$E(C) = \sum_{i=1}^{k} \sum_{X \in X_i} \sum_{j=1}^{d} \left(x_j - c_{ij}\right)^2 \quad (1)$$

where $C_i = (c_{i1}, c_{i2}, \dots, c_{id})$ is the average of the data points in subset X_i and $C = \{C_1, C_2, \dots, C_k\}$. The K-Means clustering approach focuses on discovering a k-partition of X that can minimize E. In this objective function, all dimensions take part in the computation. As a result the objective function implies an assumption that all clusters locate in original space.

The distance measurement that is controlled by local and non-local information is given as

$$(x_j - c_{ij})^2 = \left(1 - \lambda_j\right) d_l^2\left(x_j, c_{ij}\right) + \lambda_j d_{nl}^2\left(x_j, c_{ij}\right) \quad (2)$$

Using the above equation, the squared-error function can be modified as,

$$E(C) = \sum_{i=1}^{k} \sum_{X \in X_i} \sum_{j=1}^{d} \left(1 - \lambda_j\right) d_l^2\left(x_j, c_{ij}\right) + \lambda_j d_{nl}^2\left(x_j, c_{ij}\right) \quad (3)$$

where d_l denotes the distance measurement controlled by local information, and d_{nl} represents the distance measurement controlled by non-local information, λ_j from zero to one, is the weighting factor which manages the tradeoff between them.

For a subspace cluster $\boldsymbol{X_i}$, the values of data points on an irrelevant dimension should not be used to recognize the cluster. As a result the squared-error of the data points in $\boldsymbol{X_i}$ on the irrelevant dimensions should not be incorporated in the objective function. In order to recognize the clusters which are embedded in various subspaces, a weight vector $W_i = (w_{i1}, w_{i2}, \dots, w_{id})$ is integrated with each cluster X_i, where

$$w_{ij} = \begin{cases} 1, if\ dimension\ j\ is\ relevant\ to\ cluster\ X_i \\ 0, if\ dimension\ j\ is\ irrelevant\ to\ cluster\ X_i \end{cases}$$

Thus a new objective function is obtained as below,

$$E'(W, C) = \sum_{i1}^{k} \sum_{X \in X_i} \sum_{j=1}^{d} w_{ij} \begin{bmatrix} \left(1 - \lambda_j\right) d_l^2\left(x_j, c_{ij}\right) \\ +\lambda_j d_{nl}^2\left(x_j, c_{ij}\right) \end{bmatrix} \quad (4)$$

where $W = \{W_1, W_2, \dots, W_k\}$. It is clear that the value of this function will decrease when any weight w_{ij} is transformed from 1 to 0, without considering whether dimension j is applicable to cluster X_i. With the intention of solving this difficulty, an idea of virtual dimensions for each cluster is introduced.

For a cluster X_i, consider D_i' as the dimension subset of its irrelevant dimensions. The dimensions in D_i' are replaced by virtual dimensions in set VD_i, in which $|VD_i| = |D_i'|$. The dimensions in VD_i are called as virtual dimensions because they will not be present in the subspace that contains X_i and the values of data points in X_i on them are allotted artificially. The chief purpose of these virtual dimensions is to guarantee that the objective function attains its minimum when the clusters are exactly identified in their respective subspaces.

Thus the new objective function is obtained as follows,

$$\begin{aligned} E(W, C) = \sum_{i=1}^{k} \sum_{X \in X_i} \sum_{j=1}^{d} \Big(& w_{ij} \left[\left(1 - \lambda_j\right) d_l^2\left(x_j, c_{ij}\right) \right. \\ & \left. + \lambda_j d_{nl}^2\left(x_j, c_{ij}\right) \right] \\ & + \left(1 - w_{ij}\right) \left[\left(1 - \lambda_j'\right) d_l'^2\left(x_j', c_{ij}'\right) \right. \\ & \left. + \lambda_j' d'^2_{nl}\left(x_j', c_{ij}'\right) \right] \Big) \end{aligned} \quad (5)$$

where x_j' of all X is a virtual value that is used to avoid eliminating relevant dimensions. Set $x_j = c_{ij} = 0$, when $w_{ij} = 1$. In fact, this will not produce any consequence in the results of the algorithm.

The function of the second item in $E(W, C)$ is to increase the value of E every time an appropriate dimension of a cluster is taken into account as an irrelevant one. The procedure that reduces this objective function to discover accurate relevant dimensions for each cluster, $E(W, C)$ should satisfy two requirements as in [15].

1) When dimension D_j is relevant to cluster X_i,

$$\sum_{X\in X_i}(1-\lambda_j)d_l^2(x_j,c_{ij})+\lambda_j d_{nl}^2(x_j,c_{ij}) \le \sum_{X\in X_i}(1-\lambda_j^{'})d_l^{'2}(x_j^{'},c_{ij}^{'})+\lambda_j^{'} d_{nl}^{'2}(x_j^{'},c_{ij}^{'})$$

2) When dimension D_j is irrelevant to cluster X_i,

$$\sum_{X\in X_i}(1-\lambda_j)d_l^2(x_j,c_{ij})+\lambda_j d_{nl}^2(x_j,c_{ij}) > \sum_{X\in X_i}(1-\lambda_j^{'})d_l^{'2}(x_j^{'},c_{ij}^{'})+\lambda_j^{'} d_{nl}^{'2}(x_j^{'},c_{ij}^{'})$$

$E(W,C)$ is a general objective function as it can be used to discover both clusters embedded in subspaces and clusters embedded in the original space. This cannot be realized by the existing projected clustering algorithms.

3.2 Modified Projected K-Means Clustering Algorithm

3.2.1 Setting Virtual Values

As perfect irrelevant dimensions of a cluster are represented as dimensions on which the projections of data points in the cluster distribute uniformly in the complete value range, the irrelevant dimensions are given as follows. Suppose already there is a k-partition, then for every cluster X_i and each dimension D_j, compute

$$\sum_{X\in X_i}(1-\lambda_j)d_l^2(x_j,c_{ij})+\lambda_j d_{nl}^2(x_j,c_{ij})$$

and

$$\sum_{X\in X_i}(1-\lambda_j^{'})d_l^{'2}(x_j^{'},c_{ij}^{'})+\lambda_j^{'} d_{nl}^{'2}(x_j^{'},c_{ij}^{'})$$

where x_j denotes the value of data point X on dimension j, c_{ij} represents the mean of data points in cluster X_i on dimension D_j, $x_j^{'}$ indicates the virtual value of data point X on dimension j , $c_{ij}^{'}$ denotes the virtual mean of data points in cluster X_i on dimension D_j. It is unnecessary to set virtual values for each data point, only the sum should be calculated. If n_i data points distribute uniformly on dimension D_j in the range of [min,max], the sum is:

$$\sum_{X\in X_i}(1-\lambda_j^{'})d_l^{'2}(x_j^{'},c_{ij}^{'})+\lambda_j^{'} d_{nl}^{'2}(x_j^{'},c_{ij}^{'}) = \begin{cases} 2\left(\frac{max-min}{n_i-1}\right)^2\left[0.5^2+1.5^2+\cdots+\left(\frac{n_i-1}{2}\right)^2\right], \\ when\ n_i\ is\ an\ even\ number \\ 2\left(\frac{max-min}{n_i-1}\right)^2\left[1^2+2^2+\cdots+\left(\frac{n_i-1}{2}\right)^2\right], \\ when\ n_i\ is\ an\ odd\ number \end{cases}$$

If $\sum_{X\in X_i}(1-\lambda_j)d_l^2(x_j,c_{ij})+\lambda_j d_{nl}^2(x_j,c_{ij}) < \sum_{X\in X_i}(1-\lambda_j^{'})d_l^2(x_j^{'},c_{ij}^{'})+\lambda_j^{'} d_{nl}^2(x_j^{'},c_{ij}^{'})$, dimension D_j is relevant to cluster X_i, and if $\sum_{X\in X_i}(1-\lambda_j)d_l^2(x_j,c_{ij})+\lambda_j d_{nl}^2(x_j,c_{ij}) = \sum_{X\in X_i}(1-\lambda_j^{'})d_l^2(x_j^{'},c_{ij}^{'})+\lambda_j^{'} d_{nl}^2(x_j^{'},c_{ij}^{'})$, dimension D_j is irrelevant to cluster X_i. However, in real dataset, the allocation of the projections of data points on irrelevant dimension will be more complicated than uniform distribution. A parameter ε is introduced to solve this difficulty. If $\sum_{X\in X_i}(1-\lambda_j)d_l^2(x_j,c_{ij})+\lambda_j d_{nl}^2(x_j,c_{ij}) < \varepsilon.\sum_{X\in X_i}(1-\lambda_j^{'})d_l^2(x_j^{'},c_{ij}^{'})+\lambda_j^{'} d_{nl}^2(x_j^{'},c_{ij}^{'})$, dimension D_j is appropriate to cluster X_i, and if $\sum_{X\in X_i}(1-\lambda_j)d_l^2(x_j,c_{ij})+\lambda_j d_{nl}^2(x_j,c_{ij}) \ge \varepsilon.\sum_{X\in X_i}(1-\lambda_j^{'})d_l^2(x_j^{'},c_{ij}^{'})+\lambda_j^{'} d_{nl}^2(x_j^{'},c_{ij}^{'})$, dimension D_j is irrelevant to cluster X_i. The range of ε is (0,1) . It is simple to select ε. When the sum of squared-errors of virtual values is set, an appropriate ε will guarantee that $E(W,C)$ satisfies the two requirements mentioned previously.

3.2.2 The Modified Algorithm

Similar to conventional K-Means clustering approach, the proposed algorithm arbitrarily selects k data points in X as the primary cluster centers. Each cluster center C_i is related to a vector W_i whose components equal to one. Then the proposed algorithm repeats the following two steps to optimize the objective function $E(W,C)$.

1) Allocate each data point in X to the nearest cluster. This results in a k-partition. The distance between a data point X and a cluster X_i is given as below,

$$dis(X,X_i) = \sqrt{\sum_{j=1}^{d} w_{ij}\left[(1-\lambda_j)d_l^2(x_j,c_{ij})+\lambda_j d_{nl}^2(x_j,c_{ij})\right]/\sum_{j=1}^{d} w_{ij}}$$

2) Update C_i and W_i for $X_i, 1\le i\le k$. $c_{ij}=\sum_{X\in X_i} x_j/|X_i|$, $1\le j\le d$, in which $|X_i|$ is the number of data points in X_i. $w_{ij}=1$, when $\sum_{X\in X_i}(1-\lambda_j)d_l^2(x_j,c_{ij})+\lambda jdnl2xj,cij<\varepsilon.X\in Xi1-\lambda j' dl2xj',cij'+\lambda j' dnl2xj',cij'$; $w_{ij}=0$, when $\sum_{X\in X_i}(1-\lambda_j)d_l^2(x_j,c_{ij})+\lambda jdnl2xj,cij\ge\varepsilon.X\in Xi1-\lambda j' dl2xj',cij'+\lambda j' dnl2xj',cij'$.

After getting a k-partition, initially the means of the clusters are computed. Subsequently the weight vector of each cluster is found based on these means. The above mentioned two steps are continuously repeated until the partition does not undergo any change.

4. EXPERIMENTAL RESULTS

To evaluate the proposed algorithm (Modified Projected K-Means with effective distance measure), experiments were carried out using UCI Machine Learning Repository. There are 211 data sets available in the UCI Machine Learning Repository. For the purpose of evaluating the proposed algorithm against the standard K-Means and General K-Means algorithm, few standard data sets were selected from the available UCI machine learning repository data sets. Following are the data sets used for the evaluation purpose; Iris and Wine dataset.

The performance of the proposed algorithm is evaluated against the Standard K-Means and General K-Means based on the following parameters: clustering accuracy and execution time. The confusion matrices of all the three approaches are also provided.

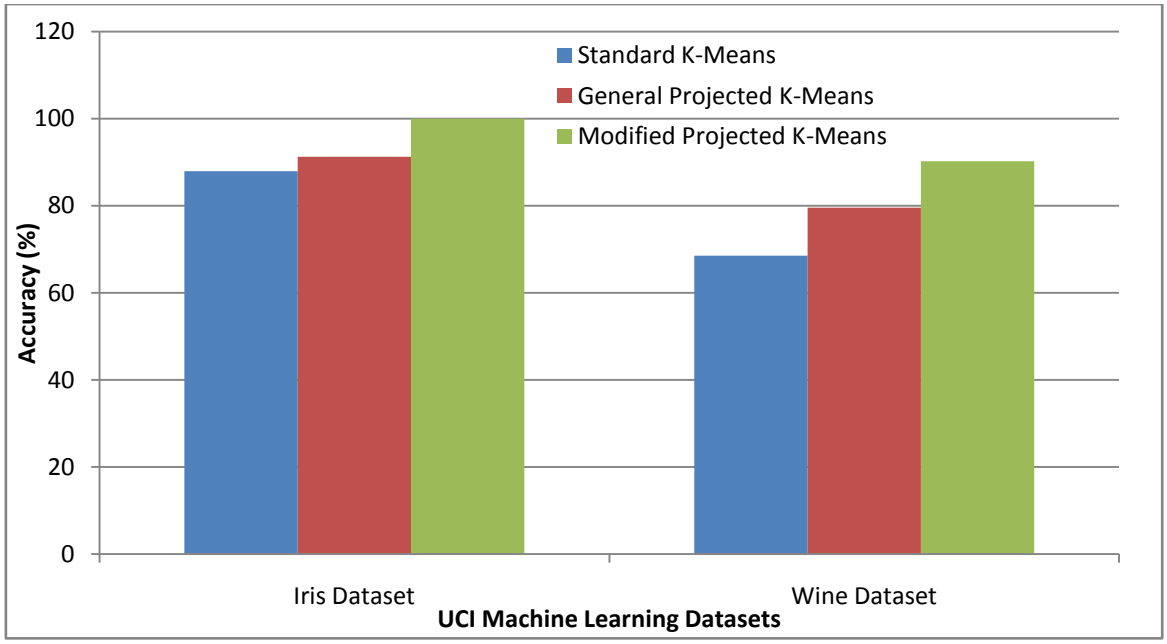

Figure. 1 Comparison of Clustering Accuracy

4.1 Clustering Accuracy

Clustering accuracy is calculated for Standard K-Means, General K-Means and the Modified Projected K-Means using iris and wine datasets. Figure.1 shows the comparison of the accuracy of clustering results of the three approaches in iris and wine datasets. From the figure, it can be observed that for both the datasets, the accuracy of clustering results using standard K-Means and General K-Means is very low than that of the Modified Projected K-Means which obtains 100% accuracy in iris dataset and 90.249% in wine dataset. It is clear from the figure that Modified Projected K-Means with effective distance measure is better than the other two approaches.

4.2 Execution Time

Figure. 2 show the execution time taken by the Standard K-Means, General K-Means and the Modified Projected K-Means using iris and wine datasets. It can be observed that the time required for execution using the Modified Projected K-Means for iris data set is 0.146 seconds and for wine dataset is 0.177 seconds, whereas, more time is required for other two clustering techniques for execution. From Figure.2, it is obvious that the proposed algorithm takes very less execution time than the Standard K-Means and General K-Means algorithms.

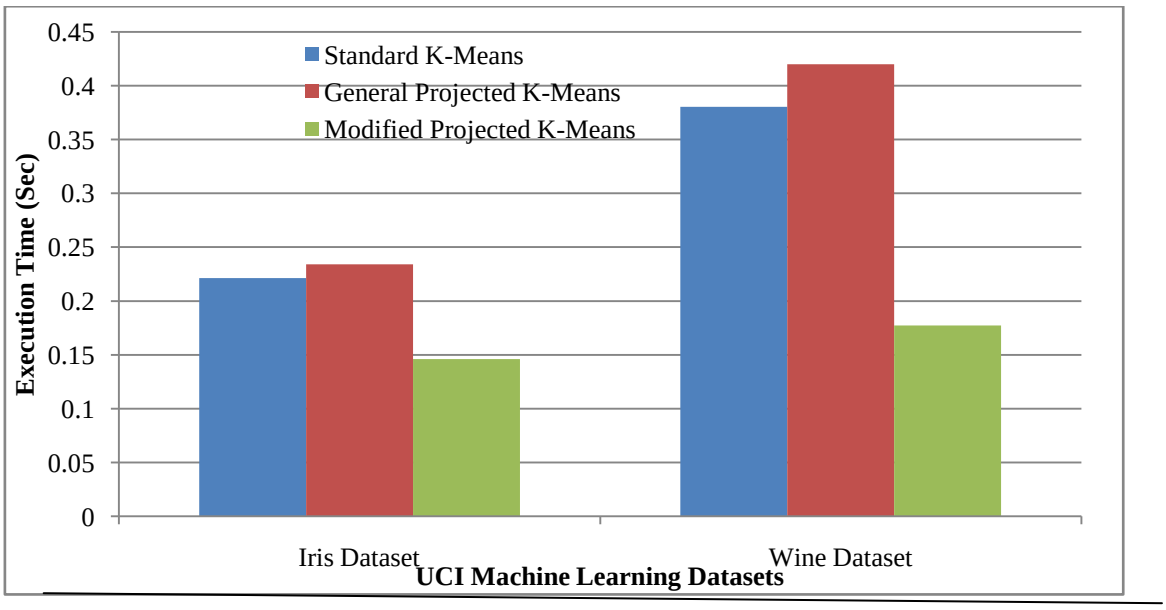

Figure. 2 Comparison of Execution Time

4.3 Confusion Matrix

The confusion matrices of the Standard K-Means, General K-Means and the Modified Projected K-Means with effective distance measure for both iris and wine dataset are provided in Table 1.

Table 1. Confusion matrices of three approaches for iris and wine datasets

	Standard K-Means	General K-Means	Modified Projected K-Means
Iris Dataset	$\begin{bmatrix} 0 & 50 & 0 \\ 48 & 0 & 2 \\ 14 & 0 & 36 \end{bmatrix}$	$\begin{bmatrix} 50 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 11 & 39 \end{bmatrix}$	$\begin{bmatrix} 50 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 0 & 50 \end{bmatrix}$
Wine Dataset	$\begin{bmatrix} 13 & 0 & 46 \\ 20 & 50 & 1 \\ 29 & 19 & 0 \end{bmatrix}$	$\begin{bmatrix} 47 & 12 & 0 \\ 0 & 50 & 21 \\ 0 & 0 & 48 \end{bmatrix}$	$\begin{bmatrix} 59 & 0 & 0 \\ 0 & 68 & 0 \\ 0 & 11 & 37 \end{bmatrix}$

From the table, it is clear that the confusion matrix of the new approach seems to be better than the Standard K-Means and General K-Means. From the above results, it is very clear that the Modified Projected K-Means provides better clustering results with quick execution time.

5. CONCLUSION

In this paper, a Modified Projected K-Means clustering algorithm with effective distance measure has been proposed that can identify natural clusters in high dimensional space, regardless of whether they are embedded in the original space or subspaces. The experimental results confirm that the proposed algorithm is an efficient algorithm with better clustering accuracy and very less execution time than the Standard K-Means and General K-Means algorithms. Modified Projected K-Means clustering algorithm with effective distance measure makes use of local and non-local information and hence provides better clustering results.

6. REFERENCES

[1] Ali Alijamaat, Madjid Khalilian and Norwati Mustapha, "A Novel Approach for High Dimensional Data Clustering", Third International Conference on Knowledge Discovery and Data Mining, pp. 264-267,

2010.

[2] Witten, Ian H and Eibe Frank, "Data Mining–Practical Machine Learning Tools and Techniques", 2nd Edition, Morhan Kaufmann, San Fransisco, 2005.

[3] Tan, Pang Nin, Michael Steinbach and Vipin Kumar, "Introduction to Data Mining", Pearson International Edition, Boston, 2006.

[4] Poncelet, Pascal, Maguelonne Teisseire and Florent Masseglia, "Data Mining Patterns: New Method and Application", London, 2008.

[5] K. Beyer, J. Goldstein, R. Ramakrishnan and U. Shaft, "When is nearest neighbor meaningful?", Lecture Notes in Computer Science, Vol. 1540, Pp. 217-235, 1999.

[6] Gabriela Moise and Jorg Sander, "Finding Non-Redundant, Statistically Significant Regions in High Dimensional Data: A Novel Approach to Projected and Subspace Clustering", Proceedings of the 14th ACM SIGKDD International Conference on Knowledge

Discovery and Data Mining, Las Vegas, Nevada, USA,

DOI: 10.1145/1401890.1401956, 2008.

[7] Charu C. Aggarwal, J. Han, J. Wang, and S. Philip Yu, "A Framework for Projected Clustering of High Dimensional Data Streams", Proc. Very Large Data Base (VLDB '04), Pp. 852-863, 2004.

[8] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications", Proc. ACM SIGMOD International conference on Management of data (SIGMOD 1998), Pp. 94-105, 1998.

[9] L. Parsons, E. Haque, and H. Liu, "Subspace Clustering for High Dimensional Data: A Review", ACM SIGKDD Explorations Newsletter, Vol. 6, Pp. 90-105, 2004.

[10] C. Böhm, K. Kailing, H. Kriegel, and P. Kröger, "Density Connected Clustering with Local Subspace Preferences," Proc. IEEE International Conference on Data Mining (ICDM 2004), Pp. 27-34, 2004.

[11] G. Moise, J. Sander and M. Ester, "P3C: A Robust Projected Clustering Algorithm", Sixth International Conference on Data Mining (ICDM '06), Pp. 414 – 425,

2006.

[12] K.Y. Yip, D.W. Cheung and M.K. Ng, "HARP: a practical projected clustering algorithm", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 11, Pp. 1387 – 1397, 2004.

[13] Ada Wai-chee Fu and E. Ng Ka Ka, "Efficient algorithm for projected clustering", Proceedings. 18th International Conference on Data Engineering, DOI: 10.1109/ICDE.2002.994727, 2002.

[14] Man Lung Yiu and N. Mamoulis, "Iterative projected clustering by subspace mining", IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 12, Pp. – 189, 2005.

[15] Yufen Sun, Gang Liu and Kun Xu, "A k-Means-Based Projected Clustering Algorithm", Third International Joint Conference on Computational Science and Optimization (CSO), Vol. 1, Pp. 466 – 470, 2010.

[16] Nguyen, Ngoc-Thanh; Trawinski, Bogdan; Jung, Jason J. (Eds.), "New Challenges for Intelligent Information and Database Systems", Vol. 251, 1st Edition, 2011.