International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume I, Issue VI, OCT - 2014

HIGH DIMENSIONAL DATA WITH SUBSPACE AND OUTLIER ANALYSIS USING MODEL BASED CLUSTERING ALGORITHM

Ms K.Abinaya
M.E- CSE, Nandha College of Technology, Erode.
abinayaks92@gmail.com
Mr.Dr.M.Vijaya kumar, M.E.Ph.D.,
Professor, Nandha College of Technology, Erode
tovijayakumar@gmail.com

ABSTRACT

Data mining methods are applied for knowledge extraction on databases. The data clustering schemes are used to group up the relevant records based on the similarity. Similarity measures are used to analyze the relationship between the transactions. Vector based similarity models are suitable for low dimensional data values. High dimensional data values are clustered using subspace clustering methods.High dimensional data partitioning process requires an efficient similarity analysis mechanism. Projective clustering attempts to find projected clusters in subsets of the dimensions of a data space. Probability model describes projected clusters in high-dimensional data space. Model-based algorithm for fuzzy projective clustering that discovers clusters with overlapping boundaries in various projected subspaces. Model Based Projective Clustering (MPC) algorithm is used in the system. Subspace clustering methods are used to cluster the high dimensional data values. The model based projective clustering algorithm is a subspace clustering technique. Non-axis-subspaces are used with similarity analysis. Anomaly transactions are partitioned with projected clusters. The proposed system is designed to perform clustering on high dimensional spaces. Non access subspaces are included in the similarity analysis. Anomaly data values are verified with similarity under the clustering process. The subspace selection process is optimized.

1. Introduction

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets, so that the data in each subset share some common trait - often proximity according to some defined distance measure. Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Besides the term data clustering, there are a number of terms with similar meanings, including cluster analysis, automatic classification, numerical taxonomy, botryology and typological analysis.

Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation. Additional analyses using standard analytical and other data mining techniques can determine the characteristics of these segments with respect to some desired outcome. For example, the buying habits of multiple population segments might be compared to determine

which segments to target for a new sales campaign.

Data clustering algorithms can be hierarchical or partitional. Hierarchical algorithms find successive clusters using previously established clusters, whereas partitional algorithms determine all clusters at once. Hierarchical algorithms can be agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

Two-way clustering, co-clustering or bi-clustering are the names for clusterings where not only the objects are clustered but also the features of the objects, i.e., if the data is represented in a data matrix, the row and columns are clustered simultaneously. Another important distinction is whether the clustering uses symmetric or asymmetric distances. A property of Euclidean space is that distances are symmetric.

2. High Dimensional Clustering

In model-based methods, data are thought of as originating from various possible sources, which are typically modeled by Gaussian mixture. The goal is to identify the generating mixture of Gaussians, that is, the nature of each Gaussian source, with its mean and covariance. Examples include the classical k-means and its variants. However, such methods would suffer from the curse of dimensionality problem for high dimensional data.

In a high-dimensional space, clusters may exist in different subspaces comprised of different combinations of features. In many real-world applications, in fact, some points are correlated with a given set of dimensions, and others are correlated with different dimensions [2]. For example, in document clustering, clusters of documents on different topics are characterized by different subsets of keywords. The keywords for one cluster may not occur in the documents of other clusters. To address the above challenges, projective clustering has been defined to find clusters in different subspaces of the same data set.

A projected cluster is an ensemble of subsets of points, each of which is associated with a subset of attributes. Two different projected clusters are illustrated for a set of data points in 3-dimensional space. There are two clusters in this example; however, they are associated with two different low-dimensional subspaces. The first cluster corresponds to the data in group C_1, which are close to each other when projected into the subspace consisting of the dimensions A_1 and A_2, while the second one corresponds to the data in group C_2 projected onto the A_1 - A_3 plane.

In this paper, a new model-based method for projective clustering proposed. The first contribution is the proposal of a probability model to describe projected clusters in a high-dimensional space. In contrast to existing models for high-dimensional data clustering, our extended Gaussian model is designed for projective clustering, and by analysis is able to explain the general assumptions used in popular projective methods. Second, system derive an objective function for projective clustering based on the probability model and propose an EM-type, parameter-free algorithm, named MPC, for optimizing the objective function. The performance of MPC has been evaluated on synthetic data sets and some widely used real-world data sets, and the experimental results show its effectiveness. The method presented in this paper is very different from the one in our previous work [4]. Although the basic density function of the projected cluster is reused, the probability model for projected clusters has been changed. This results in a different algorithm which is no more dependent on any user-defined parameter for updating the dimension weights. The new algorithm has been much better

motivated, analyzed, and experimentally evaluated.

3. Related Work

3.1 Techniques for High-Dimensional Clustering

Techniques for dimensionality reduction have been used in high-dimensional data clustering. Feature transformation techniques, such as PCA and SVD, attempt to summarize the data set in a smaller number of new dimensions created via linear combination of the original attributes, while feature selection methods select only the most relevant attributes for the clustering task. Because these traditional techniques are performed in the entire data space, they may encounter difficulties when clusters are found in different subspaces. Local Dimensionality Reduction (LDR) attempts to create a new set of dimensions for each cluster. The difficulties with such method include the determination of dimensionality for each subspace associated with the clusters. Additionally, LDR often has high computational complexity.

Biclustering also referred to as coclustering, has been proposed for simultaneous clustering on the data points and dimensions of high-dimensional data. One of its typical applications is in the analysis of gene expression data, where the task is to find subgroups of genes and subgroups of conditions such that the genes exhibit highly correlated activities for every condition. Finally, two related terms occur in the literature: subspace clustering and projective clustering. According to Parsons et al., projective clustering algorithms constitute a particular category of the subspace clustering techniques. However, different views are put forward elsewhere in the literature: see for instance [3]. System adopt the taxonomy and make a distinction between the two terms based on the ideas behind them. The idea of subspace clustering is to identify all dense regions in all subspaces, whereas in projective clustering the main focus is on discovering clusters that are projected onto particular spaces. In the subspace clustering field, CLIQUE was the pioneering approach, followed by a number of algorithms such as ENCLUS and MAFIA and SUBCLU. The major concern of this paper is projective clustering. In the following pages, system will focus only on such techniques.

3.2 Projective Clustering Methods

Projective clustering is typically based on feature weighting. Each dimension of each cluster is assigned a weighting value, indicating to what extent the dimension is relevant to the cluster. Usually, the weighting values of a given dimension may be different for different clusters. Based on the way the weights are determined, projective clustering algorithms can be divided into two categories: hard subspace clustering and soft subspace clustering.

In the first category, the dimensions are assigned weights with values of either 0 or 1, resulting in hard feature weighting for the subspaces. PROCLUS, which is based on the traditional k-medoids approach, is a representative algorithm using this weighting scheme. PROCLUS samples the data, then selects a set of medoids and iteratively improves the clustering, with the goal of minimizing the average within cluster dispersion. For each medoid, a set of dimensions is chosen whose average distances to the medoid are small compared to statistical expectation. After the subspaces have been identified, an average Manhattan segmental distance is used to assign points to medoids. PROCLUS requires users to provide the average number of relevant dimensions per cluster, which is usually unknown to users.

FINDIT, which uses a distance measure called the Dimension-Oriented Distance (DOD), is similar in structure to PROCLUS. As a hierarchical clustering algorithm, HARP automatically determines the relevant attributes

International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume I, Issue VI, OCT - 2014

of each cluster without requiring user-defined parameters. HARP is based on the assumption that two data points are likely to belong to the same cluster if they are very similar to each other along many dimensions. DOC also defines the subspace as a subset of attributes on which the projection of points in a partition is contained within a segment. DOC computes projected clusters using a randomized algorithm to minimize a certain quality function. MINECLUS improves on DOC by transforming the problem of finding the projected clusters into the problem of mining the frequent item set.

PROCLUS and the other algorithms mentioned above search for axis-aligned subspaces for the clusters, while some other methods search more general subspaces, termed nonaxis-aligned, where the new features are linear combinations of the original dimensions. ORCLUS is a generalization of PROCLUS that can discover clusters in arbitrarily oriented subspaces. By covariance matrix diagonalization, ORCLUS selects the eigenvectors corresponding to the smallest eigenvalues of the matrix of the set of points. ORCLUS inherits the weaknesses of PROCLUS mentioned above. KSM, a k-means type projective clustering algorithm, determines the non-axis-aligned subspaces by SVD computations, while EPCH performs non-axisaligned projective clustering by histogram construction.

Instead of identifying hard subspaces for clusters, the algorithms in the second category assign weights in the range [0, 1]. Since the weights can be any real number in [0, 1], system can call these soft projective clustering algorithms. Typically, the weight value for a dimension in a cluster is inversely proportional to the dispersion of the values from the center in the dimension of the cluster. In other words, a high weight indicates a small dispersion in a dimension of the cluster. Virtually all of the existing algorithms in this category are based on the following general assumptions: 1) the data project along a significant dimension onto a smaller range of values than on the other dimensions; 2) the data are more likely to be uniformly distributed along each irrelevant dimension. Proposed system will examine the capabilities of our projective clustering model, presented below, with respect to these two general assumptions.

A number of soft projective clustering algorithms have been reported recently. In [8], an algorithm making use of particle swarm optimization is presented. Since a heuristic global search strategy is used, the near-optimal feature weights could be obtained by this algorithm; however, it would run more slowly than other algorithms. To build an efficient soft projective clustering algorithm, the k-means type structure has been widely adopted. Based on the classical k-means clustering process, an additional step for computing the weighting values is added in each iteration in these algorithms, which include EWKM, FWKM, LAC and FSC [5], etc. Algorithm 1 shows a typical structure for these algorithms.

Input: the dataset and the number of clusters K;
Output: the partition C and the associated weights W;
Begin
  Find the intial cluster V and set W with equal v values;
  Report
    1. Re-group the dataset into C according to V and W;
    2. Re-compute V according to C;
    3. Re-compute W according to C;
  Until convergence is reached;
end

From Algorithm 1, the common projective clustering algorithm can be thought of as an EM-based process for estimating the unknown parameters C, V, and W of a model F(C, V, W) from which the data originate. However, the underlying F(C, V, W) is generally neglected in the above methods. The lack of such a model makes derivation of more

International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume I, Issue VI, OCT - 2014

effective clustering algorithms difficult. This has led us to work on projected cluster modeling, since system are convinced this type of modeling process allows us to benefit from the full potential of cluster analysis: for example, in describing the underlying mechanism that generates the cluster structure and addressing cluster validity problems.

In a typical model-based clustering analysis, one tries to find a mixture of multivariate distributions to approximate the data. Due to the empty space phenomenon and the property of projective clustering, as mentioned above, cluster modeling on high-dimensional data is a difficult problem. In one of the few attempts to use model-based high-dimensional data clustering, Hoff [7] proposed a model of "clustering shifts in mean and variance" based on a nonparametric mixture of sequences of independent normal random variables. The model is learned by a Markov chain Monte Carlo process; however, its computational cost is prohibitive. Harpaz et al. [1] presented a nonparametric density estimation modeling technique, where the data are described as a mixture of linear manifolds. A Bayesian approach is used to identify groups of points that fit or are embedded in lower dimensional linear manifolds. The low dimensional subspaces associated with the individual clusters are computed by PCA. The problems with this method lie in its inflexibility in determining the dimensionality of the subspaces, and its inefficient clustering process.

4. A Probability Model For Projective Clustering

The attributes of a non-axis-aligned subspace are typically combinations of the dimensions of the original data space. Since they are difficult to interpret, often making the clustering results less useful for many real applications, such as document clustering, only projected clusters in axis-aligned subspaces are formalized in the following presentation.

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 [6]. 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. However, Gaussian functions are not appropriate in high-dimensional space due to the curse of dimensionality.

$X_i=<x_{i1},x_{i2},\ldots\ldots x_{iD}>$	Ith data point R^D, i=1,2,……..,N
$DB=\{x_1,x_2,\ldots\ldots x_N\}$	The data set
K	Number of clusters
$c_1,c_2,\ldots\ldots c_k$	K clusters of DB
u_{ki}	Membership degree of x_i in c_k,k=1,2,….K
$U=\{ u_{ki} \}k\times N$	Membership matrix, where k=1,2,…..,k and i=1,2,…….N
$v_k =<v_{k1}, v_{k2,\ldots\ldots\ldots}, v_{kD} >$	Cluster center vector of c_k
$V=\{ v_{kj} \}k\times D$	Cluster center matrix, where k=1,2,……k and j=1,2,……..,D
$w_k =<w_{k1}, w_{k2,\ldots\ldots\ldots}, w_{kD} >$	A weight vector associated with c_k
$W=\{ w_{kj} \}k\times D$	Weight matrix, where k=1,2,…….,k and j=1,2,…….,D

TABLE 1: Notation Used throughout the Paper

In order to learn the underlying structure of clusters in a high-dimensional space, the distribution on each dimension examined. Consider the projections of the data points of the cluster k onto the jth dimension. It is reasonable to describe the projections using a 1D Gaussian function. The probability density function is

$$G(y_j|\mu_{kj};\sigma_k)=\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{1}{2\sigma_k^2}(y_j-\mu_{kj})^2\right),$$

International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume I, Issue VI, OCT - 2014

where μ_{kj} and σ_k denote the mean and covariance of the Gaussian. The above expression thus becomes

$$G(x_j|v_{kj}, w_{kj}; \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{w_{kj}}{2\sigma_k^2}(x_j - v_{kj})^2\right). \quad (5)$$

The major difference between (5) and the standard Gaussian is the introduction of the weighting value wkj, indicating the contribution of the jth dimension to c_k. Curves of (5) with different values of w_{kj} and a fixed σ_k. As system can see, the smaller the weighting value, the more uniformly distributed the data points. With a large weighting value, the data points would distribute within a small range. Note that the characteristic of this extended Gaussian meets the general requirements of projective clustering.

The probability model is created based on the following two assumptions. First, it is assumed that the distribution of points on each of the dimensions spanning the subspace is independent of the others. Although this assumption may not be realistic in some applications, it is a common assumption in many qualitative models, which allows us to approximate a joint distribution of the set of uncorrelated variables by the product of their marginals. Second, it is assumed that variations of points are independent of each other. Because

$$\int G(x_j|v_{kj}, w_{kj}; \sigma_k)\mathrm{d}x_j = \frac{1}{\sqrt{w_{kj}}},$$

The N inputs $x_1, x_2, \ldots, x_N$ are independently and identically distributed from the following mixture density population:

$$F(\mathbf{x}; \Theta) = \sum_{k=1}^{K} \alpha_k \prod_{j=1}^{D} \sqrt{w_{kj}} G(x_j|v_{kj}, w_{kj}; \sigma_k)$$

with

$$\sum_{k=1}^{K} \alpha_k = 1,\ \alpha_k \geq 0,\ k = 1, 2, \ldots, K, \quad (6)$$

where $\theta = \{(\alpha_k, v_k, w_k, \sigma_k) | 1 \leq k \leq K\}$ is the set of parameters, and α_k denotes the mixing weight of the kth component of the model.

5. A Model-Based Algorithm For Projective Clustering

This section presents our algorithm, MPC for projective clustering by minimizing subject to the constraints of (1), (2), and (6), which is a constrained nonlinear optimization problem. Using the Lagrangian multiplier technique, this can be transformed into an unconstrained optimization problem

$$\begin{aligned} \min J_1(U, V, W, Z) &= J(U, V, W, Z) \\ &+ \sum_{k=1}^{K} \lambda_k \left(\sum_{j=1}^{D} w_{kj} - 1\right) + \xi\left(\sum_{k=1}^{K} \alpha_k - 1\right) \\ &+ \sum_{i=1}^{N} \zeta_i \left(\sum_{k=1}^{K} u_{ki} - 1\right), \end{aligned} \quad (7)$$

where λ_k(k = 1, 2, . . .,K), ξ, and ζ_i(i = 1, 2, . . .,N) are the Lagrange multipliers corresponding to the constraints defined in (1), (2), and (6).

5.1. The Optimization Method

To achieve a local minimum of the objective function, the usual method is to use the partial optimization for each parameter in the function. Following this method, minimization of J_1 in (7) can be performed by optimizing U, V, W and Z in a sequential structure analogous to the mathematics of the EM algorithm. In each iteration, system first fix V = $\hat{v}$, W = $\hat{w}$, and Z = $\hat{Z}$, and solve U as $\hat{U}$ to minimize J_1(U, $\hat{v}$, $\hat{w}$, $\hat{Z}$). Next, system fix U = $\hat{U}$, W = $\hat{w}$, and Z = $\hat{Z}$ and solve V as $\hat{v}$ to minimize J_1(V, V ,

International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume I, Issue VI, OCT - 2014

$\hat{w}$, $\hat{Z}$). Then, U = $\hat{U}$, V = $\hat{v}$, and W = $\hat{w}$ are fixed and the optimal Z, say $\hat{Z}$, is solved to minimize J1($\hat{U}$, $\hat{v}$, $\hat{w}$, Z). Afterward, system fix U = $\hat{U}$, V = $\hat{v}$, and Z = $\hat{Z}$ to obtain $\hat{w}$ by minimizing J1($\hat{U}$, $\hat{v}$, W, $\hat{Z}$). The four partial optimization problems are solved according to the following theorems.

5.2 The MPC Algorithm

The MPC algorithm, as outlined by Algorithm 2, performs projective clustering by minimizing the objective function. Actually, this solution can also be regarded as an extension to the classical FCM algorithm by adding an additional step in each iteration to compute W for each cluster, an approach which is commonly adopted in existing soft subspace clustering algorithms.

Input: DB, K and a termination criterion which is a small positive number $\in$;
Output: U, V and the associated weights W;
begin
Let *p* be the number of iteration, *p*=0
1. Initialization
1.1 Randomly choose K cluster centers. Denote V as $V^{(0)}$;
1.2 Set all the weights of W to $\frac{1}{D}$, and denote W as $W^{(0)}$;
1.3 Set all the $\alpha_k s$ to $\frac{1}{K}$ and $\sigma_k s$ to a non- zero constant and denote them by $Z^{(0)}$;
2. repeat
2.1 Let $\hat{V} = V^{(p)}$, $\hat{W} = W^{(p)}$ and $\hat{Z} = Z^{(p)}$, compute $U^{(p+1)}$;
2.2 Let $\hat{U} = U^{(p+1)}$;
2.3 Let $\hat{V} = V^{(p+1)}$ **s** to compute $\hat{\alpha}k$ and $\hat{\sigma}k$ for k=1,2…….,K, respectively and obtain $Z^{(p+1)}$;
2.4 Let $\hat{Z} = Z^{(p+1)}$, to determine $\hat{\lambda}_k$ for k=1, 2,……, K;
2.5 Compute $W^{(p+1)}$;
2.6 p=p+1.
until $\left|J(U^{(p)},V^{(p)},W^{(p)},Z^{(p)}) - J(U^{(p-1)},V^{(p-1)},W^{(p-1)},Z^{(p-1)})\right| < \varsigma$
3. Output $U^{(p)}$ as U, $V^{(p)}$ as V and $W^{(p)}$ as W.
end

It is important to note that MPC does not require user defined parameters for feature weighting, whereas most of the existing projective clustering algorithms do: for instance, l in PROCLUS, β in FWKM, γ in EWKM, etc. The only pending coefficient, say $\hat{\lambda}_k$, in the weight updating formula MPC can be determined by numerically solving. Step 2.4 of Algorithm 2 is designed for this purpose. All the variables except $\hat{\lambda}_k$ are given and thus can be considered as constants with respect to $\hat{\lambda}_k$. Consequently, system can resolve $\hat{\lambda}_k$ using a numerical method, such as the Newton-Raphson and bisection method.

6. Subspace and Outlier Analysis

The proposed system is designed to perform clustering on high dimensional spaces. Non access subspaces are included in the similarity analysis. Anomaly data values are verified with similarity under the clustering process. The subspace selection process is optimized. The system is designed to perform data clustering on high dimensional data values. The model based projective clustering is improved with anomaly analysis. The system also enhanced with attribute alignment process. The system is divided into six major modules. They are data cleaning process, subspace selection, subspace alignment, clustering with MPC, MPC with outliers and clustering with attribute and anomaly analysis.

The data cleaning module is designed to correct noise transactions. The sub space

selection module is designed to select attribute subsets. The attribute alignment is performed under subspace alignment module. The clustering is performed with model based projective clustering technique. The outlier analysis is integrated with MPC model. The attribute and anomaly analysis is applied in the enhanced MPC model.

7. Conclusion

The projective clustering techniques are used to cluster the high dimensional data. The model based projective clustering algorithm is a subspace clustering technique. Non-axis-subspaces are used with similarity analysis. Anomaly transactions are partitioned with projected clusters. Cluster accuracy is improved in the system. Features space selection is optimized to handle non aligned attribute subspace. Outlier analysis is provided in clustering process. Cluster initialization is improved with subspace selection process.

REFERENCES

[1] R. Harpaz and R. Haralick, "Linear Manifold Clustering in High Dimensional Spaces by Stochastic Search," Pattern Recognition Letters, vol. 40, pp. 2672-2684, 2007.

[2] Life Chen, Qingshan Jiang and Shengrui Wang," Model-Based Method for Projective Clustering " IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 7, July 2012.

[3] G. Moise, J. Sander, and M. Ester, "Robust Projected Clustering," Knowledge Information System, vol. 14, no. 3, pp. 273-298, 2008.

[4] L. Chen, Q. Jiang, and S. Wang, "A Probability Model for Projective Clustering on High Dimensional Data," Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 755-760, 2008.

[5] G. Gao, J. Wu, and Z. Yang, "A Fuzzy Subspace Clustering Algorithm for Clustering High Dimensional Data," Proc. Int'l Conf. Advanced Data Mining and Applications (ADMA), pp. 271-278, 2006.

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

[7] P.D. Hoff, "Model-Based Subspace Clustering," Bayesian Analysis, vol. 1, no. 2, pp. 321-344, 2006.