Efficient High Dimension Data Clustering using Constraint-Partitioning K-Means Algorithm

Aloysius George
Grian Technologies Private Limited, Research and Development Company, India

Abstract: *With the ever-increasing size of data, clustering of large dimensional databases poses a demanding task that should satisfy both the requirements of the computation efficiency and result quality. In order to achieve both tasks, clustering of feature space rather than the original data space has received importance among the data mining researchers. Accordingly, we performed data clustering of high dimension dataset using Constraint-Partitioning K-Means (COP-KMEANS) clustering algorithm which did not fit properly to cluster high dimensional data sets in terms of effectiveness and efficiency, because of the intrinsic sparse of high dimensional data and resulted in producing indefinite and inaccurate clusters. Hence, we carry out two steps for clustering high dimension dataset. Initially, we perform dimensionality reduction on the high dimension dataset using Principal Component Analysis (PCA) as a preprocessing step to data clustering. Later, we integrate the COP-KMEANS clustering algorithm to the dimension reduced dataset to produce good and accurate clusters. The performance of the approach is evaluated with high dimensional datasets such as Parkinson's dataset and Ionosphere dataset. The experimental results showed that the proposed approach is very effective in producing accurate and precise clusters.*

Keywords: *Clustering, dimensionality reduction, PCA, COP-KMEANS algorithm, clustering accuracy, parkinson's dataset, ionosphere dataset*.

Received July 14, 2011; accepted December 30, 2011; published online August 5, 2012

1. Introduction

In recent times, majority of the data available throughout the world are warehoused in databases. Data mining that has received immense attention from the research community because of its importance is the process of detecting patterns from extremely huge quantities of data collection [1]. Knowledge Discovery in Databases (KDD) is the other name for data mining which has been identified as a potential field for database research [53]. Classification or bunching of these data into a set of categories or clusters is one of the essential methods in manipulating these data [10, 63]. Clustering is a delineative task that attempts to detect similar category of objects based on the implications of their features dimensions [26, 28]. One can detect the predominant distribution patterns and interesting correlations that exist among data attributes by clustering which can determine dense and sparse areas [4, 5].

Easier and faster data collection due to technology advances has given rise to larger and more complicated datasets with several objects and dimensions. Adaptations to existing algorithms are necessary to maintain cluster quality and speed as the datasets become larger and more diverse. By considering all of the dimensions of an input dataset, conventional clustering algorithms attempt to find out as much as possible about each described object. But, normally majority of the dimensions are irrelevant in high dimensional data. These irrelevant dimensions can hide clusters in noisy data and confuse the clustering algorithms. Also, differentiating similar data points from dissimilar ones becomes difficult if the distance between any two data points is approximately equal [44]. In extremely high dimensions, all of the objects are normally nearly equidistant from each other and such objects completely hide the clusters. In addition, curse of dimensionality is another factor associated with high dimensional data that makes the clustering algorithms to struggle. The distance measure also, becomes increasingly meaningless as the number of dimensions in a dataset is increased [16, 27, 47].

Creating clustering algorithm that can effectively deal with high dimensional data is not an easy task [5, 20]. By decreasing the dimension of the data, some of the researchers have recently solved the high-dimensional problem [33, 42]. As, poor classification efficiency due to high dimension of the feature space is the bottleneck of the classification task, dimensionality reduction is of great significance for the quality and efficiency of a classifier [15, 37], particularly for large-scale real-time data. Conventional and modern dimensionality reduction techniques can be broadly classified into Feature Extraction (FE) [39, 40, 46] and Feature Selection (FS) [23, 36] methods. FE methods are normally more effective than the FS techniques (except for a few specific cases) and their high effectiveness for real-world dimensionality reduction applications has been verified already [17, 23, 39, 40].

Linear and nonlinear algorithms [11] are two broad categories of the classical FE algorithms. Transforming

high-dimensional data to a lower-dimensional space by employing linear transformations along the lines of certain criteria is the objective of linear algorithms, for example like Principal Component Analysis (PCA) [29], Linear Discriminant Analysis (LDA) [45, 60], and Maximum Margin Criterion (MMC) [40]. Conversely, transforming the original data without altering selected local information by means of nonlinear transformations consistent with certain criteria [65], is the objective of nonlinear algorithms [7], for example Locally Linear Embedding (LLE) [50], ISOMAP [55], and Laplacian Eigenmaps. High dimensional data can be transformed into a low dimensional space with minimum reconstruction error by means of the unsupervised linear feature reduction method called PCA [29]. The direction of maximum variance in the data is identified by PCA [13, 31, 58, 66].

Here, we have proposed an approach for data clustering of high dimension dataset such as Parkinson's dataset and Ionosphere dataset using Constraint-Partitioning K-Means (COP-KMEANS) clustering algorithm. We at first tried to generate clusters from whole high dimension dataset but it proved ineffective because of the intrinsic sparse of high dimensional data. Hence, we are proposing an approach; where we will carry out two steps for producing clusters of high dimension dataset. At the start, we reduce the dimension i.e., attributes of the original dataset by performing the dimensionality reduction on the high dimension dataset using PCA as a preprocessing step to data clustering. Later, we integrate the COP-KMEANS clustering algorithm to the reduced dataset to produce good and accurate clusters.

The paper is organized as follow: section 2 deals with some of the recent research works related to the approach. Section 3 explains the data clustering of raw high dimension data. Section 4 describes the need for dimension reduction. Section 5 describes PCA. Section 6 illustrates the proposed data clustering for high dimension datasets with necessary mathematical formulations. Section 7 discusses the experimentation and evaluation results. Section 8 concludes the research work.

2. A Survey of Recent Research in the Field

Our approach concentrates on efficient data clustering for high dimension datasets. Several researchers have developed numerous approaches for clustering high dimension datasets. Here, we have offered some of the noteworthy researches for clustering high dimension datasets.

Amorim [5], has presented a method for clustering by employing two pair-wise rules (must link and cannot link) and a single-wise rules (cannot cluster) single-wise rule that uses extremely restricted quantity of labeled data. They have demonstrated that the precision of results could be improved by including these rules in the intelligent k-means algorithm and verified the same by means of experiments where the actual number of clusters in the data has not been previously known to the method. Jun and Xiong [30], have proposed a high-dimensional data clustering approach based on genetic algorithm, called GA-HD clustering. Their clustering approach has identified effective feature subspaces by searching the feature subspace using genetic algorithm. Binary encoded candidate features and cluster centers have been utilized and the extent of feature subspace contribution to subspace clustering has been proposed as the fitness function. The practicability and efficiency of the GA-HD clustering algorithm have been demonstrated by experimental results.

Khalilian *et al.* [33], have discussed that dimension reduction by means of vertical data reduction performed before employing clustering methods for exceedingly large and high dimensional data sets has the main disadvantage of reducing the quality of results. Still, extra carefulness has been recommended because dimensionality reduction methods unavoidably cause some loss of information or may impair the comprehensibility of the results, even disfiguring the real clusters. They have proposed a method for use in high dimensional datasets that improves the performance of the K-Means clustering method by employing divide and conquer technique with equivalency and compatible relation concepts. The proper precision and speed up of their proposed method have been proved by experimental results.

Rajput *et al.* [49], proposed a basic framework by integrating the hypothesis of rough set theory (reduct) and k-means algorithm for efficient clustering of high dimensional data. First, by discarding the superfluous attributes by means of the reduct concept of rough set theory, it has identified the low dimensional space in the high dimensional data set. Then, it has identified suitable clusters by employing the k-means algorithm on this low dimensional data reduct. The fact that the framework increases the efficiency of the clustering process and the precision of the resultant clustering has been proved by their experiment on test dataset.

Tajunisha and Saravanan [54] have discussed that the initial centroid selected as well as the dimension of the data substantially impact the quality of the resulting clusters in the computationally costly k-means clustering algorithms utilized for several practical applications. The precision of the resultant value may have not been up to the mark when the dimensions of the dataset are high because they are not sure that the selected dataset is free from noise and defect. So, efficiency and precision improvement necessitates decreasing the dimensionality of the given dataset. They have proposed a new method that identifies initial centroid and also, decreases the

dimension of the data by employing PCA to improve the precision of the cluster results.

Anaissi *et al.* [32], have proposed a framework based on FS, linear dimensionality reduction and non-linear dimensionality reduction for very high dimensional data reduction. They have proposed mutual information based FS for screening features and identifying the most appropriate features with least redundancy. The potential variables have also, been extracted from a high dimensional dataset by means of a kernel linear dimensionality reduction method. Also, the dimension has been reduced and the data has been visualized using a local linear embedding based non-linear dimensionality reduction. Outputs of each step and the efficiency of this framework have been demonstrated by means of experimental results.

Dash *et al.* [14], have discussed that preprocessing the data by means of an efficient dimensionality reduction technique is essential to improve the efficiency and accuracy of the mining task on high dimensional data. They have simplified the analysis and visualization of multi dimensional data set by using a proposed PCA method as the first phase for K-means clustering. They have also, made the algorithm more effective and efficient by identifying the initial centroids using a new proposed method. By comparing the results of their proposed approach with that of the original approach, they have proved that their proposed approach obtains more precise, simple to understand results and most importantly takes considerably less processing time.

3. Data Clustering of High Dimension Dataset using K-Means Algorithm and Constraint-Partitioning K-Means Algorithm

Clustering is process with grouping together objects which are similar to each other and dissimilar to the objects that belong to other clusters [4]. Cluster is used to assemble items that appear to fall naturally simultaneously [52, 61]. In general, the data clustering works as shown in Figure 1.

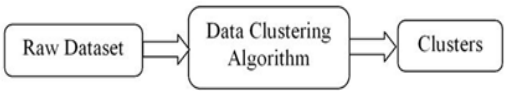

Figure 1. Data clustering of raw dataset.

Therefore, let us take the whole high dimension dataset and apply COP-KMEANS algorithm which is proved to be a clustering algorithm as described in section 3.1 to form precise clusters.

3.1. Constraint-Partitioning K-Means (COP-KMEANS)

There has been numerous wide-ranging works integrating instance-level constraints into clustering methods [6, 8, 35, 49, 59]. Two objects must be positioned into related cluster Must-Link (ML) or unrelated clusters Cannot-Link (CL) indicate instance-level constraints. Then the complete set of resulted constraints is then offered to the K-Means clustering algorithm [61]. This semi-supervised approach has led to better concert on noun phrase co reference resolution and GPS-based map refinement [47], person identification from surveillance camera clips [4] and landscape detection from hyper spectral data [41] and quite a lot of real world purposes [34].

COP-KMEANS [59] uses labeled instances as constraints to limit the K-Means clustering process [43]. All pairs of labeled instances are marked as either 'must-link' or 'cannot-link' based on a threshold value. Must-link constraints indicate that two objects have to be in the same cluster. Cannot-link constraints indicate that two objects must not be positioned in the same cluster. A set of must-link constraints $Con_{=}$, and a set of cannot-link constraints $Con_{6=}$ are generated by initializing $Con_{=}$ and $Con_{6=}$ to ϕ. Then, randomly choose two distinct points a and b from labeled data D.

$$If\ (Label(a) = Label(b))\ then\ Con_{=} = Con_{=} \cup \{(a,b)\}$$
$$else\ Con_{6=} = Con_{6=} \cup \{(a,b)\}$$

The COP-KMEANS algorithm in general takes in a data set (D), a set of must-link constraints $Con_{=}$, and a set of cannot-link constraints $Con_{6=}$ and returns a partition of the objects in D that satisfies all specified constraints. The steps of the algorithm for COP-KMEANS algorithm is explained as follows:

- *Step 1:* Consider the initial cluster centers as $C1, C2, \ldots C_k$.
- *Step 2:* For each data point d_i in D, check the constraints violation.
  a. If constraints are not violated, assign object to cluster 'k'.
  b. If constraints are violated, check for next nearest cluster. If any cluster nearby, check condition a. else return with failure.
- *Step 3:* For each cluster C_i, update its center by averaging all of the data points d_j that have been assigned to it.
- *Step 4:* Iterate the above two steps until convergence.

We applied COP-KMeans algorithm which is used as a tool for data mining to the whole high dimension dataset but observed that it did not fit properly to cluster the raw high dimensional data sets, because of the intrinsic sparse of high dimensional data. When a

high dimension dataset is taken for example, UCI data sets like Parkinson's dataset and Ionosphere dataset and so on, this algorithm frequently converges with one or more clusters which are either empty or summarize a small amount of data points (i.e., one data point) leading to generate imprecise and inaccurate clusters. Hence, a solution is very much in need for handling the problem of high dimensionality dataset to form accurate clusters.

4. Necessity of Dimension Reduction

Clustering in high-dimensional spaces for data mining is a common problem in several fields of science [45]. Original K-means algorithm and algorithms based on K-Means have extremely high computational complexity, particularly for large data sets. Also, as the dimensionality of the data is increased the number of distance calculations increases exponentially [14].

Generally, only a few dimensions are related to some clusters as the dimensionality increases, but the large amount noise that may be produced by the data in the unrelated dimensions may hide the actual data to be discovered. Moreover the distance measure fundamentally becomes meaningless for cluster analysis as all the data points located at different dimensions can be regarded as equally distanced because data normally becomes sparse as the dimensionality increases. Hence, cluster analysis of datasets consisting of a huge no. of features/attributes includes attribute reduction or dimensionality reduction as an essential data-preprocessing task [14].

Several methods overcome the problems due to high dimensionality by utilizing global dimension reduction techniques. Minimizing dimensionality of data prior to utilizing classical clustering method is a commonly utilized solution. The most popular one among these techniques is the frequently employed data mining technique called PCA [29]. But, PCA being a linear technique considers only the linear dependencies between variables. In recent times, several non-linear techniques like Kernel PCA [51], non-linear PCA [19, 21] and techniques based on neural networks [22, 50, 55, 56, 57] have been proposed.

5. Principal Component Analysis (PCA)

PCA [2, 3, 12, 18, 24, 38, 67], which is mathematically described as an orthogonal linear transformation transforms data to a new coordinate system in such a way that the largest variance by any projection of the data exists on the first coordinate (known as the first principal component), the second largest variance on the second coordinate, and so on. Several possibly correlated variables are transformed by it into fewer uncorrelated variables called principal components. Hence, key variables in a high dimensional data set that describe the discrepancy in the observation can be determined by the statistical technique PCA which can also, be used to facilitate the analysis and visualization of high dimensional data set, without considerable loss of information.

5.1. Principal Component (PC)

Theoretically, a principal component can be described as a linear association of optimally weighted monitored variables which increases the variance of the linear association and which have zero covariance with the preceding PCs. PC's are calculated by Eigen value decomposition of a data covariance matrix or singular value decomposition of a data matrix, usually after performing data centering for each attribute.

5.2. Elimination Methods of Unnecessary PCs

A number of PCs equivalent to the number of original variables are generated when the dataset is transformed to the new principal component axis. As most of the variances are described by many of the first PCs the remaining can be discarded with negligible loss of information. The number of PCs that must be preserved for interpretation is determined using the following several criteria:

- Scree Diagram automatically discards the PCs with extremely low variances by plotting the variances in percentage pertaining to the PCs.
- PCs having variance less than a specified threshold value can be discarded.
- PCs having Eigen values less than a specified fraction of the mean Eigen value can be eliminated.

6. Proposed High Dimension Data Clustering using Constraint-Partitioning K-Means Algorithm

Dimension reduction is performed as a preprocessing step in majority of the applications. Clustering, classification, and several other machine learning and data mining applications frequently employ dimension reduction. It usually reduces computational cost by removing the noisy dimensions (irrelevant attributes) while retaining the most essential dimensions (attributes). Thus, in our proposed approach, we are performing the dimensionality reduction on the high dimension dataset using PCA. The dimensionality of the high dimension dataset is reduced in such a way that dataset containing i attributes are approximately reduced to j attributes, where $i>j$. Later, we integrate the COP-KMEANS clustering algorithm to the reduced dataset to produce good and accurate clusters. The following are the steps incorporated for data clustering of raw high dimension dataset and are shown in Figure 2.

1. Perform dimensionality reduction using PCA.

2. Apply COP-KMEANS algorithm on reduced dataset.

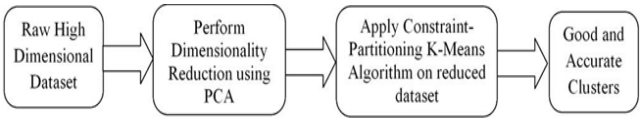

Figure 2. Proposed data clustering for high dimension dataset.

6.1. Perform Dimensionality Reduction using PCA

PCA method [2, 3, 12, 18, 24, 38, 67] performs covariance analysis between factors decrease the dimensionality of the data. By itself, it is fit for data sets in multiple dimensions, such as UCI datasets for instance Parkinson's dataset and Ionosphere dataset, and so on. For our proposed work, we are performing PCA using the covariance method. Following is a comprehensive explanation of PCA using the covariance method:

1. Organize the original dataset in a matrix form.
2. Subtract the mean from each of the data dimensions.
3. Find the data covariance matrix.
4. Find the eigenvectors and Eigen values of the covariance matrix.
5. Rearrange the eigenvectors and Eigen values in decreasing order.
6. Remove weaker components from PCs and form transformation matrix consisting of significant PC's.
7. Find the reduced data set using the reduced PCs.

Suppose we have a sample data X comprising of 'A' records and each having 'B' attributes, and we want to reduce the data such that each record will have only 'C' attributes in such a way that $C<B$.

1. Organize the high dimensional dataset X in a matrix S: Arrange data as a set of 'N' data vectors $S_1, S_2, \ldots, S_N$ where each S_n represents a single grouped record of the 'B' attributes.
   a. Write $S_1, S_2, \ldots, S_N$ as column vectors, each of which has B rows.
   b. Place the column vectors into a single matrix S of dimensions $B \times N$.
2. Subtract the mean from each of the data dimensions: For PCA to work properly, we have to subtract the mean from each of the data dimensions. Find the mean along each dimension $b=1,2,\ldots,B$. Then, place the calculated mean values into mean vector of dimensions $B \times I$. Later, subtract the mean vector from each S_n values of the data matrix S. Now, store the mean-subtracted data in the $B \times N$ matrix U. This produces a data set whose mean is zero.
3. Find the data covariance matrix C: Find the $B \times B$ covariance matrix C from the outer product of matrix U with itself.

$$C = \frac{1}{N} \sum UU^* \tag{1}$$

4. Find the eigenvectors and Eigen values of the covariance matrix C: Compute the matrix V of eigenvectors which diagonalizes the covariance matrix C

$$V^{-1} C V = D \tag{2}$$

   a. D is the diagonal matrix containing Eigen values of C which will take the form of $B \times B$ diagonal matrix.
   b. Matrix V is also, of dimension $B \times B$ contains B column vectors corresponding to the B eigenvectors of the covariance matrix C.
   c. For a covariance matrix, the eigenvectors correspond to principal components and the eigenvalues to the variance explained by the principal components.
5. Rearrange the eigenvectors and Eigen values in decreasing order: The eigenvector with the highest Eigen value is the principle component of the data set. Thus, assemble the columns of the eigenvector matrix V and Eigen value matrix D in the decreasing order of Eigen values. This gives us the components in order of significance.
6. Remove weaker components from PCs and form transformation matrix consisting of significant PC's: Selecting the number of PC's is a significant question. The largest eigenvalues correspond to the principal-components that are related with a large amount of the covariability amongst a number of observed data. Hence, we will remove weaker principal components from the set of components obtained. For the removal, perform any one of the three suitable methods explained in section 5.2. And generate the transformation matrix P with reduced PCs is formed.
7. Find the reduced data set using the reduced PCs: The transformation matrix P is applied to the original data set X to produce the new reduced projected dataset H which we can make use for data clustering.

Now, we integrate the COP-KMEANS algorithm [47] as explained in section 3.1 to the reduced dataset H.

7. Results and Discussion

The experimental results of the proposed high dimension data clustering using COP-KMEANS algorithm described in this section. The comparative analysis of the proposed algorithm with the original COP-KMEANS algorithm [56], is presented for synthetic datasets and the real world datasets.

7.1. Experimental Design and Set Up

The proposed high dimension data clustering using Constraint-Partitioning K-Means Algorithm is implemented in Java (jdk 1.6) and the experimentation is carried out on a 3.0GHz Pentium PC machine with 2GB main memory. For experimentation of the proposed work, we have used two datasets from the UCI machine repository such as Parkinson's dataset and Ionosphere dataset. Parkinson's dataset [56] contains 197 instances that are described by 23 attributes and Ionosphere dataset [57] contains 351 instances that are described by 34 attributes.

7.2. Evaluation Metrics

The performance of the proposed high dimension data clustering using constraint- partitioning K-Means algorithm is evaluated by means of three evaluation measures. They are:

1. Reduced number of attributes.
2. Clustering Accuracy.
3. Clustering Error (C_e).

We have used the clustering accuracy explained in [22, 25] for estimating the performance of the proposed approach. The evaluation metric used in the proposed approach is given below,

$$Clustering\ Accuracy,\ C_a = \frac{1}{N_d}\sum_{i=1}^{N_c} X_{cc} \quad (3)$$

$$Clustering\ Error,\ C_e = 1 - CA \quad (4)$$

where, $N_d \rightarrow$ Number of data points in the dataset.
$N_c \rightarrow$ Number of resultant cluster.
$N_{cc} \rightarrow$ Number of data points occurring in both cluster i and its corresponding class.

7.3. Experimental Results

Consider a sample dataset containing 15 data objects that are described by 10 attributes using which the initial experimentation performed. At first, we organized the original sample dataset in a matrix form as shown in Table 1. Then, we subtracted the mean from each of the data dimensions. Next, we found the data covariance matrix. Later from the covariance matrix, we computed the eigenvectors that correspond to principal components and the eigenvalues to the variance explained by the principal components. After rearranging eigen values in decreasing order, we will calculate the mean of eigen values from Table 2 and will eliminate the PCs having Eigen values less than mean Eigen value which helps in removing the weaker principal components. Thus, from the highest values of eigen values, we obtain the reduced dataset as shown in Table 3 where after applying dimension reduction, we obtained only 4 attributes. The result we obtained is the reduced data set using the reduced PCs shown in Table 3. Now, we applied COP-KMEANS algorithm to the reduced dataset shown in Table 3. Initially, we have to compute the must link constraints and cannot link constraints using Euclidean distance metric between all the points and keeping threshold as $\lambda=3$. We provided these must-link and cannot-link constraints along with the dataset as an input to the algorithm to produce clusters.

Table 1. The original data matrix X with 15 data objects having 10 attribute values.

	a	b	c	d	e	f	g	h	i	j
Data1	1	5	1	1	1	2	1	3	1	1
Data2	2	5	4	4	5	7	0	3	2	6
Data3	3	3	1	1	19	2	2	3	1	1
Data4	4	6	8	8	1	3	4	5	7	1
Data5	5	4	1	1	3	2	1	3	1	7
Data6	6	8	0	10	8	7	10	9	17	1
Data7	7	1	1	1	1	2	10	3	1	1
Data8	8	2	1	2	1	2	1	3	1	3
Data9	9	2	5	1	21	6	1	1	1	5
Data10	10	4	2	1	7	2	1	2	1	1
Data11	11	3	1	1	1	1	3	3	1	1
Data12	12	2	1	21	1	2	1	2	1	1
Data13	13	2	1	1	1	2	1	2	1	5
Data14	14	5	5	3	3	2	13	4	4	1
Data15	15	1	2	2	1	2	3	3	1	3

Table 2. shows PCs and corresponding Eigen value.

PC	Eigen Value
1	0.08940
2	0.66158
3	0.86841
4	2.9814
5	5.2712
6	6.7445
7	20.77251
8	22.95552
9	39.38671
10	49.2402

Table 3. The reduced data set.

	g	h	i	j
Data1	1	3	1	1
Data2	0	3	2	6
Data3	2	3	1	1
Data4	4	5	7	1
Data5	1	3	1	7
Data6	10	9	17	1
Data7	10	3	1	1
Data8	1	3	1	3
Data9	1	1	1	5
Data10	1	2	1	1
Data11	3	3	1	1
Data12	1	2	1	1
Data13	1	2	1	5
Data14	13	4	4	1
Data15	3	3	1	3

In the following Table 4, we provided the attributes in the input data and attributes reduced after applying the PCA algorithm. The result is taken for both the datasets for analysis. From the table, we have seen that

number of attributes is significantly reduced after applying PCA algorithm.

Table 4. Original number of attributes and reduced number of attributes using PCA are listed for Parkinson's dataset and Ionosphere dataset.

Parkinson's Dataset		**Ionosphere Dataset**	
Original no. of attribute	Reduced no. of attributer using PCA	Original no. of attribute	Reduced no. of attributer using PCA
22	4	34	7

7.4. Comparartive Analysis

This section presents the comparative analysis of the proposed approach with the COP-KMeans algorithm [47]. The clustering accuracy of the two algorithms are computed for both the datasets and given in the Table 5. Similarly, the clustering error is also, computed for both the datasets and given in Table 6. From the figures, it is very clear that the proposed approach is comparatively effective with the original COP-KMeans algorithm in producing clusters.

Table 5. Comparing the clustering accuracy of proposed approach and Original COP K means approach.

Parkinson's Dataset		**Ionosphere Dataset**	
Proposed Approach	Original COP- K means	Proposed Approach	Original COP- K means
73.84	67.89	71.51	68.67

Table 6. Comparing the clustering error of proposed approach and Original COP K means approach.

Parkinson's Dataset		**Ionosphere Dataset**	
Proposed Approach	Original COP- K means	Proposed Approach	Original COP- K means
26.16	32.31	28.49	31.33

8. Conclusions

Here, we have performed COP-KMEANS algorithm to the original high dimension dataset proved that clustering the raw high dimensional data sets did not produce precise clusters due to intrinsic sparse of high dimensional data. Hence, in our proposed approach, we have carried out two steps for clustering high dimension dataset. Initially, we have performed the dimensionality reduction on the high dimension dataset using PCA as a preprocessing step to data clustering. Later, we have integrated the COP-KMEANS clustering algorithm to the reduced dataset to produce precise clusters. The performance of the proposed approach is evaluated with high dimension UCI datasets such as Parkinson's dataset and Ionosphere dataset. The experimental results showed that the proposed approach is very effective in producing precise clusters compared with the original COP-KMEANS algorithm.

References

[1] Abbas O., "Comparison Between Data Clustering Algorithm," *The International Arab Journal of Information Technology*, vol. 5, no. 3, pp. 320-325, 2008.

[2] Aguilera A., Gutierrez R., Ocana F., and Valderrama M., "Computational Approaches to Estimation in the Principal Component Analysis of a Stochastic Process," *Applied Stochastic Models and Data Analysis*, vol. 11, no. 4, pp. 279-299, 1995.

[3] Ali A., Clarke G., and Trustrum K., "Principal Component Analysis Applied to Some Data from Fruit Nutrition Experiments," *The Statistician*, vol. 34, no. 4, pp. 365-369, 1985.

[4] Alijamaat A., Khalilian M., and Mustapha N., "A Novel Approach for High Dimensional Data Clustering," *in Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, Phuket Iran, pp. 264-267, 2010.

[5] Amorim R., "Constrained Intelligent K-Means: Improving Results with Limited Previous Knowledge," *in Proceedings of the 2nd International Conference on Advanced Engineering Computing and Applications in Sciences*, London, pp. 176-180, 2008.

[6] Bar-Hillel A., Hertz T., Shental N., and Weinshall D., "Learning a Mahalanobis M etric from E quivalence Constraints," *Journal of Machine Learning Research*, vol. 6, pp. 937-965, 2005.

[7] Belkin M. and Niyogi P., "Using Manifold Structure for Partially Labelled Classification," *in Proceedings of Conference on Advances in Neural Information Processing*, pp. 929-936, 2002.

[8] Bilenko M., Basu S., and Mooney R., " Integrating Constraints and Metric Learning in Semi-Supervised Clustering," *in Proceedings of the 21st International Conference on Machine Learning*, USA, pp. 11-18, 2004.

[9] Blum A. and Langley P., "Selection of Relevant Features and Examples in Machine Learning," *Journal of Artificial Intelligence*, vol. 97, no. 1-2, pp. 245-271, 1997.

[10] Bouveyrona C., Girarda S., and Schmid C., "High-Dimensional Data Clustering," *Journal of Computational Statistics and Data Analysis*, vol. 52, no. 1, pp. 502-519, 2007.

[11] Brian S. and Dunn G., *Applied Multivariate Data Analysis*, Edward Arnold, 2001.

[12] Croux C. and Haesbroeck G., "Principal Component Analysis Based on Robust Estimators of the Covariance or Correlation Matrix," *Influencefunctions and Efficiencies*, Biometrika, vol. 87, no. 3, pp. 603-618, 2000.

[13] Dash R., Dash R., and Mishra D., "A Hybridized Rough-PCA Approach of Attribute Reduction for High Dimensional Data Set," *European Journal of Scientific Research*, vol. 44, no. 1, pp. 29-38, 2010.

[14] Dash R., Mishra D., Rath A., and Acharya M., "A Hybridized K-Means Clustering Approach for High Dimensional Dataset," *International Journal of Engineering, Science and Technology*, vol. 2, no. 2, pp. 59-66, 2010.

[15] Demartines P. and H´erault J., "Curvilinear Component Analysis: A Selforganizing Neural Network for Non Linear Mapping of Data Sets," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 148-154, 1997.

[16] Dinga C., Hea X., Zhab H., and Simona H., "Adaptive Dimension Reduction for Clustering High Dimensional Data," *in Proceedings of the IEEE International Conference on Data Mining*, pp. 147-154, 2002.

[17] Fan W., Gordon M., and Pathak P., "Effective Profiling of Consumer Information Retrieval Needs: A Unified Framework and Empirical Comparison," *Journal of Decision Support Systems*, vol. 40, no. 2, pp. 213-233, 2004.

[18] Farmer S., "An Investigation into the Results of Principal Component Analysis of Data Derived from Random Numbers," *Statistician*, vol. 20, no. 4, pp. 63-72, 1971.

[19] Girard S., "A Nonlinear PCA Based on Manifold Approximation," *Computational Statistics*, vol. 15, no. 2, pp. 145-167, 2000.

[20] Hasan Y., Hasan M., and Ridley M., "Incremental Transitivity Applied to Cluster Retrieval," *The International Arab Journal of Information Technology*, vol. 5, no. 3, pp. 311-319, 2008.

[21] Hastie T. and Stuetzle T., "Principal Curves," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502-516, 1989.

[22] He Z., Xu X., and Deng S., "Clustering Mixed Numeric and Categorical Data: A Cluster Ensemble Approach," *Technical Report*, Cornell University Library, 2005.

[23] Hoch R., "Using IR Techniques for Text Classification in Document Analysis," *in Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, USA, pp. 31-40, 1994.

[24] Horgan G., "Principal Component Analysis of Random Particles," *Journal of Mathematical Imaging and Vision*, vol. 12, no. 2, pp. 169-175, 2000.

[25] Huang Z., "Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283-304, 1998.

[26] Jain A., Murty M., and Flynn P., "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, 1999.

[27] Jianhong W. and Guojun G., "Subspace Clustering for High Dimensional Categorical Data," *SIGKDD Explorations*, vol. 6, no. 2, pp. 87-94, 2004.

[28] Jiawei H. and Kamber M., *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

[29] Jolliffe I., *Principal Component Analysis*, Springer-Verlag, 1986.

[30] Jun H. and Xiong L., "Genetic Algorithm-Based High-Dimensional Data Clustering Technique," *in Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery*, Tianjin, vol. 1, pp. 485-489, 2009.

[31] Jun Y., Benyu Z., Ning L., Shuicheng Y., Qiansheng C., Weiguo F., Qiang Y., Wensi X., and Zheng C., "Effective and Efficient Dimensionality Reduction for Large-Scale and Streaming Data Preprocessing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 3, pp. 320-333, 2006.

[32] Anaissi A., Kennedy P., and Goyal M., "A Framework for High Dimensional Data Reduction in the Microarray Domain," *in Proceedings of the 5th IEEE International Conference on Bio-Inspired Computing: Theories and Applications*, China, pp. 903-907, 2010.

[33] Khalilian M., Mustapha N., Suliman M., and Mamat D., "A Novel K-Means Based Clustering Algorithm for High Dimensional Data Sets," *in Proceedings of the International MultiConference of Engineers and Computer Scientists*, Hong Kong, vol. 1, pp. 503-507, 2010.

[34] Kiri L., Wagstaff S., and Davidson I., "When is Constrained Clustering Beneficial, and Why?," *in Proceedings of the 21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference*, USA, pp. 1-2, 2006.

[35] Klein D., Kamvar D., and Manning C., "From Instance-Level Constraints to Space-Level Constraints: Making the Most of Prior Knowledge in Data Clustering," *in Proceedings of the 19th International Conference on Machine Learning*, USA, pp. 307-313, 2002.

[36] Kohavi R. and John G., "Wrappers for Feature Subset Selection," *Journal of Artificial Intelligence*, vol. 97, no. 1-2, pp. 73-324, 1997.

[37] Kohonen T., *Self-Organizing Maps*, Springer-Verlag, New York, 1995.

[38] Konishi S. and Rao C., "Principal Component Analysis for Multivariate Familial Data," *Biometrika*, vol. 79, no. 3, pp. 631-641, 1992.

[39] Lewis D., "Feature Selection and Feature Extraction for Text Categorization," *in Proceedings of Workshop Speech and Natural Language*, USA, pp. 212-217, 1992.

[40] Li H., Jiang T., and Zhang K., "Efficient and Robust Feature Extraction by Maximum Margin Criterion," *in Proceedings of Conference on Advances in Neural Information Processing Systems*, pp. 97-104, 2004.

[41] Lu Z. and Leen T., " Semi-Supervised Learning with Penalized Probabilistic Clustering," *in Proceedings of Advances in Neural Information Processing Systems*, vol. 17, pp. 849-856, 2005.

[42] Maaten L., Postma E., and Herik H., "Dimensionality Reduction: A Comparative Review," *Technical Report*, University of Maastricht, 2007.

[43] MacQueen J., "Some Methods for Classification and Analysis of Multivariate Observations," *in Proceedings of the 5th Berkley Symposium on Mathematical Statistics and Probability*, Statistics, vol. 1, pp. 281-297, 1967.

[44] Moise G. and Sander J., "Finding Non-Redundant, Statistically Significant Regions in High Dimensional Data: A Novel Approach to Projected and Subspace Clustering," *in Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, USA, pp. 533-541, 2008.

[45] Martinez A. and Kak A., "PCA Versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 228-233, 2001.

[46] Oja E., *Subspace Methods of Pattern Recognition*, Research Studies Press, England, 1983.

[47] Parsons L., Haque E., and Liu H., "Subspace Clustering for High Dimensional Data: A Review," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 90-105, 2004.

[48] Prasad G., Dhanalakshmi Y., Vijayakumar V., and Rameshbabu I., "Mining for Optimised Data using Clustering Along with Fuzzy Association Rules and Genetic Algorithm," *International Journal of Artificial Intelligence & Applications*, vol. 1, no. 2, pp. 30-41, 2010.

[49] Rajput D., Singh P., and Bhattacharya M., "An Efficient and Generic Hybrid Framework for High Dimensional Data Clustering," *in Proceedings of International Conference on Data Mining and Knowledge Engineering*, *World Academy of Science, Engineering and Technology*, Rome, pp. 174-179, 2010.

[50] Roweis S. and Saul L., "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, vol. 290, no. 5500, pp. 2323-2326, 2000.

[51] Scholkopf B., Smola A., and Muller K., "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation*, vol. 10, no. 5, pp. 1299-1319, 1998.

[52] Sembiring R., Zain J., and Embong A., "Clustering High Dimensional Data using Subspace and Projected Clustering Algorithms," *International Journal of Computer Science & Information Technology*, vol. 2, no. 4, pp. 162-170, 2010.

[53] Srikant R. and Agrawal R., "Mining Quantitative Association Rules in Large Relational Tables," *in Proceedings of the ACM SIGMOD International Conference Management Data*, vol. 25, no. 2, pp. 1-12, 1996.

[54] Tajunisha N. and Saravanan V., "An Increased Performance of Clustering High Dimensional Data Using Principal Component Analysis," *in Proceedings of the 1st International Conference on Integrated Intelligent Computing*, Bangalore, pp. 17-21, 2010.

[55] Tenenbaum J., Silva V., and Langford J., "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, vol. 290, no. 5500, pp. 2319-2323, 2000.

[56] UCI Machine Learning Repository, available at: http://archive.ics.uci.edu/ ml/datasets/Parkinsons, last visited 2008.

[57] UCI Machine Learning Repository, available at: http://archive.ics.uci.edu/ ml/datasets/Ionosphere, last visited 1989.

[58] Valarmathie P., Srinath M., and Dinakaran K., "An Increased Performance of Clustering High Dimensional Data through Dimensionality Reduction Technique," *Journal of Theoretical and Applied Information Technology*, vol. 13, no. 7, pp. 271-273, 2009.

[59] Wagstaff K., Cardie C., Rogers S., and Schrödl S., "Constrained K-Means Clustering with Background Knowledge," *in Proceedings of the 18th International Conference on Machine Learning*, USA, pp. 577-584, 2001.

[60] Webb A., *Statistical Pattern Recognition*, John Wiley, England, 2002.

[61] Witten I., Frank E., and Hall M., *Data Mining-Practical Machine Learning Tools and Technique*, Morhan Kaufmann, 2005.

[62] Xing E., Ng A., Jordan M., and Russell J., "Distance Metric Learning, with Application to Clustering with Side-Information," in *Proceedings of Advances in Neural Information Processing Systems*, vol. 15, pp. 505-512, 2003.

[63] Xu R. and Wunsch D., "Survey of Clustering Algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645-678, 2005.

[64] Yager R., "Fuzzy Summaries in Database Mining," *in Proceedings of the Artificial Intelligence for Applications*, Los Angeles, pp. 265-269, 1995.

[65] Yan J., Zhang B., Liu N., Yan S., Cheng Q., Fan W., Yang Q., Xi W., and Chen Z., "Effective and Efficient Dimensionality Reduction for Large-Scale and Streaming Data Preprocessing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 3, pp. 320-333, 2006.

[66] Yee Y. and Walter L., "An Empirical Study on Principal Component Analysis for Clustering Gene Expression Data," *Technical Report,* University of Washington, 2000.

[67] Zelmat M., Kouadri A., and Albarbar A., "Prediction of Boiler Output Variables through the PLS Linear Regression Technique," *The International Arab Journal of Information Technology*, vol. 8, no. 3, pp. 260-264, 2011.

Aloysius George received his PhD degree from Prescott University in 2007. Currently, he is the manager of Grian Technologies Private Limited, a research and development company. He has more than 13 years of research experience and has publications in various international journals and conferences. His research areas of interest include biometrics, datamining, image processing and database management systems.