DECISION TREE STATE TYING BASED ON PENALIZED BAYESAIN INFORMATION CRITERION

Wu Chou and Wolfgang Reichl

Bell Laboratories, Lucent Technologies,
600 Mountain Ave.,
Murray Hill, NJ 07974, USA

ABSTRACT

In this paper, an approach of penalized Bayesian information criterion (pBIC) for decision tree state tying is described. The pBIC is applied to two important applications. First, it is used as a decision tree growing criterion in place of the conventional approach of using a heuristic constant threshold. It is found that original BIC penalty is too low and will not lead to compact decision tree state tying model. Based on Wolfe's modification to the asymptotic null distribution, it is derived that two times BIC penalty should be used for decision tree state tying based on pBIC. Secondly, pBIC is studied as a model compression criterion for decision tree state tying based acoustic modeling. Experimental results on a large vocabulary (Wall Street Journal) speech recognition task indicate that compact decision tree could be achieved with almost no loss of the speech recognition performance.

1. INTRODUCTION

Recently, decision tree based statistical approach has become increasingly popular in speech recognition. As a powerful statistical framework, decision tree is extremely suitable for applications involving clustering and classification. It has a wide range of applications in acoustic modeling, speaker adaptation, pronunciation modeling and etc. In acoustic modeling, the classification and predication power of decision tree makes it possible to synthesis model units or contexts which do not occur in the training data. Moreover, the splitting procedure in decision tree is a model selection process. It provides a way to balance the model complexity and the limited amount of training data upon which the parameters of the model will be estimated.

In general, special care must be taken to avoid an over grown decision tree that cannot be robustly supported by the data. The problem of finding model with correct order and complexity is treated in statistics as the problem of model identification. Although maximum likelihood estimation (MLE) is asymptotically efficient under certain regularity conditions, it does not provide a solution to the problem of model identification. MLE tends to over estimate the order of the true model and favor the model with more parameters as discovered by Huber[1]. This bias in MLE is further characterized under the hypothesis testing framework, and it is shown that under certain regularity conditions, the asymptotic null distribution of $-2\log\lambda$ is a χ_d^2 distribution with degree of freedom d equals to the difference of number parameters in two models, where λ is the likelihood ratio test statistics. Since then, many approaches for model identification are proposed, including the popular AIC by Akaike [2] and BIC (Bayesian Information Criterion) by Schwarz [3]. Recently, Chen [4] reported results of using BIC for speaker clustering and channel change detection and showed that BIC as a model identification criterion was applicable for various applications.

In this paper, we propose an approach of using a special penalized Bayesian information criterion to determine the decision tree state tying. In our experiments on a large vocabulary speech recognition (Wall Street Journal) task, it is found that original BIC will not lead to a compact representation of the decision tree. The decision tree is often overgrown, suggesting a modification to original BIC is needed. In fact, applying BIC type model identification criterion in decision tree state tying involves the complicated problem of tests for number of components in a mixture. Unfortunately, the classical asymptotic theory cannot apply to this situation, because it does not satisfy the regularity conditions required in order for asymptotic null distribution to be a χ_d^2 distribution. Based on a simulation study comparing single and two mixture models, which is typical in decision tree state tying, Wolfe proposed that the asymptotic null distribution would be approximated by χ_{2d}^2 distribution [5,6]. We found that using two times penalty of BIC in decision tree clustering indeed leads to similar high recognition performance as using a pre-selected threshold, a popular method with a root in AIC type model identification criterions. Moreover, it is found that penalized BIC can be used as an effective model compressing method. This may due in part to the close relation between BIC and other data compression criterions, such as minimum description length (MDL), in information theory. Model compression using penalized BIC seems more robust than the threshold based methods. In penalized BIC approach, the available training data at

tree nodes has a much stronger influence on the final shape of the decision tree and it favors more to models with low dimension. In our experiments, compact decision tree state tying models were constructed with almost no loss of the speech recognition performance.

2. STATISTICAL MODEL IDENTIFICATION IN DECISION TREE STATE TYING

In decision tree based state tying, a set of phonetic questions $\{Q(i)|i=1,...,M\}$ characterizing the phonetic properties of the context is selected. These phonetic questions are related to acoustic phonetic properties of the phonemes, such as a front vowel, nasal, fricative etc. Each question $Q(i)$ divides the acoustic phonetic space into two parts $A_{Q(i)}$ and $A_{\bar{Q}(i)}$ depending on the yes/no answer to the question. The acoustic space partition based on the phonetic questions is formed by the finite intersects of $\{A_{Q(i)}, A_{\bar{Q}(i)}|i=1,...,M\}$. The phonetic decision tree based state tying is to find a decision tree whose leaf nodes form a partition of the acoustic phonetic space, and under certain constraints, the log likelihood of the tree is maximized.

The standard CART one-step greedy growing algorithm is a top down process. It grows the terminal nodes of the tree one-step at a time. At each step, it searches for the best terminal node to grow and the best question to apply so that it leads to a maximum increase of the log likelihood by splitting the node into two children nodes. In other words, it is to find $(\bar{t},\bar{q})$ such that

$$(\bar{t},\bar{q}) = \text{argmax}_{(t,q)}[L(t,q,y)+L(t,q,n)-L(t)],$$

where $L(t,q,y)$ and $L(t,q,n)$ are the log likelihood of yes/no split of node *t* according to question *q*. In addition, model identification in decision tree state tying is applied by requiring the node split satisfies the condition:

$$L(\bar{t},\bar{q},y)+L(\bar{t},\bar{q},n)-L(t) \;>\; \Delta,$$

where Δ is a constant threshold determined by experiments. The tree splitting process in decision tree state tying can be viewed from the statistical hypothesis testing framework for testing the number of components in the mixture:

$$H_0 : \{x_i \in \bar{t} \,|\, i=1,...,N\} \;\sim\; N(\mu,\Sigma)$$

versus

$$H_1 : \; \{x_i \in \bar{t}_L\} \sim N(\mu_L,\Sigma_L),\; \{x_i \in \bar{t}_R\} \sim N(\mu_R,\Sigma_R),$$

where $\bar{t}_L$ and $\bar{t}_R$ are left and right children nodes from tree node $\bar{t}$. It is important to note that in decision tree state tying, single mixture Gaussians with diagonal covariance matrices are used in order to reduce the computational complexity and to derive the cluster Gaussians directly from their members without going back to data for re-estimation. We assume from now on the covariance matrices Σ, Σ_L, and Σ_R are all diagonal. The constant threshold approach has a relation to the AIC criterion, but it does not depend explicitly on the number of data samples at the tree node.

The Bayesian information criterion (BIC) is to select model j that maximizes

$$\log M_j(x_{1,...,}x_N) \;-\; \frac{1}{2}k_j \log N,$$

where k_j is the number of parameters in model j and $\log M_j(x_{1,...,}x_N)$ is the log likelihood of model j given data sample $\{x_{1,...,}x_N\}$. BIC criterion is derived for a special family of distributions as an asymptotic Bayesian factor. It is considered as a more conservative criterion than AIC and leans more than AIC towards lower dimension models.

The log maximum likelihood ratio statistics between the single Gaussian mixture distribution model at tree node $\bar{t}$ and the model formed by its children nodes $\bar{t}_L$ and $\bar{t}_R$ is given by:

$$\log\lambda\;(t) \;=\; \frac{N}{2}\log|\Sigma| - \frac{N_L}{2}\log|\Sigma_L| - \frac{N_R}{2}\log|\Sigma_R|.$$

The BIC criterion for selecting node $\bar{t}$ split versus node $\bar{t}$ becomes

$$B(t) = \log\lambda\;(t) \;-\; P_{BIC}(t) \;>\; 0, \quad (1)$$

where $P_{BIC}(t) = \frac{1}{2}(k+k)\log N$, assuming each model to be a k-dimensional single mixture Gaussian distribution with diagonal covariance matrix. It is important to point out that hypothesis testing for mixture model is a very complicated problem and the usual asymptotic distribution theory does not apply to the case of tests of number of components in a mixture distribution. As pointed in [6], the regular asymptotic procedures for testing of the presence of a mixture or for the number of components perform very poorly. A reason for this difficulty is that the particular geometry of the parameter space is non-Euclidean. If hypothesis H_1 that $f(x|\theta)$ is parameterized as a m-mixture Gaussians depends on parameters $(p_1,...,p_m,\theta_{1,...,}\theta_m)$ and the null hypothesis H_0 that $f(x|\theta)$ is a $(m-1)$-mixture Gaussians, the null hypothesis H_0 is covered in H_1 through different parameter values. In other words, the parameter space of H_0 does not correspond to a particular subspace and it makes χ^2_d asymptotic distribution quite unrealistic. In addition, criterions of AIC

and BIC are asymptotic results, describing the asymptotic behavior of the estimator as the sample size N goes to infinity. For finite samples, significant deviations can occur and the error terms dropped in the asymptotic analysis can become a factor. Correction terms are also added based on more precise error term estimation.

3. PENALIZED BAYESIAN INFORMATION CRITERION FOR DECISION TREE STATE TYING

The asymptotic distribution for likelihood ratio test of mixture model is still a topic of active research. There is no theoretical justification for taking null distribution of $-2\log\lambda$ to be a chi-squared distribution at the first place, let alone for its subsequent refinement. These criterions are used as a rough guide and penalty terms are used to make them suitable for various applications. The penalized Bayesian Information Criterion (pBIC) is a criterion that multiplies a constant penalty term on the BIC value $P_{BIC}(t)$ and the pBIC criterion for selecting node $\bar{t}$ split versus node $\bar{t}$ becomes

$$pB(t) = \log\lambda\ (t)\ -\ p_b * P_{BIC}(t)\ >\ 0\,, \qquad (2)$$

where p_b is a penalty factor. BIC is a special case of pBIC with $p_b = 1$. In pBIC, p_b can be used as a design parameter to control the model complexity, although it may depart from its original theoretically justified value. For applications, where original theoretical justification does not apply, p_b is an important model parameter and must be selected with care. In our experiments of using pBIC for decision tree state tying, the original theoretical value of $p_b = 1$ did not lead to a compact decision tree state tying model with good recognition performance. The tree is overgrown indicating a penalty factor greater than original theoretical value $p_b = 1$ is needed. Wolfe's study on null distribution of single versus two mixture Gaussian distributions further indicates that the degree of freedom would be approximated by twice the difference in the number of parameters in the two hypotheses. Based on Wolfe's modification, we derive that the penalty factor p_b for decision tree state tying using pBIC is around two, which is two times the original BIC value. In our proposed decision tree state tying algorithm using pBIC, the tree node splitting is based on equation (2) with penalty factor $p_b \geq 2$. This is a significant departure from the original BIC approach and it is the first justification of directly using pBIC in decision tree state tying for large vocabulary speech recognition. This approach is further confirmed by the experimental results presented in the next section, and the penalty factor of $p_b > 2$ is used as model compression factor for compact decision tree state tying model.

4. EXPERIMENTAL RESULTS

The proposed approach of decision tree state tying using penalized BIC was tested on the Wall Street Journal (WSJ) speech recognition task. 12 mel-cepstral coefficients and the normalized energy plus their 1st and 2nd order time derivatives were used as acoustic features. Phonetic decision tree state tying was used to cluster equivalent sets of context dependent states and to construct unseen triphones[7]. The final triphone HMMs were built based on the tied states from the clustering. The number of mixtures for each tied state depends on the amount of training data assigned and varies from 4 to 12. Decoding was done using a one-pass N-gram decoder [9], in which the search was conducted on a layered self-adjusting decoding graph using the cross-word triphone models. The standard SI-84 and SI-284 training data sets were used to train the WSJ models. The language models used in the experiments were the standard trigram language models provided by NIST for the WSJ corpus.

4.1 Decision Tree State Tying based on pBIC with $p_b = 2$

First set of experiments was designed to verify the proposed approach with $p_b = 2$, a BIC penalty factor value derived in Section 3. Experiments were based on gender independent models with cross-word acoustic model units and tagged position dependent acoustic model units [8]. The baseline model is an optimized model G_Δ obtained using a constant threshold Δ in tree node splitting. The value of Δ was determined experimentally.

Model		**WSJ-92**	
SI-84	#states	5k-closed	20k-open
G_Δ	4400	4.43 %	11.57%
$Gp_b = 2$	3566	4.35 %	11.56%
$Gp_b = 1$	6033	4.48 %	11.90%

Table 1: Word error rates for different state tying methods acoustic models (SI-84 training data).

Table 1 tabulates the word error rates on the WSJ-92 test data set. Model $Gp_b = 2$ was obtained with $p_b = 2$ as proposed in Section 3, and $Gp_b = 1$ model was obtained with $p_b = 1$ which corresponds to the original BIC criterion. The experimental results confirmed that using

pBIC penalty factor $p_b = 2$ resulted in a very compact decision tree state tying model. The speech recognition performance even slightly better than model G_Δ which is based on a node splitting criterion with a heuristic constant threshold Δ. Model $Gp_b = 2$ also has the least number of individual states after decision tree state tying. It has 19% less number of states than the best constant threshold G_Δ model and 42% less number of states than $Gp_b = 1$ model. It is also clear that the decision tree in $Gp_b = 1$ model is overgrown. It has a large increase of the number of individual sates without visible improvement in speech recognition performance.

4.2 Model Compression in Decision Tree State Tying Using pBIC

The second set of experiments were designed to use the proposed pBIC approach as a method for model compression in decision tree state tying based acoustic modeling. Different Bayesian penalty factor p_b values were used and the experiments were based on the full SI-284 training data set. Table 2 lists the word error rate and number of individual states after decision tree state tying.

Model		**WSJ-92**	
SI-284	#states	5k-closed	20k-open
G_Δ	8363	3.0%	9.5%
$Gp_b = 2$	8141	3.0%	9.7%
$Gp_b = 3$	6489	3.2%	9.4%
$Gp_b = 4$	5364	3.2%	9.5%

Table 2: Word error rates for different state tying methods acoustic models (SI-284 training data).

As it is indicated in Table 2, there is no significant speech recognition performance change even when the number individual states were reduced by more than 36% from 8363 states to 5364 states. In fact, the compact decision tree models $Gp_b = 3$ and $Gp_b = 4$ performed very well on the 20k-open test set and seems more robust to out-of-vocabulary words.

5. SUMMARY

In this paper, an approach of decision tree state tying based on penalized Bayesian information criterion pBIC is described and applied to two applications. First, it is used as a decision tree growing criterion in place of the conventional approach of using a heuristic constant threshold. It was found that original BIC penalty is too low and did not lead to a compact decision tree state tying model. Based on Wolfe's modification to the asymptotic null distribution, it was derived that two times BIC penalty should be used for decision tree state tying with pBIC. Secondly, pBIC is studied as a model compression criterion for model compression. Experimental results indicated that significant reduction of individual states could be achieved without loss of the speech recognition performance.

REFERENCES

[1] P. J. Huber, "The Behavior of Maximum Likelihood Estimation under Nonstandard Conditions", *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, pp. 221-233, 1967.

[2] H. Akaike, "A New Look at the Statistical Model Identification", *IEEE Trans. On Automatic Control*, vol. AL-19, No. 6, pp. 716 – 723.

[3] G. Schwarz, "Estimating the Dimension of A Model", Annals of Statistics, vol. 6, No. 2, pp. 461 – 464.

[4] S. Chen, et al, "Clustering via the Bayesian Information Criterion with Applications in Speech Recognition", *Proc. ICASSP'98*, vol. II, pp. 645 – 649, 1998.

[5] J. H. Wolfe, " A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of multinomial distributions", *Technical Bulletin STB 72-2*, San Diego, U.S. Navel Research and Training Research Lab.

[6] G. J. McLachlan and K. E. Basford, "Mixture Models", Marcel Dekker, Inc, 1988.

[7] W. Reichl and W. Chou, "Decision Tree State Tying Based on Segmental Clustering for Acoustic Modeling" ICASSP 98, Seattle, May 1998.

[8] W. Reichl and W. Chou, "A Unified Approach of Incorporating General Features in Decision Tree based Acoustic Modeling", submitted to *ICASSP'99*.

[9] Q. Zhou, and W. Chou, "An Approach to Continuous Speech Recognition Based on Layered Self-Adjusting Decoding Graph", ICASSP 97, Munich, Germany, April 1997.

[10] K. Shinoda et al, "Speaker Adaptation with Autonomous Model Complexity Control by MDL Principle", *Proc. ICASSP'96*, pp. 717 – 720, 1996.