Volume 3, Issue 6, June 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Projective Clustering Method for the Detection of Outliers in Non-Axis Aligned Subspaces

N.Devambika[*]
PG Scholar,
Department of CSE
P.B.Collge of Engineering
India

S.Anbu
Professor & HOD,
Department of CSE
P.B.College of Engineering
India

G.Thiyagarajan
Assistant Professor,
Department of CSE
P.B.College of Engineering
India

Abstract*— Clustering the case of non-axis-aligned subspaces and detection of outliers is a major challenge due to the curse of dimensionality. The normal clustering was efficient in axis-aligned subspaces only. To solve this problem, projective clustering has been defined as an extension to traditional clustering that attempts to find projected clusters in subsets of the dimensions of a data space. 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.Fuzzy Logic is mainly used to find the empty space. In model-based methods, data are thought of as originating from various possible sources, which are typically modelled by Gaussian mixture.

***Keywords*— high dimensions, projective clustering, and probability model, fuzzy logic and clustering.**

I. Introduction

In model-based methods, data are thought of as originating from various possible sources, which are typically, modelled by Gaussian mixture [3], [4], [5]. 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 [1], [2] and its variants. However, such methods would suffer from the curse of dimensionality problem for high dimensional data [6]. Many types of real-world data, such as the documents represented in the Vector Space Model (VSM) used in text mining and the microarray gene expression data of bioinformatics, consist of very high dimensional features. The data are inherently sparse in high-dimensional spaces [7], [8], making the Gaussian function inappropriate in this case.

In other words, most of the volume of a Gaussian function is contained in the tails instead of near the center in high-dimensional space: the so called "empty space phenomenon" [7]. 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 clusterings. Non-axis-aligned subspace clustering is a generalization of feature extraction; instead of defining a single set of features for the whole data.

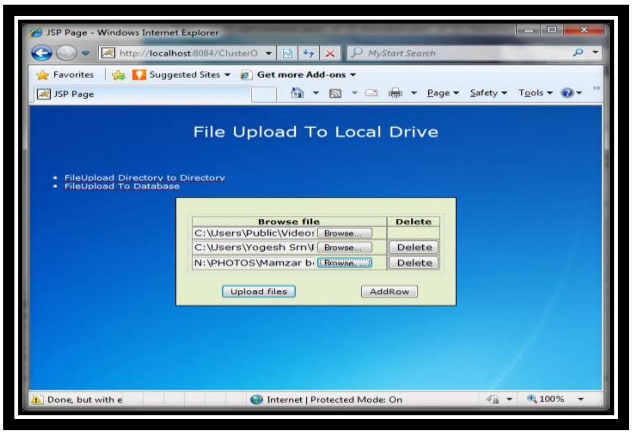
II. Literature Survey

L. Chen et al proposed a Clustering high dimensional data is a big challenge in data mining due to the curse of dimensionality [2]. To solve this problem, projective clustering has been defined as an extension of traditional clustering that seeks to find projected clusters in subsets of dimensions of a data space. In this project, the problem of modeling projected clusters is first discussed, and an extended Gaussian model is proposed. Second, a general objective criterion used with k-means type projective clustering is presented based on the model. Finally, the expressions to learn model parameters are derived and then used in a new algorithm named FPC to perform fuzzy clustering on high dimensional data. The experimental results on document clustering show the effectiveness of the proposed clustering model.

III. Related Works

A. Analysis of high dimensional data

Transferring a group of data from one location to another. In that analysis a destination folder capacity to receive data from the sender. 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. Two related terms occur in the literature: subspace clustering and projective clustering.



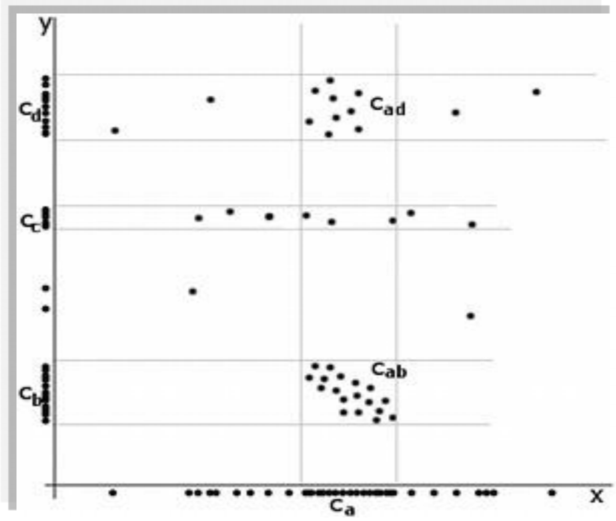
File Upload To Local Drive

- FileUpload Directory to Directory
- FileUpload To Database

Browse file	Delete
C:\Users\Public\Video:	
C:\Users\Yogesh Srn\	Delete
N:\PHOTOS\Mamzar b	Delete

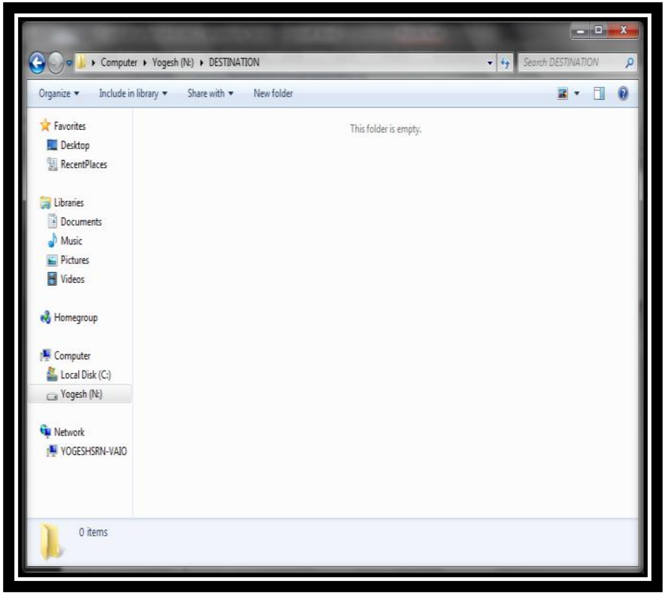
B. Probability Model

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 modelled by a Gaussian.



C. Optimization method

In optimization method mainly used to optimize the free empty space in local PC. To achieve a local minimum of the objective function, the usual method is to use the partial optimization for each parameter in the function.



This folder is empty.

0 items

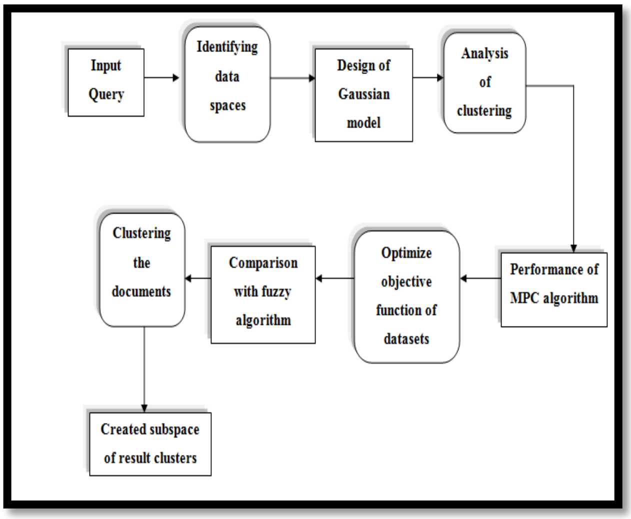
D. Efficiency of Model based algorithm

To Performs projective clustering by minimizing the objective function.Model predictive control which advances method of process control in data transfer between sender and receiver.

E. Evaluation with real-world datasets

Transfer a data from one mobile to another, if receiver mobile has less memory capacity and the remaining data will automatically stored in their phone memory.

F. Architecture Overview


Input Query
Identifying data spaces
Design of Gaussian model
Analysis of clustering
Performance of MPC algorithm
Optimize objective function of datasets
Comparison with fuzzy algorithm
Clustering the documents
Created subspace of result clusters

IV. Experimental Results

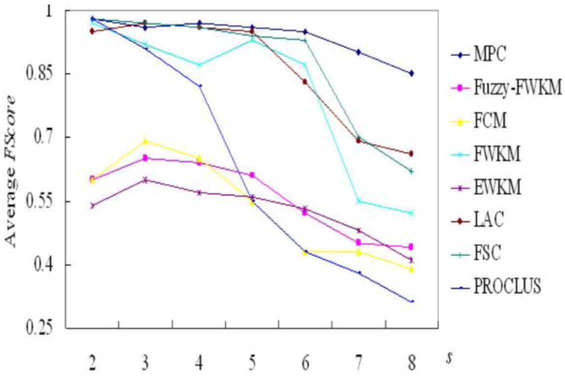
Six clustering algorithms, OUTLIER, PROCLUS, EWKM, LAC, FSC, and FWKM, were tested on the real world data sets. Since FCM and Fuzzy-FWKM are not projective clustering algorithms, we left them out of this set of experiments. The performances of the algorithms were measured in terms of FScore.

A. Proposed method

Clustering the case of non-axis-aligned subspaces and detection of outliers is a major challenge due to the curse of dimensionality. The first contribution is the proposal of a probability model to describe projected clusters in a high-dimensional space. A model-based algorithm for fuzzy projective clustering that discovers clusters with overlapping boundaries in various projected subspaces. To extend MPC in the case of non-axis-aligned subspaces. Another interesting extension would be for the detection of possible outliers in a data set. The computational expressions for calculating the optimal values of the parameters automatically and save the information without any damage.The experiments were conducted on synthetic data sets, UCI data sets, and email corpora widely 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.

B. 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.



V. Conclusion

In this paper, the problem of providing a probability model to describe projected clusters in high dimensional data. This problem becomes difficult due to high-dimensional data and the fact that only a small number of the dimensions may be considered in the clustering process. This will be proposed an extended Gaussian model which meets the general requirements of projective clustering well. It also derived an objective clustering criterion based on the model, allowing the use of a k-means type paradigm. Outlier to the case of non-axis-aligned subspaces. Another interesting extension would be for the detection of possible outliers in a data set. Our future efforts will also be directed toward developing techniques to build a robust initial condition for the clustering algorithm. the Ling-Spam data set, the maximal FScore of LAC was large (0.98), but the average FScore was only 0.88 and the standard deviation reached 0.10. In general, outlier is more robust than the other algorithms. This can be explained by the observation of Jain et al. that fuzzy clustering is usually better than hard clustering at avoiding local minima.

References

[1] Y. Lu, S. Wang, S. Li, and C. Zhou, “Particle Swarm Optimizer for Variable Weighting in Clustering High-Dimensional Data,” Proc. IEEE Swarm Intelligence Symp., pp. 37-44, 2009.

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

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

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

[5] S. Wang and H. Sun, “Measuring Overlap-Rate for Cluster Merging in a Hierarchical Approach to Colour 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 Neighbour in High Dimensional Spaces,” Proc. Int’l Conf. Very Large Databases (VLDB), pp. 506-515, 2000.

[9] 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.

[10] J.Z. Huang, M.K. Ng, H. Rong, and Z. Li, “Automated Variable Weighting in k-Means Type Clustering,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 5, pp. 657-668, May 2005.

[11] C. Domeniconi et al., “Locally Adaptive Metrics for Clustering High Dimensional Data,” Data Mining and Knowledge Discovery, vol. 14, pp. 63-97, 2007.

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

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

[14] M.J. Zaki and M. Peters, “Clicks: Mining Subspace Clusters in Categorical Data via Kpartite Maximal Cliques,” Proc. Int’l Conf. Data Eng. (ICDE), sspp. 355-356, 2005.