Feature Selection in Clustering Problem

Volkar Roth
vroth@inf.ethz.ch

Tilman Lange
tilman.lange@inf.ethz.ch

ETH Zurich, Institute f. Computational Science
Hirschengraben

Clustering is a significant task which group together similar objects, based on the similarity between a pair of data in different features. Different features have their own effects on clustering. Some features are irrelevant for clustering and may hurt the results. To handle all these issues, feature selection is an efficient way [1].The advantages of feature selection are: performance of clustering rises, provides fast and cost-efficient solutions an better understanding for the process generating data[2].Feature selection methods in supervised learning can be divided in filter methods and Wrapper methods. Use if the classifier differentiates these methods. Wrapper method uses classifier, while filter method does not. In unsupervised learning, it is much harder problem.

A strategy used is in “ad hoc” manner i.e. combining the clustering step and the relevance determination step. In the clustering step, set of hypothetical partitions is extracted. In relevance determination step, features are scored according to relevance. In this paper. we show a similar kind of approach, combining Gaussian Mixture Model with a Bayesian feature selection principle. Efficient optimization algorithm with guaranteed convergence to a local optimum for model is presented

In this paper, we have used classical Expectation-Maximization Algorithm, the M-step is re-formulated as Linear Discriminant Analysis Problem, which uses fuzzy labels in preceding E-step. Linear regression is then used to restate M-step, allowing to regularize the estimation. This distribution has a Automatic Relevance Determination (ARD) prior containing a free hyper parameter, which helps in encoding the relevance of corresponding variable in linear regression. These hyper parameters are then integrated out to achieve integrated feature selection mechanism at M-step.

A collection of N samples x_i in real space of d dimension. Consider Gaussian mixture of 2 components with identical covariance matrix ∑.Then log-likelihood can be give as

$$l^{mix} = \sum_{i=1}^{N} \log\left(\sum_{\nu=1}^{2} \pi_\nu \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_\nu, \Sigma)\right)$$

Where π_v sums to one and φ denotes Gaussian density. The EM Algorithm for maximizing l^{mix} can be given as

$$\textbf{E-step: } \text{set } p_{\eta i} = \text{Prob}(\boldsymbol{x}_i \in \text{class } \eta) = \frac{\pi_\eta \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_\eta, \Sigma)}{\sum_{\nu=1}^{2} \pi_\nu \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_\nu, \Sigma)}.$$

$$\textbf{M-step: } \text{set } \boldsymbol{\mu}_\nu = \frac{\sum_{i=1}^{N} p_{\nu i} \boldsymbol{x}_i}{\sum_{i=1}^{N} p_{\nu i}}, \quad \Sigma = \frac{1}{N} \sum_{\nu=1}^{2} \sum_{i=1}^{N} p_{\nu i} (\boldsymbol{x}_i - \boldsymbol{\mu}_\nu)(\boldsymbol{x}_i - \boldsymbol{\mu}_\nu)^\top.$$

After iterated application of the E- and M-step, an observation xi is finally assigned to the class v with highest probability of membership p_{vi}.

Further, incorporating the automatic feature selection mechanism into EM Algorithm [3]. The following algorithm can help in solving 2-class LDA problem:

1. Choose an initial N vector of scores, orthogonal to k vector of ones
2. Run a linear regression of X

Feature selection can be incorporated by using some constraints on linear regression.

In the Bayesian ARD formalism, this combinatorial explosion of the search space is overcome by relaxing the binary selection variable to a positive real-valued variance of a Gaussian prior over each component of the coefficient vector. Following the Bayesian inference principle, we introduce hyper priors 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 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 LASSO algorithm maximizes the likelihood.

While selecting the model, we have only one free parameter l1-constraint κ. We will observe the stability of data partitions for selecting κ. The stability measuring concept of solution is an efficient way of model selection unsupervised learning problems. Selecting one feature will lead to many competing hypotheses for splits. On the other hand, if we select more features it will lead to problem of finding structure in high dimensional datasets. For the model selection, quantitative measure of stability, for different values of the l1–constraint κ, following steps are followed:

(i) compute m noisy replications of the data
(ii) run the class discovery algorithm for each of these datasets
(iii) compute the m x m matrix of pair wise Hamming distances between all partitions
(iv) cluster the partitions into compact groups and score the groups by their frequency
(v) select dominant groups of partitions and choose representative partitions

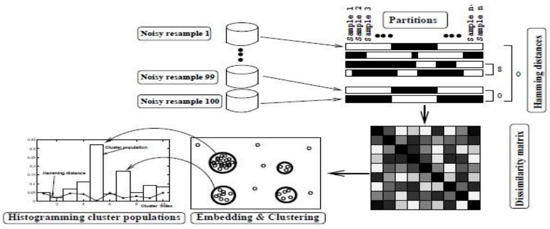

Fig. model selection-schematic work-flow for one fixed value of the l1-constraint κ

We test model for the task of clustering digits from the handwritten database. Set of random images were extracted and model selection procedure was followed, stability of the solution for different values of constraint values were observed. Highly stable solution is found using the model, then the representative partition of the dominating partition cluster is chose. A comparison with the selected feature leads us to the conclusion.

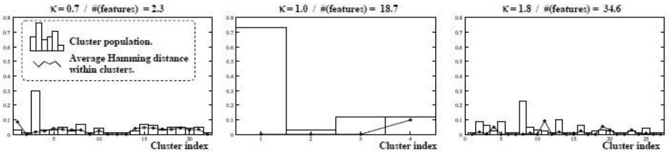

Fig. three different choices for l1-constraint κ

Second experiment was performed on the clustering of the faces. Some leading eigen faces are chosen and selection of an optimal model is done. Choosing the representative cluster of the dominating partition helps us getting to the results.

In this paper, simultaneous clustering objects and extracting subset of the features which are most discriminative for the object partition.

[1] Dash M., Liu H. (2000) Feature Selection for Clustering. In: Terano T., Liu H., Chen A.L.P. (eds) Knowledge Discovery and Data Mining. Current Issues and New Applications. PAKDD 2000

[2] Yuanhong Li, Ming Dong, Jing HUa(2008), Localized Feature Selection in Clustering,In Pattern Recognition Letters, Volume 29,Issue 1,1 January 2008

[3] T. Hastie, R. Tibshirani, and A. Buja. Flexible discriminant analysis by optimal scoring. *J. Am.Stat. Assoc.*, 89:1255–1270, 1994.

[4] T. Lange, M. Braun, V. Roth, and J.M. Buhmann. Stability-based model selection. In *Advance in Neural Information Processing Systems*, volume 15, 2003.