CLUSTERING VIA THE BAYESIAN INFORMATION CRITERION WITH APPLICATIONS IN SPEECH RECOGNITION

Scott Shaobing Chen & P.S. Gopalakrishnan
IBM T.J. Watson Research Center
email: schen@watson.ibm.com

ABSTRACT

One difficult problem we are often faced with in clustering analysis is how to choose the number of clusters. In this paper, we propose to choose the number of clusters by optimizing the Bayesian information criterion (BIC), a model selection criterion in the statistics literature. We develop a termination criterion for the hierarchical clustering methods which optimizes the BIC criterion in a greedy fashion. The resulting algorithms are fully automatic. Our experiments on Gaussian mixture modeling and speaker clustering demonstrate that the BIC criterion is able to choose the number of clusters according to the intrinsic complexity present in the data.

1. INTRODUCTION

Clustering methods have been widely used in statistical data analysis to model a complex data set. Globally the data set might be inhomogeneous and difficult to be understood. However, if we cluster the data into homogeneous regions, then each cluster is much simpler, for which various models can be constructed. Many clustering algorithms have been developed in the literature, ranging from hierarchical methods such as bottom-up (or agglomerative) methods and top-down (or divisive) methods, to optimization methods such as the k-means algorithm.

One difficult problem we often encounter in clustering analysis is how to choose the number of clusters. The common practice is to pre-determine the number of clusters, then run the clustering algorithm. In hierarchical methods, clustering trees are often constructed; according to the desired number of clusters, one can go down the tree to obtain desired clustering. In the k-means algorithm, according to the desired number of clusters, one picks that many initial seeds. Another common practice is to threshold the distance measures during the hierarchical process; the thresholding level is tuned on a training set. However, all these methods are *heuristic*.

Ideally the number of clusters should be chosen automatically according to the complexity of the data set: the higher the complexity, the more clusters are needed. For example, in speech recognition, we use Gaussian mixtures to model the output distribution for sub-phonetic units (or HMM states) [2]; each Gaussian then corresponds to a cluster in the speech feature space. Different phonetic units may have different degrees of complexity. Figure 1 shows the samples corresponding to the first sub-phonetic units of phone KD-3 and ER-3 from the WSJ SI-284 corpus, plotted along the directions of their first 2 principle components. It appears that KD-3 has a simpler structure than ER-3; therefore fewer Gaussians are needed in order to have an adequate model for KD-3.

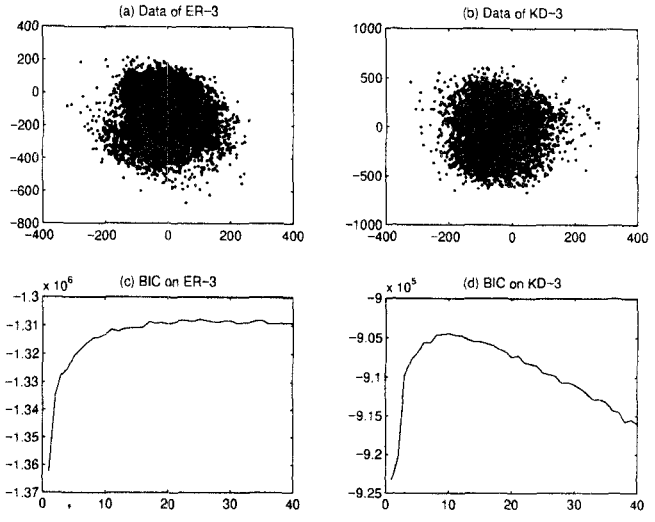

Figure 1. Different degrees of complexity in phone ER-3 and KD-3

In this paper, we propose to choose the number of clusters by optimizing the Bayesian information criterion (BIC), a model selection criterion in the statistics literature. The resulting algorithm is fully automatic. In general, one can run the clustering algorithm to obtain a clustering $\mathcal{C}_n$ which has n clusters. Each clustering $\mathcal{C}_n$ is evaluated at its BIC value. Then the clustering with the highest BIC value is chosen. It is often rather slow to perform such global searching on the BIC values. However, for hierarchical methods, we propose a termination criterion which optimizes the BIC criterion in a greedy

0-7803-4428-6/98 $10.00 © 1998 IEEE

fashion: in the bottom-up methods, two nodes are merged only if the merging increases the BIC value; similarly in the top-down methods, a node is split only if the splitting increases the BIC value. Our experiments on Gaussian mixture modeling and speaker clustering demonstrate that the BIC criterion is able to choose the number of clusters according to the intrinsic complexity present in the data set.

This paper is organized as follows: section 2 describes model selection criterions in the statistics literature; section 3 presents clustering via BIC; section 4 explains hierarchical clustering via greedy BIC; we present our experiments on speaker clustering and Gaussian mixture modeling in section 5 and section 6; in section 7, we compare our speaker clustering algorithm with other recent works in the literature [5, 7].

2. MODEL SELECTION CRITERIONS

The problem of model identification is to choose one among a set of candidate models to describe a given data set. We often have candidates of a series of models with different number of parameters. It is evident that when the number of parameters in the model is increased, the likelihood of the training data is also increased; however, when the number of parameters is too large, this might cause the problem of *overtraining*. Several criteria for model selection have been introduced in the statistics literature, ranging from nonparametric methods such as cross-validation, to parametric methods such as the Bayesian Information Criterion (BIC) [8].

BIC is a likelihood criterion penalized by the model complexity: the number of parameters in the model. In detail, let $\mathcal{X} = \{x_i : i = 1, \cdots, N\}$ be the data set we are modeling; let $\mathcal{M} = \{M_i : i = 1, \cdots, K\}$ be the candidates of parametric models. Assuming we maximize the likelihood function separately for each model M, obtaining, say $L(\mathcal{X}, M)$. Denote $\#(M)$ as the number of parameters in the model M. The BIC criterion is defined as:

$$BIC(M) = \log L(\mathcal{X}, M) - \frac{1}{2}\#(M) \times \log(n) \quad (1)$$

The BIC procedure is to choose the model for which the BIC criterion is maximized. This procedure can be derived as a large-sample version of Bayes procedures for the case of independent, identically distributed observations and linear models [8].

The BIC criterion has been widely used for model identification in time series [9], linear regression [3], etc. BIC is closely related to other penalized likelihood criterions such as AIC [1] and RIC [3]; BIC has theoretical advantages because of its connection with Bayesian procedures.

3. CLUSTERING VIA BIC

In this section, we describe how to apply BIC in clustering analysis.

Let $\mathcal{X} = \{x_i \in \mathcal{R}^d : i = 1, \cdots, N\}$ be the data set we wish to cluster. Let $\mathcal{C}_k = \{c_i : i = 1, \ldots, k\}$ be the clustering which has k clusters. We model each cluster c_i as a multi-variate Gaussian distribution $N(\mu_i, \Sigma_i)$, where μ_i can be estimated as the sample mean vector and Σ_i can be estimated as the sample covariance matrix. Thus the number of parameters for each cluster is $d + \frac{1}{2}d(d+1)$. Let n_i be the number of samples in cluster c_i. One can show that

$$BIC(\mathcal{C}_k) = \sum_{i=1}^{k}\{-\frac{1}{2}n_i \log|\Sigma_i|\} - Nk(d + \frac{1}{2}d(d+1)) \quad (2)$$

We choose the clustering which maximizes the BIC criterion.

We applied the BIC criterion to cluster the data sets of phone KD-3 and ER-3 shown in Figure 1. For a given number of clusters, k-means algorithm was employed to generate the clustering. As in shown Figure 1 (a) and (b), KD-3 seems to have a lower degree of complexity than ER-3. Indeed the BIC criterion chose 25 clusters for ER-3 versus 10 clusters for KD-3. Figure 1 (c) and (d) shows how the BIC value evolves as the number of clusters increases: for KD-3, the BIC value increases initially, then declines rapidly as the likelihood saturates; for ER-3, the BIC value increases in the beginning, then declines slowly, indicating a higher degree of complexity.

We comment that we could model each cluster as simpler distributions, such as multivariate Gaussians with diagonal covariance matrices, or as more complex distributions, such as Gaussian mixtures. In either case, we will still be able to derive a BIC criterion.

4. HIERARCHICAL CLUSTERING VIA GREEDY BIC

As one can imagine, it is often very costly to search globally for the best BIC value, since clustering has to be performed to obtain different numbers of clusters. However, for hierarchical clustering methods, it is possible to optimize the BIC criterion in a greedy fashion.

Let $\mathcal{X} = \{x_i \in \mathcal{R}^d : i = 1, \cdots, N\}$ be the data set. Bottom-up methods start with each data sample as initial nodes, then successively merge two nearest nodes according to a distance measure. Let $\mathcal{S} = \{s_1, \ldots, s_k\}$ be the current nodes; suppose s_1 and s_2 are the candidate pair for merging, and the merged new node is s. Thus we are comparing the current clustering tree $\mathcal{S}$ with a new clustering tree $\mathcal{S}' = \{s, s_3, \ldots, s_k\}$. We model each node s_i as a multivariate Gaussian distribution $N(\mu_i, \Sigma_i)$. It is clear from (2) that the increase of the BIC value by merging s_1 and s_2 is

$$-n \log|\Sigma| + n_1 \log|\Sigma_1| + n_2 \log|\Sigma_2| + N(d + \frac{1}{2}d(d+1)) \quad (3)$$

where $n = n_1 + n_2$ is sample size of the merged node and Σ is the covariance matrix of the merged node.

Our BIC termination procedure is that two nodes should not be merged if (3) is negative. Since the BIC value in increased at each merge, we are searching for an "optimal" clustering tree by optimizing the BIC criterion in a greedy fashion.

Note that we merely use our criterion (3) for termination. It is possible to use our criterion (3) as the distance measure in the bottom-up process. However, in many applications, it is probably better to use more sophisticated distance measures. It is also clear that our criterion can be applied to top-down tree methods.

5. APPLICATION IN SPEAKER CLUSTERING

Suppose we have a collection of sentences; each sentence is from a certain unknown speaker. We are interested in clustering the sentences according to speaker identities. Most speaker clustering algorithms in the literature are hierarchical clustering methods with various distance measures. In this section, we present our speaker clustering experiment utilizing the BIC termination criterion (3).

The data set consists of the clean prepared and the clean spontaneous portion of the HUB4-96 evaluation data [2], hand-segmented into 824 short segments. Cepstra coefficients were extracted as feature vectors x_i for each segment. We used the log likelihood ratio distance measure proposed in Gish et. al. [4]; Bottom-up clustering was performed with the maximum linkage, with the BIC termination criterion (3).

The true number of speakers is 28; the BIC termination criterion chose 31 clusters. For each cluster, we define the purity as the ratio between the number of segments by the dominating speaker in that cluster and the total number of segments in that cluster. Figure 2 shows the purities of each cluster. Clearly our algorithm results in not only clusters with high purity, but also the appropriate number of clusters.

Speaker clustering can enhance the performance of unsupervised adaptation. The reason is that most of the 824 segments here are quite short, around $2 \sim 3$ seconds. Without speaker clustering, unsupervised adaptation techniques such as MLLR [6] has small improvements due to lack of data. Good speaker clustering can bring the segments of the same speaker together thus improves the performance of unsupervised adaptation. We started from a baseline system which had about 90k Gaussians. The decoding results were scored according to two conditions: clean prepared and clean spontaneous. As shown in Table 1, the baseline error rates were 18.8% and 27.0% for the two conditions respectively; without clustering, MLLR reduced the error rates by only 0.1%; with our clustering, MLLR reduced the error rates by 1.3% for the clean condition and by 2.4% for the spontaneous condition. Table 1 also shows the error rates of MLLR using the ideal clustering by the true speaker identities. It is clear that our speaker clustering enhanced the performance of MLLR as much as the ideal clustering.

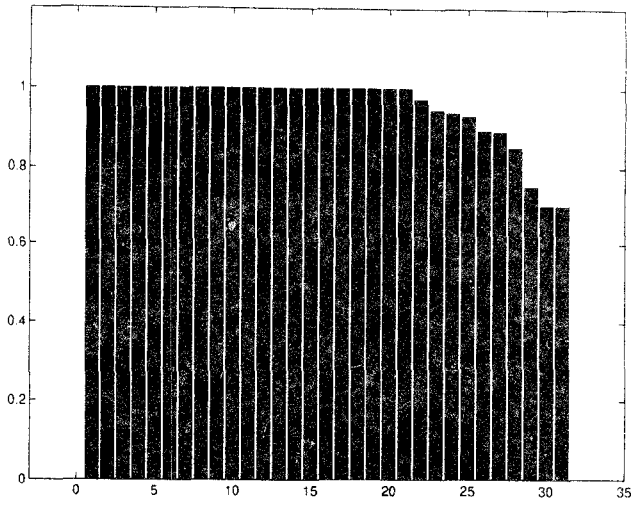

Figure 2. Clustering Purities

	Prepared	Spontaneous
Baseline	18.8%	27.0%
MLLR w/o clustering	18.7%	26.9%
MLLR w/ ideal clustering	17.5%	24.8%
MLLR w/ BIC clustering	17.5%	24.6%

Table 1. MLLR adaptation enhanced by BIC clustering

6. APPLICATION IN GAUSSIAN MIXTURE MODELING

Suppose we have a data set of high dimensional continuous observations $\mathcal{X} = \{x_i \in \mathcal{R}^d, i = 1, \cdots, n\}$, and we are interested in modeling the data set as a Gaussian mixture distribution. When the sample size n is large, top-down clustering methods are often used to cluster the data set into clusters, and each cluster results in a Gaussian in the mixture. It is evident that our BIC criterion (3) can be applied here to choose the number of clusters.

Gaussian mixtures are used in the IBM large vocabulary speech recognition systems to model the HMM states [2]. Normally the number of Gaussians are chosen by a threshold method: for example, if the total number of samples in the entire training data set is 10 million, and if the number of samples in each cluster is required to be at least 100, then there will be 100k Gaussians in the system.

We compare two recognition systems: one has 70k Gaussians, chosen by the thresholding method; the other has 38k Gaussians, chosen by the BIC criterion. The two systems were tested on the WSJ 92 development test set; they gave similar error rates, as shown Table 2. Thus by picking the right number of Gaussians, the BIC crite-

rion delivers a system with smaller number of Gaussians without sacrificing the performance.

	70k baseline system	38k BIC system
Error Rate	8.85%	8.92%

Table 2. Choosing the number of Gaussians via BIC

7. DISCUSSION

In this section, we compare our application in speaker clustering with some other recent works in the literature.

Jin et al. of BBN [5] proposed a similar automatic speaker clustering algorithm. They also used the log likelihood ratio distance measure proposed in Gish et al. [4], however, with the distances between consecutive segments scaled down by a parameter α. They performed hierarchical clustering; for any given number k, the clustering tree was pruned to obtain k tightest clusters. An heuristic model selection criterion

$$\sum_{j=1}^{k} n_j^{\alpha} |\Sigma_j^{\alpha}| * \sqrt{k} \tag{4}$$

was then used to search through the space of (α, k) for the best clustering. They applied this algorithm to cluster the HUB4-96 evaluation data for the purpose of unsupervised adaptation. Similarly, this automatic clustering enhanced the unsupervised adaptation as much as the cheating clustering according the the true speaker identities.

This heuristic model selection criterion (4) resembles the BIC criterion (2): they both penalize the likelihood by the number of clusters. However, the BIC criterion has a solid theoretical foundation and seems more appropriate. Indeed the number of speaker clusters found by their algorithm is considerably less than the truth. Moreover, extra information such as the adjacency of the segments was utilized in their algorithm.

Siegler et al. of CMU [7] proposed another speaker clustering algorithm. They chose the symmetric Kullback-Leibler metric as the distance measure, and performed hierarchical clustering. The clusters were obtained by thresholding the distances. Unlike our method and the BBN clustering, this clustering is not fully automatic: the thresholding level was tuned in a delicate fashion: it had to be small enough such that the clusters created were made up of segments from only one speaker and yet large enough to improve the performance of the unsupervised adaptation.

8. CONCLUSION

We proposed an automatic algorithm to choose the number of clusters in clustering analysis via the BIC criterion. We developed a termination criterion for the hierarchical methods which optimizes the BIC criterion in a greedy fashion. Our experiments in speaker clustering and Gaussian mixture modeling demonstrated that the BIC criterion is able to choose the number of clusters according to the intrinsic complexity present in the data set.

REFERENCES

[1] H. Akaike, "A new look at the statistical identification model", IEEE Trans. Auto. Control, vol 19, pp 716-723, 1974.

[2] R. Bakis, S. Chen, P.S. Gopalakrishnan, R. Gopinath, S. Maes, L. Polymenakos, "Transcription of broadcast news shows with the IBM large vocabulary speech recognition system", Proceedings of the Speech Recognition Workshop, pp 67-72, 1997.

[3] D. Foster and E. George, "The risk inflation factor in multiple linear regression", Technical Report, Univ. of Texas, 1993.

[4] H. Gish and N. Schmidt, "Text-independent speaker identification", IEEE Signal Processing Magazine, pp 18-21, Oct. 1994.

[5] H. Jin, F. Kubala and R. Schwartz, "Automatic speaker clustering", Proceedings of the Speech Recognition Workshop, pp 108-111, 1997.

[6] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMM's", Computer Speech and Language, vol. 9, no. 2, pp 171–186.

[7] M. Siegler, U. Jain, B. Ray and R. Stern, "Automatic segmentation, classification and clustering of broadcast news audio", Proceedings of the Speech Recognition Workshop, pp 97-99, 1997.

[8] G. Schwarz, "Estimating the dimension of a model", The Annals of Statistics, vol. 6, pp 461-464, 1978.

[9] W.S. Wei, Time Series Analysis, Addison-Wesley, 1993.