ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

PROJECTIVE CLUSTERING APPROACH FOR THE DETECTION OF OUTLIER AND NON-AXIS-ALIGNED SUBSPACES

J.Ghayathri[1] N.Surya[2]

Professor, Computer Science Department, Kongu Arts and Science, Erode, India[1]

M.PHIL (CS), Kongu Arts and Science, Erode, India[2]

Abstract: Clustering the case of non-axis-aligned subspaces and detection of outliers is a major challenge due to the curse of dimensionality. To solve this problem, the proposed implementation is extension to traditional clustering and finds subsets of the dimensions of a data space .In this project, a probability model is proposed to describe in hidden views and the detection of possible selection of relevant views. A projective clustering is proposed for Outlier Detection in High Dimensional Dataset that discovers the detection of possible outliers and non-axis–aligned subspaces in a data set and to build a robust initial condition for the clustering algorithm it improves the parameters in the connection between L∞ corsets and sensitivity that is made in Lemma and improve clustering in the case of non-axis-aligned subspaces and detection of outliers in datasets. The suitability of the proposal demonstrated is done with synthetic data set and some widely used real-world data set.

Keywords: Clustering, high dimensions, projective clustering, probability model.

I. INTRODUCTION

DATA clustering has a wide range of applications and has been studied extensively in the statistics, data mining, and database communities. Many algorithms have been proposed in the area of clustering [1][2]. One popular group of such algorithms, the model-based methods, has sparked wide interest becauseof their additional advantages, which give them the capacity to describe the underlying structures of populations in the data [6].

In the model-based method the data are originated from the various sources [9] and that data are modeled by the Gaussian mixture. The aim is to find the mixture of Gaussians that is nature of the Gaussian source. However, such methods would suffer from the curse of dimensionality problem for high dimensional data. The goal of the clustering is to group the data based on the relations between the data. The grouped data is fetched through the projective clustering method which fetches the related data from different cluster groups. Projective clustering is a class of problems in which the input consists of high-dimensional data, and the goal is to discover those subsets of the input that are strongly correlated in subspaces of the original space. Each subset of correlated points, together with its associated subspace, defines a projective cluste[1]r. Thus, although all cluster points are close to each other when projected on the associated subspace, they may be spread out in the full-dimensional space. This makes projective clustering algorithms particularly useful when mining or indexing datasets.

Outlier Detection

The outlier detection is the process of detecting the unclustered data from the dataset. The data [7] in the dataset is clustered according to the relations between the data in the dataset. The data present in the dataset that cannot be grouped according to the relations is identified as outliers [8].

II. RELATED STUDIES

Basic Concepts of Subspace Clustering

A subspace clustering is a collection of subspace clusters. The first2 subspace clustering algorithm CLIQUE was published in 1998 and was soon followed by many related methods [3]. The algorithms have been applied for instance to clustering gene expression data: it is often the case that a group of genes behaves similarly only in a subset of experiments (i.e. in a subspace0 [10]. Reviews of some of the existing subspace clustering algorithms can be found. Other names that have been used for the same or a closelyIn high dimensional spaces1, traditional clustering methods suffer from the curse of dimensionality, which is why their application is often preceded by feature selection and extraction.

For instance, a practitioner might apply Principal Component Analysis (PCA) to project the data onto a low-dimensional subspace before trying to cluster the data points. However, it is sometimes unrealistic to assume that all clusters of points lie in the same subspace of the data

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

space. Subspace clustering methods address this issue by assigning a distinct subspace to each group of data points.

Before proceeding, let us introduce our notation. The data matrix X consists of elements xij 2 R, where i 2 {1, 2, . . . ,m} and j 2 {1, 2, . . . , p}. We denote the m rows by {r1, r2, . . . , rm}, where ri = (xi1, xi2, . . . , xip), and the p columns by {c1, c2, . . . , cp}, where ci = (x1i, x2i, . . . , xmi)T . We will often refer to the rows as data points and to the columns as attributes. A clus- ter Ci _ {r1, r2, . . . , rm} is a subset of the data points. A clustering2 C is a partitioning of the set of m data points into clusters C1,C2, . . . ,CK of sizes m1,m2, . . . ,mK.

III. NON-AXIS-ALIGNED SUBSPACE CLUSTERINGS

A non-axis-aligned subspace cluster S is a pair (R,W), where R _ {r1, r2, . . . , rm} is a subset of the rows and W is a collection of vectors {w1,w2, . . . ,wD}, where wi 2 Rp. The vectors in W form a basis for an arbitrary subspace of the original p-dimensional data space. We use W also to denote this subspace. Naturally, an axis-aligned subspace cluster is a special case of a non-axis aligned subspace cluster. In the case of an axis-aligned subspace cluster, W is a subset of the original basis vectors {e1, e2, . . . , ep}, where e1 = (1 0 0 . . . 0),e2 = (0 1 0 0 . . . 0), etc.

A non-axis-aligned subspace clustering S is a collection {S1, S2, . . . , SK} of K non-axis aligned subspace clusters. The algorithms ORCLUS, KSM , and 4C produce these kinds of clustering. Non-axis-aligned subspace clustering is a generalization of feature extraction; instead of defining a single set of features for the whole data.

Meta-Clustering

Meta-clustering refers to investigating the structure of a set of clustering. Meta-clustering discards the idea of trying to derive a single good clustering for a data set; instead, it is acknowledged that the data can be well represented in several different, complementary ways. For instance, assume that a given data set has been clustered several times by different algorithms. A meta clustered might now observe that these clustering form two tight groups of clustering, and give the user a representative of each of these groups, instead of a single 'best' clustering.

There are various ways to produce different clustering for a data set: we could use different algorithms, a single algorithm with various parameter values and initializations, change metrics, use various dimensionality reduction schemes, or sample the data. Meta-clustering may be used to investigate whether some of these clustering form tight groups, whether some of the clustering are outliers, whether the effect of the parameter values is strong or weak, etc. For instance, it has been empirically shown by means of meta clustering that only a small number of clustering algorithms is enough to represent a large number of clustering criteria.

IV. A NEAR-LINEAR ALGORITHM FOR PROJECTIVE CLUSTERING INTEGER POINTS

A near-linear algorithm for integer (j; k) projective clustering in the L1 sense when the dimension is part of the input. Recall that in this problem we are given a set P of n points in Rm and integers j _ 1, k _ 0, and the goal is to find j k-subspaces so that the sum of the distances of each point in P to the nearest subspace is minimized; the point coordinates are integers of magnitude polynomial in m and n. Our randomized algorithm , for any parameter " > 0, runs in time O(mn polylog(mn)) and outputs a solution that with constant probability is within (1 + ") of the optimal solution.

A. Probability Model

It is important to note that the Gaussian mixture is a fundamental hypothesis that many model-based clustering algorithms make regarding the data distribution model. In this case, data points are thought of as originating from various possible sources, and the data from each particular source is modeled by a Gaussian.

B. Chenet al.: model-based method for projective clustering

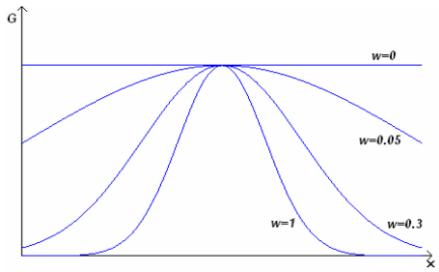

Fig.1.Model-based method for projective clustering

Changes in probability density with different weighting values

C. j-flat Fitting Using Lemma Concepts:

In this section, we consider the j-at fitting problem. We first introduce the concept of shape kernel and then use it to derive PTAS for the j-at fitting problem. To solve the j-at setting problem, one way is to use the concept of kernel set introduced by Agarwal et al. in. For a set P of Rd points, its kernel set is a new set of Rd points of size O (1 (__)(d□1)=2) which can be constructed through an _-net inside a unit sphere, where _ is a measure of the fatness of P. Kernel set captures the structure and extent of P and is rather powerful for solving many problems. Despite the obvious advantages provided by kernel set, there are also some issues when used for solving the RPC problem, which leads us to adopt a different structure called shape kernel. One issue is that the value of _ could be large for some point sets.
Although as pointed out in [1], it can be reduced by using some linear transform on the point set. However, this seems to be difficult to extend to the case of k _ 2 (i.e., multiple j-ats as in the RPC problem), as there may not exist a single linear transform for all j-ats. Another issue is

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

that kernel set maintains more than succinct information for RPC. For RPC, it is actually succinct to maintain a small set of points which jointly approximate the mean of the original point set. One consequence of the redundant information in the kernel set is that its size could still be relatively large, making it di_cult to further improve the total running time of kernel set based algorithms. To resolve the aforementioned issues, we use a different strategy to construct the kernel.

V. AVERAGE VPC OF THE THREE FUZZY CLUSTERING ALGORITHMS

Average FScore of the algorithms, with increment of variances on the relevant dimensions algorithms choose their initial cluster centers via some random selection methods, and thus the clustering results may vary depending on the initialization. Figs. 3 and 4 show the average results of the algorithms on these data sets, in terms of VPC and FScore, respectively. Detailed clustering results on the data set with s ¼ 8, which is the most difficult case of the seven data sets (as shown in Table 2), are illustrated in Table 3. The values in the max columns correspond to the best results of the algorithms, and the average results are reported in the format average _ 1 standard deviation in the table. Figs. 3 and 4 show that outlier is able to achieve high quality overall results, especially when the clusters overlap considerably, whereas FCM, Fuzzy-FWKM, and EWKM perform poorly, and the other algorithms encounter difficulties when the cluster overlapping becomes significant, i.e., when s > 6. Examining these results in more detail, we can see that the values of VPC yielded by FCM and Fuzzy-FWKM are close to 1 K , which indicates that these two algorithms tend to assign each point to all the clusters with approximately equal membership degrees. This is due to the fact that FCM measures the similarity between data points by considering all features of a data set. With the high-dimensional data used in the experiments, such a similarity measurement in the entire data space would be less meaningful due to the empty space phenomenon . Fuzzy-FWKM employs a feature weighting mechanism in the clustering process; however, each dimension is assigned the same weight for different clusters in this algorithm.

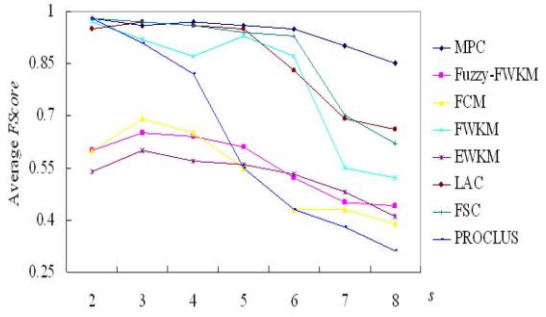

Fig 2. Average vpc of the three fuzzy clustering algorithms

Our result is a near-linear algorithm for integer (j; k) projective clustering in the L1 sense when the dimension is part of the input. Recall that in this problem we are given a set P of n points in Rm and integers j _ 1, k _ 0, and the goal is to find j k-subspaces so that the sum of the distances of each point in P to the nearest subspace is minimized; the point coordinates are integers of magnitude polynomial in m and n. Our randomized algorithm, for any parameter " > 0, runs in time O(mn polylog(mn)) and outputs a solution that with constant probability is within (1 + ") of the optimal solution. To obtain this result, we observe that in a fairly general sense, shape setting problems that have small core sets in the L1 setting also have small corsets in the L1 setting. Using this observation, and the corset construction of for the L1 setting in axed dimension, we are able to obtain a small core set for the L1 setting in axed dimension. To solve the problem when the dimension is part of the input, we use a known dimension reduction result.

VI. PROPOSED METHOD

The detection of outliers in the high dimensional dataset is major challenge because of its dimensionality. To solve this problem, the proposed implementation is extension to traditional clustering and finds subsets of the dimensions of a data space .In this project, a probability model is proposed to describe in hidden views and the detection of possible selection of relevant views.

In this paper, we first discussed the problem of providing a probability model to describe projected clusters in high dimensional data. The experiments were conducted on cancer datasets, airline datasets which used in real-world applications and the results show the effectiveness of outlier.

There are many directions that are clearly of interest for future exploration. One avenue of further study is to extend outlier to the case of non-axis-aligned subspaces. Another interesting extension would be for the detection of possible outliers and the subspaces of the low dimensional data in a data set. Our future efforts will also be directed toward developing techniques to build a robust initial condition for the clustering algorithm.

Experimental Results

The algorithms are implemented to detect the Outlier and subspace dimensionality of the low dimensional data in the cancer and airline dataset.

The Fuzzy-FWKM is not a projective clustering algorithm which employs the weighting mechanism in which each dimension is assigned same for different clusters. The graphs were drawn to find out the effectiveness of the outlier detection and subspace clustering in the concert. The graph compares the two fields in the dataset and shows the comparison result of the dataset**.**

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 4, April 2013

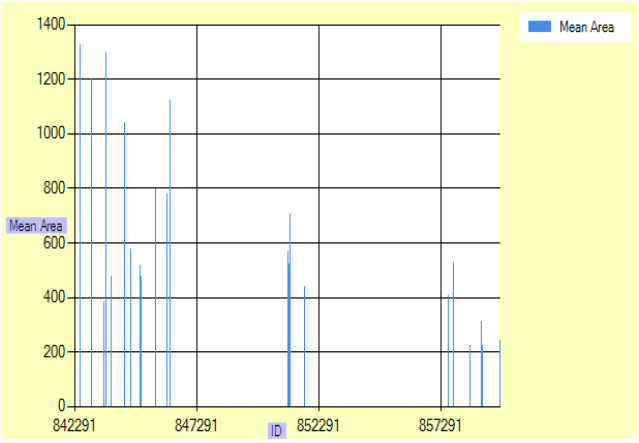

Fig 3. Comparison of Mean area vs. ID

The rate of the mean varies from one ID to another ID and the result of the comparison is shown as a graph. The cancer dataset consist of values in it and the graph is drawn according to that graph. In the cancer dataset the ID and the mean values are compared, the values of the graph are listed according to the values in the dataset. The every field in the dataset is compared and the graph can be drawn according to their values in it.

VII. CONCLUSION

In this paper, we first discussed the problem of providing a probability model to describe projected clusters in high dimensional data. This problem becomes difficult due to the sparsity of high-dimensional data and the fact that only a small number of the dimensions may be considered in the clustering process. In this model method to detect the outliers that exist in the database without the data clustering will be detected.

The subspace clustering finds the low dimensional data in the clustered data set that has been used in our experiment. The result of the Cancer data set will result in both the outlier detection and subspace clustering of the data in the database. The Gaussian model that satisfies all the criteria that is accepted by the projective clustering. The experiments show that outlier is suitable for clustering real-world data especially for e-mail documents. To confirm the suitability of our algorithm for document clustering, the capability of outlier in identifying the keywords of document categories is analyzed below. From the subspaces of resulting clusters, we can obtain the relevant dimensions that represent important keywords by sorting the dimension weights in descending order.

REFERENCES

[1] R.K. Agarwal and N.H. Mustafa, "K-Means Projective Clustering," Proc. ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems (PODS), pp. 155-165, 2004
[2] M.J. Zaki and M. Peters, "Clicks: Mining Subspace Clusters in Categorical Data via Kpartite Maximal Cliques," Proc. Int'l Conf. Data Eng. (ICDE), pp. 355-356, 2005.
[3] M. Dutta, A.K. Mahanta, and A.K. Pujari, "QROCK: A Quick Version of the ROCK Algorithm for Clustering of Categorical Data," Pattern Recognition Letters, vol. 26, pp. 2364-2373, 2005.
[4] K. Jain and R. C. Dubes. "*Algorithms for Clustering Data*." Prentice Hall, 1988.
[5] T. Zahn. Graph-theoretic methods for detecting and describing gestalt clusters. *IEEE Transactions on Computing*, 20(31):68–86, 1971.
[6] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review," ACM Computing Survey, vol. 31, no. 3, pp. 264-323, 1999.
[7] F. Angiulli and C. Pizzuti. Fast outlier detection in high dimensional spaces. In PKDD '02: Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery, pages 15–26, 2002.
[8] F. Angiulli and C. Pizzuti. Outlier mining in large high dimensional data sets. IEEE Transactions on Knowledge and Data Engineering, 17:203–215, 2005.
[9] Lifei Chen, Qingshan Jiang, and Shengrui Wang, "Model-based Method for Projective clustering". IEEE Transactions on Knowledge and Data Engineering,
[10] Chu, Yi-Hong, "Density Conscious Subspace Clustering for High- Dimensional Data". IEEE Transactions on Knowledge and Data Engineering

BIOGRAPHY

J.Ghayathri, Professor, Department of Computer Science, Kongu Arts and Science College, Erode, Tamil Nadu, India.

N.Surya, Pursuing MPhil, Kongu Arts and Science College, Erode, Tamil Nadu, India