HMM TOPOLOGY OPTIMIZATION FOR HANDWRITING RECOGNITION

Danfeng Li, Alain Biem[1] *and Jayashree Subrahmonia*[1]

University of Illinois at Urbana-Champaign
Dept. of Electrical and Computer Engineering, 405 N. Mathews Ave., Urbana, IL 61801, USA
dli1@ews.uiuc.edu

[1]IBM T. J. Watson Research Center, P. O. Box 218, Yorktown Heights, NY 10598, USA
{biem,jays}@us.ibm.com

ABSTRACT

This paper addresses the problem of Hidden Markov Model (HMM) topology estimation in the context of on-line handwriting recognition. HMMs have been widely used in applications related to speech and handwriting recognition with great success. One major drawback with these approaches, however, is that the techniques that they use for estimating the topology of the models (number of states, connectivity between the states and the number of Gaussians), are usually heuristically-derived, without optimal certainty. This paper addresses this problem, by comparing a couple of commonly-used heuristically-derived methods to an approach that uses Bayesian Information Criterion (BIC) for computing the optimal topology. Experimental results on discretely-written letters show that using BIC gives comparable results to using heuristic approaches with a model that has nearly 10% fewer parameters.

1. INTRODUCTION

This paper addresses the problem of HMM topology selection in the context of on-line handwriting recognition. With the fast growth of PDAs (personal digital assistants) and hand-held PCs, as convenient alternatives to standard computers with bulky keyboards, the problem of accurate on-line handwriting recognition is gaining a lot of interest. Hidden Markov models, which have been successfully used in speech recognition, are currently being used to model on-line handwriting as well with great success [1, 2]. Hidden Markov models are a type of Markov modeling where the sequencing of states modeling the formation of a letter is abstracted from observations.

In on-line handwriting, each letter shape is typically modeled by one or more left-to-right HMMs, as shown in Figure 2. Solid lines indicate transitions between states when a feature vector is observed. Typically, a mixture of Gaussians is used to model the distribution of feature vectors at the state-level. Estimating the topology of the HMM, defined by the number of states, the connectivity between the states and the number of Gaussians used to model the distribution of feature vectors in each state, will be the focus of this paper.

There are two commonly-used heuristic approaches for selecting the topology of a HMM in on-line handwriting. The first assumes the same topology for all the HMMs. This however cannot be the best topology for hand-written letters, where there is large variation between letters. For example, a HMM for a '.' will not be appropriate for a 'W' because a 'W' will typically be longer than a '.' and hence should be modeled by a longer state sequence. Also, the number of Gaussians used to model distribution of feature-vectors might vary depending on the complexity of the shapes. Hence, in common practice, HMM-based handwriting recognition systems use non-uniform topology across letters, with each letter being modeled by a HMM with varying number of states and Gaussians. A commonly used heuristic approach for determining the number of states for a HMM is to make it the average number of feature-vectors for the letter modeled by the HMM, or as the mode of the feature-vector-count histogram for the letter. This gives better performance compared to systems with the same number of states for all HMMs. Other parameters of the topology are typically estimated as follows: the number of Gaussians is estimated as the minimum number required to efficiently model the training data, while also generalizing to unknown data sets, and the connectivity between the states is picked so as to allow for all possible variations of forming the character.

Various approaches to optimal model selection have been proposed by statisticians and information theorists [3, 4]. A common approach uses the Bayesian Information Criterion (BIC) for selecting the model order. This approach uses the sum of the likelihood of the data and a penalty term on the number of parameters of the model as the optimization criterion for computing the "best" model. Such a criterion is desirable, as it obeys the Occam's razor principle which states that we should select the simplest model that best fits the data, among competing complexities. It is the principle of parsimony: A model should be simple enough to allow for less computation and complex enough to be able to capture data specifics. The BIC's penalty term also depends on the number of data points, which makes it attractive for model selection as it may prevent model over-fitting [4].

This paper investigates the use of the Bayesian Information Criterion for estimating the number of HMM states, the connectivity between states and the number of Gaussian mixtures per state. Experimental results on a database of discretely-written letters by various writers show that the BIC criterion can be used to compute simpler models that perform as well as the ones developed using heuristic approaches.

The rest of the paper is organized as follows. Section 2 gives an overview of the overall HMM-based on-line handwriting recognition system. Section 3 gives details of two topology estimation schemes compared in this paper: the histogram-based method and the BIC-based selection. Section 4 is dedicated to experimental

investigations. Finally, Section 5 provides discussion and conclusions.

2. SYSTEM DESCRIPTION

2.1. Pre-processing

The recognizer uses a electronic tablet or a digitizer, which captures the pen-tip position (x_t,y_t) as a function of time during writing. This is in contrast to off-line handwriting recognition which treats the input ink as an image and applies typical image-processing techniques. The (x_t,y_t) coordinates collected by the digitizer (the "digital ink") are passed through the preprocessor, where denoising, segmentation, normalization, resampling and feature extraction are performed [5].

To increase robustness to writer style variability, each letter is modeled using a set of lexeme models. Lexemes are letter allographs which are derived manually or through data clustering techniques. For illustration, examples of lexemes are shown in Figure 1. Typically, each lexeme is modeled by an HMM.

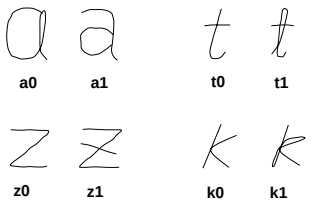

Fig. 1. Different writing styles modeled by different lexemes.

2.2. HMM topology

Figure 2 shows a fully connected left-to-right HMM.

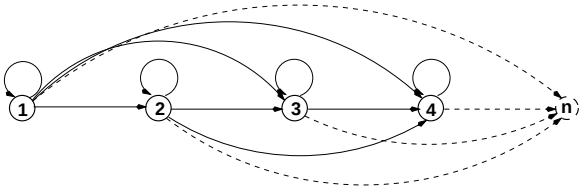

Fig. 2. A 4-state lexeme model.

The model is a fully continuous left-to-right HMM with two kinds of states: a set of emitting states (solid circle) and a non-emitting state (dashed circle). The emitting states are associated with observations and are modeled by a mixture of Gaussians, while the non-emitting state is not associated with any observation. The transitions between emitting states are denoted by solid lines and the transition from an emitting state to the non-emitting state is denoted by dashed lines.

The sequence of observations must end at the non-emitting state when the last feature-vector has been processed. This is particularly useful to discriminate between two letters, when one letter is a subset of the other. The non-emitting state acts like a penalty term on the larger model, hence enhancing discrimination.

Another characteristic of this model is state skipping, where forward jumping from one emitting state to any other emitting state is allowed. This is a useful feature in handwriting as it helps to deal with missing strokes.

2.3. Training

We used the standard Baum-Welch [6] algorithm to estimate the parameters of the models. However, modifications had to be made to allow for non-emitting states. This was done by a slight modification of the transition probability matrix and distribution probability matrix as follows.

Consider a standard transition matrix for a HMM with n states

$$A_o = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ & a_{22} & \cdots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{bmatrix}, \quad (1)$$

where a_{ij} is the transition probability from state i to state j. By adding a non-emitting state at the last state, the transition matrix becomes

$$A_n = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & a_{1(n+1)} \\ & a_{22} & \cdots & a_{2n} & a_{2(n+1)} \\ & & \ddots & \vdots & \vdots \\ & & & a_{nn} & a_{n(n+1)} \\ & & & & 1 \end{bmatrix}, \quad (2)$$

where $a_{i(n+1)}$ is the transition probability from the emitting state i to the non-emitting state $n+1$.

Similarly, the standard observation matrix is usually of the form

$$B_o = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{T1} & b_{T2} & \cdots & b_{Tn} \end{bmatrix}, \quad (3)$$

where b_{ti} is the probability of observing feature t when in state i. By taking into account the non-emitting state, the new observation matrix is

$$B_n = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} & 0 \\ b_{21} & b_{22} & \cdots & b_{2n} & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ b_{T1} & b_{T2} & \cdots & b_{Tn} & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \quad (4)$$

This is equivalent to assigning a virtual feature to state $n+1$, at the end of the observation sequence and disallowing any other feature at that state. Baum-Welch training is then applied as usual using the transformed matrices.

3. TOPOLOGY OPTIMIZATION

3.1. Heuristic topology estimation

There are a number of heuristic approaches for computing the number of states for each model :

1. Compute the number of states for a lexeme model as the minimum number of feature-vectors observed for the lexeme.
2. Compute the number of states as the average number feature-vectors for the lexeme.
3. Compute the number of states as the mode of the feature-vector-count histogram. This approach is the most commonly used approach and is the heuristic approach used for comparison in this paper [2].

Other parameters of the topology are typically estimated as follows: the number of Gaussians is estimated as the minimum number required to model the training data well, while also generalizing to other data sets. The connectivity between the states is picked to allow for all possible variations of forming the letter. These heuristic approaches provide a simple method for estimating the topology of the model, but are not optimal.

3.2. Bayesian Information Criterion

Let D denote a data set $(X_1, \ldots, X_N)$ which is specified by a vector of d unknown parameters $\theta = (\theta_1, \ldots, \theta_d)$. Before observing the data, our belief in θ is described by the prior probability $p(\theta)$.

Likelihood of the data can be written as

$$p(D) = \int p(D \mid \theta) p(\theta) d\theta. \tag{5}$$

Assume that data D contains n independent and identically distributed observations. Then, using a Taylor series approximation and assuming that the prior contains the same information as an average observation, it can be shown [4, 7] that

$$\log p(D) = \log p(D \mid \hat{\theta}) - \frac{d}{2} \log n + O(n^{-\frac{1}{2}}) \tag{6}$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ.

Hence, model parameter election from a pool of models using the BIC criterion can be done as follows:

$$\theta^* = \max_{\hat{\theta}_i} \{ \sum_{n=1}^{N} \log P(X_n | \hat{\theta}_i) - \underbrace{\frac{d_i}{2} \log N}_{penalty} \}, \quad i = 1, \ldots, I \tag{7}$$

where X_n is the n-th observation, $\hat{\theta}_i$ the maximum likelihood estimate of the parameter set of the i-th model, d_i is the number of parameters in the i-th model, N is the number of observations and I is the total number of models. BIC has been proposed as a model selection criterion and has been widely and successfully applied to filter order estimation or clustering [8]. It uses a combination of two terms: a maximum likelihood term and a penalty term that depends on the number of parameters of the model and the number of data. The penalty term acts as measure of complexity that penalizes large-size models. The likelihood of the model tends to increase as the model size increases, whereas the penalty term increases with the size of the model. The sum of the two helps pick the optimal model from a set of available models.

Figure 3 shows the log likelihood of the data, the penalty term and the sum of the two as a function of the number of states for models corresponding to the letters 'E', 'F', 'G', 'H'. All the models were trained using 5 Gaussian mixtures per state. As seen from the plots, the maxima of the curves can easily be identified.

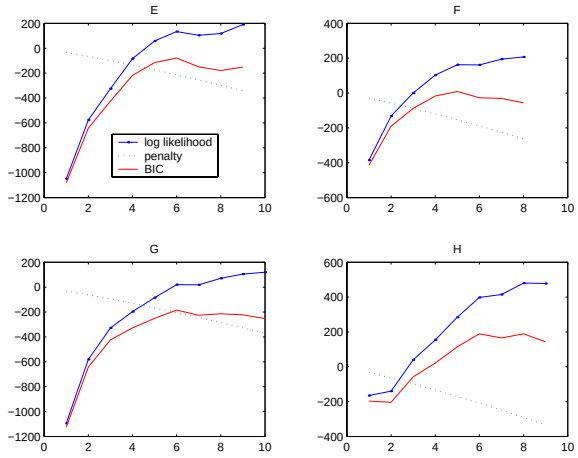

Fig. 3. Log likelihood, penalty and BIC vs. number of states (fixed 5 Gaussian mixtures per state).

4. EXPERIMENT RESULTS AND DISCUSSION

The database used in this paper, collected by the IBM Pen Technologies Group, consists of isolated letters, written by 174 writers. The data set contains 106,395 tokens. We used 96,563 tokens for training (corresponding to 157 writers) and 9,832 tokens for testing, (corresponding to 17 writers). The set of 92 characters were manually pre-clustered into 366 lexemes according to similarity between writing styles. The set of characters includes English letters, digits, and keyboard symbols. The HMM structure proposed in Section 2 was used to build the lexeme models. Model training was done at the lexeme level and the recognition rates were computed at the character level; a recognition result was considered to be in error only if it recognized a lexeme as being a lexeme of a different character.

Two sets of experiments were performed. In the first, the number of states was varied while the number of Gaussians was fixed at 5. Table 1 shows a summary of the comparative results of the histogram-based heuristic approach for computing the number of states with the approach using BIC.

Table 1. Comparison between histogram-based and BIC methods

Method	Histogram	BIC
Number of test tokens	9832	
Number of lexemes	366	
Number of Gaussians	5	
Accuracy	89.6%	90.0%
Total number of states	2304	2111

The results indicate that the histogram-based heuristic method and the BIC method give similar recognition accuracy. However, using BIC picks simpler models that give comparable accuracies to the histogram-based approach. The models have 10% fewer parameters.

Table 2 shows some examples where using the BIC approach gives significant improvement over the histogram-based approach. Lex denotes the lexeme index, $Token$ denotes the total number of testing tokens, Hc denotes the number of correctly recognized

tokens using the histogram-based method, Hn denotes the number of states selected by histogram-based method. Similarly, Bc and Bn denote the number of correctly recognized tokens and the selected number of states, using BIC. In all the cases, using BIC gives comparable accuracy with a much simpler model (as shown for the lexeme 'F6' in the table) or improved accuracy with a very small increase in model complexity.

In the above experiment, the number of Gaussian mixtures was fixed while the number of states was varied and the optimal number of states computed. The next experiment is an attempt to understand whether it is more efficient to increase the number of states or the number of Gaussians when looking for more complex models. In order to understand this, the following experiment was performed. The sum of the data likelihood and the penalty term using BIC for a single lexeme model was plotted as a function of the number of states of the HMM and the number of gaussians per state as shown in Figure 4.

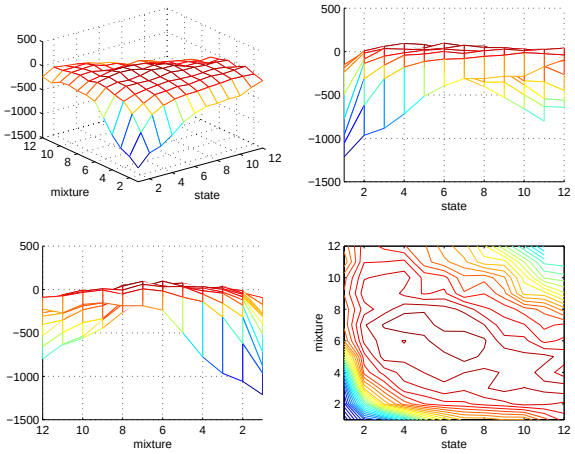

Fig. 4. sum of the data likelihood and the penalty .vs. number of states and mixtures

The plots as seen above are relatively flat on the top. There is no single maximum value that can be chosen through tuning of the number of states and the number of Gaussians. One can interpret this result to mean that increasing the number of states is equivalent to increasing the number of Gaussians when using BIC. This is intuitive because the penalty term of BIC only depends on the number of parameters of the model and does not distinguish between the types of parameters. Hence, it may be desirable to have a penalty term that evaluates the contributions of parameters more specifically to be able to better analyze the cost of increasing the number of parameters.

5. CONCLUSION

In this paper, we have investigated the use of the Bayesian Information Criterion (BIC) for estimating the topology of a HMM. Results using the BIC were compared to the histogram-based approach. Results show that using BIC selects models that give similar performance as those selected using heuristically-derived approaches, but having 10% fewer parameters. Although we have studied this technique in the context of handwriting recognition, the approach is general and can be applied to other HMM-based tasks.

Table 2. Examples of the improvement from histogram to BIC. Lexeme are labeled as a character followed by an index (see text for legend).

Lex	Token	Hc	Hn	Bc	Bn
#1	6	6	8	6	6
H4	21	21	13	21	9
F6	12	12	13	12	8
+0	62	30	4	40	5
.0	60	24	1	32	3
O0	41	18	5	25	6

6. ACKNOWLEDGEMENTS

The authors would like to thank the following members of the Pen Technologies group of IBM Watson Research Center for their assistance in this work: Ha Jin Young, Thomas Kwok, Michael Perronne, John Pitrelli and Gene Ratzlaff.

7. REFERENCES

[1] S. R Veltman and R. Prasad, "Hidden markov model applied to on-line handwritten isolated character recogntion," *IEEE Transactions on Image Processing*, vol. 3, no. 3, pp. 314–318, May 1994.

[2] K. S. Nathan, H. S. M. Beigi, J. Subramonia, G.J Clary, and H. Maruyama, "Real-time on-line unconstrained handwriting recognition using statistical methods," *ICASSP*, vol. 4, pp. 2619–2623, 1995.

[3] H. Akaike, "A new look at the statistical model indentification.," *IEEE Trans. AC*, vol. 19, pp. 716–723, 1995.

[4] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[5] C. C. Tappert, C. Y. Suen, and T. Wakahara, "The state of the art in on-line handwriting recognition," *IEEE trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 8, pp. 787–808, Aug. 1990.

[6] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey, 1993.

[7] A. E. Raftery, "Bayesian model selection in social research," *Sociological Methodology*, 1994.

[8] S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian Information Criterion with applications in speech recognition," *ICASSP*, vol. 2, pp. 645–649, 1998.