Multiple Principal Component Analyses and Projective Clustering

Nittaya Kerdprasop and Kittisak Kerdprasop
Data Engineering and Knowledge Discovery Research Unit
School of Computer Engineering, Suranaree University of Technology, Thailand
{nittaya, kerdpras}@ccs.sut.ac.th

Abstract

Projective clustering is a clustering technique for high dimensional data with the inherent sparsity of the data points. To overcome the unreliable measure of similarity among data points in high dimensions, all data points are projected to a lower dimensional subspace. Principal component analysis (PCA) is an efficient method to dimensionality reduction by projecting all points to a lower dimensional subspace so that the information loss is minimized. However, PCA does not handle well the situation that different clusters are formed in different subspaces. We propose a method of multiple principal component analysis for iteratively computing projective clusters. The objective function is designed to determine the subspace associated with each cluster. Some experiments have been carried out to show the effectiveness of the proposed method.

1. Introduction

Clustering is a widely used technique to discover homogeneous groups, or clusters, of data according to a certain similarity measure. Many algorithms have been designed [10] to compute a partition on full-dimensional data set. While these approaches work successfully on low-dimensional data sets, their efficiency decrease significantly in higher dimensional space [9, 12]. In high dimensional data, some dimensions tend to be redundant or irrelevant. Massive dimensions can confuse the clustering algorithms. It is also difficult to group similar data points in very high dimensions because the distance between any two data points becomes almost the same [5, 8]. The most difficult problem on clustering high dimensional data is that different clusters may exist in different subspaces of different dimensions [4].

A possible solution to these problems is to use dimension reduction or feature selection techniques. By means of dimension reduction, one first reduces the dimensions of the original data set by removing less important dimensions or by transforming the original data set into a lower dimensional space. The conventional clustering algorithms can then be applied to the new data set. However, an attempt to reduce dimensions of all data points results in significant information loss.

Recognizing the need for an efficient algorithm for clustering high dimensional data, the concept of *subspace clustering* or *projective clustering* has thus been proposed [1, 2, 3, 4]. The goal of projective clustering is to find clusters embedded in lower dimensional subspaces. It can minimize the information loss in the process of dimension reduction by projecting those high dimensional data points into different lower dimensional subspaces for different clusters.

Statistical methods such as Principal Component Analysis (PCA) [11] can effectively reduce the dimensionality of the original data by projecting all points on a subspace so that the information loss is minimized. Then, a standard clustering method can be used in this subspace. However, PCA does not work well when different subsets of data points embedded in different lower-dimensional subspaces. We, thus, propose the method to iteratively apply PCA aiming at transforming the original data set into various lower-dimensional subspaces. The subsequent steps compute a partitioning of data points into disjoint groups. We briefly explain the concept of PCA in Section 2. The proposed method of multiple PCA and the partitional clustering steps are then presented in Section 3. The experimental results shown in Section 4 verify the efficiency of the proposed method. Finally, conclusion remarks are presented in Section 5.

2. Dimensionality reduction with PCA

Dimensionality reduction is an important preprocessing step for unsupervised clustering, especially on high dimensional data set. There are two major

Proceedings of the 16th International Workshop on Database and Expert Systems Applications (DEXA'05)
1529-4188/05 $20.00 © 2005 **IEEE**

approaches to dimension reduction: feature selection and feature transformation. Feature selection is a process of finding a minimum subset of original features that satisfies some criteria, such as information measure, and no new feature to be generated. Feature transformation methods, on the other hand, transform data from the original *d*-dimensional feature space to a new *q*-dimensional ($q < d$) feature space through some functional mapping.

Principal Component Analysis (PCA) [11,14], sometimes called the Karhunen-Loeve (KL) transformation [4], is a widely used method for feature transformation to reduce the number of dimensions of a data set. The goal of PCA is to find basis vectors for a subspace which maximizes the least square reconstruction error. Let $X= (x_1,...,x_n)$ be a *d*-dimensional data matrix of points. PCA projects the correlated high dimensional data onto a hyperplane. This mapping uses only the first few *q* nonzero eigenvalues and the corresponding eigenvectors of the covariance matrix F, $F = U\Lambda U^T$ where Λ is a matrix that includes the eigenvalues λ_i of F in its diagonal in decreasing order, and U is a matrix that includes the eigenvectors corresponding to the eigenvalues in its column. The vector $y = W^T(x_i)$ is a *q*-dimensional reduced representation of the observed vector x_i where the W weight matrix contains the *q* principal orthonormal axes in its column $W = U_q\Lambda_q^{1/2}$.

Tipping and Bishop [13] developed a method called probabilistic PCA to associate a proper probability model for PCA. The advantage of the probabilistic PCA model is that it can be extended to mixture model where data can be viewed as arising from several populations mixed in varying proportions. The entire data set is then modeled by a Gaussian with restricted covariance matrix:

$$p(x) = \frac{1}{(2\pi)^{d/2} |\det A|^{1/2}} \exp(-\frac{1}{2}(x-\bar{x})^T A^{-1}(x-\bar{x})),$$

where $A = \sigma^2 I + WW^T$ is the modified covariance matrix and I is the identity matrix. W is found as in the original PCA algorithm, and σ^2 is found by calculating the average of the variance in the discarded dimensions:

$$\sigma^2 = \frac{1}{d-q} \sum_{i=q+1}^{d} \lambda_i.$$

3. Multiple PCA with projective clustering

Our approach aims at performing clustering in low dimensional subspaces, instead of the original high dimensional space where clusters are not well-separated. The intuition idea is to reduce the dimensions using PCA. Since the embedded clusters may lie in different subspaces, the subspace initially obtained using PCA does not necessary coincide with the subspace spanned by the *k* cluster centers. Therefore, we propose to iteratively perform PCA and clustering on the reduced subspace until the convergence criteria has been reached. The algorithm can be defined as follows.

Algorithm Clustering with multiple PCA

Input: a set of data vectors $X = [x_1, ..., x_n]$ of *d* dimensions and the number of cluster (*k*)

Output: a set of cluster centers $C_\mu = [\mu_1, ..., \mu_k]$ and the relevant attributes

Steps:

1. *Initialization:* Refine the initial points for clustering (as proposed in [7]) on sample data, and center the data matrix X so that the value of each variable is subtracted for that variable.
2. Do the first dimension reduction using PCA to obtain the *q*-dimensional subspace, $q < d$.
3. While not convergence
   - 3.1 Run clustering algorithm on the *q*-dimensional subspace to obtain clusters.
   - 3.2 Use cluster membership to construct the *k* cluster centroids in the original space.
   - 3.3 Compute the span of *k* centroids using singular value decomposition.
   - 3.4 Apply PCA to obtain a new *q*-dimensional subspace.
4. Return a set of cluster centers associated with relevant attributes

We propose the method for iterative high dimensional clustering on the basis of fuzzy k-means [6]. A prior probability of the cluster can be computed as:

$$\alpha_i = \frac{1}{n}\sum_{j=1}^{n}\gamma_{i,j},$$

where γ is the degree of membership.

The cluster centers are determined from

$$v_i^x = \frac{\sum_{k=1}^{n}(\gamma_{i,k})^m (x_k - W_i\langle y_{i,k}\rangle)}{\sum_{k=1}^{n}(\gamma_{i,k})^m}$$

where the expectation of the latent variables is

$$\langle y_{i,k}\rangle = M_i^{-1} W_i^T (x_k - v_i^x),$$

the q×q matrix $M_i = \sigma_{i,x}^2 I + W_i^T W_i$ and the fuzzy weighting component $m = 2$. The new value of W_i can be computed from

$$\tilde{W}_i = F_i W_i (\sigma_{i,x}^2 I + M_i^{-1} W_i^T F_i W_i)^{-1}.$$

The covariance matrix F_i can be computed by

$$F_i = \frac{\sum_{k=1}^{n}(\gamma_{i,k})^m (x_k - v_i^x)(x_k - v_i^x)^T}{\sum_{k=1}^{n}(\gamma_{i,k})^m}.$$

The new value of $\sigma_{i,x}^2$ is

$$\sigma_{i,x}^2 = \frac{1}{q} tr(F_i - F_i W_i M_i^{-1} \tilde{W}_i^T).$$

The fuzzy covariance matrix is

$$A_i = \sigma_{i,x}^2 I + \tilde{W}_i \tilde{W}_i^T.$$

The square distance measurement, D^2, for the fuzzy k-means is defined as the product of three terms: prior probability of the cluster, the distance between the k^{th} data point and the centroid v_i of cluster i, and the distance between the cluster prototype and the data in the subspace. The objective function, J, of the clustering is

$$J = \sum_{i=1}^{c}\sum_{k=1}^{n}(\gamma_{i,k})^m D^2 + \sum_{k=1}^{n}\lambda_i(\sum_{i=1}^{c}\gamma_{i,k} - 1).$$

In the PCA step (steps 2 and 3.4 in the algorithm), we use a simple and fast method to determine the number of principal dimensions to be retained at each level. For component *i*, the dimension q_i to be retained in the corresponding sub-components in the next level is considered from the two criteria: the proportion of variance of the first *r* components and the size of important variance.

The proportion of variance of the first *r* components can be computed from the summation of the first *r* variances divided by the sum of all variances. We set projections accounting for over 85% of the total variance as a threshold, that is,

$$\frac{\sum_{i=1}^{r}\lambda_i}{\sum_{i=1}^{d}\lambda_i} > 0.85.$$

The size of important variance is a simple measurement to discard principal components that have sample variances below $\bar{\lambda}$,

$$\bar{\lambda} = \frac{1}{d}\sum_{i=1}^{d}\lambda_i.$$

The intuitive idea is that if $\lambda_i < \bar{\lambda}$, then the i^{th} principal direction is less interesting than average.

Our method of determining how many principal components to retain is to consider from the proportion of variance and the size of importance variance,

$$(\frac{\sum_{i=1}^{r}\lambda_i}{\sum_{i=1}^{d}\lambda_i} > 0.85) \text{ OR } (\lambda_i > \bar{\lambda}).$$

With the proposed criteria, each cluster component can have potentially different dimensionality.

4. Experimental evaluation

We compare our proposed projective clustering (PKM) with the fuzzy k-means (FKM) and k-means (KM) algorithms. The experiments are performed on the Pentium IV 1.0 GHz machine with 512 MB of main memory. We generate the synthetic data sets of 50,000 points with varied dimensions up to 100. To assess the quality of a clustering algorithm we use the distance metric

$$\sqrt{\sum_{i=1}^{n}\min_{j\in\{1..k\}} \| x_i - c_j \|^2}.$$

The performance is evaluated on the clustering quality, the number of iteration toward convergence criteria, and the running time. The results are shown in Figures 1 through 3.

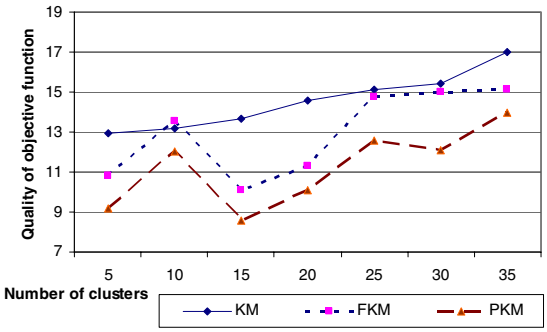

Figure 1. A comparison on clustering quality

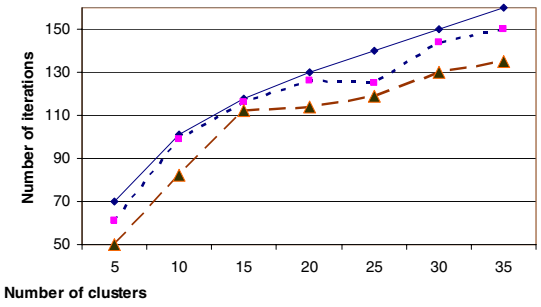

Figure 2. A comparison on number of iteration

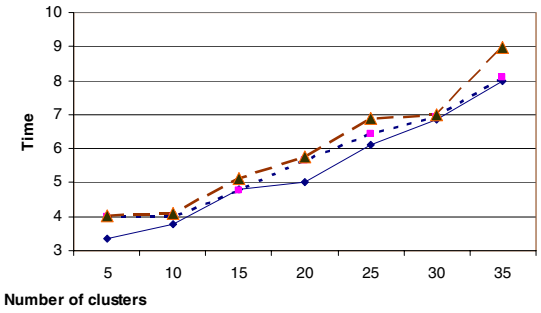

Figure 3. A comparison on running time

5. Conclusions

We have introduced a new method for clustering high dimensional data using the concept of projective clustering. The algorithm is based on the fuzzy partitional clustering algorithm. The key to the effectiveness of finding clusters on different subspaces is due to the power of PCA. The main justification of powerful dimension reduction is that PCA uses singular value decomposition (SVD) which gives the best low rank approximation to original data. We perform multiple PCA to project data into different subspaces for the effectiveness of discovering clusters embedded in different layers. The experimental results confirm the efficiency of our proposed method. Running experiments on real data is an essential step toward the practicality improvement of the method.

Acknowledgements

This research has been supported by grants from the Thailand Research Fund (TRF, MRG4780170), and the National Research Council. The Data Engineering and Knowledge Discovery Research Unit is fully supported by the research grants from Suranaree University of Technology.

References

[1] C. Aggarwal, J. Wolf, P. Yu, C. Procopiuc, and J. Park, "Fast algorithms for projected clustering", *Proceedings of ACM SIGMOD Int. Conf. on Management of Data*, 1999, pp. 61-72.

[2] C. Aggarwal and P. Yu, "Finding generalized projected clusters in high dimensional spaces", *Proceedings of ACM SIGMOD Int. Conf. on Management of Data*, 2000, pp.70-81.

[3] C. Aggarwal and P.Yu, "Redefining clustering for high-dimensional applications", *IEEE Transactions on Knowledge and Data Engineering*, 14(2), 2002, pp.210-225.

[4] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications", *Proceedings of ACM SIGMOD Int. Conf. on Management of Data*, 1998, pp.94-105.

[5] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is nearest neighbor meaningful?", *Proceedings of 7th Int. Conf. on Database Theory*, 1999, pp.217-235.

[6] J. Bezdek and J. Dunn, "Optimal fuzzy partitions: A heuristic for estimating the parameters in a mixture of normal distributions", *IEEE Transactions on Computers*, 1975, pp.835-838.

[7] P. Bradley and U. Fayyad, "Refining initial points for k-means clustering", *Proceedings of 15th Int. Conf. on Machine Learning*, 1988, pp.91-99.

[8] A. Hinneburg, C. Aggarwal, and D. Keim, "What is the nearest neighbor in high dimensional spaces?", *Proceedings of 26th Int. Conf. on Very Large Data Bases*, 2000, pp.506-515.

[9] A. Hinneburg and D. Keim, "Optimal grid-clustering: Towards breaking the curse of dimensionality in high dimensional clustering", *Proceedings of 25th Int. Conf. on Very Large Data Bases*, 1999, pp.506-517.

[10] A. Jaine and R. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, NJ, 1988.

[11] I. Jolliffe, *Principal Component Analysis*, Springer Verlag, 1986.

[12] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: A review", *SIGKDD Explorations*, 6(1), 2004, pp.90-105.

[13] M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analyses", *Neural Computation*, 11(2), 1999, pp.443-482.

[14] K. Yeung and W. Ruzzo, "Principal component analysis for clustering gene expression data", *Bioinformatics*, 17(9), 2001, pp.763-774.