REVIEW OF ICL CRITERION FOR CLUSTERING IN MIXTURE MODELS

Nusrat Jahan

Department of Mathematics and Statistics
Mississippi State University
Mississippi State, MS 39762 USA
email:njahan@ra.msstate.edu

ABSTRACT

Statistical analysis of mixture models are of interest because it is an alternative to nonparametric density estimation and it is a powerful way of modelling in cluster analysis. In case of density estimation, optimization of the Bayesian Information criterion (BIC) generally results in a good approximation of the density to be estimated. But in case of cluster analysis, the BIC tends to overestimate the number of clusters when the data is a poor fit to the mixture model. In this context a modification of BIC, integrated completed likelihood (ICL) criterion has been investigated. In the ICL approach, the integrated completed likelihood is maximized to select both a relevant form of model and relevant number of clusters. In the BIC approach, only the observed likelihood is maximized. Where as the integrated completed likelihood includes the estimated (using maximum a posteriori function) missing data. The ICL criterion penalizes for the complexity of the mixture model, thus ensuring the partitioning of data with the greatest evidence. This paper will focus on the computation of ICL, effectiveness and drawbacks of ICL in the context of cluster analysis. The differences between ICL and BIC will also be investigated.

1. Introduction

Mixture model based clustering provides a less restrictive framework for partitioning of data into relevant number of clusters - an alternative to nonparametric modeling [2]. In this context, selection of the right model is characterized by the model form and the number of mixture components in the model. Bayesian-based methodologies for identifying a mixture model for clustering data is widely available in literature ([4],[6],[8]).

In the Bayesian framework, a model is selected among several competing models because it has the highest posterior probability given the data set. If all models have same prior probabilities, the selected model with highest posterior probability is equivalent to selecting the model with the largest integrated likelihood (which is the marginal likelihood of the model). Under regularity conditions, Bayesian information criterion (BIC) is a reliable approximation to this integrated likelihood [8]. But the regularity conditions do not hold for estimating the number of clusters ([5], [7]).

The authors [2] claim that the traditional Bayesian Information Criterion (BIC) works well when the mixing proportions are restricted to be equal, without this restriction BIC over estimates the number of clusters. There is also a lack of theoretical justification for BIC approximation of the intigrated likelihood ([1],[2]). The algorithm proposed in [2] identifies a mixture model for clustering data by maximizing the integrated completed likelihood (ICL) with a penalizing function for the complexity of the model. Which is essentially a modification of the BIC criterion.

In this article, ICL criterion is discussed in section 2. Also the difference between BIC and ICL is highlighted in this section. Section 3 discusses and compares the results from simulated and real data sets obtained in [2]. An overall discussion section ends this paper.

2. ICL Algorithm

A finite mixture model with n independent random variables $x_1....,x_n$, from k-component mixture can be represented by

$$f(x_i \mid m, K, \theta) = \sum_{k=1}^{n} p_k \phi(x_i \mid a_k) \quad i = 1, \ldots, n \qquad k = 1, \ldots, K$$

Where m denotes the model and $\theta = (p_1, ...p_k, a_1, ..., a_k)$ is the parameter vector. p_k's are the mixing proportions ($0 < p_k < 1$, for all k=1,...,K and $\sum_k p_k = 1$) and a_k's are model parameters.

In the Bayesian context for clustering the data, a model M_l is selected among H different models because it has the highest posterior probability given the data. The posterior probability is defined by,

$$P(M_l \mid x) = \frac{f(x \mid M_l)P(M_l)}{\sum_{r=1}^{H} f(x \mid M_r)P(M_r)}$$

For the model M_l, $f(x \mid M_l)$ is the integrated (marginal) likelihood and $P(M_l)$ is the prior probability. The integrated likekihood can be approximated by BIC under regularity conditions. But When it comes to estimating the number of components in a mixture model or in other words identifying the number of relevant clusters, the regularity conditions for BIC no longer holds ([5],[7]). The reason for this is that, if the mixture model(K) has more components than the true model(K^1), then some of the mixing proportions ($K - K^1$) will be close to zero as the sample size increases and the corresponding estimated proportions will be on the boundary of the parameter space [1].

Mixture model can also be expressed in terms of incomplete data structure [3], where the complete data $y = (y_1, ..., y_n) = ((x_1, z_1), ..., (x_n, z_n))$, $z = (z_1, ..., z_n)$ represents the missing data. For a better clustering structure of the data instead of integrated likelihood, the integrated completed likelihood (which includes the missing data) has been proposed in ([1],[2]).

$$f(X, Z \mid m, k) = \int f(X, Z \mid m, k, \theta)\pi(\theta \mid m, k)d\theta$$

This is approximated by a modified BIC approximation, referred to as ICL criterion.

$$ICL(m, k) = \max_{\theta} \log f(x, \hat{z} \mid m, k, \hat{\theta}) - \frac{\nu_{m,k}}{2} \log n$$

where $\hat{\theta}$ is m.l. estimate of θ (mixture vector parameter) obtained by EM algorithm, the missing data Z is replaced by maximum a posteriori (MAP) operator from $\hat{\theta}$ and $\nu_{m,k}$ is the number of free parameters in the model m with k components. For each situation, the number of times EM algorithm is initiated with random centers, depends on the features of the data to be classified, this number increases with sample size and space dimension [1]. In each situation, the solution providing the largest observed likelihood is selected [2].

It should be noted that by isolating the contributions of missing data Z to the $f(X, Z \mid m, k)$ (the intigrated completed likelihood),the problem of regularity conditions has been avoided in ICL criterion.

The only difference between ICL and BIC is the penalty factor - substraction of the estimated mean entropy from the log-likelihood. Mean entropy is a measure of the ability of the k-component mixture model to provide a relevant partition of data [3]. In other words, ICL penalizes for the number of parameters in the model (in other words model compexity) and thus ensuring that data is partitioned with the greatest evidence.

3. Discussion

In [2] the authors compared the results for five experiments, obtained using both BIC and ICL criterions. Two of these experiments were Monte Carlo experiments with simulated data and others were real data sets. For each experiment, maximum likelihood estimate of parameter vector was obtained via EM algorithm. To obtain sensible maxima, for each case EM algorithm was initiated 20 times with random centers [2]. The model providing the largest observed likelihood was selected.

The first experiment used simulated two types of three-component Gaussian mixtures. The mixtures only differ by the second component of the varaince matrix. Fifty samples were generated for each type of simulated data. In this experiment alltogether 28 models were considered with number of clusters one to seven. For the first type of three-component Gaussian mixture, components were well separated. Both ICL and BIC selected the right model most of the times (BIC 92 percent and ICL 88 percent of the simulations).

For the second type of three-component Gaussian mixtures, first two components were overlapping. In this situation BIC outperformed ICL. BIC selected the right model in 92 percent of the simulations, where as ICL selected the right model only 8 percent of the cases.

These two results indicate for a Gaussian mixture, performance of BIC is as good as ICL and in some cases better than ICL.

The second experiment considered in [2], is a simulated mixture of uniform and gaussian clusters. This experiment considered only one model with number of clusters varying from one to five. In this case performance of BIC is inferior than that of ICL. The ICL criterion chose the right model in 100 percent and BIC chose the right model in 60 percent of the simulations.

Experiments with three real data sets were also conducted [2]. In the case of *Old Faithful Geyser* data, 28 models with cluster numbers one to six were considered. For almost every model, ICL selected a cluster size of two and BIC selected a cluster size of 3. ICL result clearly distinguished two clusters where as BIC result pointed to model deviations from normality.

In the *French Departments* data, 28 models with cluster number from one to five were considered. Both criterias favored the same model and results for cluster number identification were comparable.In this case performances of BIC and ICL were equivalent.

In the *Acoustic Emission Control* data, Gaussian mixture model (with equal proportions and different volumes) and a uniform distribution was considered for the modeling purpose of the data. Cluster size varied from two to twenty. In this case BIC overestimated the number of clusters. ICL identified ten clusters with strong evidence.

The results obtained in [2] suggests that whenever the true model is Gaussian, performance by BIC is very close to that of ICL. But if the true model is not Gaussian, BIC tends to overestimate. These experiments support the contentions of [2], that is a modification of BIC criterion - which accounts for model complexity is required.

4. Summary

Finite mixtures can be employed as a powerful modelling way in cluster analysis ([4],[5]). In this context, selecting a relevant form of model and assessing a sensible number of clusters is of utmost importance. The

results presented in [2] is very impressive. It appears that if the mixture data is not Gausssian, ICL performs better than the BIC criterion in terms of identifying the correct model and also choosing the correct number of clusters. But if the mixture comes from a Gaussian distribution, the performance of BIC is equivalent to that of ICL. The difference between ICL and BIC is the entropy term - the additional criteria included in ICL, which ensures well-separated clusters.

References

[1] C. Biiernacki, G.Celeux and G.Govaert,"Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood",Technical report 3,521,Inria,1998.

[2] C. Biiernacki, G.Celeux and G.Govaert,"Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood",IEEE transactions on Pattern Analysis and Machine Intelligence,Vol.22,No.7,July 2000,pp.719-725.

[3] A.P.Dempster, N.M.Laird, and D.B. Rubin, "Maximum Likelihood for Incomplete Data via the EM Algorithm"(with discussion), J.Royal Statistical Soc.B,Vol.39,pp.1-38,1977.

[4] J.Diebolt and C.P.Robert,"Estimation of Finite Mixture Distributions through Bayesian Sampling", J. Royal Statistical Soc.B,Vol.56,pp.363-375 ,1994.

[5] C.Fraley and A.E.Raftery,"How Many Clusters? Which Clustering Method? Answers via model-Based Cluster Analysis,"Computer J.,Vol.41,pp.578-588, 1998.

[6] S.J.Roberts, D.Husmeier,I.Rezek, and W.Penny,"Bayesian Approaches to Gaussian Mixture Modelling",IEEE transactions on Pattern Analysis and Machine Learning,Vol.20,pp.1,133-1,142,1998.

[7] K.Roeder and L.Wasserman,"Practical Bayesian Density Estimation Using Mixtures of Normals", J.Am.Statistical Assoc.,Vol.92,pp.894-902,1997.

[8] G.Schwarz,"Estimating the Dimensions of a Model",Annals of Statistics, Vol.6,pp.461-464,1978.