Bayesian Context Clustering Using Cross Valid Prior Distribution for HMM-Based Speech Recognition

Kei Hashimoto, Heiga Zen, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda

Department of Computer Science and Engineering
Nagoya Institute of Technology, Nagoya, Japan

Abstract

This paper proposes a prior distribution determination technique using cross validation for speech recognition based on the Bayesian approach. The Bayesian method is a statistical technique for estimating reliable predictive distributions by marginalizing model parameters and its approximate version, the variational Bayesian method has been applied to HMM-based speech recognition. Since prior distributions representing prior information about model parameters affect the posterior distributions and model selection, the determination of prior distributions is an important problem. However, it has not been thoroughly investigated in speech recognition. The proposed method can determine reliable prior distributions without tuning parameters and select an appropriate model structure dependently on the amount of training data. Continuous phoneme recognition experiments show that the proposed method achieved a higher performance than the conventional methods.

Index Terms: variational Bayes, cross validation, context clustering, continuous phoneme recognition

1. Introduction

In speech recognition systems, context dependent models (e.g. triphone HMMs) are widely used [1]. Although a large number of triphones can capture variations in speech data, too many model parameters lead to the over-fitting problem. Therefore, maintaining the balance between model complexity and the amount of training data is quite important for obtaining high generalization performance. The decision tree based context clustering is an important method for context dependent HMM estimation to deal with the problem of training data insufficiency, not only for robust parameter estimation but also for predicting probability distributions for unseen contexts [2]. This method constructs a parameter tying structure which can assign a sufficient amount of training data to each HMM state. The tree is grown step by step, choosing questions that divide the context using a greedy strategy to maximize some objective function.

The maximum likelihood (ML) criterion is inappropriate as a model selection criterion since the ML criterion increases monotonically as the number of states increases. Some heuristic thresholding is therefore necessary to terminate the dividing nodes in context clustering. To solve this problem, the minimum description length (MDL) criterion has been employed to select the model structure [3]. However, the MDL criterion is based on an asymptotic assumption, therefore it is ineffective when the amount of training data is small. On the other hand, since the Bayesian approach does not use an asymptotic assumption, unlike the MDL criterion, it is available even in the case of small amount of training data. In the Bayesian approach, an appropriate model structure can be selected by maximizing the marginal likelihood [4].

The Bayesian approach uses prior information which is represented by the prior distribution. Since prior distributions affect the estimation of posterior distributions and model selection, the determination of prior distributions is an important problem for estimating appropriate acoustic models. However, it has not been thoroughly investigated in speech recognition, and the determination technique of prior distribution has not been developed. This paper proposes a prior distribution determination technique using the cross validation and apply it to the context clustering. The cross validation is known as a straightforward and useful method for model structure optimization. The context clustering based on cross validation has been proposed, and it can select the more appropriate model structure than the conventional ML criterion [5]. The proposed method can be regarded as an extension of context clustering using cross validation to the Bayesian approach. By using cross valid prior distributions, it is expected that the higher generalization ability is achieved and an appropriate model structure can be selected in the context clustering without tuning parameters.

The rest of this paper is organized as follows. Section 2 describes the speech recognition based on variational Bayesian method, and Section 3 describes the prior distribution determination technique using cross validation and apply it to the context clustering. In Section 4, results of the continuous phoneme recognition experiments are presented. Finally, conclusions and future works are drawn.

2. Speech recognition based on variational Bayesian method

2.1. Bayesian approach

The output distribution is obtained based on a left-to-right HMM which has been widely used to represent an acoustic model for speech recognition. Let \( O = (o_1, o_2, \ldots, o_T) \) be a set of training data of \( D \) dimensional feature vectors, and \( T \) is used to denote the frame number. The output distribution is represented by:

\[
P(O \mid \Lambda) = \prod_{t=1}^{T} a_{z_{t-1} z_t} N(o_t \mid \mu_{z_t}, S_{z_t}^{-1}).
\]

where \( Z = (z_1, z_2, \ldots, z_T) \) is a sequence of HMM states, \( z_t \in \{1, \ldots, N\} \) denotes a state at frame \( t \) and \( N \) is the number of states in an HMM. A set of model parameters \( \Lambda = \)
\{a_{ij}, \mu_i, S_i\}_{i,j=1}^N\) consists of the state transition probability \(a_{ij}\) from state \(i\) to state \(j\), the mean vector \(\mu_i\) and the covariance matrix \(S_i^{-1}\) of a Gaussian distribution \(\mathcal{N}(\cdot | \mu_i, S_i^{-1})\).

The Bayesian approach assumes that a set of model parameters \(\Lambda\) is probabilistic variables, while the ML approach estimates constant model parameters. The posterior distribution for a set of model parameters \(\Lambda\) is obtained with the famous Bayes theorem as follows:

\[
P(\Lambda | O) = \frac{P(O | \Lambda) P(\Lambda)}{P(O)} ,
\]

where \(P(\Lambda)\) is a prior distribution for \(\Lambda\). Once the posterior distribution \(P(\Lambda | O)\) is estimated, the predictive distribution for input data \(X\) is represented by:

\[
P(X | O) = \int P(X | \Lambda) P(\Lambda | O) d\Lambda .
\]

The model parameters are integrated out in Eq.(3) so that the effect of over-fitting is mitigated, and robust classification is obtained. However, it is difficult to solve the integral and expectation calculations. Especially, when a model includes latent variables, the calculation becomes more complicated. Posterior distributions with latent variables are represented as follows:

\[
P(\Lambda | O) = \sum_Z \frac{P(O, Z | \Lambda) P(\Lambda)}{P(O)} .
\]

To overcome this problem, the variational Bayesian (VB) method has been proposed as a tractable approximation method of the Bayesian approach and shown good generalization performance in many applications [6].

2.2. Variational Bayesian method

The variational Bayesian method maximizes a lower bound of log marginal likelihood \(\mathcal{F}\) instead of the true likelihood. A lower bound of log marginal likelihood is defined by using Jensen’s inequality:

\[
\mathcal{L}(\Lambda) = \log \sum_Z P(O, Z | \Lambda) P(\Lambda) d\Lambda = \log \sum_Z Q(Z) Q(\Lambda) P(O, Z | \Lambda) P(\Lambda) Q(Z) Q(\Lambda) d\Lambda \geq \sum_Z Q(Z) Q(\Lambda) \log \frac{P(O, Z | \Lambda) P(\Lambda)}{Q(Z) Q(\Lambda)} d\Lambda = \mathcal{F}
\]

In the VB method, VB posterior distributions \(Q(\Lambda)\) and \(Q(Z)\) are introduced to approximate the true posterior distributions. The optimal VB posterior distributions can be obtained by maximizing the objective function \(\mathcal{F}\) with the variational method as follows:

\[
Q(\Lambda) = C_\Lambda P(\Lambda) \exp \left\{ \sum_Z Q(Z) \log P(O, Z | \Lambda) \right\} , \tag{6}
\]

\[
Q(Z) = C_Z \exp \left\{ \int Q(\Lambda) \log P(O, Z | \Lambda) d\Lambda \right\} , \tag{7}
\]

where \(C_\Lambda\) and \(C_Z\) are the normalization terms of \(Q(\Lambda)\) and \(Q(Z)\), respectively. These optimizations can be effectively performed by iterative calculations as the Expectation and Maximization (EM) algorithm, which increases the value of objective function \(\mathcal{F}\) at each iteration until convergence.

2.3. Bayesian context clustering

The decision tree based context clustering is a top-down clustering method to optimize the state tying structure for robust model parameter estimation. A leaf of the decision tree corresponds to a set of HMM states to be tied. The decision tree growing process begins with a root node that may have all HMM states, or all states associated with a particular phone, etc. Then, a question which divides the set of states into two subsets assigned respectively to two child nodes, “Yes” node and “No” node, is chosen so that the corresponding new HMM has the largest value of objective function for training data. The decision tree is grown in a greedy fashion, successively splitting nodes by selecting the pair of a question and node which maximize the gain of objective function at each step.

In the Bayesian approach, an optimal model structure can be selected by maximizing the objective function \(\mathcal{F}\). When a node is split into “Yes” node and “No” node by the question \(q\), the gain \(\Delta \mathcal{F}_q\) is defined as the difference of \(\mathcal{F}\) before and after splitting:

\[
\Delta \mathcal{F}_q = \mathcal{F}_q^{\text{yes}} + \mathcal{F}_q^{\text{no}} - \mathcal{F}_q,
\]

where \(\mathcal{F}_q^{\text{yes}}\) and \(\mathcal{F}_q^{\text{no}}\) are the value of objective function \(\mathcal{F}\) of split nodes by a question \(q\), and \(\mathcal{F}_q\) is the value before a splitting. The question \(q\) for splitting a node is chosen from the question set as follows:

\[
\hat{q} = \arg \max_q \Delta \mathcal{F}_q .
\]

By stopping the node splitting when the condition \(\Delta \mathcal{F}_q \leq 0\) is satisfied, an optimal model structure is obtained.

3. Bayesian context clustering using cross validation

In the Bayesian approach, prior distributions are usually determined heuristically. However, hyper-parameters (parameters of prior distributions) affect the model selection as tuning parameters. Therefore, to automatically select an appropriate model structure, a determination technique of prior distribution is required. One possible approach is to optimize the hyper-parameters using training data so as to maximize the marginal likelihood. However, it still needs tuning parameters which control influences of prior distributions, and often leads to the over-fitting problem as the ML criterion. In this paper, we propose the prior distribution determination technique using cross validation and apply it to the context clustering.

3.1. Cross validation

For \(K\)-fold cross validation, the training data is randomly divided into \(K\) different groups. Then, a model is trained using \(N - 1\) groups of data. and the objective function is computed for the group excluded in the training. This process is repeated for \(K\) times with different combinations of \(K - 1\) groups. The value of objective function is accumulated and the accumulated value is used for evaluation of model structure.

3.2. Bayesian approach using cross validation

Let \(O = \{O^{(1)}, O^{(2)}, \ldots, O^{(K)}\}\) be a set of training data and \(O^{(k)}\) be a partition for \(K\)-fold cross validation. For the \(k\)-th evaluation, \(O^{(k)} = \{O^{(j)} | j \neq k\}\) is used for the determination of prior distributions and \(O^{(k)}\) is used for the estimation of posterior distributions. Then, the Bayesian approach
using cross validation calculates the log marginal likelihood:

\[ L^{(k)}(O) = \log P(O^{(k)} | O^{(k)}) \].

(10)

Using Jensen’s inequality, the lower bound of log marginal likelihood \( F^{(k)} \) is defined as Eq.(5). For the \( k \)-th evaluation, the optimal VB posterior distributions of model parameters can be obtained by maximizing \( F^{(k)} \) with respect to \( Q(\Lambda^{(k)}) \) with the variational method as follows:

\[
Q(\Lambda^{(k)}) = \mathcal{C}_{\Lambda^{(k)}} P(\Lambda^{(k)} | O^{(k)})
\]

\[ \exp \left\{ \sum_k Q(Z^{(k)}) \log P(O^{(k)}, Z^{(k)} | \Lambda^{(k)}) \right\} , \]

(11)

where \( P(\Lambda^{(k)} | O^{(k)}) \) is a prior distribution which represents prior information \( O^{(k)} \) and \( C_{\Lambda^{(k)}} \) is a normalization term.

In the Bayesian approach, a conjugate prior distribution is widely used as a prior distribution \( P(\Lambda) \). When the output probability distribution is a Gaussian distribution, the conjugate prior distribution becomes a Gaussian-Wishart distribution:

\[
P(\mu, S) = N(\mu | \nu, (\xi S)^{-1}) W(S | \eta, B),
\]

(12)

where \((\xi, \eta, \nu, B)\) is a set of hyper-parameters. Moreover a Gaussian distribution is proportional to Gaussian-Wishart distribution as follows:

\[
\prod_{t=1}^{T} N(o_t | \mu, S^{-1})
\]

\[ \propto N(\bar{o}_t | \bar{\mu}, (TS)^{-1}) W(S | T + D, (TC)^{-1}) , \]

where \( \bar{o} = \frac{1}{T} \sum_{t=1}^{T} o_t \) and \( C = \frac{1}{T} \sum_{t=1}^{T} o_t o_t^\top - \bar{o} \bar{o}^\top \) are sufficient statistics of training data. Thus, the prior distribution can be determined by sufficient statistics of the prior information. The prior distribution of the \( k \)-th cross validation model parameters \( P(\mu^{(k)}, S^{(k)} | O^{(k)}) \) is obtained from Eq.(13):

\[
P(\mu^{(k)}, S^{(k)} | O^{(k)}) = N(\mu^{(k)} | \bar{o}^{(k)}, (T^{(k)}S^{(k)})^{-1}) \times W(S^{(k)} | T^{(k)} + D, (T^{(k)}C^{(k)})^{-1}) ,
\]

(14)

where \( \bar{o}^{(k)} \) and \( C^{(k)} \) are sufficient statistics of a subset of training data \( O^{(k)} \). The cross valid prior distribution can be determined without tuning parameters.

3.3. Bayesian context clustering using cross valid prior distribution

The objective function of the Bayesian approach using cross validation \( F^{(CV)} \) is obtained by summing \( F^{(k)} \) for each fold:

\[
F^{(CV)} = \sum_{k=1}^{K} F^{(k)},
\]

(15)

In the proposed method, an optimal model structure can be selected by maximizing the objective function \( F^{(CV)} \). The question \( \hat{q} \) for splitting the node is chosen from the question set as follows:

\[
\hat{q} = \arg \max_q \Delta F_q^{(CV)} ,
\]

(16)

where \( \Delta F_q^{(CV)} \) is the gain in the value of objective function \( F_q^{(CV)} \) when a node is split by the question \( q \). The gain \( \Delta F_q \) is obtained by:

\[
\Delta F_q^{(CV)} = F_q^{(CV)} - F_q^{CV,n} - F_q^{CV,p} .
\]

(17)

where \( F_q^{CV,n} \) and \( F_q^{CV,p} \) are the value of objective function \( F_q^{CV} \) of split nodes by a question \( q \), and \( F_q^{CV} \) is the value before a splitting. By stopping the node splitting when the condition \( \Delta F_q^{(CV)} \leq 0 \) is satisfied, an optimal model structure can be obtained.

4. Experiments

To evaluate the effectiveness of the proposed method, speaker independent continuous phoneme recognition experiments were performed.

4.1. Experimental conditions

The experimental conditions are summarized in Table 1. The training data consisted of about 20,000 Japanese sentences and 1,000 Japanese sentences were prepared from Japanese Newspaper Article Sentences (JNAS). The test data consisted of 100 Japanese sentences were prepared from JNAS. Three-state left-to-right HMMs were used to model 43 Japanese phonemes. 144 questions were prepared in decision tree context clustering. Each state output probability distribution was modeled by a single Gaussian distribution with a diagonal covariance matrix.

In these experiments, the following three algorithms were compared.

- “MDL” : Acoustic models trained by ML criterion and model structures selected by MDL criterion.
- “CV-ML” : Acoustic models trained by ML criterion and model structures selected by cross validation with ML criterion.
- “CV-Bayes” : Acoustic models trained by the Bayesian criterion and model structures selected by cross validation with the Bayesian criterion.

In “CV-ML” and “CV-Bayes”, the cross validation uses 10 folds.

4.2. Experimental results

Figure 1, 2 show the phoneme accuracy of acoustic models trained by 20,000 and 1,000 utterances, respectively. To evaluate the performance of model selection, the phoneme accuracy for several model structures which are selected by varying a
threshold also is shown. In these figures, the lines represent the phoneme accuracy for each model structure and the points represent the phoneme accuracy of the model structure selected automatically by each method. These figures clearly show that the proposed method “CV-Bayes” outperforms the “MDL” and “CV-ML” baselines. In Figure 1, “CV-Bayes” achieved 8.48% relative error reductions over “MDL”.

Figure 3, 4 show the lower bound of log marginal likelihoods using cross validation $F^{(CV)}$ for the training data, $F$ for the test data and the phoneme accuracy. In these figures, a similar tendency between $F^{(CV)}$ and $F$ was observed. The model structure which gives the highest $F^{(CV)}$ also achieved the highest $F$. However, the phoneme accuracy was not proportional to $F$, and the proposed method could not select the model structure which achieved the highest phoneme accuracy. This means that although the proposed method could select the model structure which accurately predict acoustic features for each HMM state, it is not identical to with the performance in the classification problem. This result suggests that a Bayesian criterion which represent the classification performance directly is required.

5. Conclusions

This paper proposes a prior distribution determination technique using cross validation and apply it to the context clustering for speech recognition based on the Bayesian approach. The results on continuous phoneme recognition experiments demonstrated that CV-Bayes outperform both MDL and CV-ML. The proposed method could determine prior distributions without tuning parameters, and select the model structure which accurately predict acoustic features for each HMM state. As a future work, we will applied a Bayesian criterion which represent the classification performance directly to context clustering.

6. References