Improving Clustering with Hidden Markov Models using Bayesian Model Selection

C. Li and G. Biswas
Box 1679, Station B
Department of EECS, Vanderbilt University
Nashville, TN 37235, USA

Abstract

This paper presents a Bayesian clustering methodology that partitions temporal data into homogeneous groups, and constructs state based profiles for each group in the hidden Markov model (HMM) framework. We propose a Bayesian HMM clustering methodology that improves upon existing HMM clustering algorithm by incorporating HMM model size selection into the clustering control structure. Experimental results indicate the effectiveness of our methodology.

1 Introduction

Unsupervised classification, or 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. Categorization and interpretation of structure are achieved by analyzing the models constructed in terms of the feature value distributions within each group. In many real applications, the dynamic characteristics, i.e., how a system interacts with the environment and evolves over time, are of interest. Such behavior or characteristic of these systems is best described by temporal features whose values change significantly during the observation period. Our goal for temporal data clustering is to construct profiles of dynamic processes by constructing and analyzing well defined, parsimonious models of data.

We assume that the temporal data sequences that define the dynamic characteristics of the phenomenon under study satisfy the Markov property, and the data generation may be viewed as a probabilistic walk through a fixed set of states. We characterize dynamics of objects in individual clusters using hidden Markov models.

Our ultimate goal is to use the extracted HMM models as an accurate explainable representation of the system dynamics. 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. We approach these tasks by (i) developing an explicit HMM model size selection procedure that dynamically modifies the size of the HMMs during the clustering process, and (ii) casting the HMM model size selection and partition selection problems in terms of a Bayesian model selection problem.

2 HMM Definitions

A HMM is a non-deterministic stochastic Finite State Automata. The basic structure of a HMM consists of a connected set of states, $S = (S_1, S_2, ..., S_n)$. We use first order HMMs, where the state of a system at a particular time t is only dependent on the state of the system at the immediate previous time point, i.e., $P(S_t|S_{t-1}, S_{t-2}, ..., S_1) = P(S_t|S_{t-1})$. In addition, we assume all the temporal feature values are continuous, therefore, we use the continuous density HMM (CDHMM) representation where all temporal features have continuous values. A CDHMM of n states for data having m temporal features can be characterized[1] in terms of three sets of probabilities [4]: the initial state probabilities, the transition probability, and the emission probabilities. The initial state probabilities, $\vec{\pi}$ of size n, defines the probability of any of the given states being the initial state of the given sequence. The transition probability matrix, A of size $n \times n$, defines the probability of transition from state i at time t, to state j at the next time step. And the emission

[1] We assume the continuous features are sampled at a predefined rate, and the temporal feature values are defined as a sequence of values.

0-7803-6583-6/00/$10.00 © 2000 IEEE

probability matrix, B of size $n \times m$, defines the probability of generating feature values at any given state. For CDHMM, the emission probability density function of each state is defined by a multivariate Gaussian distribution.

3 Clustering with HMMs

A clustering algorithm for temporal data that incorporates HMM model size selection can be described in terms of a search procedure with four nested loops:
loop 1: derive the number of clusters in a partition;
loop 2: the object distribution to clusters in a given partition size;
loop 3: the HMM model sizes for individual clusters in the partition; and
loop 4: the HMM parameter configuration for the individual clusters.

A primary limitation of the earlier work on clustering with HMM models ([5], [1], [6]) is that for search step 1, no objective criterion measure is used to automatically select the cluster partition based on data. A pre-determined threshold value on data likelihood, or a post-clustering Monte-Carlo simulation, is used instead. Another limitation is that they assume a prespecified and uniform HMM size for all models in the intermediate and final clusters in a partition. Therefore, search step 3 does not exist in those systems.

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 chosen criterion. We use the well known Maximum Likelihood (ML) parameter estimation method, the *Baum-Welch* procedure [4] to iteratively guide the parameter search process to the locally maximum values.

4 The Bayesian Clustering Methodology

4.1 Bayesian Model Selection

From Bayes theorem, the posterior probability of the model, $P(M|X)$, 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. For the purpose of comparing alternate models, we have $P(M|X) \propto P(M)P(X|M)$. Assuming none of the models considered is favored a priori, $P(M|X) \propto P(X|M)$. 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, as in the case of CDHMM, the integration computation often becomes too complex to express in a closed analytic form. In this paper, we look at one efficient approximation methods: the Bayesian information criterion (BIC) [2], where in log form, marginal likelihood is approximated by:

$$logP(M|X) \approx logP(X|M, \hat{\theta}) - \frac{d}{2}logN.$$

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

4.2 Bayesian Clustering

In model-based clustering, it is assumed that data is generated by a mixture of underlying probability distributions. 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 observations $X = (x_1, \cdots, x_N)$, let $f_k(x_i|\theta_k, \lambda_k)$ be the density of an observation x_i from the kth component model, λ_k, where θ_k is the corresponding parameters of the model. The likelihood of the mixture model given data is expressed as: $P(X|\theta_1, \cdots, \theta_K, P_1, \cdots, P_K) = \prod_{i=1}^{N} \sum_{k=1}^{K} P_k \cdot f_k(x_i|\theta_k, \lambda_k)$, where P_k is the probability that an observation belongs to the kth component($P_k \geq 0, \sum_{k=1}^{K} P_k = 1$). Bayesian clustering casts the model-based clustering problem into the Bayesian model selection problem. Given partitions with different component clusters, the goal is to select the best overall model, M, that has the highest *posterior probability*, $P(M|X)$.

5 Bayesian HMM Clustering

We have adapted Bayesian clustering to the CDHMM clustering problem, so that: (i) components of a

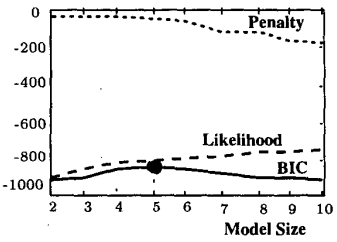

Figure 1: HMM model size selection

Bayesian mixture model are represented by CDHMMs, and (ii) $f_k(X_i|\theta_k, \lambda_k)$ in data likelihood computation computes the likelihood of a multi-feature temporal sequence given a CDHMM.

First, we describe how the general Bayesian model selection criterion is adapted for the HMM model size selection and the cluster partition selection problems. Then we describe how the characteristics of these criterion functions are used to design our heuristic clustering search control structure.

5.1 Criterion Functions

5.1.1 Criterion for HMM Size Selection

The HMM model size selection process picks the HMM with the number of states that best describe the data. We use Bayesian model selection criterion to select the best HMM model size given data.

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

$$logP(X_k|\lambda_k) \approx \sum_{j=1}^{N_k} logP(X_{kj}|\lambda_k, \hat{\theta_k}) - \frac{d_k}{2} logN_k,$$

where N_k is the number of objects in cluster k, d_k is the number of parameters [2] in λ_k, and $\hat{\theta_k}$ is the ML parameters in λ_k.

Figure 1 illustrates how BIC works for HMM model size selection. Data generated on a 5-state HMM is modeled using HMMs of sizes ranging from 2 to 10. BIC values corresponding to the different HMM sizes are plotted. The dashed lines show the likelihood of data for the different size HMMs. The dotted lines show the penalty for each model. And the solid lines show BIC as a combination of the above two terms. We observe that as the size of the model increases, the model likelihood also increases and the model penalty and parameter prior decreases monotonically. BIC has its highest value corresponding to the size of the original HMM for data.

[2] Significant parameters include all the parameters for emission probability definitions and only the initial probabilities and transition probabilities that are greater than a threshold value t, t is set to 10^{-6} for all experiments reported in this paper.

5.1.2 Criterion for Partition Selection

In the Bayesian framework, the best clustering mixture model, M, has the highest *partition posterior probability* (PPP), $P(M|X)$. We approximate PPP with the marginal likelihood of the mixture model, $P(X|M)$.

For partition with K clusters, modeled as $\lambda_1, \cdots, \lambda_K$, the PPP computed using the BIC approximation is:

$$logP(X|M) \approx \sum_{i=1}^{N} log[\sum_{k=1}^{K} P_k \cdot P(X_i|\hat{\theta_k}, \lambda_k)] - \frac{K+\sum_{k=1}^{K} d_k}{2} logN,$$

where $\hat{\theta_k}$ and d_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. When computing the data likelihood, we assume that the data is complete, i.e., each object is assigned to one known cluster in the partition. Therefore, $P_k = 1$ if object X_i is in cluster k, and $P_k = 0$ otherwise. The best model is the one that balances the overall data likelihood and the complexity of the entire cluster partition.

Figure 2 illustrates how BIC works for cluster partition selection: given data consisting of an equal number of objects from four randomly generated HMMs, the BIC scores are measured when data is partitioned into 1 to 10 clusters. At first when the number of clusters is small, because the improvements of data likelihood dominates the change of the BIC values, i.e., the BIC values monotonically increase as the size of the partition increases. It reaches the peak value when the size of the partition corresponds to the true partition size, four. Subsequently, the improvements of data likelihood becomes less and less significant, the penalty on model complexity term dominate the change of the BIC measure, and it decreases monotonically as the size of the partition continues to increase.

5.2 The Clustering Control Structure

Given the characteristics of the BIC criterion in partition selection (step 1) and HMM model size selection (step 3), we employ a sequential search strategy for both search steps 1 and 3. We start with the simplest model, i.e., a one cluster partition for step 1 and a

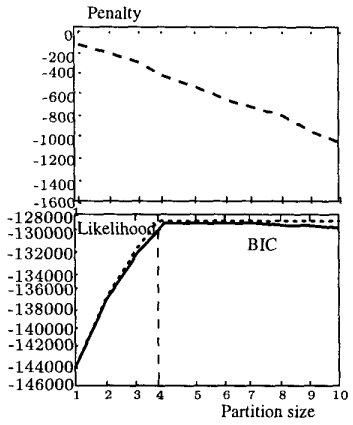

Figure 2: Cluster partition selection

one state HMM for step 3. Then, we gradually increase the size of the model, i.e., adding one cluster to the partition or adding one state to the HMM, and re-estimate the model. After each expansion, we evaluate the model using BIC. If the score of the current model decreases from that of the previous model, we may conclude that we have just passed the peak point, and accept the previous model as our final model. Otherwise, we continue with the model expansion process.

To find the best partition structure, i.e., object distribution in clusters, for a given size (K) partition, we first select K seeds that are likely to form the centroids of the K clusters in the partition. A seed includes a set of k objects, (k = 3 for all experiments shown here). The purpose of including more than one object in each seed is to ensure that there is sufficient data to build a reliable initial HMM. The first object in a seed is selected by choosing an object that has the least likelihood given all cluster models in the current partition (the first object in the first seed is selected randomly from the data set). The remaining objects in the seed are the ones have the highest likelihood given the HMM model built based on the first object. We apply HMM model size selection for each chosen seed for find the best model size for each cluster.

Next, search step 2 distributes objects to individual clusters such that the overall data likelihood given the partition is maximized. We assign object, x_i, to cluster, $(\hat{\theta_k}, \lambda_k)$, based on its sequence-to-HMM likelihood measure [4], $P(x_i|\hat{\theta_k}, \lambda_k)$. Individual objects are assigned to clusters whose model provides the highest data likelihood. If after one round of object distribution, any object changes its cluster membership, models for all clusters are updated 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. After the objects are distributed into clusters, for a HMM model size, step 4 estimates the model parameters for each cluster using the Baum-Welch procedure. Table 1 gives the complete description of the Bayesian HMM clustering (BHMMC) algorithm.

Table 1: The BHMMC control structure

```
K=1
do
   Select K seeds
   Apply HMM model 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 model size selection on the final clusters.
```

6 Experimental Results

In this section, we experimentally validate the BHMMC algorithm with artificially generated data. First, we describe how synthetic models and data are generated for the experiments. Then, we give the performance indices we use to evaluate the experimental results. Finally, we analyze the experimental results using the proposed performance indices.

6.1 Data

To construct HMM models of different sizes, first, we assign state definitions by randomly selecting mean and variance values from value ranges [0, 100] and [0,25] respectively. Then we assign state transition probabilities and initial probabilities by randomly sampling from value range [0, 1], and then normalize the probabilities.

Based on each model, we randomly generated 40 objects, each object is described by two temporal features, and the sequence length of each feature is set to 100. For experiment 1, we generated five different HMMs for each of the three model sizes: 5, 10, and 15 states. Then, a separate data set is created based on each of these 15 HMMs. For experiment 2, we constructed three groups of data sets. Individual data sets in each group contain three models, each of

Table 2: BIC for HMM model size selection

Criterion measure	True HMM model size		
	5	10	15
BIC	5(0)	10(0)	13.2(2.2)

a different size, i.e., 4, 6, and 8 states. In group 1, the three models all have a pairwise model distance between [-100, 0] (distance level 1), models in group 2 have pairwise model distance between [-400, -300] (level 2), and models in group 3 have pairwise distance between [-700, -600] (level 3). Models in group 1 are more difficult to separate than those in group 2, which are more difficult to separate than those in group 3. Five data sets, i.e., five sets of three HMM models, are constructed for each group according to the pairwise model distance requirement. Given each set of three HMM models, one data set is created by combining data objects generated from the 3 different HMMs. For these combined data sets, we know the number of models involved, and the model size and parameter configuration of each. The goal of clustering here is to rediscover the correct partition model based on data.

6.2 Performance Indices

In addition to the partition posterior probability, we use two other performance indices to evaluate the quality of the cluster partitions generated: (i) **Partition Misclassification Count (PMC)**, which computes the number of object misclassified when compared to the true object to cluster assignment in the generative partition model. The smaller the sum of the misclassification counts for all objects in a partition, the better quality the partition is in comparison; and (ii) **Between Partition Similarity (BPS)** which measures the similarity between the derived partition and the generative partition in terms of the likelihood of temporal sequences generated by one partition given the other partition, and vice versa. the larger the BPS, the more similar the partition in comparison is to the true partition, thus the better quality is the partition. Details on these evaluation criterions can be found in [3].

6.3 Experiments

The first experiment studies the effectiveness of BIC in selecting HMM model sizes based on data. Table 2 shows average model sizes and the standard deviations of the HMMs derived from data. For 5-state and 10-state HMMs, BIC selected HMMs that have sizes identical to the generative HMMs. For 15-state generative HMMs, the sizes of the derived models differ among trials, and have an average size smaller than that of the true HMMs. This is 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.

Table 3: Cluster partition size rediscovered given a three cluster partition

Data Set	Fixed HMM size clustering			Varying HMM size clustering
	3	6	10	
level 3	3(0)	3(0)	2.8(0.45)	3(0)
level 2	3(0.7)	3.2(0.45)	2.6(0.54)	3(0)
level 1	2.8(0.83)	2.8(0.45)	2.2(0.45)	3(0)

Table 4: PMC on results obtained from data three cluster partition model

Data Set	Fixed HMM size clustering			Varying HMM size clustering
	3	6	10	
level 3	0(0)	0(0)	16(35)	0(0)
level 2	17(35.28)	1 (2.23)	32(43.8)	0(0)
level 1	39.6(40)	16(35)	64(35.8)	0(0)

The second experiment studies the effect of the HMM model size selection on cluster partition generation. Two different clustering methods are compared: (1) the BHMMC which performs dynamic HMM model size selection, and (2) a clustering algorithm that uses a pre-determined, fixed size HMM throughout clustering. Table 3 shows the mean and standard deviation of the partition size and the PMC score for partitions generated for different data sets. BHMMC with HMM model size selection significantly outperforms clustering with fixed HMM sizes. When HMM model size is applied, the BHMMC algorithm completely rediscovered the correct partition models on all trials. 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 four

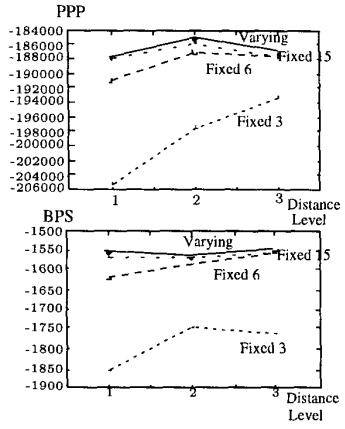

Figure 3: HMM cluster using HMM model size selection vs. using fixed HMM model size

generative HMMs.

When the size of the HMMs are fixed and small, they do not possess the ability to discriminate among objects that are generated from multiple, more complex HMMs. Therefore, objects from different generative HMMs are grouped into the same cluster in the final partition. On the other hand, when using fixed size HMMs that are too big, adding one new cluster to the partition incurs a large model complexity penalty that sometimes can not be offset by the data likelihood gain. 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. These lead to better quality cluster partitions. Figure 3 compares the partitions in terms of their PPP and BPS scores. The solid lines represent the BHMMC with model size selection, and the three dashed lines represent clustering with fixed 3-state HMM, 6-state HMM, and 10-state HMM. For all trials, partitions generated with HMM model size selection have higher posterior model probability and larger between partition similarity than those obtained from clustering with the fixed size HMMs.

7 Summary

We have presented a Bayesian temporal data clustering methodology using HMMs. Bayesian model selection criterion has been successfully applied to solve both HMM model size selection and cluster partition selection problems. The incorporation of the HMM model size selection procedure not only generates more accurate model structure for individual clusters, but also improves the quality of the partitions generated.

Because of the computational complexity of clustering algorithms involving HMMs, for future work, we would like to incorporate incremental clustering strategies where we start with a cluster partition built based on small data, and gradually revise the size and structure of the partition as more data is collected. Also, we would like to look into partition evaluation criterion based on model prediction accuracy. Even though the purpose of our HMM clustering is not prediction per se, how well the set of cluster models can predict may be used to evaluate the quality of the partition.

References

[1] DERMATAS, E., AND KOKKINAKIS, G. Algorithm for clustering continuous density hmm by recognition error. *IEEE Transactions on Speech and Audio Processing 4*, 3 (May 1996), 231–234.

[2] HECKERMAN, D., GEIGER, D., AND CHICKERING, D. M. A tutorial on learning with bayesian networks. *Machine Learning 20* (1995), 197–243.

[3] LI, C., AND BISWAS, G. A bayesian approach to temporal data clustering with hidden markov model representation. In *Proceedings of the Seventeenth International Conference on Machine Learning* (2000), P. Langley, Ed.

[4] RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*, 2 (Feb. 1989), 257–285.

[5] RABINER, L. R., LEE, C. H., JUANG, B. H., AND WILPON, J. G. Hmm clustering for connected word recognition. In *Proceedings of the Fourteenth International Conference on Acoustics, Speech, and Signal Processing* (1989), pp. 405–408.

[6] SMYTH, P. Clustering sequences with hidden markov models. In *Advances in Neural Information Processing*, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. Cambridge:MA, MIT Press, 1997, pp. 648–654.