FEATURE SELECTION IN CLUSTERING PROBLEM

CRITICAL REVIEW

In this review,I will discuss the paper on Feature Selection in Clustering Problem(2003) by Volker Roth and Tilman Lange. According to the authors, the goal of Clustering is to discover structures among individuals described using several features.In this paper they have discussed one of the clustering methods, called model-based method, where model is selected for clusters, then optimizing the fit between the data and the model.The advantage of this method is to provide rigorous framework to assess the number of clusters and the role of each feature in clustering process. I shall agree with the authors that it is important to select relevant features for cluster analysis point of view.

INTRODUCTION

This paper proposes an algorithm to deal with the feature selection procedure using a gaussian mixture model in an iterative way. This is called an "ad hoc" strategy. Here the algorithm iterates the clustering and feature selection. Most discriminative feature for the object partition is automatically extracted from simultaneous clustering . In our paper the Gaussian Mixture Model combines clustering method with Bayesian Feature Selection Mechanism for selecting relevant features.Further an optimization algorithm with guaranteed convergence to local optimum is applied to the model, having one free parameter, K. A stability based model selection procedure is followed for K.The only free model parameter is selected by resampling based stability analysis.

ALGORITHM

- GAUSSIAN MIXTURE AND LDA

  A collection of N samples xi in real space of d dimension. Consider a Gaussian mixture of 2 components with identical covariance matrix Σ.

  The likelihood equation in the M step can be viewed as a weighted mean. One replicates the N-observation 2 times,with the v-th such replication having the observation weights pvi. M-step can be carried out via augmented LDA, which can further be represented as an optimal scoring problem. The point of optimal scoring is to turn categorical values into quantitative ones. Let class membership of N-data vectors be coded as a matrix Z, the i,v-th entry of which equals one if the i-th observation belongs to the class v.The score vector θ assigns the real number θ_v to

the entries in the v-th column of Z. The simultaneous estimation of scores and regression coefficients β constitute the optimal scoring problem.
After integrated application of the E and M step , an observation xi is finally assigned to the class v with highest probability of the membership pvi.

- LDA AND AUTOMATIC RELEVANCE DETERMINATION
  Further, incorporating the automatic feature selection mechanism into EM Algorithm [3]. The following algorithm can help in solving 2-class LDA problem:
  - Choose an initial N vector of scores, orthogonal to k vector of ones.
  - Run a linear regression of X

  Feature selection can be incorporated by using some constraints on linear regression.Following the Bayesian inference principle, we introduce hyperpriors and integrate out these variances, and we finally arrive at the l1–constrained LASSO problem If M-step globally converges, we can guarantee the convergence to a local maximum of the constrained likelihood of our Expectation maximization Model.
  For an unconstrained solution, the algorithm simply reduces to standard EM procedure and continuously iterates the likelihood. In case l1-constraint is active, then in every iteration the LASSO algorithm maximizes the likelihood.

CONCLUSION : PROS AND CONS

The problem tackled in the paper consists of simultaneously clustering objects and automatically extracting the subsets of features which are most discriminative for object partition. In our paper the Gaussian Mixture Model combines clustering method with Bayesian Feature Selection Mechanism for selecting relevant features.Further an optimization algorithm with guaranteed convergence to local optimum is applied to the model.Gaussian Mixture Model can be more accurate as they model more information.

Gaussian Mixture models are effective for modeling arbitrary distribution, it appears to be advantageous in our paper in Clustering USPS digits and clustering faces.Clusters can be characterized by a small number of parameters. Also the results may satisfy the statistical assumptions of the generative models.

We choose an EM algorithm for optimization,which guarantees convergence and requires no parameter tuning. It is one the primary tools for training gaussian mixture models. Nevertheless, EM needs to pre-assign an appropriate number of density components, that is, the number of clusters. Roughly, the mixture may overfit the data if too many

components are utilized, whereas a mixture with too few components may not be flexible enough to approximate the true underlying model. Subsequently, the EM almost always leads to a poor estimate result if the number of components is misspecified. Unfortunately, from the practical viewpoint, it is hard or even impossible to know the exact cluster number in advance.

But gaussian mixture models can be computationally expensive if the number of distributions are large or the dataset contains very few observed data points.

FUTURE DIRECTION

Selecting the appropriate set of features from the input data determines how effectively Machine Learning Algorithms(MLA) can perform..Our method currently only implements partitions of the object set into two clusters. For finding multiple clusters, we propose to iteratively split the dataset. Such iterative splits have been successfully applied to the problem of simultaneously clustering gene expression datasets and selecting relevant genes.

REFERENCES

[1] Maugis, C., Celeux, G., & Martin-Magniette, M. (2009). Variable Selection for Clustering with Gaussian Mixture Models. Biometrics, vol.65(no.3), 701-709
[2] T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. J. R. Stat. Soc. B, 58:158–176, 1996
[3] Yiu-ming Cheung, "Maximum weighted likelihood via rival penalized EM for density mixture clustering with automatic model selection," in IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 750-761, June 2005
[4] T. Hastie, R. Tibshirani, and A. Buja. Flexible discriminant analysis by optimal scoring. J. Am. Stat. Assoc., 89:1255–1270, 1994.
[5] F. Meinecke, A. Ziehe, M. Kawanabe, and K.-R. M¨uller. Estimating the reliability of ICA projections. In Advances in Neural Information Processing Systems, volume 14, 2002.