A Model Selection Criterion for Classification: Application to HMM Topology Optimization

Alain Biem
IBM T. J. Watson Research Center
P.O Box 218, Yorktown Heights, NY 10549, USA
biem@us.ibm.com

Abstract

This paper proposes a model selection criterion for classification problems. The criterion focuses on selecting models that are discriminant instead of models based on the Occam's razor principle of parsimony between accurate modeling and complexity. The criterion, dubbed Discriminative Information Criterion (DIC), is applied to the optimization of Hidden Markov Model topology aimed at the recognition of cursively-handwritten digits. The results show that DIC-generated models achieve 18% relative improvement in performance from a baseline system generated by the Bayesian Information Criterion (BIC).

1. Introduction

Model selection, that is, the process of choosing a structure or an order of a model remains a critical part of any signal processing technique. Until now, the Occam's razor principle has been the guiding principle for model selection. The Occam's razor principle proposes the selection of models that best fit the data among competing complexities. It is the principle of parsimony: a model should be simple enough for efficient computation and complex enough to be able to capture data specifics. Also, it is argued that the Occam's razor principle helps generalization: a complex model is likely to over-fit the data while a simpler model can smooth out noisy characteristics of the source distribution.

The Occam's razor principle derives naturally from information theoretic consideration and the Bayesian framework of pattern recognition. The Information Theory viewpoint has proposed such criterion such as the Minimum Description Length (MDL), which minimizes the codelength of a message [1]. The Bayesian framework maximizes the posterior of the model structure, given a body of data. A typical instance of the Bayesian framework is the Bayesian Information Criterion (BIC) [9], which maximizes the likelihood of the data while penalizing large-size models. BIC has been applied to various tasks in model selection including estimation of finite mixture densities and filter order estimation. It has been argued that both the Information Theory viewpoint and the Bayesian viewpoint are equivalent [1]. Both viewpoints has led to model selection criteria that are the sum of two terms: a model-fitting term and a term that measures the complexity of the models. In this paper, we consider the Bayesian framework as the representative of the Occam's razor principle.

Although Occam's razor principle has been quite successful in a wide range of tasks in pattern recognition, experimental evidences show that Occam's razor principle does not necessarily select the best performing models when used within a classification task [6]. In particular, classical Bayesian-based model selection criteria focus on estimating models using within-class statistics, without regard to competing classification categories. For classification problems, Bayesian-based model selection may not be appropriate.

This paper introduces a model selection criterion aimed at classification problems. The Occam's razor principle is replaced by the discriminative principle. The goal is not to select the simplest model that best explains the data, but to select the model that is the less likely to have generated data belonging to competing classification categories. The proposed model selection criterion is discriminative in the sense that the model is selected in regard to the classification task by making use of data that belong to competing classes, thus introducing knowledge of the classification task in the model selection process. Since the criterion focuses on determining discriminant models, the performance of the resulting models is increased when compared to a system based on Occam's razor principle.

The proposed criterion is tested within a complex modeling architecture represented by a Hidden Markov Model (HMM) and is compared to BIC. The goal is to select both the number of states and the number of mixtures of Gaussians in a Continuous Hidden Markov Model-based online

0-7695-1960-1/03 $17.00 © 2003 **IEEE**

handwriting recognition system.

We first review classical Bayesian-based model selection techniques, and then introduce the derivation of the proposed discriminative model selection technique. Finally, we describe an application of the proposed criterion to optimizing the topology of an HMM aimed at cursively-written online handwritten digit recognition.

2. The Model Selection Problem in the Classification Context

We are given a set of M categories or classes $\{C_i : i = 1, ..., M\}$. The classification task is to assign an incoming pattern x to one the classes. Decoding is based on the Bayes classification theory, which chooses the class C_i that yields the maximum value of $P(C_i \mid x)$, the posterior probability of the class, given the data. By using the Bayes rule, the Bayes decoding process is equivalent to maximizing the product of the class-conditional probability $P(x \mid C_i)$ and the prior probability of the class $P(C_i)$. That is,

$$x \text{ belongs to } \hat{C} \quad \text{if} \quad \hat{C} = \arg\max_{C_i} P(x \mid C_i)P(C_i) \quad (1)$$

where x is the input data to the system, typically a sequence of training vectors. The Bayes classification scheme is guaranteed to minimize the probability of error [3].

For each class C_i, there is a set of L_i candidate models, $\{M_{il} : l = 1, ..., L_i\}$; each model M_{il}, viewed as the union of the model structure (or topology) $\mathcal{T}_{il}$ and the parameter of the model θ_{il}, implements the class-conditional probability $P(x \mid C_i)$ as $P(x \mid \mathcal{T}_{il}, \theta_{il})$. Also, for each class C_i, we assume the availability of a representative data set X_i.

The model selection problem consists of selecting a single topology $\mathcal{T}_{il}$ as sole representative of the class C_i. This is done by devising a selection criterion $\mathcal{C}(\cdot)$ such that for each class C_i,

$$\text{choose } \mathcal{T}_{il} \quad \text{if} \quad \mathcal{T}_{il} = \arg\max_{\mathcal{T}_{ik}} \mathcal{C}(\mathcal{T}_{ik}). \quad (2)$$

Various model selection paradigms are devised by a judicious choice of the criterion function $\mathcal{C}(\cdot)$.

2.1. Bayesian Model Selection

The Bayesian model selection criterion chooses the topology that yields the highest value of the posterior probability of the topology, given the data, $P(\mathcal{T}_{il} \mid X_i)$, which by taking the log, and making use of the Bayes theorem, is equivalent to setting

$$\mathcal{C}(\mathcal{T}_{il}) = \log(P(X_i \mid \mathcal{T}_{il})) + \log(P(\mathcal{T}_{il})). \quad (3)$$

The Bayesian selection method is based on the joint probability $P(\mathcal{T}_{il})P(X_i \mid \mathcal{T}_{il})$, composed of the prior of the model structure $P(\mathcal{T}_{il})$, which describes our preference for a particular topology, and the term $P(X_i \mid \mathcal{T}_{il})$, which is the probability of the topology $\mathcal{T}_{il}$ generating the given data set X_i; this term is sometimes referred to as the evidence [7] or the integrated likelihood [4]. It has been argued that the dual effect of the prior and the evidence implements the Occam's razor principle [7, 9].

A common practice in Bayesian model selection is to ignore the prior over the structure $P(\mathcal{T}_{il})$ (which means assuming equal prior across all the topologies of the class) and using the evidence $P(X_i \mid \mathcal{T}_{il})$ as the criterion for model selection. The evidence is computed by integrating over the entire parameter set as follows:

$$p(X_i \mid \mathcal{T}_{il}) \quad = \quad \int p(X_i \mid \mathcal{T}_{il}, \theta_{il}) p(\theta_{il} \mid \mathcal{T}_{il}) d\theta_{il}. \quad (4)$$

2.2. Bayes Factor

Let us consider two competing topologies $\mathcal{T}_{il}$ and $\mathcal{T}_{ik}$ of class C_i. Using the Bayes theorem, the ratio of their posterior probability can be broken down as

$$\frac{P(\mathcal{T}_{il} \mid X_i)}{P(\mathcal{T}_{ik} \mid X_i)} = \frac{P(X_i \mid \mathcal{T}_{il})}{P(X_i \mid \mathcal{T}_{ik})} \frac{P(\mathcal{T}_{il})}{P(\mathcal{T}_{ik})}. \quad (5)$$

The quantity

$$B^i_{lk} = \frac{P(X_i \mid \mathcal{T}_{il})}{P(X_i \mid \mathcal{T}_{ik})} \quad (6)$$

is called the Bayes factor [4] for topology $\mathcal{T}_{il}$ against topology $\mathcal{T}_{ik}$ within class C_i and is used in statistics to select one topology over the other when all topologies are all equally likely a priori.

2.3. Bayes Factor Criterion

Bayes factors have been widely used in hypothesis testing, where the goal is to test the validity of the null hypothesis against the alternative. In this section, we embed the Bayes factors into a model selection criterion that can be used straightforwardly in pattern recognition applications.

Let

$$\mathcal{B}^i_l = \left\{ \prod_{k=1, k\neq l}^{L_i - 1} B^i_{lk} \right\}^{\frac{1}{L_i - 1}} \quad (7)$$

be the geometric mean of the Bayes factors of the topology $\mathcal{T}_{il}$ against competing topologies of class C_i. We define the *Bayes Factor Criterion* of model or topology $\mathcal{T}_{il}$ as the logarithm of $\mathcal{B}^i_l$:

$$\log \mathcal{B}^i_l = \log P(X_i \mid \mathcal{T}_{il}) - \frac{\sum_{k=1, k\neq l}^{L_i - 1} \log P(X_i \mid \mathcal{T}_{ik})}{L_i - 1} \quad (8)$$

The Bayes Factor Criterion of model or topology $\mathcal{T}_{il}$ is the difference of two terms: the first term is the evidence of the model, given the data X_i. The second term is the average of the evidence for competing models to $\mathcal{T}_{il}$. For each class C_i, it is straightforward that:

$$\hat{l} = \arg\max_{l} P(X_i \mid \mathcal{T}_{il}) \quad \Longleftrightarrow \quad \hat{l} = \arg\max_{l} \mathcal{B}^i_l. \quad (9)$$

The model that yields the highest evidence also yields the highest value of the Bayes Factor Criterion and vice-versa. Hence, the Bayes Factor Criterion directly implements the Bayesian principle of model selection.

2.4. Bayesian Information Criterion

A key problem in Bayesian-based model selection is the computation of the evidence $P(X_i \mid \mathcal{T}_{il})$. The integral in Eq. (4) is often intractable when complex architectures are involved and must be evaluated by means of numerical methods, such as Markov Chain Monte-Carlo [5] or through an approximation such as the Laplacian approximation.

The Laplace method of integral approximation [9, 8], which has been extensively used in computing the integral in Eq. (4), assume that the function

$$g(\theta) = p(X \mid \mathcal{T}, \theta) p(\theta \mid \mathcal{T})$$

is strongly peaked at the most probable parameter-set θ_{MP}. A Taylor expansion around the optimum of the logarithm of this function leads to a tractable form of the evidence:

$$\begin{aligned} p(X \mid \mathcal{T}) \approx\ & p(X \mid \mathcal{T}, \theta_{MP}) \\ & p(\theta_{MP} \mid \mathcal{T})(2\pi)^{\frac{K}{2}} \det(A)^{-\frac{1}{2}} \end{aligned} \quad (10)$$

where K is the number of free parameters in the model and $A = -\nabla^2 \log P(\theta \mid X, \mathcal{T})|_{\theta=\theta_{MP}}$. This approximation introduces a relative error of order $O(N^{-1})$, where N is the size of the data set. As N increases, $\det(A)$ tends to $N^k \det(I)$, where I is the Fisher information matrix for a single observation. θ_{MP} is approximated by the maximum likelihood (ML) estimate θ_{ML}, as the function $g(\theta)$ is dominated by the likelihood term $p(X \mid \mathcal{T}, \theta)$. Using these conditions and taking the logarithm of Eq. (10) lead to the following approximation of the evidence:

$$\begin{aligned} \log p(X \mid \mathcal{T}) \approx\ & \log p(X \mid \mathcal{T}, \theta_{ML}) + \log p(\theta_{ML} \mid \mathcal{T}) \\ & + \frac{k}{2}\log(2\pi) - \frac{k}{2}\log N - \frac{1}{2}\log(\det(I)). \end{aligned} \quad (11)$$

The above approximation, which introduces an $O(N^{-\frac{1}{2}})$ relative error, can be straightforwardly computed over the original integral form of Eq. (4) and is commonly the first step in the derivation of most Bayesian selection criteria.

Asymptotically, the prior of the parameters $p(\theta_{ML} \mid \mathcal{T})$ tends to multivariate Gaussian densities with means θ_{ML} and covariance I^{-1}. These conditions lead to the BIC, defined as

$$\begin{aligned} \log P(X_i|\mathcal{T}_{il}) &\approx BIC(\mathcal{T}_{il}) \quad (12) \\ &= \log p(X_i \mid \mathcal{T}_{il}, \hat{\theta}_{il})) - \frac{K_{il}}{2}\log N_i \end{aligned}$$

where $\hat{\theta}_{il}$ is the Maximum Likelihood estimate of the parameter of the model M_{il}, K_{il} is the number of free parameters in the model, and N_i is the size of the data set X_i. BIC is the sum of the likelihood and the term $\frac{K_{il}}{2}\log N_i$, which can be viewed as a penalty on the number of free parameters in the model. To account for the fact that probability estimates are not accurate, a regularizing parameter is introduced, leading to the following form of the BIC:

$$BIC(\mathcal{T}_{il}) = \log p(X_i \mid \mathcal{T}_{il}, \hat{\theta}_{il})) - \alpha\frac{K_{il}}{2}\log N_i \quad (13)$$

where $\alpha > 0$ is the regularizing term.

3. Discriminative Model Selection

The derivation of the discriminative model selection follows a path similar the derivation of the Bayes Factor Criterion described in section 2.2 and section 2.3. Unlike, BIC, the proposed discriminative model selection criterion takes into account the goal of the models, which is the classification task. This task-oriented model selection criterion is more adapted to the classification problem.

3.1. Discriminant Factors

We define the discriminant factor of model $\mathcal{T}_{il}$ against class C_j as

$$D^i_{lj} = \frac{P(X_i \mid \mathcal{T}_{il})}{P(X_j \mid \mathcal{T}_{il})}. \quad (14)$$

The quantity $P(X_j \mid \mathcal{T}_{il})$ is an evidence-like term computed by making use of data set X_j belonging to the competing class C_j. We refer to this term as the anti-evidence of model $\mathcal{T}_{il}$ against C_j. The anti-evidence measures the capacity of the corresponding model to generate data belonging to competing classes. The ratio of the evidence and the anti-evidence is thus a measure of the model capacity to discriminate data from the two competing classes. Unlike the Bayes factor, which compares within-class models, the discriminant factor compares a model against a competing class by making use of the data set X_j generated by models of the competing class C_j.

Now, let

$$\mathcal{D}^i_l = \left\{ \prod_{j=1, j\neq i}^{M-1} D^i_{lj} \right\}^{\frac{1}{M-1}} \quad (15)$$

be the geometric mean of the discriminant factors of model $\mathcal{T}_{il}$. The Discriminant Factor Criterion (DFC) is simply the logarithm of $\mathcal{D}_l^i$. That is,

$$DFC(\mathcal{T}_{il}) = \log P(X_i \mid \mathcal{T}_{il}) - \frac{\sum_{j=1,j\neq i}^{M-1} \log P(X_j \mid \mathcal{T}_{il})}{M-1}. \tag{16}$$

Unlike the Bayes Factor criterion, the Discriminant Factor Criterion is the difference between the evidence of the model, given the corresponding data set, and the average over anti-evidences of the model. By choosing the model which maximizes the evidence, and minimize the anti-evidences, the result is the best generative model for the correct class and the worst generative model for the competitive classes; this scheme thus selects the most discriminant models, resulting in an improved accuracy in regard to the classification task.

3.2 Discriminative Information Criterion

Application of the DFC as represented in Eq. (16) requires the estimation of the evidence and the anti-evidence terms. We define the Discriminative Information Criterion (DIC) as an approximation of the DFC, where evidences and anti-evidences are replaced by their BIC approximation as defined in Eq. (12). That is,

$$DIC(\mathcal{T}_{il}) = \log P(X_i \mid \mathcal{T}_{il}, \hat{\theta}_{il}) - \frac{\sum_{j=1,j\neq i}^{M-1} \log P(X_j|\mathcal{T}_{il},\hat{\theta}_{il})}{M-1} + \frac{K_{il}}{2(M-1)} \sum_{j=1,j\neq i}^{M-1} \log \frac{N_j}{N_i} \tag{17}$$

The DIC is the sum of two terms. The first term is a difference between the likelihood of the data, $\log P(X_i \mid \mathcal{T}_{il}, \hat{\theta}_{il})$, and the average of anti-likelihood terms, $\log P(X_j \mid \mathcal{T}_{il}, \hat{\theta}_{il})$, where the anti-likelihood of the data X_j against model M_{il} is a likelihood-like quantity in which the data and the model belong to competing categories. The second term $\frac{K_{il}}{2(M-1)} \sum_{j=1,j\neq i}^{M-1} \log \frac{N_j}{N_i}$ is zero when all data sets are of the same size. Considering that the first term contributes the most to the discriminative capabilities, we used in this paper the following approximated version of the criterion:

$$\begin{aligned} DIC(\mathcal{T}_{il}) &= \log P(X_i \mid \mathcal{T}_{il}, \hat{\theta}_{il}) \qquad (18) \\ &\quad - \frac{\alpha}{M-1} \sum_{j=1,j\neq i}^{M-1} \log P(X_j \mid \mathcal{T}_{il}, \hat{\theta}_{il}). \end{aligned}$$

Again, the parameter $\alpha > 0$ acts as a regularizer, a necessary term to compensate for the non-optimality of likelihood and anti-likelihood estimates.

4. HMM Topology Selection

In recent days, Hidden Markov modeling has enjoyed a widespread use in both on-line and off-line handwriting recognition. This success of HMM is explained by its high flexibility in modeling variable-length sequence and the existence of easy-to-use learning algorithms, such as the Expectation-Maximization (EM) -algorithm, that enables the re-estimation of model parameters, given an a priori chosen HMM topology. The HMM topology, defined in this context as the number of states, the number of mixtures per state and the transition between states, directly influences the modeling capacity of the model. Its choice is crucial for achieving a high-performing system. The HMM topology, however, is usually chosen heuristically, leading to a sub-optimality of the model, where the optimal model is the one that yields the smallest error-rate. Previous work on HMM topology optimization has been tackled through various approaches, ranging from state-splitting techniques [10] or state-reducing techniques [11], using the likelihood as the optimization criterion. As has been argued, the likelihood criterion is a poor choice for model selection as the likelihood increases incrementally with the number of parameters. Although Bayesian techniques have been applied successfully to HMM topology optimization [6, 2], there is no guarantee that the resulting model are optimal in regard to error minimization.

The HMM is characterized by its topology $\mathcal{T}$ and its parameters θ, given this topology. $\mathcal{T}$ is defined by the number of states in the model, Q, the number of Gaussian mixtures per states, L, and the connecting architecture between the states. As the connectivity between states is preset, the topology of the HMM is uniquely characterized by the number of states and the number of mixtures per state. In this paper, we assume a left-to-right topology with a two-state transition where all states have the same number of mixtures. A model $M = \{\mathcal{T}, \theta\}$ is viewed as the union of the model structure or topology $\mathcal{T} = \{Q, L\}$ and the parameter of the model $\theta = \{A, \mu, \Sigma, \omega\}$, where A is the transition matrix probability, μ, Σ, ω are the set of means, the set of covariance matrices and the set mixing weights, respectively.

4.1 Model Selection Procedure

We assume the availability of three data sets: a training data set, used for training the HMMs, a held-out set used for computing the selection criteria and a test set. The steps for the model selection procedure are as follows. For each model selection criterion, first, train various HMMs configurations (obtained by varying the number of states and the number of mixtures per state), using the ML criterion on the training data set. Second, use the held-out set to compute the values of the selection criterion for all configurations and select the configuration that yields the highest value of the model selection criterion.

4.2 Tasks and Database

The task is the recognition of freely-written digits. A training set of 9796 tokens, written by 100+ writers was used for training various HMM configurations using the EM algorithm. The held-out set comprises 4554 tokens written by 20+ writers. The test set was a subset of the Unipen database, comprising 2603 tokens.

At the front-end of the recognizer, the input handwriting signal is segmented, normalized, re-sampled and then feature extraction is performed from a moving window along the time-series. A nine-dimensional vector is created for each frame using PCA (Principal Component Analysis).

4.3. Experimental Results

Table 1 illustrates the best recognition accuracy for DIC and BIC across a wide range of values of the parameter α. Clearly, DIC outperforms BIC in terms of performance: more than 18% relative decrease in error rate. As expected, the DIC system exhibits a higher number of parameters than BIC.

Table 1. Recognition rate of BIC and DIC in the digit recognition task

Criterion	Rec. Rate	#states	#parameters
BIC	93.97	87	1692
DIC	95.08	101	4299

Figure 1 shows the recognition rate versus the parameter α. Clearly, DIC exhibits higher performance than BIC across a wide range values of the regularizing parameter. This result shows that DIC, as expected, is more oriented to the goal of classification that BIC is.

All experimental results indicate that DIC is better than BIC for achieving a high-performing system. This is done, however, at the cost of larger system, confirming that the Occam's razor principle does not necessarily lead to the best system for classification problems.

5. Conclusion

We have proposed the Discriminative Information Criterion (DIC) as a criterion suited for classification task. The criterion was applied to optimizing the topology of a Continuous Hidden Markov Model. Application to the classification of cursively-written digits shows that the DIC-generated system realizes more than 18% relative error rate reduction when compared to classical BIC criterion.

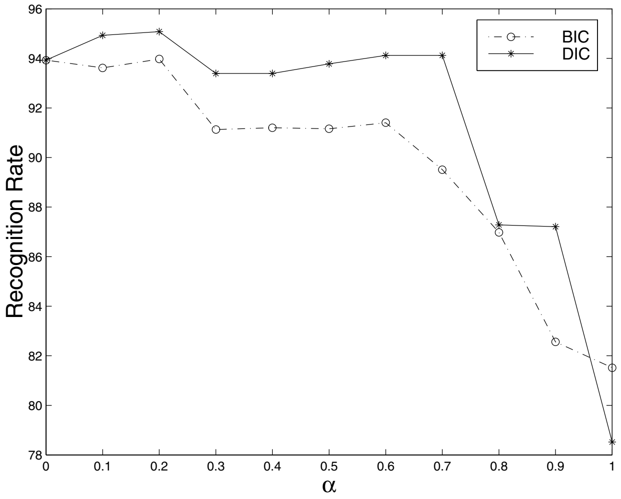

Figure 1. Recognition rate for DIC and BIC as function of the parameter α.

References

[1] A. Barron, J. Rissanen, and B. Yu. The Minimum Description Length Principle in Coding and Modeling. *IEEE trans. Inform. Theory*, 44(6):2743–2760, Oct. 1998.

[2] A. Biem, J.-Y. Ha, and J. Subrahmonia. Bayesian Model Selection Criterion for HMM Topology Optimization. In *Proc. of ICASSP*, volume 1, pages 989 –992, 2002.

[3] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley Interscience Publications, 1973.

[4] R. E. Kass and A. E. Raftery. Bayes factors. *J. Am. Statitical Association*, 90:773 –795, 1994.

[5] R. E. Kass and A. E. Raftery. Bayes factors. Technical Report 254, University of Washington, Department of Statistics, 1994.

[6] D. Li, A. Biem, and J. Subrahmonia. HMM Topology Optimization for Handwriting Recognition. In *Proc. of ICASSP*, volume 3, pages 1521 –1524, 2001.

[7] D. MacKay. Bayesian Interpolation. *Neural Computation*, 4(3):415–447, 1992.

[8] J. Olivier and R. Baxter. MML and Bayesianism: similarities and differences (Introduction to Minimum Encoding Inference - part ii. Technical Report 206, Monash University, Australia, 1994.

[9] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[10] H. Singer and M. Ostendorf. Maximum Likelihood Successive State Splitting. In *ICASSP*, pages 601–604, 1996.

[11] A. Stolcke and S. Omohundro. Hidden Markov Model induction by Bayesian Model Merging. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in NIPS*, volume 5, pages 11–18. Morgan Kaufmann, San Mateo, CA, 1993.