AN ENHANCED APPROACH FOR PROJECTING CLUSTERS IN HIGH DIMENSIONAL SPACES

Dr.K.PRASADH
Principal,
MookambikaTechnicalCampus
Muvattupuzha
Kerala,India

ARUN SHALIN.L.V
Research Scholor,
Manonmanium Sundarnar university
Tamil Nadu,India

Abstract **Clustering high-dimensional data has been a major challenge due to the inherent sparsity of the points. Most existing clustering algorithms become substantially inefficient if the required similarity measure is computed between data points in the full-dimensional space. To address this problem, a number of projected clustering algorithms have been proposed. However, most of them encounter difficulties when clusters hide in subspaces with very low dimensionality. To overcome the difficulties, a robust partitioned distance-based projected clustering algorithm [PCKA] is presented in the previous work [1] describes the process of identifying the low dimensional values in high dimensional space and avoids the computation of the distance in the full dimensional space. But the existing PCKA algorithm analyzed only about the attribute relevance and does not discussed the analysis of redundancy in the database. To enhance the process of projecting clusters of high dimensional subspaces, an enhanced framework is presented in this work which solves the projected clustering problem. Even though the number of dimensions for all the clusters specific subspace varies, the process of identifying the single small subset of dimensions for all the clusters is achieved efficiently. An experimental evaluation is carried out to estimate the performance of the proposed enhanced approach for projecting clusters in high dimensional spaces in terms of average cluster dimensionality, outlier immunity and consumption time.**

1. INTRODUCTION

Data mining is the procedure of mining practical information from a data set. Clustering is a fashionable data mining method which is planned to assist the user ascertains and recognizes the organization or alignment of the data in the set according to a definite similarity determination. Clustering algorithms regularly utilize a distance metric (e.g., euclidean) or a resemblance evaluation to partition the database. The partition describes the data points in every division are more analogous than points in diverse partitions.

The frequently employed euclidean distance, as computationally trouble-free, needs parallel objects to contain secure values in all proportions. Nevertheless, with the high-dimensional data normally met these days, the notion of connection among objects in the full-dimensional space is frequently unacceptable and normally not useful. Current hypothetical outcome disclose that data points in a place be inclined to be more evenly spaced as the measurement of the space enhances, only if the workings of the data point are separately and similarly dispersed (i.i.d.). Even though the i.i.d. constraint is hardly ever pleased in genuine applications, it still turns into less significant to distinguish data points supported on a space or a comparison measure determined utilizing all the dimensions.

Clustering is a data mining system for diverse applications. One of the causes for its recognition is the capability to exert on datasets by lowest amount or no a priori information. This constructs clustering sensible for genuine world applications. In recent times, high dimensional data has stimulated the attention of database researchers owing to its novel challenges carried to the district. In high dimensional space, the space from a record to its adjacent neighbor can come close to its space to the furthest record. In the framework of clustering, the crisis acts as the distance among two records of the similar cluster to come near the distance among two records of diverse clusters.

Subspace clustering has been examined widely as conventional clustering algorithms frequently fall short to identify significant clusters in high-dimensional data spaces. Many newly planned subspaces grouping techniques has two rigorous troubles:

- At first, the algorithms naturally extent the data dimensionality or subspace dimensionality of the clusters.

- Second, for presentation motivations, several algorithms utilize a worldwide density threshold for clustering, since clusters in subspaces of considerably assorted dimensionality provide varying densities.

High dimensional data subspace clustering has been main concern owing to sparsity of data points. Nearly the entire clustering algorithm turns into ineffective if the vital distance comparison measure is determined for low dimensional spatial distance of high dimensional data with sparsity of data point beside diverse proportions. In high dimensional datasets, clusters can be created in sub-spaces. Only a separation of attributes is applicable to every cluster, and every cluster can contain a diverse deposit of pertinent attributes. An attribute is appropriate to a cluster if it assists to recognize the constituent records of it. This provides the values at the appropriate attributes are dispersed about some precise values in the cluster, whereas the records of further clusters are likely to have low values. Identifying clusters and its appropriate attributes from a dataset is termed as the projected (subspace) clustering problem. For every cluster, an estimated clustering algorithm establishes a set of attributes that it imagines to be the most appropriate to the cluster.

2. LITERATURE REVIEW

Projected clustering develops the reality that in high dimensional data sets, diverse sets of data points might be connected beside diverse sets of proportions. A number of estimated clustering algorithms have been anticipated in current years. Even though these prior algorithms have been thriving in ascertaining clusters in diverse subspaces, they bump into obscurity in recognizing very low-dimensional expected clusters entrenched in high-dimensional space. The author in [1] presented clustering method supported on the K-Means algorithm, with the calculation of space constrained to subsets of features where object values are intense.

The most instructive dimensions are chosen by abolishing inappropriate and unnecessary ones. Such methods get moving on clustering algorithms and progress their performance [2]. However, in several applications, diverse clusters might survive in diverse subspaces extent by diverse dimensions. A different hypercube method called recurrent prototype based Clustering (FPC) is planned in [3] to progress the effectiveness of the clustering process. FPC reinstates the randomized section of the dimensions with methodical exploration for the finest cluster defined by an arbitrary medoid point. A mixture of strategies has been projected to analyze the number of entities in a data set [4]. In this system, a fine probability measure, termed as Bayesian Information Criterion is presented.

With respect to the incidence of inappropriate dimensions, high dimensional data are also differentiated by the occurrence of outliers. Outliers can be termed as a position of data points that are significantly different, incomparable, or contradictory as regards the outstanding data. A general method to recognize outliers is to examine the connection of every data point with the break of the data, supported on the notion of closeness [5].

Clustering is the data mining method to cluster the associated data based on similarity space determination [6]. Clustering generally utilizes distance measure for instance Euclidean, Manhattan or Minkowski etc. The numbers of estimated clustering algorithms have been planned in current years but they do not deal with the low dimensional clusters on high dimensional space. Feature collection method can get faster the clustering procedure but nevertheless there is considerable information loss [7]. In [8], a methodical approach is engaged to estimate the main paradigms in a widespread construction. An envoyed clustering algorithm is employed to differentiate the diverse features of every pattern and provide a comprehensive evaluation of their properties. For comparability, realize projected algorithms in a widespread framework [9]. By expanding the projected clustering structure on an extensively used data employed for repeatable and flexible experiments [10]. Several subspace clustering algorithm is employed for dimensionality unbiased subspace [11] and it has been utilized in several unique applications. Indexing subspace clustering technique is implemented based on the process removal of redundancy [12] and evaluated it with the experiments [13].

3. PROPOSED ENHANCED APPROACH FOR PROJECTING CLUSTERS IN HIGH DIMENSIONAL SPACES

The proposed work is efficiently designed for projecting the clusters in high dimensional spaces by adapting the enhanced approach in k-mediods algorithm. The proposed enhanced approach for projecting clusters in high dimensional spaces is processed under three different processes. The first phase achieves attribute relevance and redundancy analysis by identifying sparse and dense regions and their position in every attribute. With the outcomes of the first phase, the objective of the second process is to eradicate outliers, whereas the third process efficiently identifies clusters in diverse subspaces. The clustering procedure is based on the K-mediods algorithm, with the addition of distance controlled to

subsets of attributes wherever entity values are dense. The architecture diagram of the proposed enhanced approach for projecting clusters in high dimensional spaces is shown in fig 3.1.

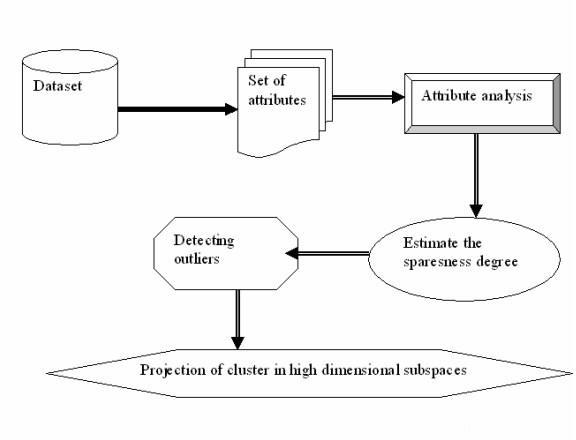

Fig 3.1 Architecture diagram of the proposed EAPC

From the above figure (fig 3.1), it is being observed that the proposed EAPC is processed under four different phases. The outcome of every phase acts as n input for the next phase. The final outcome would be the projected clusters of high dimensional subspaces.

To describe the proposed EAPC algorithm, consider *DB* be a dataset of n-dimensional points, where the attributes is denoted by $A = \{A_1, A_2, ..., A_n\}$ be the set of *N* data points, where $x_i = (x_{i1}, x_{i2}, .., x_{in})$. Each $x_{ij} where i = \{0,1,..., n\}; j = \{0,1,...., N\}$ communicates to the assessment of data point x_i on attribute A_j. x_{ij} is termed as a one dimensional point.

Each data point x_i fits in either to one estimated cluster or to the position of outliers *OUT*. For the given the number of clusters *nc*, acts as an input parameter, a estimated cluster $C_s, s = 1,2,..., nc$ is termed as a pair (SP_s, SD_s) where SP_s is a division of data points of database and SD_s is a division of dimensions of set of attributes A, such that the projections of the data points in SP_s beside every dimension in SD_s are directly grouped. The proportions in SP_s are termed as significant dimensions for the cluster C_s. The lasting proportions, i.e., $A - SD_s$ are termed as inappropriate dimensions for the cluster C_s. The cardinality of the position SD_s is indicated by d_s, where $d_s \cdot d$ and n_s specifies the cardinality of the set SP_s, where $n_s < N$.

The proposed EAPC is paying attention on determining axis-parallel anticipated clusters which persuade the subsequent properties:

1) Estimated clusters have got to be dense. Particularly, the expected values of the data points beside every dimension of $SD_s, s = 1,.., nc$ provide regions of high thickness in association with every dimension of $A - SD_s$
2) The separation of dimensions SD_s might not be displaced and it might contain diverse cardinalities.
3) For every expected cluster C_s the projections of the data points in SP_s beside every dimension in SP_s are analogous to all other along with a comparison utility, but divergent to further data points not in C_s.

The primary property supported the information that appropriate dimensions of the clusters control dense sections in evaluation to inappropriate ones and such a notion of ”thickness” is relatively qualified in all proportions in the given dataset. In clustering process, K-means-based clustering is used, processed under the Euclidean distance so as to determine the comparison among a data point and a cluster midpoint such that only proportions that enclose dense regions are concerned in the space calculation.

The proposed EAPC proceeds in four phases:

Attribute relevance analysis

The objective is to recognize all proportions in a dataset which display some cluster composition by determining dense regions and their position in every measurement. The fundamental supposition for the attribute relevance analysis phase is that in the situation of estimated clustering, a cluster must contain appropriate proportions in which the projection of every point of the cluster is secure to a adequate number of further expected points, and this notion of ”proximity” is qualified with all the proportions. The recognized dimensions symbolize probable aspirants for appropriate proportions of the clusters.

Sparseness Estimation

With assist of attribute significance analysis, the sparseness level y_{ij} are determined for diverse proportions. The sparseness level y_{ij} are specified by the method.

$$y_{ij} = \sum \frac{(r - c_i^j)^r}{k} \quad \text{…… eqn 1}$$

Where $r \in p_i^j(x_{ij})$

The least assessment of y_{ij} signifies solid region and highest assessment signifies thin region. Likewise

diverse y_{ij} values are determined for diverse spatial images from diverse dimensions. With facilitate of over assessment of y_{ij} for every image we can simply perceive the dense regions. The images with better principles of y_{ij} indicate the thin regions. The attribute with less values of y_{ij} signify the opaque regions. The below algorithm describes the process of estimating the sparseness value of the dataset.

1: Input: A_j, m max, k
2: Output: sparseness degree
3: compute the sparseness degree y_{ij};
4: Normalize y_{ij} in the interval] 0, 1];
5: for m = 1 to m_max do
6: if m = = 1 then
7: Estimate the parameters of the gamma distribution based on the likelihood
Formula
8: Compute the value of sparseness using eqn 1
9: else
10: Estimate the mixture parameters as the initialization of the outlier process
11: End If
12: End

Outlier handling

Based on the outcome of the first phase, the objective is to recognize and eradicate outlier points from the dataset. Similar to the preponderance of clustering algorithms, EAPC believes outliers as points that do not group fine.

Discovery of projected clusters

The objective of this phase is to recognize clusters and its appropriate dimensions. The clustering process is done based on a customized description of the K-means algorithm in which the calculation of distance is controlled to subsets in which the data point values are termed as dense. Based on the recognized groups, the results of phase 1 are processed by choosing the suitable proportions of every cluster. The process of this phase follows two steps:

1. In the first step, group the data points using K-Means algorithm, with the calculation of distance controlled to subsets of proportions where values of objects are dense.
2. Based on the groups attained in the first step, the next step is to choose the appropriate proportions of the recognized clusters by the properties of the relevant attributes.

Using above steps, we evaluate the process of discovery of projected clusters efficiently using an enhanced version of K-means algorithm.

4. EXPERIMENTAL EVALUATION

The main goal of the experiments presented in this section was to evaluate the capability of projected clustering algorithms to correctly identify projected clusters in various situations. An experimental evaluation is conducted to estimate the performance of the proposed enhanced approach for projecting clusters in high dimensional spaces. The proposed approach is implemented in Java. The first phase achieves attribute relevance and redundancy analysis by identifying sparse and dense regions and their position in every attribute. With the outcomes of the first phase, the objective of the second process is to eradicate outliers, whereas the third process efficiently identifies clusters in diverse subspaces. The clustering procedure is based on the K-Means algorithm, with the addition of distance controlled to subsets of attributes wherever entity values are dense. Our algorithm is accomplished of distinguishing projected clusters of low dimensionality entrenched in a high-dimensional distance and passes up the calculation of the space in the full-dimensional distance. In this section, we develop a sequence of experiments considered to estimate the appropriateness of the proposed algorithm in terms of

1. Average cluster dimensionality,
2. outlier immunity and
3. Consumption time.

5. RESULTS AND DISCUSSION

In this work we have seen how the clusters have been projected in high dimensional spaces. The below table and graph describes the performance of the proposed enhanced approach for projecting clusters in high dimensional spaces.

No. of clusters	Average cluster dimensionality (%)	
	Proposed EAPC	Existing PCKA
2	8	4
4	11	7
6	14	11
8	18	12
10	22	14
12	24	16
14	25	17

Table 5.1 No. of clusters vs. Average cluster dimensionality

The above table (table 5.1) describes the average cluster dimensionality based on the number of clusters partitioned with respect to the database. The dimensionality of the cluster of the proposed enhanced approach for projecting clusters in high dimensional spaces is compared with an existing distance-based projected clustering algorithm.

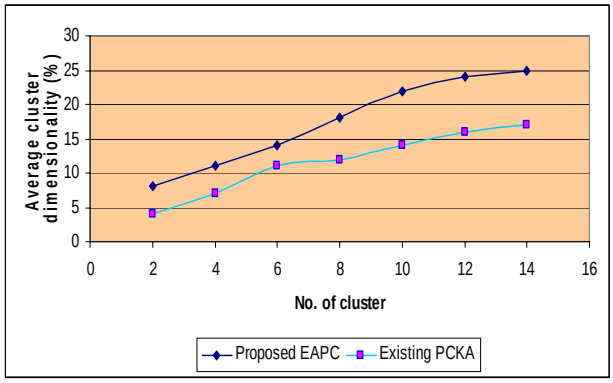

Fig 5.1 No. of clusters vs. Average cluster dimensionality

Fig 5.1 describes the average cluster dimensionality based on the number of clusters partitioned with respect to the database. The set of experiments was used here to examine the impact of cluster dimensionality. EAPC is capable to accomplish vastly precise results and its performance is normally reliable. As we can see from Fig. 5.1, EAPC is more robust to deviation of the cluster dimensionality than the existing PCKA algorithm. If the average cluster dimensionality is very low, only the proposed EAPC provides satisfactory results. Experiments showed that the proposed EAPC algorithm efficiently identifies the clusters and its dimensions precisely in a variety of situations. EAPC eradicates the choice of inappropriate dimensions in all the data sets used for experiments. This can be achieved by the fact that EAPC initiates its process by detecting dense regions and their positions in every dimension, facilitating it to control the calculation of the space to subsets of dimensions where the expected values of the data points are specified as dense. Compared to an existing PCKA, the proposed EAPC achieved better cluster dimensionality and the variance is 30-40% high.

No. of clusters	Outlier immunity (%)	
	Proposed EAPC	Existing PCKA
2	2	5
4	3	5.6
6	3.4	7
8	3.5	7.2
10	4.2	7.8
12	4	8.4
14	4.9	8.9

Table 5.2 No. of cluster vs. outlier immunity

The above table (table 5.2) describes the presence of outlier immunity based on the number of clusters partitioned with respect to the database. The outlier immunity of the cluster of the proposed enhanced approach for projecting clusters in high dimensional spaces is compared with an existing distance-based projected clustering algorithm.

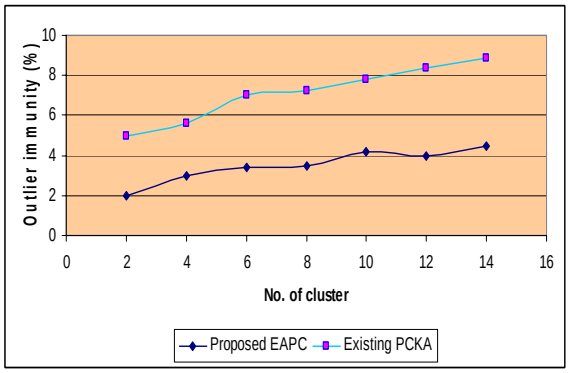

Fig 5.2 No. of cluster vs. outlier immunity

Fig 5.2 describes the presence of outlier immunity based on the number of clusters partitioned with respect to the database. As observed from the figure 5.1, EAPC exhibit reliable performance from the first set of experiments on data sets with no outliers. In tricky cases, EAPC presents much improved results than the existing PCKA algorithm. The results stated in Fig. 5.2 recommend that the proposed EAPC is less receptive to the proportion of outliers in data sets. With respect to the cluster dimensionality, differences in the occurrence of diverse percentages of outliers have no significant impact on the performance of EAPC. This can be enlightened by the real that the outlier managing system of EAPC builds a proficient employment of the information stored in the database, generous it high outlier immunity. Compared to an existing PCKA, the proposed EAPC provides less outlier immunity since it gives better cluster dimensionality result and the variance in outlier immunity is 40-50% low in the proposed EAPC.

Dataset dimensionality	Time consumption (seconds)	
	Proposed EAPC	Existing PCKA
50	110	150
100	200	240
150	220	290
200	280	340
250	340	400
300	400	450
350	435	520

Table 5.3 Dataset dimensionality vs. Time consumption

The above table (table 5.3) describes the consumption of time to perform the projection of clustered high dimensional dataset based on the dimensionality of the dataset described in the

database. The time consumption of the cluster of the proposed enhanced approach for projecting clusters in high dimensional spaces is compared with an existing distance-based projected clustering algorithm.

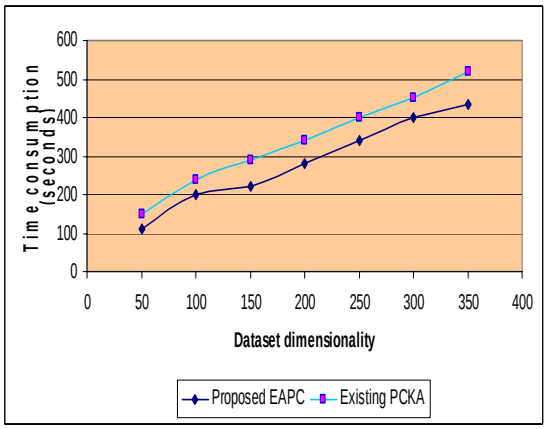

Fig 5.3 Dataset dimensionality vs. Time consumption

Fig 5.3 describes the consumption of time to perform the projection of clustered high dimensional dataset based on the dimensionality of the dataset described in the database. The proposed EAPC balances linearly with the increase in the data dimension. As specified in the scalability experiments with respect to the data set size, the execution time of EAPC is generally provides improved results than that of PCKA when the time required to project the clusters in high dimensionality subspace employed for frequent runs is also incorporated. The time consumption is measured in terms of seconds. Compared to the existing PCKA, the proposed EAPC consumes less time since it gives better cluster dimensionality result and the variance in time consumption is 35-50% low in the proposed EAPC.

6. CONCLUSION

We have proposed an enhanced approach for projecting clusters in high dimensional spaces for the demanding problem of high dimensional clustering, and exemplified the appropriateness of the proposed EAPC algorithm in tests and compared with previous work PCKA. Experiments show that EAPC gives significant results and considerably develops the superiority of clustering when the dimensionalities of the clusters are much lower than that of the data set. Furthermore, our EAPC algorithm provides precise results when managing data with outliers. The performance of EAPC on real data sets proposes that EAPC approach could be an fascinating tool. The accuracy achieved by EAPC results from its constraint of the distance calculation to subsets of attributes, and its process for the original collection of these subsets.

REFERENCES

[1] Mohamed Bouguessa and Shengrui Wang, "Mining Projected Clusters in High-Dimensional Spaces", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 4, APRIL 2009

[2] H. Liu and L. Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 4, pp. 491-502, Apr. 2005.

[3] M. Lung and N. Mamoulis, "Iterative Projected Clustering by Subspace Mining," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 2, pp. 176-189, Feb. 2005.

[4] M. Bouguessa, S. Wang, and H. Sun, "An Objective Approach to Cluster Validation," Pattern Recognition Letters, vol. 27, no. 13, pp. 1419-1430, 2006.

[5] F. Angiulli and C. Pizzuti, "Outlier Mining in Large High-Dimensional Data Sets," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 2, pp. 369-383, Feb. 2005.

[6] A. Gnanabaskaran et. Al., "An Efficient Approach to Cluster High Dimensional Spatial Data Using K-Mediods Algorithim", European Journal of Scientific Research ISSN 1450-216X Vol.49 No.4 (2011), pp. 617-624

[7] C.C. Aggarwal, C. Procopiuc, J. L. Wolf, P.S. Yu, and J.S. Park, "Fast Algorithm for Projected Clustering," Proc. ACM SIGMOD Conf.,pp.61–72,1999. doi:10.1109/ICDE. 2009.188. 1152

[8] Emmanuel M¨ uller et. Al., "Evaluating Clustering in Subspace Projections of High Dimensional Data", Proceedings of the VLDB Endowment Volume 2 Issue 1, August 2009 Pages 1270-1281

[9] E. Muller, I. Assent, et. Al., " OpenSubspace: An open source framework
for evaluation and exploration of subspace clustering algorithms in WEKA", In Open Source in Data Mining Workshop at PAKDD, pages 213, 2009.

[10] E. Muller, I. Assent, and T. Seidl. HSM:Heterogeneous subspace mining in high dimensional data. In SSDBM, pages 497{516, 2009.

[11] I. Assent, R. Krieger, E. M uller, and T. Seidl. DUSC: Dimensionality unbiased subspace clustering. In ICDM, pages 409{414, 2007.

[12] I. Assent, R. Krieger, E. Muller, and T. Seidl. INSCY: Indexing subspace clusters with in-process-removal of redundancy. In ICDM, pages 719{724, 2008.

[13] I. Assent, E. M uller, R. Krieger, T. Jansen, and T. Seidl. Pleiades: Subspace clustering and evaluation. In ECML PKDD, pages 666{671, 2008.

AUTHORS PROFILE

Dr.k.prasadh He is the Principal of Mookambika technical campus muvattupuzha Kerala,india.He had completed his MTECh from MS University and PhD from VM university.His area of intrest include Data mining,Computer Networks ,Image processing

ARUN SHALIN.L.V He had done his BTech and ME from Anna university Chennai. His area of intrest include Database Management systems, Data mining