© 2014 IJEDR | Volume 2, Issue 4 | ISSN: 2321-9939

A Survey on Clustering High Dimensional Data Techniques

[1]R.Aarthi, [2]P.Thiyagarajan
[1]Assistant Professor (CSE), [2]Student
Nandha College of Technology, Erode

***Abstract* - Cluster analysis is the one in which uses to divide the data into groups. It mainly developed for the propose of summarization and improved understanding. The example for cluster analysis has been given below. Let we takes the group which related to document for browsing. That are in order to find the genes and proteins which has similar functionality, or as a means of data compression. The term clustering has a long history and a large no of clustering techniques which have been developed in statistics and pattern recognition. This provide a short introduction to cluster analysis, and then find the focus on challenge of clustering high dimensional data. Hereby i present a brief overview of several recent techniques , including a more detailed description of recent work of our own which uses a concept based clustering approach.**

***Keywords* – Data mining, clustering, high dimensional data, Subspace clustering, Projected clustering**

I. INTRODUCTION

Cluster analysis is the one in which is uses to divide the data into meaningful or useful groups. If meaningful clusters are the one which uses to define the goal, then the resulting clusters should capture the "natural" structure of the data. Let's we see the example for it. cluster analysis has been used to group related documents for browsing, to find genes and proteins which have the similar functionality, that are uses to provide a grouping of spatial locations prone to the earthquakes. In some other cases, cluster analyses are only a useful starting point for other uses, e.g., data compression and efficiently are uses to finding the nearest neighbors of points. Its mainly uses for understanding and utility, cluster analysis has long been used in a wide variety of fields: psychology and other social sciences, biology and statistics.

We have provided a short introduction to cluster analyses, which are focus on the challenge of clustering high dimensional data. Here we have presented a brief overview of several techniques which developed recently, that are including a more detailed description of recent work of our own which uses the concept-based approach. In all cases, the approaches to clustering having the high dimensional data which must deal with the "curse of dimensionality", which, in general terms, it is widely observed the phenomenon that data analysis techniques , which work well at lower dimensions, that often perform poorly as the dimensionality of the analyzed data increases.

II. CLUSTERING HIGH DIMENSIONS DATA TECHNIQUES

The operations of Clustering high dimensional data techniques has recently grown in advance.The popular methods as mentioned above were analyzed in detail.

A. *Gaussian mixture models using high-dimensional data*

Clustering divides a given dataset {x1 , ..., xn } of n data points into k homoge-neous groups. Popular clustering techniques use Gaussian Mixture Models (GMM),which assume that each class is represented by a Gaussian probability density. Data k{x1 , ..., xn } ∈ Rp are then modeled with the density f (x, θ) = $\sum_{i=1}^{k}$ πi $\varphi(x, \theta i)$)where φ is a multi-variate normal density with parameter θi = {μi , Σi } and πi are the mixing proportions. This model which uses to estimates full covariance matrices and therefore the number of parameters is very large in high dimensions.

However, due to the empty space phenomenon we can assume that high-dimensional data live in subspaces with a dimensionality lower than the dimensionality of the original space. We here propose to the work in low-dimensional class-specific subspaces in order to adapt classification to high-dimensional data and to limit the number of parameters to estimate.

B. *The decision rule*

Classification assigns an observation x ∈ Rp with unknown class membership to one of k classes C1 , ..., Ck known a priori. The optimal decision rule is the one which called Bayes decision rule, this affects the observation x to the class which has the maximum posterior probability P (x ∈ Ci |x) = πi φ(x, θi)/ l=1 πl φ(x, θl). Maximizing the posterior probability is equivalent to minimizing −2 log(πi φ(x, θi)). For the model [aij bi Qi di], this results in the decision rule δ + which assigns x to the class minimizing the following cost function Ki (x):

$$K_i(x)=\|\mu_i-P_i(x)\|^2_{Ai}+\frac{1}{bi}\|x-P_i(x)\|^2+\sum_{j=1}^{di}\log(a_{ij})+(p-d_i)\log(b_i)-2\log(\pi_i)$$

where $\|.\|_{\Lambda i}$ is the Mahalanobis distance associated with the matrix $\Lambda i = Q_i \Delta I\, Q_i^t$. The posterior probability can therefore be rewritten as follows: P (x ∈ C_i |x) = $1\sum_{l=1}^{k}\exp\left(\frac{1}{2}(Ki(x) - Kl(x))\right)$. It measures the probability that x belongs to Ci2 and allows to identify dubiously classified points. We can observe that this new decision rule is mainly based on two distances: the distance between the projection of x on Ei and the mean of the class; and the distance between the observation and the subspace Ei .

This rule assigns a new observation to the class for which it is close to the subspace and for which its projection on the class subspace is close to the mean of the class. If we consider the model [a_i b_i Q_i d_i], the variances ai and bi balance the importance for the both distances. The example, if the data having too much noisy, i.e. bi is large, it is natural to balance the distance $\|x - Pi(x)\|^2$ by $1/b_i$ in order to take into account the large variance in E(1/i) .Remark that the decision rule δ^+ of our models uses only the projection on E_i and we only have to estimate a d_i -dimensional subspace. Thus, our models are significantly more parsimonious than the general GMM. For example, if we consider 100-dimensional data, that are made of 4 classes with common intrinsic dimensions d_i equal to 10, the model [a_i b_i Q_i d_i] requires the estimation of 4 015 parameters whereas the full Gaussian mixture model estimates 20 303 parameters.

C. High Dimensional Data Clustering

In this section we derive the EM-based clustering framework for the model [a_{ij} b_i Q_i d_i] and the sub-models. The new clustering approach are referred to by the High-Dimensional Data Clustering, which has the lack of space, we do not need to present the proofs of the following results which can be found in .

The clustering method HDDC

Unsupervised classification organizes data in homogeneous groups using only the observed values of the p, whreas p is the explanatory variables. Normally, the parameters uses to estimated by the EM algorithm which repeats iteratively E or M steps. Suppose if we use the parameterization that presented in the previous section, that the EM algorithm for estimating the parameters $\theta = \{\pi_i , \mu_i , \Sigma_i , a_{ij} , b_i , Q_i , d_i \}$, would be written as follows:

E step: this step computes at the iteration q the conditional posterior probabilities: $t_{ij}^{(b)}=P(x_j £ C_i^{(q)}|x_j)$,from the relation , it may consider:

$$t_{ij}^{(q)}=1/\sum_{l=1}^{k} \exp(1/2(K_i^{(q-1)}(x_j)-K_l^{(q-1)}(x_i))) \qquad (1)$$

where K_i is defined

M step: this step maximizes at the iteration q has the conditional likelihood. Proportions,which means and covariance matrices of the mixture are estimated by:

$$\pi_i^{(q)}=(n_i^{(q)}/n),\ \mu_i^{(q)}=(1/n_i^{(q)})\sum_{j=1}^{n} T_{ij}^{(q)}x_j, n_i^{(q)}=\sum_{j=1}^{n} t_{ij}^{(q)} \qquad (2)$$

$$\Sigma_i^{(q)} =(1/n_i^{(q)})\sum_{l=1}^{n} t_{ij}^{(q)}(x_j-\mu_i^{(q)})(x_j-\mu_j^{(q)})^t \qquad (3)$$

The estimation of the HDDC parameters are detailed in the following subsection.

Estimation of HDDC parameters

Assuming for the moment that parameters di are known and omitting the index q of the iteration for the sake of simplicity, we obtain the following closed form estimators for the parameters of our models:–

Subspace Ei : the d_i is the first columns of Q_i, that are estimated by the eigenvectors associated with the d_i largest eigenvalues λ_{ij} of Σ_i .

Model [a_{ij} b_i Q_i d_i]: the estimators of aij that having the di largest eigenvalues λ_{ij} of Σ_i and the estimator of bi is the mean of the $(p - d_i)$ smallest eigenvalues of Σ_i and can be written as follows:

$$b_i=\frac{1}{(p-di)}(Tr(\Sigma_i)-\sum_{j=1}^{n} \lambda_{ij}) \qquad (4)$$

Model[a_i b_i Q_i d_i]: the estimator of b_i which given at (4) and the estimator of a_i is the mean of the d_i largest eigenvalues of $\sum_i$

$$Ai=(1/d_i)\sum_{j=1}^{di} \lambda_{ij} \qquad (5)$$

Model[a_i bQ_i d_i]: the estimator of a_i is given at (5) and the estimator of b is:

$$b=\frac{1}{(np-\sum_{i=1}^{k} nidi)}(nTr(W)-\sum_{i=1}^{k} n_i\sum_{j=1}^{di} \lambda_{ij}) \qquad (6)$$

Charles Bouveyron, St.Girard, and C.Schmid

Where $W=\sum_{i=1}^{k} \pi\Sigma_i$

Model[abQ$_i$d$_i$]: the estimator of b is given at(6) and the estimator of a is:

$$a=\frac{1}{\sum_{l=1}^{K} nidi}\sum_{i=1}^{k} n_i\sum_{j=i}^{di} \lambda_{ij} \qquad (7)$$

D. Intrinsic dimension estimation

We also have to estimate the intrinsic dimensions of each subclass. That is very difficult problem which has no unique technique to use. Our approach is based on the eigenvalues of the class conditional where the covariance matrix Σ_i of the class is C_i . Whreas the jth eigenvalue of Σ_i corresponds to the fraction of the full variance carried by the jth eigenvector of Σ_i .Therefore we estimated the class specific dimension di , i = 1, 2,3,4.., k, with the empirical method screen-test of Cattell [3] which analyzes the differences between eigenvalues in order to find a break in the screen. The selected dimension is the one for where the subsequent differences are smaller than the threshold. In our experiments, the threshold is chosen by the cross-validation. We also compared the probabilistic criterion BIC which gave very similar results.

III. CONCLUSION

In this survey various techniques of Cluster high dimensional data were described in detail.These techniques are most important which uses to find the similar functionality at genes and proteins.The Clustering high dimensional data techniques mentioned in this review paper are used in many advanced for summarization or improved understandings. This high dimensional data in clustering is to determine the intrinsic grouping in a set of unlabeled data

REFRENCES

[1] Bocci, L., Vicari, D., Vichi, M.: A mixture model for the classification of three-way proximity data. Computational Statistics and Data Analysis, **50**, 1625–1654 (2006).
[2] Bouveyron, C., Girard, S., Schmid, C.: High-Dimensional Data Clustering. Technical Report 1083M, LMC-IMAG, Université J. Fourier Grenoble 1 (2006).
[3] Cattell, R.: The scree test for the number of factors. Multivariate Behavioral Research, **1**, 245276 (1966).
[4] D'Alche Buc, F., Dagan, I., Quinonero, J.: The 2005 Pascal visual object classes challenge. Proceedings of the first PASCAL Challenges Workshop, Spring er (2006).
[5] Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, **39**, 1–38 (1977).
[6] Dorko, G., Schmid, C.: Object class recognition using discriminative local features. Technical Report **5497**, INRIA (2004).
[7] Fraley, C., Raftery, A.: Model-based clustering, discriminant analysis and density estima-tion. Journal of American Statistical Association, **97**, 611–631 (2002).
[8] Parsons, L., Haque, E., Liu, H.: Subspace clustering for high dimensional data: a review. SIGKDD Explor. Newsl. **6**, 90–05 (2004).
[9] Schwarz, G.: Estimating the dimension of a model. Annals of Statistics, **6**, 461–464 (1978).
[10] Tipping, M., Bishop, C.: Mixtures of probabilistic principal component analysers. Neural Computation, 443–482 (1999).
[11] Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories. Technical Repo rt **5737**, INRIA (2005).
[12] C. Aggarwal and P. Yu. ”Finding Generalized Projected Clusters in High Dimensional Space”. In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIG- MOD’00), 2000.
[13] C. C. Aggarwal and C. Procopiuc. ”Fast Algorithms for Projected Clustering”. In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD’99), 1999.
[14] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications”. In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIG- MOD’98), 1998.
[15] M. Ankerst, M. M. Breunig, H.P. Kriegel, and J. Sander.”OPTICS: Ordering Points to Identify the Clustering Structure”. In Proc. ACM SIGMOD Int. Conf. on Man agement of Data (SIGMOD’99), 1999.
[16] C. Baumgartner, C. B¨hm, D. Baumgartner, G. Marini,o K. Weinberger, B. Olgem¨ller, B. Liebl, and A. A.o Roscher. ”Supervised machine learning techniques for the classification of metabolic disorders in newborns”. Bioinformatics, 2004. in press.
[17] M. Dash, K. Choi, P. Scheuermann, and H. Liu. ”Feature Selection for Clustering – A Filter Solution”. In Proc. IEEE Int. Conf. on Data Mining (ICDM’02), 2002.
[18] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. ”A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD’96), 1996.
[19] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Academic Press, 2001.
[20] K. Kailing, H.-P. Kriegel, P. Kr¨ger, and S. Wanka.o ”Ranking Interesting Subspaces for Clustering High Di mensional Data”. In Proc. 7th European Conf. on Prin ciples and Practice of Knowledge Discovery in Databases (PKDD’03), 2003.
[21] C. M. Procopiuc, M. Jones, P. K. Agarwal, and T. M. Murali. ”A Monte Carlo Algorithm for Fast Projective Clustering”. In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD’02), 2002.
[22] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher ”Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces Cerevisiae by Mi croarray Hybridization.”. Molecular Biolology of the Cell, 93273–3297, 1998.

Author’s Profile

Thiyagarajan is a student of M.E(Computer Science and Engineering) at Nandha College of Technology,Erode and completed her Bachelor degree from Excel Engineering College.Her Areas of interest are Data Mining and Cloud Computing.
Aarthi received the M.E degree in Computer Science and engineering.She is currently working at the Department of computer science.