High Dimensional data clustering using Partition-Constraint Algorithm for knowledge Discovery

N.SreeRam

Department of CSE, Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India -522302
sriramnimmagadda@gmail.com

DOI : 10.48047/IJFANS/11/ISS4/131

Abstract. Computational efficiency and result quality both necessitate that huge data-bases have been clustered. The data mining experts believe that feature space clustering over the original data space is necessary in order to reach both of the aforementioned goals. As a consequence, we used COP-KMEANS (Con-straint-Partitioning K-Means) clustering on our high-dimensional dataset. This method did not successfully group the data into effective and efficient clus-ters, because of the inherent sparseness of our high-dimensional dataset, and therefore produced erroneous and indeterminate clusters. Dimensionality re-duction can apply on original dataset as a preparatory step to high dimen-sional data clustering. Once we have successfully clustered the dimensions reduced dataset with the COP-KMEANS method, we will do that task again, but this time on the resulting clusters. We use two artificial high-dimensional datasets to test the working of the proposed technique. According to the ex-perimental results, the suggested method is highly successful in built-up ac-curate and exact clusters.

Keywords: constraint partitioning k-means, Dimensionality Reduction, Feature space clustering, and High dimensional dataset

1. Introduction

The bulk of data found around the world is stored in databases these days. Most focus in the research community has been paid to data mining, which denotes to the procedure of finding intrinsic forms from high dimensional data [1]. Bunching or categorization of this data into several groups is one of the most fund-mental tools for processing these types of data [10, 63]. Categorization is a delineative exercise that aims to identify items that have comparable qualities based on their implications. Useful analyses include grouping data to define dense and sparse zones, which reveal the prominent distribution patterns and intriguing relationships found in that data [4, 5]. Data gathering is now easier and faster because to technological improvements, which has led to the appearance of big datasets with multiple dimensions. Cluster quality and speed must be maintained, especially as datasets get larger and more varied. Clustering algorithms like traditional ones seek to find out about each item reported in the given dataset. In multidimensional data, the huge attributes are not important as that are completely irrelevant.

Since all of the items in extremely high dimensions are nearly in between from each other, they are invisible, and entirely mask the clusters. In addition, the curse of dimensionality also causes difficulty for clustering algorithms when faced with high-dimensional data. When the number of dimensions in a dataset increases, the distance metric loses significance [16, 27, and 47]. Clustering algorithms designed to handle high-dimensional data are difficult to implement [5, 20]. Some academics have recently overcome the high-dimensional challenge by lowering the data's dimension. Classification tasks have two major issues: one is low classification efficiency owing to the huge feature space, and the other is the need for dimensionality reduction. Feature extraction methods (FE) focus on extracting relevant features from data, whereas feature selection methods (FS) are focused on finding relevant features from the data.

FE approaches are typically more successful than FS (except a few of special circumstances) and have previously been proven in terms of their high efficiency for real-world dimensional reductions [17, 23, 39, and 40]. Two basic groups of traditional FE algorithms are linear and nonlinear algorithms [11]. Here we have suggested a strategy to high-dimensional data clustering such as Parkinson's dataset and Ionosphere dataset utilizing the Clustering algorithm Constraint-Partitioning K- Means (COP-KMEANS). We originally tried to build clusters from high-dimensional datasets, but because of the intrinsically scant size of the high-dimensional data it was not effective. Therefore, we propose a technique; where we will take two stages to produce high-size dataset clusters. At the beginning, we lower the dimension, i.e., by reducing the dimensionality of the high-dimensional dataset by making the pre-processing step to data clustering using PCA. Later, we integrate the clustering method COP-KMEANS in the reduced data set in order to obtain good, precise clusters.

2. High-dimensional Data Clustering using K-Means Algorithm and K-Means Algorithm Clustering

2.1. Limited K-Means (COP-KMEANS) partitioning

Many works have been done incorporating instance-level restrictions into clustering algorithms [6, 8, 35, 49, 59]. In the associated cluster Must-Link (ML) or unrelated clusters two items must be positioned Cannot-Link (CL) indicates limitations of instance-level. Then the whole set of restrictions is supplied to the algorithm of the K-Means cluster [61]. This semi-monitored technique has led to greater concentration on the reference phrase and GPS-based map refinement [47], proof of identity by persons with camera clips [4], landscape recognition from hyper-spectral data [41], and soon. COP-KMEANS [59] employs categorized cases to control the k-means clustering [43]. Every pair of identified occurrences is labeled 'must-link' or 'cannot-link' on the basis of a threshold value. Must-link constraints show two items in the same cluster. Cannot-link restrictions indicate that two items should not be in the same cluster.

COP-KMEANS algorithm described as follows:

```
Step 1: Consider C1, C2, Ck, initial cluster centers.
Stage 2: Check the breach of limits for each data point di in D.
    a. Allot object to cluster 'k' if restrictions are not violated.
    b. Formed for the closest cluster if limitations are broken. If you have any cluster nearby, verify condition a.
Step 3: Average all the dj data points assigned to each cluster Ci to update its center.
Step 4: Reiterate the two stages mentioned above until convergence.
```

Tried the Cop-K Means technique, which is utilized as a tool for data extraction for the complete data set, but we noted that as of the inherent small amount of high-dimensional data, it did not fit correctly to cluster the raw high-dimensional data sets. For example, the high dimensional dataset, UCI data sets such as the Parkinson dataset and the Ionosphere dataset, sometimes converge with one or more empty or summary clusters, which might create imprecise and erroneous clusters, i.e. one dataset. A solution is therefore very important to manage the high dimensionality data set problem to construct precise clusters

3. Dimension need Reduction

High-dimensional data mining space clustering is a typical challenge in numerous scientific fields [45]. Original K-means and K-means algorithms had very high computational cost, especially for huge data sets. Furthermore, the complex-ity of data rises exponentially in the number of distance calculations [14]. As the dimensionality rises, just a few dimensions have to do with particular clusters, but the vast quantitative noise that data in unrelated dimensions make might mask the real data to be identified. In addition, the distance measurement has little significance for cluster analysis, because all data points in various dimen-sions may be considered to be equally distant as the data is usually sparse as dimensionality rises. Thus, the analysis of clusters of data sets comprising a large number of features/attributes entails a reduction of attributes or dimensionality as a vital preprocessing task. Various ways resolve the challenges caused to their high dimensionality by using strategies of global dimension reduction. A typically used technique is to minimize the dimension of data before using the conventional cluster approach. The most prominent of these strategies is the PCA data mining methodology, which is often used [29]. However, PCA is a linear method-ology that only evaluates the linear dependence of variables. More recently, other non-linear algorithms have been developed such as the Kernel PCA [51], the non-linear PCA [19, 21] and neural network-based technology [22].

4. **Proposed Data clustering method for High Dimensional Data using k-Means**

Reduction of dimensionality is carried out in most applications as a preprocessing step. Clustering, classification and many other applications of ML and DM often use reduced dimensions. It generally decreases computer cost by reducing the noisy measurements (irrelevant qualities) while maintaining the key aspects (attributes). Thus, we use PCA to reduce dimensionality of the high-dimensional dataset in our technique. The dimensionality of the high dimension dataset is therefore reduced to approximately j properties where I > j is con-tained. Next, we incorporate the clustering method COP-KMEANS in the reduced dataset to find good, precise clusters. The processes taken for the grouping of raw high-dimensional datasets are illustrated in Figure 2.

1. Reduce dimensionality using PCA.
2. Apply COP-KMEANS Reduced C <1> algorithm to the dataset UU * (1).

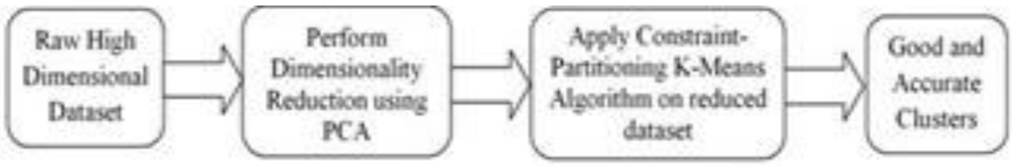
Raw High Dimensional Dataset
Perform Dimensionality Reduction using PCA
Apply Constraint-Partitioning K-Means Algorithm on reduced dataset
Good and Accurate Clusters

Fig.1. Proposed high-dimension data clustering

5. **Results and Discussion**

The investigational results of the proposed COP-KMEANS algorithm high-dimensional data clustering presented in this part. For synthetic datasets and ac-tual data sets, we give a comparison study of the proposed approach with the COP-KMEANS method [56].

A. Experimental setup

In Java the suggested High-Dimensional Data Clustering utilizing the K-Means algorithms is developed and the tests are conducted on a 2GB main-memory 3.0GHz Pentium PC computer. We have used two UCI repository data sets, such as the Parkinson dataset and the Ionosphere dataset, to test the suggested approach. Parkinson's [56] dataset has 197 occurrences defined as 23, whereas Ionosphere [57] has 351 occurrences characterized as 34 characteristics.

B. Metrics of evaluation

Three assessment metrics analyze the working of the proposed high dimen-sional data clustering utilizing constraint-partitioning K-Means method. They are the following:

- Reduced attribute number.
- Accuracy clustering.

- Error clustering (Ce).

We utilized the accuracy of cluster [22, 25] to estimate the performance of the methodology provided. The assessment metric utilized in the technique suggested is shown below, Clustering Accuracy, N d I − clustering mistake, Ce = 1 − CA Where, Nd quo Number of dataset data points. Nc Footnote Number of clusters resulting.

Ncc/Complete Number of data points in both cluster I and its respective classes

C. Experimental findings

Consider a sample data collection having 15 data items defined by 10 attributes for the beginning experiment. First, the original sample data set was structured in a matrix form as presented in Tab.1. We distant the mean from each of the dimensions of the data as next step. Later, we found the matrix for data co-variance. Later using the covariance matrix, we calculated the vectors that match the major components and their unique values with the variance described by the main components.

Dataset	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
DS1	12	10	22	32	21	32	33	43	18	13
DS2	7	15	24	42	51	27	10	32	27	61
DS3	36	21	21	21	17	32	19	22	19	14
DS4	14	16	87	53	18	33	42	57	7	18
DS5	15	42	18	27	36	28	18	36	17	72
DS6	16	81	98	19	82	71	10	92	174	19
DS7	17	19	19	36	17	27	10	36	186	16
DS8	1	26	17	11	19	28	19	38	134	33
DS9	11	28	53	52	21	61	12	18	123	54
DS10	19	43	27	19	73	29	72	23	121	12
DS11	9	37	19	9	19	18	36	34	19	19
DS12	20	38	36	20	28	34	14	18	23	17
DS13	13	112	11	13	13	29	15	29	1	53
DS14	18	59	52	18	36	27	16	43	4	17
DS15	15	32	26	15	29	42	17	32	19	28

Table 1 Matrix structure from original Dataset

Dataset	**PC-1**	**PC-2**	**PC-3**	**PC-4**
DS1	11	13	18	16
DS2	10	32	33	26
DS3	24	32	33	34
DS4	14	15	17	16
DS5	16	32	1	17
DS6	10	19	17	15
DS7	10	13	1	12

DS8	17	13	14	13
DS9	11	14	1	15
DS10	11	12	13	18
DS11	8	12	11	13
DS12	17	16	14	15
DS13	12	21	22	16
DS14	13	14	14	11
DS15	13	12	18	17

Table 2 Dataset after Dimensionality Reduction

After rearranging our own size in a decreasing order, we compute the mean of our own size in Table 2 and remove PCs with Eigen size below the Eigen size which helps to remove the weaker main size. Thus, we get the reduced data set as shown in Table 3, from the greatest value of our own values, where we received just 4 characteristics after application of dimension reduction. The outcome is a reduced data set with the decreased PCs shown Table 3. Now we used the COP-KMEANS method in Table 3 for the smaller dataset. Initially, we must calculate the limitations and we must not connect limitations using the Euclidean distance metric between all locations with a μ=3 threshold. We supplied these connection restrictions together with the data set as an input to the algorithm to construct clusters.

Artificial Dataset-1		Artificial Dataset-2	
Actual number of attributes	Number of attributes after Dimensionality Reduction	Actual number of attributes	Number of attributes after Dimensionality Reduction
27	5	46	9

Table 3 Actual number of attributes and Number of attributes after Dimensionality Reduction

D. Comparative Analysis

Let's assume that the dataset DS consisting of the records 'A' and each one has 'B' characteristics, which means that each record will have just 'C' attributes so that Table 4 will be reduced. Similarly, the error rate in clustering is calculated for both datasets and for Table 5. From that it is quite evident that the proposed method effective rather COP-K Means method

Artificial Dataset-1		Artificial Dataset-2	
Proposed Approach	COP-K means	Proposed Approach	COP-K Means
78.35	62.17	77.15	69.01

Table 4 Comparision of the proposed approach with COP K means approach in teems of clusatering Accuracy

Artificial Dataset-1		Artificial Dataset-2	
Proposed Technique	COP-K means	Proposed Technique	COP-K Means
25.08	36.39	24.89	37.14

Table 5 Comparing the proposed technique with COP K means in terms of error rate

6. Conclusions

Here, we implemented a COP-KMEANS data technique to the original high-dimensional dataset that, thanks to the intrinsic sparse of high dimensional data, clustering of raw high dimensional data sets could not yield exact clusters. There-fore, we have taken two steps to cluster the high-dimensional dataset in our suggested technique. Initially, we reduced dimensionality on the high-dimensional data set using “PCA” as a preparation step for clustering given dataset. Later, we merged the clustering technique COP-KMEANS into the reduced data set to obtain precision clusters. With high dimensional UCI data sets, such as Parkinson's data and Ionosphere dataset, the performance of the suggested methodology is examined. The experimental findings demonstrated that the methodology provided is extremely efficient in constructing accurate clusters in comparison with the 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, 2008Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016).
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, 1995Author, F.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010).
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 Dimension-al 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 Previ-ous 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 Metricfrom 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 Classifica-tion,” 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, 2000474 The International Arab Journal of Information Technology, Vol. 10, No. 5, September 2013
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