MDL-Based Selection of the Number of Components in Mixture Models for Pattern Classification

Hiroshi Tenmoto, Mineichi Kudo and Masaru Shimbo

Division of Systems and Information Engineering
Graduate School of Engineering
Hokkaido University, Sapporo 060-8628, Japan
{tenmo, mine, shimbo}@main.eng.hokudai.ac.jp
<http://ips6.main.eng.hokudai.ac.jp/>

Abstract. A new method is proposed for selection of the optimal number of components of a mixture model for pattern classification. We approximate a class-conditional density by a mixture of Gaussian components. We estimate the parameters of the mixture components by the EM (Expectation Maximization) algorithm and select the optimal number of components on the basis of the MDL (Minimum Description Length) principle. We evaluate the goodness of an estimated model in a trade-off between the number of the misclassified training samples and the complexity of the model.

1 Introduction

In pattern recognition, we often apply clustering techniques over the training samples to understand the spatial structure of the samples and to reduce the computational costs of design of classifiers.

In approximating a class-conditional density by a mixture of Gaussian components, we encounter the situation of finding initial components for an iterative procedure called *EM algorithm* [1]. In this situation, two problems arise: 1) How should we select the number of components? and 2) How should we construct initial components? The first problem is crucial but difficult, and there have been many efforts made to resolve this problem (these efforts are discussed in reference [2]). For example, Ichimura [3] has proposed a method for selecting the number of components based on information criteria. In this method, the optimal number of components is selected on the basis of the *MDL principle* [4] in a trade-off between the likelihood of the model to the training samples and the complexity of the model itself.

All of these studies have been carried out as unsupervised learning; that is, individual class-conditional density has been approximated without taking into account the relationship between classes. However, our main concern is in approximating boundaries between classes. In the case of a limited number of training samples, a better approximation of the class-conditional density of

individual classes does not necessarily mean a better classification with the Bayes classifier based on the approximated densities. Therefore, we try to select the optimal number of components by evaluating both the classification error of the training samples and the complexity of the classifier in a supervised MDL principle.

2 Classifiers Based on Mixture Models

We describe a class-conditional density as follows:

$$f(\mathbf{x}) = \sum_{k=1}^K p_k N(\mathbf{m}_k, \Sigma_k)(\mathbf{x}),$$

where K is the number of the components, $N(\mathbf{m}_k, \Sigma_k)$ is a Gaussian density with a mean vector $\mathbf{m}_k$ and a covariance matrix Σ_k , and p_k ($\sum_{k=1}^K p_k = 1$) is the mixture ratio of the k th component. For M classes, $f_1(\mathbf{x}), \dots, f_M(\mathbf{x})$ will be constructed.

In the *EM algorithm* [1], the mixture parameters are estimated so as to maximize the following log-likelihood:

$$L = \sum_{i=1}^N \sum_{k=1}^K z_{ik} \log(p_k N(\mathbf{m}_k, \Sigma_k)(\mathbf{x}_i)), \quad (1)$$

where N is the number of training samples $\{\mathbf{x}_i\}$, and z_{ik} ($\sum_{k=1}^K z_{ik} = 1$) is the membership value of $\mathbf{x}_i$ to the k th component. As initial membership values $\{z_{ik}\}$ to be given to the algorithm, we use the fuzzy *c*-means method with $m = 1.6$, m being the fuzziness factor [5].

We estimate the class-conditional densities for each class, then we construct a classifier on the basis of Bayes rule and classify unknown samples by the classifier.

Too many components in the case of a limited number of training samples may cause singularity of the covariance matrices. Thus, in this paper, we adopt only the number of components whose covariance matrices are not singular.

3 Selection of the Number of Components Based on the MDL Principle

We consider two approaches based on the *MDL principle* [4]: 1) one approach is in an unsupervised situation (i.e., clustering), and 2) the other approach is in a supervised situation (i.e., classification).

3.1 Unsupervised Approach

Ichimura [3] proposed a method for selecting the number of components in each class on the basis of the MDL principle. This method selects the number of

components that minimizes the following formula:

$$\text{MDL} = -L + \frac{1}{2}m \log N,$$

$$m = (K - 1) + K \left\{ D + \frac{1}{2}D(D + 1) \right\},$$

where L is the log-likelihood in Eq. (1), K is the number of the components for the class, and D is the number of the features. For each class, according to this formula, we select the optimal number of components.

3.2 Supervised Approach (Proposed Method)

To take into consideration information of the classification results of the training samples, we use the other MDL calculation method proposed by Kudo *et al.* [6].

In this approach, both factors, the classification ability for the training sample set S and the complexity of the classifier c , are evaluated together in bit length. We select a classifier that minimizes the following formula:

$$\text{MDL} = L(S|c) + L(c),$$

where $L(S|c)$ is the description length for encoding the labels of the misclassified training samples of S by a classifier c , and $L(c)$ is the description length of c .

We calculate the first term as follows:

$$L(S|c) = N^- H(N_1^-, \dots, N_M^- | N^-) + \frac{1}{2}(M-1) \log_2 N^- + N H(N^-, N^+ | N) + \frac{1}{2} \log_2 N,$$

where M is the number of classes; N^+ and N^- are the number of the training samples classified correctly and incorrectly, respectively;

$$H(N_1, \dots, N_M | N) = - \sum_{i=1}^M \frac{N_i}{N} \log_2 \frac{N_i}{N}$$

is an entropy; and N_i^+ (N_i^-) is the number of correctly (wrongly) classified training samples of the i th class. This term, $L(S|c)$, decreases according to the classification accuracy for the training samples.

In addition, we calculate the second term as

$$L(c) = \alpha K D^\beta \log_2 N,$$

where K is the total number of the components for all classes, and α and β are the parameters for suppressing the overabundant effect of this term. These parameters depend on the representation ability of the classifier but are independent of individual problems. We use $\alpha = 0.088$ and $\beta = 1.212$ reported in reference [6]. This term decreases according to the simplicity of the classifier.

To select the optimal combination with respect to the numbers of components of all classes, we use a simple greedy algorithm. Starting with one component for every class, we increase by one the number of components of only the class for which $L(S|c)$ takes the minimum. This procedure is repeated as long as every covariance matrix of the components is nonsingular. From all examined combinations, we select the combination by which the MDL attains the minimum.

4 Experiments

We carried out experiments on two artificial datasets called *Cross data* and *Ring data* and on two real datasets called *Vowel data* [7] and *Ship data* [8].

Cross data is a two-class problem in 2-dimensional space: the first class forms a cross and the second class is of four squares surrounding the cross. These classes are perfectly separable. In this dataset, 200 samples per class were used for training and 1000 samples per class were used for test. Ring data is also a two-class problem in 2-dimensional space: the first class forms a ring with an inner radius of 2.147 and an outer radius of 3.030, and the second class is a Gaussian density with a mean at the origin and a covariance matrix of unity. The Bayes recognition rate is 95.5%. In this dataset, 100 samples per class were used for training and 1000 samples per class were used for test.

Vowel data is a Japanese five-vowel dataset [7]. We used six features including formant frequencies. The number of training samples is 200 per class, and the number of test samples is 1000 per class. Ship data aims to distinguish eight types of military ships with eleven features. For details, see reference [8]. We split the dataset evenly into two sets for training and test. In advance, for these real datasets, we normalized all features to a mean of 0 and a variance of 1 in order to avoid singularity of the covariance matrices as far as possible.

We show the results in Figs. 1 – 4 with recognition rates. In these figures, each mixture component is represented by an ellipse drawn as points that are one standard deviation from the mean. For the artificial datasets, an exhaustive search was carried out in addition to the simple greedy search.

In Cross data, the unsupervised method was superior to our supervised approach, but in Ring data, our approach was superior to the unsupervised approach, especially when the exhaustive search was used (Figs. 1 and 2). In the two real datasets, it can be seen that there exists a better solution than that of the unsupervised approach (Figs. 3 and 4). This supports our basic insistence that we should consider separately the two problems of approximation of class-conditional densities and of construction of classifiers based on approximated densities.

Fig. 1. Mixture components for Cross data

Fig. 2. Mixture components for Ring data

Fig. 3. Variations of MDL values and recognition rates for Vowel data

Fig. 4. Variations of MDL values and recognition rates for Ship data

The greedy search was not sufficient for finding the best solution from all possible solutions. For example, in Ring data, the greedy search selected too many components for the first class. Therefore, we should improve the search algorithm so as to examine a wider range of solutions without requiring much computation time. Several feature selection methods, including the *Sequential Floating Forward (Backward) Search* method [9], may be useful.

In all of the experiments, our supervised MDL had a tendency to overestimate the number of components, because the second term $L(S|c)$ of our MDL was too suppressed. Thus, the coefficients in the second term of our MDL criterion should be reexamined.

5 Conclusion

We tried to select the optimal number of components in mixture models for pattern classification using a classification-oriented MDL criterion.

In this study, we examined every combination of the numbers of components assigned to classes, for example, (1,1), (1,2), (1,3), (2,1), (2,2), (2,3), ... for two classes. However, we can keep the results for each class and later combine them. This is useful for economizing computation costs.

To confirm the effectiveness of our approach, experiments on more datasets are needed.

Acknowledgement

The authors are grateful to Professor Dr. Sklansky for providing the Ship data.

References

1. Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Royal Stat. Soc. Series B* **39** (1977) 1–38
2. McLachlan, G. J., Basford, K. E.: *Mixture Models*. Marcel Dekker, Inc., New York (1988) 21–29
3. Ichimura, N.: Robust Clustering Based on a Maximum Likelihood Method for Estimation of the Suitable Number of Clusters. *Trans. Inst. Electron. Info. Comm. Eng. D-II* **J78** 8(1995) 1184–1195 (in Japanese)
4. Rissanen, J.: A Universal Prior for Integers and Estimation by Minimum Description Length. *Ann. Stat.* **11** (1983) 416–431
5. Bezdek, J. C.: Cluster Validity with Fuzzy Sets. *J. Cybern.* **3** 3(1974) 58–73
6. Kudo, M., Shimbo, M.: Selection of Classifiers Based on the MDL Principle Using the VC Dimension. *Proc. ICPR '96* (1996) 886–890
7. Hayamizu, S. *et al.*: Generation of VCV/CVC Balanced Word Sets for Speech Database. *Bullet. Electro. Lab.* **49** 10(1985) 803–834
8. Park, Y., Sklansky, J.: Automated Design of Multiple-Class Piecewise Linear Classifiers. *J. Classification* **6** (1989) 195–222
9. Pudil, P., Novovičová, J., Kittle, J.: Floating Search Methods in Feature Selection. *Pattern Recogn. Letters* **15** (1994) 1119–1125