Review on "Improving Clustering with Hidden Markov Models Using Bayesian Model Selection"

Peng Peng

Department of Electrical and Computer Engineering
Mississippi State University
Mississippi State, MS 39762 USA
email: peng@isip.mstate.edu

ABSTRACT

Here is a review of the paper "Improving Clustering with Hidden Markov Models using Bayesian Model Selection" by C. Li and G. Biswas published in the Proceedings of 2000 IEEE International Conference on Systems, Man, and Cybernetics, 2000 [1]. Hidden Markov Model clustering procedure is an effective approach to improve the speed and accuracy of recognizers based on Hidden Markov Models. However, in the earlier work on clustering with Hidden Markov models ([2], [3], [4]) there is no objective criterion measure to determine the number of clusters in a partition. The HMM size for all models in the final clusters in a partition is also pre-specified and uniform. All these pre-determined factors will decrease the accuracy of HMM clustering if they are not correctly set.

This paper focuses on a Bayesian HMM clustering methodology that improves existing HMM clustering algorithm by incorporating Hidden Markov Model size selection into the HMM clustering procedure. The criterion for HMM size selection and partition selection are investigated and the complete clustering control structure for the Bayesian HMM clustering algorithm is shown in this paper. Experiments are performed on the Bayesian HMM clustering algorithm with artificially generated data. From the experiments, the new method not only generates more accurate model structure for individual clusters, but also improves the quality of the partitions generated, i.e., reduces Partition Misclassification Count (PMC) and increases Between Partition Similarity (BPS) [6]. It seems Bayesian HMM clustering algorithm does work in improving accuracy of the existing HMM clustering algorithm.

1. INTRODUCTION

In many real applications, the dynamic characteristics, i.e., how a system interacts with the environment and evolves over time, are of interest. Such dynamic characteristics can be assumed to satisfy the Markov property and such behavior or characteristic of these systems can be best described by temporal features whose values change significantly during the observation period. These dynamic characteristics are assumed to satisfy the Markov property, and may be viewed as a probabilistic walk though a fixed set of states. Thus we can characterize dynamics of objects in individual clusters using hidden Markov models.

For speaker-independent speech recognition, template training by clustering is required to achieve high word recognition accuracy for practical tasks. Clustering derives structure from data by objectively partitioning data into homogeneous groups so that the within group object similarity and the between group object dissimilarity are optimized simultaneously. It is important for our clustering system to determine the best partitions of the data, and the best model structure, i.e., the number of states in a model, to characterize the dynamics of the homogeneous data within each cluster. These tasks can be approached by (1) developing an explicit HMM model size selection procedure that dynamically modifies the size of the HMMs during the clustering process, and (2) casting the HMM model size selection and partition selection problems in terms of a Bayesian model selection problem.

2. METHODOLOGY

The clustering algorithm for temporal data that incorporates HMM mode size selection can be described in terms of a search procedure with four nested loops:

Loop1: derive the number of clusters in a partition;

Loop2: the object distribution to clusters in a given partition size;

Loop3: the HMM model sizes for individual clusters in the partition; and

Loop4: the HMM parameter configuration for the individual clusters.

Once a model size(i.e., the number of states in the HMM model) is selected, step 4 is invoked to estimate model parameters that optimize a chose criterion. We use the well known Maximum Likelihood (ML) parameter estimation method, the Baum-Welch procedure [7] to iteratively guide the parameter search process to the locally maximum values.

2.1. Bayesian Criterion for HMM Size Selection

From Bayes theorem, the posterior probability of the model, $P\langle M|X\rangle$, is given by

$$P(M|X) = \frac{P(M)P(X|M)}{P(X)}$$

where $P(X)$ and $P(M)$ are prior probabilities of the data and the model respectively, and $P(X|M)$ is the marginal likelihood of the data. Assuming none of the models considered is favored a priori, we have $P(M|X) \propto P(X|M)$. That is, the posterior probability of a model is directly proportional to the marginal likelihood. Therefore, the goal is to select the mixture model that gives the highest marginal likelihood.

Given the parameter configuration, θ, of a model M, the marginal likelihood of the data is computed as

$$P(X|M) = \int_{\theta} P(X|\theta, M)P(\theta|M)d\theta$$

When parameters involved are continuous valued, the above computation often becomes too complex to express in a closed analytic form. One efficient approximation method is the Bayesian Information Criterion (BIC) [5], where in log form, marginal likelihood is approximated by:

$$\log P(M|X) \approx \log P(X|M, \hat{\theta}) - \frac{d}{2}\log N$$

d is the number of parameters in the model, N is the number of data objects, and $\hat{\theta}$ is the ML parameter configuration of model M. $\log P(X|M, \hat{\theta})$, the data likelihood, tends to promote larger and more detailed models of data, whereas the second term, $-\frac{d}{2}\log N$, is the penalty term which favors smaller models with less parameters. BIC selects the best model for the data by balancing these two terms.

Applying the BIC approximation, marginal likelihood of the HMM for cluster k is computed as:

$$\log P(X_k|\lambda_k) \approx \sum_{j=1}^{N_k} \log P(X_{kj}|\lambda_k, \hat{\theta}_k) - \frac{d_k}{2}\log N_k$$

where N_k is the number of objects in cluster k, d_k is the number of parameters in λ_k, and $\hat{\theta}_k$ is the Maximum Likelihood parameters in λ_k. The HMM size can chosen according to the highest BIC value.

2.2. Bayesian Criterion for Partition Selection

In model-based clustering, the mixture model, M, is represented by K component models and a hidden, independent discrete variable C, where each value i of C represents a component cluster, modeled by λ_i. Given observation $X = (x_1, \ldots, x_N)$, the density of an observation x_i from the kth component model can be represented by $f_k(x_i|\theta_k, \lambda_k)$, where θ_k is the corresponding parameters of the model. Therefore, the likelihood of the mixture model given data is expressed as:

$$P(X|\theta_1, \ldots, \theta_K, P_1, \ldots, P_K) = \prod_{i=1}^{N} \sum_{k=1}^{K} P_k \cdot f_k(x_i|\theta_k, \lambda_k)$$

where P_k is the probability that an observation belongs to the kth component. Bayesian clustering algorithm casts the model-based clustering problem into the Bayesian model selection problem.

According to Bayesian theorem, the best clustering mixture model has the highest partition posterior probability (PPP), $P(M|X)$. Here PPP can be approximated with the marginal likelihood of the mixture model, $P\langle X|M\rangle$. Therefore, given partitions with different component clusters, the goal is to select the best overall model, M, that has the highest marginal likelihood of the mixture model, $P\langle M|X\rangle$.

For partition with K clusters, modeled as λ_k $k = 1, \ldots, K$, the PPP computed using BIC approximation is:

$$\log P(X|M) \approx \sum_{i=1}^{N} \log\left[\sum_{k=1}^{K} P_k \cdot P(X_i|\hat{\theta}_k, \lambda_k)\right]$$

$$-\frac{K + \sum_{k=1}^{K} d_k}{2}\log N$$

where $\hat{\theta}_k$ and λ_k are the ML model parameter configuration and the number of significant model parameters of cluster k, respectively. P_k is the likelihood of data given the model for cluster k. Each object is assigned to one know cluster in the partition. Therefore, $P_k = 1$ if object X_i is in cluster k, and $P_k = 0$ otherwise. The best model clustering is the one that balances the overall data likelihood and the complexity of the entire cluster partition.

2.3. Bayesian HMM Clustering Control Structure

Given the characteristics of the BIC criterion in partition selection and HMM size selection, the paper employ a sequential search strategy for both selection search. Table 1 gives the complete description of control structure for the Bayesian HMM clustering (BHMMC) algorithm.

Table 1: BHMMC control structure

```
K=1
do
  Select K seeds
  Apply HMM size selection on each seed
  Object redistribution:
    do
      Distribute objects to clusters with the highest likelihood
      Apply HMM parameter estimation for all clusters
    while there are objects change cluster memberships
  Compute PPP of the current partition
  K = K+1
while Current PPP > PPP of the previous partition
Accept the previous partition as the final cluster partition
Apply HMM size selection on the final clusters.
```

In this control structure, four steps in Bayesian HMM clustering algorithm are implemented. Given a partition size K, step 1 selects K seeds that are likely to form the centroids of the K clusters in the partition. HMM size selection is applied for each chosen seed to find the best model size for each cluster. Next, step 2 distributes objects to individual clusters such that the overall data likelihood given the partition is maximized. If any object changes its cluster membership in step 2, step 3 would update models for all clusters to reflect the current data in the clusters. Then, all objects are redistributed based on the set of new models. Otherwise, the distribution is accepted. Then step 4 estimates the model parameters for each cluster using the Baum-Welch procedure. The clustering procedure is finished.

3. EXPERIMENT ANALYSIS

The experiments in this paper are performed on synthetic models and data. There are totally two experiments. For experiment 1, five different HMMs are generated for each of the three model sizes: 5, 10, 15 states. Then a separate data set is created based on each of these 15 HMMs. For experiment 2, three groups of data sets are constructed. Individual data sets in each group contain three models. Each model has a different size, i.e., 4, 6, and 8 states. There are different pairwise model distances in the three groups. Five data sets are constructed for each group according to the pairwise model distance requirement. Each data set is created by combining data objects generated from the three different HMMs. Therefore, for these combined data sets, the number of models involved, the model size and parameter configuration are known.

However, the test set only containing artificially generated data, which is hard to convince us that if these experiments were performed on the practical training data sets, the results would still be satisfying.

In experiment 1, the effectiveness of BIC in selecting HMM sizes was tested. As can be seen in experiment 1, BIC selected HMMs that have sizes identical to the generative HMMs for 5-state and 10-state HMMs. For 15-state generative HMMs, the sizes of the derived models differ among trials and have an average size 13.2, which is smaller than that of the true HMMs. This should be attributed to the well known problem with the Baum-Welch ML parameter estimation procedure. It sometimes converges to a locally maximum parameter configuration, which prematurely terminates the sequential HMM model size search process. It is a reasonable explanation about the experiment result.

In experiment 2, thee effect of the HMM size selection on cluster partition generation was studied. Two different clustering methods are compared: (1) the Bayesian Hidden Markov Model Clustering (BHMMC) which performs dynamic HMM size selection, and (2) a clustering algorithm that uses a pre-determined, fixed size HMM throughout clustering. When model size selection is not applied, the partitions generated with too small a fixed HMM, i.e., a 3-state HMM, are considered better than those generated with too big a fixed HMM, i.e., a 15-state HMM. Partitions of better quality are generated when the fixed HMM size equals the average size of the generative HMMs. When the HMM model selection procedure is applied, individual clusters are modeled with HMMs of appropriate sizes to best fit data, and the complexity of all HMMs in the partition and the overall data likelihood are carefully balanced. For all trials, partitions generated with HMM size selection have higher posterior model probability and larger between partition similarity than those obtained from clustering with the fixed size HMMs.

From these experiments, two ideas were verified. The BIC algorithm works well when the size of true HMMs are not too large. Incorporating HMM size selection into HMM clustering algorithm leads to better quality cluster partitions than those generated by fixed HMM size clustering algorithm.

4. CONCLUSION

This paper presented a review of an Bayesian HMM Clustering algorithm. The paper by C. Li and G. Biswas applies Bayesian criterion for HMM size selection and partition selection in HMM clustering algorithm. The new clustering method is called Bayesian temporal data clustering methodology using HMMs.

According to the review, the incorporation of the HMM size selection procedure not only generates more accurate model structure for individual clusters, but also improves the quality of the partitions generated.

As we point out in this paper, the testing set should be practical to make sure that the clustering algorithm can be really verified on the data set. We hope we can see more experimental results on the effect of BHMMC method when the method is applied to the true speech data sets.

REFERENCES

[1] C. Li, G. Biswas, “Improving Clustering with Hidden Markov Models Using Bayesian Model Selection,” Proceedings of 2000 IEEE International Conference on Systems, Man, and Cybernetics, vol. 1, pp. 194-199, 2000.

[2] L.R. Rabiner, C.H. Lee, B.H. Juang, J.G. Wilpon, “HMM Clustering for Connected Word Recognition,” Proceedings of the Fourteenth International Conference on Acoustics, Speech, and Signal Processing, pp. 405-408, 1989.

[3] E. Dermatas, G. Kokkinakis, “Algorithm for Clustering Continuous Density HMM by Recognition Error,” IEEE Transactions on Speech and Audio Processing, vol. 4, no. 3, pp. 231-234, May 1996.

[4] P. Smyth, “Clustering Sequences with Hidden Markov Models,” Advances in Neural Information Processing, M.C. Mozer, M.I. Jordan, T. Petsche, Eds. Cambridge, MA, MIT Press, pp. 648-654, 1997.

[5] D. Heckerman, D. Geiger, D.M. Chickering, “A Tutorial on Learning with Bayesian Networks,” Machine Learning, vol. 20, pp. 197-243, 1995.

[6] C. Li, G. Biswas, “A Bayesian Approach to Temporal Data Clustering with Hidden Markov Model Representation,” Proceedings of the Seventeenth International Conference on Machine Learning, P. Langley, Ed, 2000.

[7] L.R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, February 1989.