Analyzing The Performance Of Model Based Projective Clustering

[1] *Sathya Sharmila.V , PG Scholar, Department of Computer Science and Engineering , Sri Shakthi Institute of Engineering and Technology Coimbatore-641 062*
[2] *Kannammal.K.E ,Head of the Department, Department of Computer Science and Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore-641 062*

Abstract - *Clustering aims to find the structure in the collection of unlabelled data. The data objects of same cluster will have certain relationships among them. Therefore, the cluster relationship between the data objects must be assumed by each clustering technique. But handling the data items of larger amount is difficult because of the high dimensionality of datasets and the time required to cluster those data items. So high dimensional data clustering is a challenging area in data mining technique. The Projective clustering method is appropriate to cluster high dimensional data items that are projected in various subspaces. The model based projective clustering, which is based on feature weighting is an extension to traditional clustering technique that attempts to cluster the data items with overlapping subspace along the various subset of attributes. The performance of the model based method for projective clustering is evaluated with the real world data set.*

Keywords- Data mining, Clustering, Projective clustering, High dimension, Feature weighting

1. INTRODUCTION

The explosive growth of data in this scientific era requires powerful tools to discover useful information from huge amount of data. This has lead to the advent of DataMining. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data collected from various sources. The sources of data may consists of databases, web or data warehouses. There are number of data mining functionalities such as Classification, Regression, Clustering, etc,. These functionalities uses class labels to analyze the given data set. But, there are situations where the data items may not be characterized by the class labels. So the clustering is a method which is used to analyze the data items having no class labels. It efficiently generates a class label for the data items where it is applied on. Thus clustering plays a major role in data mining technology.

Clustering is a division of data into group of similar objects. Each group, called a cluster, consists of objects that are similar between themselves and dissimilar compared to objects of other groups. Its main aim is to maximize the intra-cluster similarity and to minimize the inter-cluster similarity. The analysis of cluster can be performed by different kinds of algorithms that differ considerably in the way they analyze the data items and they generate the cluster formation. Clustering is characterized by variety of concepts such as the distance between the objects of same cluster in such a way that smaller the distance between the objects larger the probability to be in the same cluster, density of the data items projected in the original data space and statistical distributions. So based on the data set to be considered and the required results the specific algorithms can be formulated. These algorithms may require initial parameter settings which is based on the domain knowledge. Cluster analysis is an iterative task to discover the knowledge. In each iteration the parameters are modified appropriately until it obtains the desirable solution. The model parameters may vary for each of the clustering technique and in some cases it must be predefined by the user.

The requirements that clustering algorithm should satisfy are:

1. To discover the clusters in the arbitrary shape
2. To determine the input parameters, minimal requirements for domain knowledge is required
3. To deal with noise and outliers
4. High dimensionality
5. Interpretability and usability

2. RELATED WORKS

Clustering techniques can be broadly classified into two categories: Partitional clustering and Hierarchical clustering. Given a dataset to be clustered and a clustering criterion, partitional clustering technique obtains a division of the objects into clusters such that the objects of a cluster are similar to each other and dissimilar to objects in different clusters. Given a database of n objects, it

constucts k partitions such that k<=n where each partition represents a cluster. This method has two main objectives: Each group must contain atleast one object.Each object must belong to exactly one group. The representative for the k clusters is determined by the popular k-means and k-medoid methods. They allocate each object to the cluster having the closest representative to the object such that the sum of the squared distances between the objects and their representatives is reduced. The Hierarchical clustering produces a set of nested clusters that can be represented as a hierarchical tree.It is visualized as a dendrogram which is a tree like structure that records the merges and splits. It does not need to assume any particular number of clusters. Any desired number of clusters can be formed by 'cutting' the dendrogram at any level and each branch forms a cluster. It requires a termination condition. A decision made to merge or split the clusters cannot be undone. It can be a top-down splitting or bottom-up merging. BIRCH, ROCK and CURE are some of the hierarchical methods. Both the technique requires the user to specify the initial parameters. The Density based clustering finds the dense clusters without considering the representative objects. The Grid based clustering deals with the data space instead of the data objects.

The similarity between any two objects is measured using sum of squared error, absolute error criterion or maximum likelihood estimation depending upon the clustering criterion. The dissimilarity matrix or proximity matrix can be used to record the estimated similarity measures between the objects. The discussed methods are useful only for the dataset having low dimensions. The Feature Transformation Techniques and Feature Selection Techniques are useful to reduce the number of dimensions to be handled but they are performed in the entire data space which encounters difficulties to find the clusters in the different subspaces. Local Dimensionality Reduction,LDR attempts to create a new set of dimensions for each cluster. But determining the required dimensions for each subspace associated with the clusters is a difficult task. LDR has high computational complexity. The Hierarchical clustering can be performed but it requires that the clusters to be merged must have same dimensions.

3. MODEL BASED PROJECTIVE CLUSTERING

Formulating a desired similarity measure is very difficult for data objects lying in different subspaces. So the high dimensional data analysis requires a valuable clustering technique. The traditional methods uses the Euclidean distance between the data objects which is not enough for high dimensional objects. Several analysis shows that the similarity measures based on all the dimensions is ineffective. Therefore, the dimensions are weighted equally to compute the distance between the clusters. It is supported for the clusters that are projected in Non-axis aligned subspace.

A. Algorithm

The algorithm takes the database and the number of clusters as input. It outputs the membership matrix, cluster center matrix and the weight matrix.

Input: DB, Cluster K
Output: Membership matrix U, Vector Matrix V, Weight Matrix W
Begin
Let number of iteration be p

1. Initialization
   Randomly choose K cluster center
   Assign V =V(0),W=W(0)
2. Repeat
   Let V=V(p), W=W(p) and Z=Z(p)
   Compute U(p+1)
   Let U=U(p+1), Compute V(p+1)
   Let V=V(p+1), Compute W(p+1)
   p=p+1
3. Output
   U ,V and W
   End

B. System Model

A database DB = {x1;x2;...;xN} which contains K clusters. x_i where i=1;2;...;N are called data points in the D-dimensional space. It is assumed that the data set has been normalized such that each x_{ij} = [0, 1] where j=1;2;...;D. The membership degree of xi with respect to the kth cluster c_k, where k = 1;2;...;K, is denoted as u_{ki}. The cluster c_k is associated with a weight vector w_k = <wk1;wk2;...;wkD>. Here, the weight w_{kj} is the measure of the relevance of the j^{th} dimension to c_k. If the dimension is highly relevant, then the higher weights will be assigned.

4. DATASET COLLECTION

A dataset is a collection of data. A single database table, corresponds to the contents of a dataset, or a single statistical data matrix, In the table each column, represents a particular variable, and each row in the dataset is corresponds to a given member. Each variables in the dataset, list the values, such as weight and height of the object, for each member in the dataset. The each value in dataset is known as a datum. The Datasets are characterized by the parameters used in the data generation process, including N (the number of data points), D (the data set dimensionality) and K (the number of clusters).

A.WINE Dataset:

The title of the Database is Wine recognition data. It was updated on Sept 21, 1998 by C.Blake. The data items are the chemical characteristics of wines obtained during the analysis. They are derived from three cultivars of same region in Italy. The analysis focused on determining the quantities of thirteen constituents found in each of the wine. The number of attributes in this dataset is thirteen.

The properties of the attributes of the dataset are,

- All the attributes are continuous.
- No Missing Attribute Values

The attributes defined in this WINE dataset includes,

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

5. SYSTEM IMPLEMENTATION

The first step is to initialize cluster centers and to assign weights with equal values for each dimensions. Assign the data points to respective clusters based on the membership degree of each data points with respect to each cluster center and the associated weights for each dimension. Recalculate centers and weight and reassign the data points iteratively. The assignment of weights can be done in two ways. The Hard clustering assigns weights with values either 0 or 1.The soft clustering assigns weights in the range [0,1] based on the probability that the dimension is appropriate to the respective cluster. The larger the weights,the larger its appropriateness.

A.Preprocess:

In Membership Function, data's are bound to each cluster. This represents the fuzzy 743ehavior of the algorithm. The appropriate matrix named U, want to simply build, the factors are numbers between 0 and 1, and that represents the membership of degree between the data center of cluster. . Uki is the membership degree of Xi in all the clusters where k=1,2,..K and i=1,2,..N. This process is performed as a preprocess step.

B. Cluster Center Marix Generation:

The cluster centers can be initially specified with a K-by-p matrix, or else chosen from the matrix X with a predefined number of clusters. Through the observation select the Initial Cluster Centers, The cluster centers are assigned as first K observations. Second, place the current cluster centers for loop and check each remaining observation to see whether it can replace current cluster centers. If the distance between the closest cluster center and observation and its is greater than the distance between the two closest cluster centers (Cluster I and Cluster j), the observation will replace the cluster center of I or j cluster depending on which one is closer to the observation. Allocate observations to the closest cluster. If the distance between the observation and the second closest cluster center is greater than the closest distance between the closest cluster center and other cluster centers, then the observation will replace with the closest cluster center.

C. Weight Calculation:

The distribution of feature weights may thus be rigorously unfair when most of the dimensions are irrelevant to the clusters. As discussed previously, weights correspond to cluster shapes, in the model-based approach of MPC. The Irrelevant dimensions will result in a very large dimension-dependent variances. The trade-off of all the dimensions, the dimension independent cluster size parameter are determined, the dimension weights w_{kj} corresponding to large w_{kj} are forced to become very

small, which, in turn, shrink the effect of these irrelevant dimensions in cluster formation.

6. RESULTS EVALUATION

To measure the quality of the results of the clusters formed, the partition coefficient $V_{PC,}$ the cluster validity index can be used. The value of V_{PC} is related to the uncertainty of the memberships which depends upon the overlapping of two different clusters. If the obtained value of V_{PC} is larger, then the memberships are less fuzzy. It is defined as:

$$V_{PC} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=0}^{N} U_{ki}^2$$

K=number of clusters

N=number of data points

U_{ki}=membership degree of x_i with respect to the cluster k

The value of V_{PC} obtained by the MPC algorithm for clustering the WINE Dataset having 13 dimensions is 0.86 which is a good result comparing with other well known and proven projective clustering algorithms.

7. CONCLUSION

The problem is to describe the projected cluster in high dimensional data. It became difficult due to the sparsity of high dimensional data and the fact that only small number of dimension is to be considered in the clustering process. The MPC model which is suitable for all criteria and it is more robust projective clustering algorithm. The experiments were conducted on the real world dataset and result shows the effectiveness of MPC in the case of high dimensional dataset.The future work will also be directed towards developing techniques based on the robust initial condition for the clustering algorithms. The MPC algorithm can also be extended to handle the data items in non-axis aligned subspaces.

REFERENCES

[1] 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.

[2] S.B. Kotsiantis and P.E. Pintelas, "*Recent Advances in Clustering: A Brief Survey*" WSEAS Trans. Information Science and Applications,vol. 11, no. 1, pp. 73-81, 2004.

[3] T. Hastie, R. Tibshirani, and J. Friedman, "*The Elements of Statistical Learning: Data Mining*", Inference, and Prediction, second ed. Springer-Verlag, 2001.

[4] S. Dasgupta, "*Learning Mixtures of Gaussians,* "Proc.Ann. Symp .Foundations of Computer Science, pp. 634-644, 1999.

[5] S. Wang and H. Sun, "*Measuring Overlap-Rate for Cluster Merging in a Hierarchical Approach to Color Image Segmentation,"* Int'l J. Fuzzy Systems, vol. 6, no. 3, pp. 147-156, 2004.

[6] M. Steinbach, L. Ertoz, and V. Kumar, "*The Challenges of Clustering High Dimensional Data,"* http: //www-users. cs.umn.edu /ertoz/papers/ clustering_ chapter. pdf, 2003.

[7] M. Verleysen, "*Learning High-Dimensional Data*", Limitations and Future Trends in Neural Computation,pp. 141-162, IOS Press,2003.

[8] A. Hinneburg, C.C. Aggarwal, and D.A. Keim, "*What Is the Nearest Neighbor in High Dimensional Spaces,"*Proc. Int'l Conf.Very Large Databases (VLDB),pp. 506-515, 2000.

[9] C.C. Aggarwal, C. Procopiuc, J.L. Wolf, P.S. Yu, and J.S. Park,"*Fast Algorithm for Projected Clustering,"*Proc. ACM SIGMODInt'l Conf. Management of Data,pp. 61-71, 1999.

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

[11] C. Bohm, K. Kailing, H.P. Kriegel, and P. Kroger, "*Density Connected Clustering with Local Subspace Preferences*", Proc.IEEEInt Conf. Data Mining (ICDM), pp.27-34, 2004.

[12] L. Jing, M.K. Ng, J. Xu, and J.Z. Huang, "*A Text Clustering System Based on k-means Type Subspace Clustering*", Int'l J. Intelligent Technology, vol. 1, no. 2, pp. 91-103,2006.

[13] M. Bouguessa, S. Wang, and H. Sun, "An *Objective Approach to Cluster Validation*",Pattern Recognition Letters, vol. 27, pp. 1419-1430, 2006.

[14] D. Lowd and P. Domingos, "*Naive Bayes Models for Probability Estimation*", Proc. Int'l Conf. Machine Learning (ICML), pp. 529-536,2005.

[15] Y.M. Cheung, "*K-Means: A New Generalizedk-Means Clustering Algorithm*", Pattern Recognition Letters, vol. 24, pp. 2883-2893,2003.

