

Incremental Learning of Mixture Models for Simultaneous Estimation of Class Distribution and Inter-Class Decision Boundaries

Dwi Sianto Mansjur, Biing Hwang Juang
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA, USA 30332
{dwi,juang}@ece.gatech.edu

Abstract

In this paper, we propose a novel design of high performance Bayes classifier from a small number of observations. The two main challenges to obtain the classifier are the lack of the true functional form of the class-conditional density and the lack of enough data to estimate the parameters of the classifiers. Incremental learning of Gaussian Mixture Model (GMM) is used to mitigate the lack of the true functional form. Moreover, the classifier uses the training samples from all classes to evaluate the goodness of a particular mixture to be used as the classifier for a specific class. This selection process eases the difficulty of the accurate parameter estimation. Thus, the important trait of the proposed classifier is being able to estimate simultaneously class-conditional density and inter-class boundaries to arbitrary precision. Our experimental results show that the proposed classifier not only has better performance than the conventional classifiers but also requires fewer parameters.

1 Introduction

Bayes decision theory is the foundation of statistical pattern recognition. According to the theory, the maximum *a posteriori* (MAP) decision rule is guaranteed to minimize the probability of error. Thus, the conventional classifier attempts to estimate the class-conditional densities as accurate as possible. In practice, these density estimations might be inaccurate either due to the lack of information on the true functional form of the class-conditional density, or due to the limited number of observations available to approximate

“quickly” approximate any target density with arbitrary precision [4]. For classifier design, this mitigates the problem due to the lack of information on true functional form. Moreover, this result still holds if the mixture component is learned incrementally, e.g., start with one mixture and add one mixture after another. To identify a new component to add to the model is equivalent to find the global maximum of a log-likelihood surface. It is proposed in [4] to perform grid search in the parameter space. In many practical applications, this searching is infeasible especially in the case of a small number of observations. We address this problem by using classification based mixture selection technique.

In this paper, we propose the design of a Bayes classifier with the objective of achieves good approximations of class-conditional densities and inter-class decision boundaries simultaneously. To achieve the objective, we make use of a mixture selection criteria and add mixture components incrementally. The main advantage of the criteria is its capability to use the contribution of the available training sample to quantify the goodness of a particular mixture as the inter-class decision boundaries. The use of incremental learning is later shown to allow the classifier design to circumvent the need to specify the initialization mixtures and still be able to approximate the true density to arbitrary precision. The experiments show that the proposed classifier has good generalization capability in addition to low complexity.

The remainder of this paper is organized as follows. Section 2 introduces the conventional and the discriminative mixture selection criteria that are used in this study. Section 3 explains the incremental learning algorithm used to obtain the low complexity and highly discriminative classifier. Experiments on a set of machine learning datasets are provided in the section 4.

2 Mixture Model Selection

For a given set of M categories or classes $\{C_i : i = 1, \dots, M\}$, the classifier task is to assign an incoming pattern x into one of the classes. Decoding based on Bayes classification theory is guaranteed to minimize the probability of error. Thus, a Bayes classifier choose the class C_i that yields the maximum value of $P(C_i|x)$, the posteriori probability of the class, given the data. By using the Bayes rule, the Bayes decoding is equivalent to maximizing the product of the class-conditional probability $p(x|C_i)$ and the prior probability of the class $P(C_i)$. That is,

$$x \in C_i \text{ if } i = \arg \max_k p(x|C_k)P(C_k) \quad (1)$$

where x is the input data to the system. Mixture model selection problem consists of selecting a single topology \mathcal{T}_{il} as sole representative of the class C_i . This process is carried out by using selection criteria such that:

$$C_i \text{ uses } \mathcal{T}_{il} \text{ if } l = \arg \max_k \mathcal{C}(\mathcal{T}_{ik}) \quad (2)$$

where we assume that there is a set of L_i candidate models, $\{\mathcal{M}_{il} : l = 1, \dots, L_i\}$ for each class C_i . Each model \mathcal{M}_{il} includes both the model structure (topology) \mathcal{T}_{il} and the parameter of the model Λ_{il} . Thus, each model \mathcal{M}_{il} implements the class-conditional probability $p(x|C_i)$ as $p(x|\mathcal{T}_{il}, \Lambda_{il})$.

We consider two approaches for mixture model selection for classifier design, i.e., Bayesian Information Criteria (BIC) and Discriminative Information Criteria (DIC). The major difference between the two approaches is in the use of class label information. BIC does not use the information while DIC does use the information.

The standard criterion to choose the number of components in each class is to use BIC [1]; which assigns score to each topology using the following definition:

$$BIC(\mathcal{T}_{il}) = \log p(X; \mathcal{T}_{il}, \Lambda_{il}) - \beta \frac{K_{il}}{2} \log N_i \quad (3)$$

where K_{il} is the number of free parameters and N_i is size of the observations for class C_i . The BIC is the sum of the likelihood and the term $\frac{K_{il}}{2} \log N_i$ which is a penalty of the number of free parameters in the model. The β term is intended as the regularizing term. In this study, β is set to be 1.

To quantify the goodness of a mixture to specify the inter-class decision boundary, we advocate the use model selection based on DIC proposed by Alain Biem [1]. This criterion takes into account not only the number of free parameters in the model but also the inter-class boundaries, which are important for classification

(decision) process. This criterion is defined as follows:

$$\begin{aligned} DIC(\mathcal{T}_{il}) &= \log p(X_i; \mathcal{T}_{il}, \Lambda_{il}) \\ &- \beta \frac{\sum_{j=1, j \neq i}^{M-1} \log p(X_j; \mathcal{T}_{il}, \Lambda_{il})}{M-1} \\ &+ \frac{K_{il}}{2(M-1)} \sum_{j=1, j \neq i}^{M-1} \log \frac{N_j}{N_i} \end{aligned} \quad (4)$$

The first term is the likelihood of the data, $\log P(X_i|\mathcal{T}_{il}, \Lambda_{il})$, and the second term is the average of the anti-likelihood terms, $\log P(X_j|\mathcal{T}_{il}, \Lambda_{il})$. The anti-likelihood of the data X_j against the model from class C_i is a likelihood-like quantity in which the data and the model belongs to competing categories. The third term accounts for the number of free-parameters in the models and equals to zero if the datasets are of the same size. In this study, the regularizing constant β is set to be 1 to indicate equal important between the target class and the competing classes.

Indeed, there are many information criteria for classification task, e.g. Classification Information Criteria. The DIC is unique because it is closely related to the Minimum Classification Error (MCE) classifier design framework [3]. Using the smooth 0-1 approximation to classification loss through the ratio of two likelihood terms, the MCE framework optimize the classifier and the DIC procedure select the best mixture to ease the difficulties to approximation of classifier parameters.

3 Incremental Mixture Learning

Conventional mixture-based classifiers use EM algorithm to approximate the class-conditional density. The main attraction of the EM algorithm is that it guarantees non-decreasing in the data log-likelihood. Nevertheless, the EM algorithm assumes the information on the number of mixtures is available for classifier design. To avoid this difficulty, we use the incremental mixture learning, in which we increase the number of mixture one by one.

For each class i , we starts the learning process from one mixture model. Using k to denote the number of mixture in the model, the model \mathcal{M}_{il} implements the class-conditional probability $p(x|C_i)$ is written as $p(x; \mathcal{T}_{il}^{(k)}, \Lambda_{il}^{(k)})$. In the case, a new mixture is added with density $\phi_{k+1}(x)$, a new class conditional density $p(x; \mathcal{T}_{il}^{(k+1)}, \Lambda_{il}^{(k+1)})$ is defined as follows:

$$\begin{aligned} p(x; \mathcal{T}_{il}^{(k+1)}, \Lambda_{il}^{(k+1)}) &= (1 - \alpha)p(x; \mathcal{T}_{il}^{(k)}, \Lambda_{il}^{(k)}) \\ &+ \alpha\phi_{k+1}(x) \end{aligned} \quad (5)$$

Datasets		Cancer	Heart	Iris	New Thyroid	Wine	Zoo
Non-Incremental	BIC	95.3(2.0)	81.0(3.7)	96.7(5.0)	96.8(5.3)	95.5(3.5)	91.7(10.2)
	DIC	96.5(2.2)	82.3(4.0)	97.3(4.5)	97.8(3.3)	96.5(4.1)	91.6(8.9)
Incremental	BIC	95.6(7.9)	80.6(2.9)	96.7(4.7)	97.0(5.9)	95.4(4.0)	91.5(9.5)
	DIC	97.2(1.9)	83.7(6.4)	98.0(4.6)	97.4(3.3)	96.6(3.5)	91.6(10.5)

Table 1. The average accuracy % and the standard deviation (in parentheses) of four type of mixture learnings using ten-fold cross-validation

where $p(x; \mathcal{T}_{il}^{(k)}, \Lambda^{(k)})$ is the current mixture model, $\phi_{k+1}(x)$ is the new component and α are the mixing weights for the new components with the range of value between 0 and 1 for $\alpha \in (0, 1)$. This is analogous to the incremental training procedure called Greedy-EM for unsupervised probability density estimation [6].

To find the new components, a set of candidate components are obtained randomly from a data partitioning techniques such as KD-tree [2]. Each candidate component will adjust its parameter using the partial EM update, where only the parameters of the new component are updated. In the expectation step of the partial EM [6], we compute the posterior probability as follows:

$$P(k+1|x_i) = \frac{\alpha \phi_{k+1}(x_j)}{(1-\alpha)p(x; \mathcal{T}_{il}^{(k)} \Lambda_{il}^{(k)}) + \alpha \phi_{k+1}(x_j)} \quad (6)$$

During the maximization step, the component parameters are updated as:

$$\alpha_{i(k+1)} = \frac{1}{N_i} \sum_{j=1}^{N_i} P(k+1|x_j), \quad (7)$$

$$\mu_{i(k+1)} = \frac{\sum_{j=1}^{N_i} P(k+1|x_j)x_j}{\sum_{j=1}^{N_i} P(k+1|x_j)} \quad (8)$$

$$\Sigma_{i(k+1)} = \frac{\sum_{j=1}^{N_i} P(k+1|x_j)(x_j - \mu_j)(x_j - \mu_j)^T}{\sum_{j=1}^{N_i} P(k+1|x_j)} \quad (9)$$

After applying the partial EM steps for each candidate components, we compute the mixture selection score for the set of candidate initial parameter for component $k+1$ using either (3) or (4). The process is then repeated for all classes. In all, the novel classifier design techniques is summarized in Algorithm (1).

4 Experiments

Experiments have been conducted on UCI machine learning datasets [5] to demonstrate the performance of the proposed classifier design. The number of patterns, features and classes for those datasets are listed in Table (3). For each dataset, in order to obtain an estimate of the generalization error, ten-fold cross-validation was used, i.e. ten experiments were conducted, with one of the folds used for testing and the remaining nine folds for training. In each experiment, nine classifiers were

Algorithm 1 Incremental classifier learning.

Require: K as the maximum number of mixture

P as the maximum number of candidates

- 1: **for** Class indicated by i such that $1 \leq i \leq M$ **do**
- 2: Start with one-mixture model computed as follows:

$$\begin{aligned} \mu_{i1} &= \frac{1}{N_i} \sum_{x \in C_i} x \\ \Sigma_{i1} &= \frac{1}{N_i} \sum_{x \in C_i} (x - \mu_{i1})(x - \mu_{i1})^T \\ w_{i1} &= 1 \end{aligned}$$

- 3: **while** the number of mixture k is less than K **do**
 - 4: Prepare a set randomly-chosen candidates of size P .
 - 5: Apply EM update equations (7), (8), (9) to each of the candidates until convergence.
 - 6: Increase the mixture to $k+1$ by combining existing k mixtures with the candidates.
 - 7: Apply conventional EM update equations for all $k+1$ mixture.
 - 8: Compute the model selection criteria for each candidate using either (3) or (4).
 - 9: **end while**
 - 10: Set the candidates with the highest scores as the model for class i
 - 11: **end for**
-

constructed using one of the nine folds as a validation set and the remaining eight folds for the incremental training process (training set). The combination of mixtures that provides the minimum expected error probability on the nine runs were selected and subsequently evaluated using the data of the test fold.

Four type of experiments were conducted in this study. The first two type of experiments are based on non-incremental (k-means) mixture learning with two different mixture selection criteria, i.e. using BIC and DIC model selection. The other two type of experiments use the same mixture selection criteria, but they are based on incremental mixture learning. Most conventional mixture-based classifier design is based on non-incremental version with BIC criteria. Incremental version with BIC criteria is somewhat common. The incremental version with DIC criteria is the technique proposed in this study. The non-incremental version with DIC criteria is included for completeness purpose. The results of this study are summarized into two tables, i.e., Table (1) and Table (2).

Datasets		Cancer	Heart	Iris	New Thyroid	Wine	Zoo
Non-Incremental	BIC	115.5(36.7)	184.3(41.1)	57.5(10.6)	51.0(14.7)	105.0(31.2)	425.3(76.9)
	DIC	97.7(32.4)	190.0(22.9)	58.3(10.4)	58.1(16.7)	106.9(12.73)	435.60(47.9)
Incremental	BIC	110.0(27.9)	173(37.5)	52.0(8.9)	58.60(10.9)	102.60(7.9)	325(50.9)
	DIC	66.0(30.9)	82.0(12.7)	42.0(22.7)	49.8(21.25)	88.5(20.9)	224.0(60.5)

Table 2. The average number of free parameter and the standard deviation (in parentheses) of four type of mixture learning using ten-fold cross-validation

Dataset	Number of Classes	Number of Samples	Number of Features
Cancer	2	683	9
Glass	6	214	9
Iris	3	150	4
New Thyroid	3	215	5
Wine	3	178	13
Zoo	7	101	16

Table 3. Characteristics of the datasets

We can observe from Table (1) that the incremental and non-incremental models with BIC selection do not show significant difference in term of accuracy. Similarly, the incremental and non-incremental models with DIC criterion do not indicate that significant difference in accuracy. However, we do observe that models with DIC criterion have higher accuracy compared to the corresponding models with BIC criterion regardless if the mixtures are obtained incrementally or not. The reason is simply that the BIC criterion for a class is computed from the model and data for that specific class. On the other hand, the DIC criterion for a class is obtained from all data including those that do not belong to the specific class.

From Table (2), we can observe that the incremental or non-incremental with BIC selection use approximately the same number of parameters in order to achieve comparable accuracy. In the case of DIC selection, although the accuracies of the incremental models are comparable to that of the non-incremental models, the incremental models consistently require fewer parameters the non-incremental models. We believe this is due to fact that the incremental version does not face the initialization problems as those faced by the non-incremental models. Thus, coupling incremental version with model selection with discriminative capability is really a good approach for classifier design.

Notice that the mixture structure \mathcal{T}_{il} is selected discriminatively in this study. However, the mixture parameter Λ_{il} are not optimized discriminatively. It is possible to optimize the mixture parameter Λ_{il} further us-

ing MCE framework [3].

5 Conclusion

In this paper, we have proposed a novel Bayes classifier design that considers simultaneous estimation of both class-conditional density and inter-class boundaries. To obtain this classifier, we have used the incremental Gaussian mixture learning to mitigate the lack of true functional form of class-conditional density. The difficulties of accurate estimation the classifier parameter (the inter-class boundaries) during the learning process is solved by using DIC mixture selection. From the experiments with UCI machine learning datasets, the proposed classifiers with incremental mixture learning and discriminative mixture selection are shown not only to have better performance but also to require fewer parameters than the conventional classifiers. Furthermore, this DIC criterion is chosen due to its close relation to the Minimum Classification Error (MCE) classifier design framework. In this future, we plan to investigate the use of the MCE discriminative training framework on the mixture model learned using this framework.

References

- [1] A. Biem. A model selection criterion for classification: Application to hmm topology optimization. In *7th International Conference on Document Analysis and Recognition*, pages 104–108, 2003.
- [2] A. G. Gray and A. W. Moore. ‘n-body’ problems in statistical learning. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *NIPS*, pages 521–527. MIT Press, 2000.
- [3] B.-H. Juang, W. Chou, and C.-H. Lee. Minimum classification error rate methods for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 5(3):257–265, 1997.
- [4] J. Q. Li and A. R. Barron. Mixture density estimation. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *NIPS*, pages 279–285. The MIT Press, 1999.
- [5] C. B. D. Newman and C. Merz. UCI repository of machine learning databases, 1998.
- [6] N. A. Vlassis and A. Likas. A greedy em algorithm for gaussian mixture learning. *Neural Processing Letters*, 15(1):77–87, 2002.