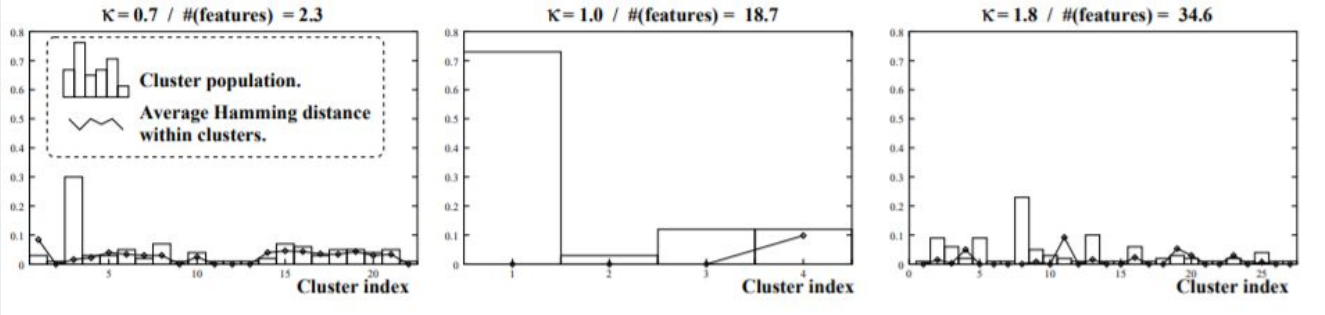
FEATURE SELECTION IN CLUSTERING PROBLEM

CSL2010 - INTRO TO MACHINE LEARNING - PROJECT

Aditya Soni

B20ME083



Overview of the project

In this project , approach to combining clustering and feature selection is presented. It implements a wrapper strategy for feature selection, in the sense that the features are directly selected by optimizing the discriminative power of the used partitioning algorithm. On the technical side, we present an efficient optimization algorithm with guaranteed local convergence property. The only free parameter of this method is selected by a resampling-based stability analysis. Experiments with real-world datasets demonstrate that our method is able to infer both meaningful partitions and meaningful subsets of features. In the given 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.

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, κ. A stability based model selection procedure is followed for κ.The only free model parameter is selected by resampling based stability analysis.

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:

- Likelihood in M-step can be viewed as weighted mean, replicates N observations 2 times, v-th replication having weight pvi.
- M-step can be carried out through augmented LDA which can be expressed as optimal scoring problem.
- Optimal scoring turns 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 θv 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 pv

Expectation maximization Algorithm

Expectation maximization Algorithm provides a general solution for the parameter estimate in density mixture models. Nevertheless, it 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.

Technical Updates Proposed

I here propose a method , which learns the model parameter via maximizing a weighted likelihood. Under a specific weight design ,we then introduce Rival Penalized EM (RPEM) algorithm for density mixture clustering.The RPEM algorithm learns the parameter by making mixture components compete each other at each time step.Comparing Rival Penalized EM algorithm to EM algorithm, the RPEM fades out the redundant densities from a density mixture during parameter learning process. Hence, RPEM automatically select an appropriate number of densities. The convergence speed of the RPEM relies on the value of the learning rate. Often, by a rule of thumb, we arbitrarily set the learning rate at a small positive constant.

Conclusion

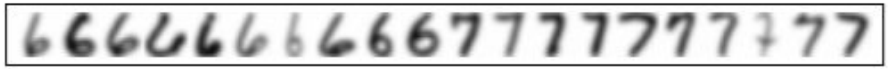


Figure 2: Sample images of digits '6' and '7' from the USPS database.

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.

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.