A Hidden Markov Model-Based Approach to Sequential Data Clustering

Antonello Panuccio, Manuele Bicego, and Vittorio Murino

Dipartimento di Informatica, University of Verona
Ca' Vignal 2, Strada Le Grazie 15, 37134 Verona, Italy
{panuccio,bicego,murino}@sci.univr.it

Abstract. Clustering of sequential or temporal data is more challenging than traditional clustering as dynamic observations should be processed rather than static measures. This paper proposes a Hidden Markov Model (HMM)-based technique suitable for clustering of data sequences. The main aspect of the work is the use of a probabilistic model-based approach using HMM to derive new proximity distances, in the likelihood sense, between sequences. Moreover, a novel partitional clustering algorithm is designed which alleviates computational burden characterizing traditional hierarchical agglomerative approaches. Experimental results show that this approach provides an accurate clustering partition and the devised distance measures achieve good performance rates. The method is demonstrated on real world data sequences, i.e. the EEG signals due to their temporal complexity and the growing interest in the emerging field of Brain Computer Interfaces.

1 Introduction

The analysis of sequential data is without doubts an interesting application area since many real processes show a dynamic behavior. Several examples can be reported, one for all is the analysis of DNA strings for classification of genes, protein family modeling, and sequence alignment.

In this paper, the problem of unsupervised classification of temporal data is tackled by using a technique based on Hidden Markov Models (HMMs). HMMs can be viewed as stochastic generalizations of finite-state automata, when both transitions between states and generation of output symbols are governed by probability distributions [1]. The basic theory of HMMs was developed in the late 1960s, but only in the last decade it has been extensively applied in a large number of problems, as speech recognition [1], handwritten character recognition [2], DNA and protein modeling [3], gesture recognition [4], behavior analysis and synthesis [5], and, more in general, to computer vision problems.

Related to sequence clustering, HMMs has not been extensively used, and a few papers are present in the literature. Early works were proposed in [6,7], all related to speech recognition. The first interesting approach not directly linked to speech issues was presented by Smyth [8], in which clustering was faced by devising a "distance" measure between sequences using HMMs. Assuming each

T. Caelli et al. (Eds.): SSPR&SPR 2002, LNCS 2396, pp. 734–743, 2002.

model structure known, the algorithm trains an HMM for each sequence so that the log-likelihood (LL) of each model, given each sequence, can be computed. This information is used to build a LL distance matrix to be used to cluster the sequences in K groups, using a hierarchical algorithm.

Subsequent work, by Li and Biswas [9,10], address the clustering problem focusing on the model selection issue, i.e. the search of the HMM topology best representing data, and the clustering structure issue, i.e. finding the most likely number of clusters. In [9], the former issue is addressed using standard approach, like Bayesian Information Criterion [11], and extending to the continuous case the Bayesian Model Merging approach [12]. Regarding the latter issue, the sequence-to-HMM likelihood measure is used to enforce the within-group similarity criterion. The optimal number of clusters is then determined maximizing the Partition Mutual Information (PMI), which is a measure of the inter-cluster distances. In the second paper [10], the same problems are addressed in terms of Bayesian model selection, using the Bayesian Information Criterion (BIC) [11], and the Cheesman-Stutz (CS) approximation [13]. Although not well justified, much heuristics is introduced to alleviate the computational burden, making the problem tractable, despite remaining of elevate complexity. Finally, a model-based clustering method is also proposed in [14], where HMMs are used as cluster prototypes, and Rival Penalized Competitive Learning (RPCL), with state merging is then adopted to find the most likely HMMs modeling data. These approaches are interesting from the theoretical point of view, but they are not tested on real data. Moreover, some of them are very computationally expensive.

In this paper, the idea of Smyth [8] has been extended by defining a new metric to measure the distance, in the likelihood sense, between sequences. Two clustering algorithms are proposed, one based on the hierarchical agglomerative approach, and the second based on a partitional method, variation of the K-means strategy. Particular care has been posed on the HMM training initialization by utilizing a Kalman filtering and a clustering method using mixture of Gaussians. Finally, and most important, the proposed algorithm has been tested using real data sequences, the electroencephalographic (EEG) signals. Analysis of this kind of signals became very important in the last years, due to the growing interest in the field of *Brain Computer Interface (BCI)* [15]. Among all we choose these signals for their temporal complexity, suitable for HMM modeling.

The rest of the paper is organized as follows. In Sect. 2, HMM will be introduced. Section 3 describes how the EEG signal has been modeled and the specific initialization phase of the proposed approach. The core of the algorithm is presented in Sect. 4, in which the definition of distances and the clustering algorithms will be detailed. Subsequently, experimental results are presented in Sect. 5, and, finally, conclusions are drawn in Sect. 6.

2 Hidden Markov Models

A discrete HMM is formally defined by the following elements [1]:

- *A set* $S = \{S_1, S_2, \cdots, S_N\}$ *of (hidden) states.*
- *A state transition probability distribution*, also called transition matrix $A = \{a_{ij}\}$, representing the probability to go from state S_i to state S_j.

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i] \qquad 1 \leq i, j \leq N, \quad a_{ij} \geq 0, \quad \sum_{j=1}^{N} a_{ij} = 1 \tag{1}$$

- *A set* $V = \{v_1, v_2, \cdots, v_M\}$ *of observation symbols.*
- *An observation symbol probability distribution*, also called emission matrix $B = \{b_j(k)\}$, indicating the probability of emission of symbol v_k when system state is S_j.

$$b_j(k) = P[v_k \text{ at time t } | q_t = S_j] \qquad 1 \leq j \leq N, 1 \leq k \leq M \tag{2}$$

  with $b_i(k) \geq 0$ and $\sum_{j=1}^{M} b_j(k) = 1$.
- *An initial state probability distribution* $\pi = \{\pi_i\}$, representing probabilities of initial states.

$$\pi_i = P[q_1 = S_i] \qquad 1 \leq i \leq N, \quad \pi_i \geq 0, \quad \sum_{i=1}^{N} \pi_i = 1 \tag{3}$$

For convenience, we denote an HMM as a triplet $\lambda = (A, B, \pi)$.

All of our discussion has considered only the case where the observation was characterized as a sequence of discrete symbols chosen from a finite alphabet. In most application, observations are continuous signals. Although it is possible to quantize such continuous signals via codebooks, it would be advantageous to be able to use HMMs with continuous observation densities. In this case the emission probability distribution B becomes

$$P(O|j) = b_j(\mathbf{O}) = \sum_{m=1}^{M} c_{jm} \mathcal{M}[\mathbf{O}, \mu_{jm}, \mathbf{\Sigma}_{jm}] \tag{4}$$

where $\mathbf{O}$ is observation vector being modeled, c_{jm} is the mixture coefficient for the mth mixture in state j and $\mathcal{M}$ is any log-concave or elliptically symmetric density (e.g. Gaussian density). The adaption of reestimation formulas of Baum-Welch procedure for the continuous case is straightforward [16].

Although the general formulation of continuous density HMMs is applicable to a wide range of problems, there is one other very interesting class of HMMs that seems to be particularly suitable for EEG signals: the autoregressive HMMs [17]. In this case, the observation vectors are drawn from an autoregression process. In the next section it is explained how these models are applied to EEG modeling.

3 EEG Signal Modeling

Electroencephalographic (EEG) signals represent the brain activity of a subject and give an objective mode of recording brain stimulation. EEGs are an useful tool used for understanding several aspects of the brain, from diseases detection to sleep analysis and evocated potential analysis. The system used to model EEG signal is largely based on Penny and Roberts paper [18]: the key idea above this approach is to train an autoregressive HMM directly on the EEG signal, rather than use an intermediate AR representation. Each HMM state can be associated with a different dynamic regime of the signal, determined using a Kalman Filter approach [19]. Kalman filter is used to preliminary segment the signal in different dynamic regimes: these estimates are then fine-tuned with HMM model. The approach is briefly resumed in the rest of this section.

3.1 Hidden Markov AR Models

This type of models differs from those defined in Sect. 2 by the definition of observation symbol probability distribution. In this case B is defined as

$$P(y_t|q_t = S_i) = N(y_t - \mathbf{F}_t\hat{\mathbf{a}}_i, \sigma_i^2) \tag{5}$$

where $\mathbf{F}_t = -[y_{t-1}, y_{t-2}, \cdots, y_{t-p}]$, $\hat{\mathbf{a}}_i$ is the (column) vector of AR coefficients for the ith state and σ_i^2 is the estimated observation noise for the i-th state, estimated using Jazwinski method [20]. The prediction for the ith state is $\hat{y}_t^i = \mathbf{F}_t\hat{\mathbf{a}}_i$. The order of AR model is p.

The HMM training procedure is fundamentally a gradient descent approach, sensitive to initial parameters estimate. To overcome this problem, a Kalman filter AR model is passed over the data, obtaining a sequence of AR coefficients. Coefficients corresponding to low evidence are discarded. Others are then clusterized with Gaussian Mixture Models [21]. The center of each Gaussian cluster is then used to initialize the AR coefficients in each state of the HMM-AR model.

The number of clusters (i.e. the number of HMM states) and the order of autoregressive model were decided by performing a preliminary analysis of classification accuracy. Varying number of states from 4 to 10, and varying order of autoregressive model from 4 to 8, we have found that best configuration was $K = 4$ and $p = 6$. The classification accuracy obtained was about 2% superior than one obtained using Neural Network [22] on same data, showing that Hidden Markov Models are more effective in modeling EEG signals.

To initialize the transition matrix we used prior knowledge from the problem domain about average state duration densities. We use the equation $a_{ii} = 1 - \frac{1}{d}$ to let HMM remain in state i for d samples. This number is computed knowing that EEG data is stationary for a period of the order of half a second [23].

4 The Proposed Method

Our approach, inspired by [8], can be depicted by the following algorithm:

1. We train an $m-$states HMM for each sequence S_i, $(1 \leq i \leq N)$ of the dataset D. These N HMM are identified by λ_i, $(1 \leq i \leq N)$ and have been initialized with a Kalman filter AR model as described in Sect. 3.
2. For each model λ_i we evaluate its probability to generate the sequence $S_j, \quad 1 \leq j \leq N$, obtaining a measure matrix L where

$$L_{ij} = P(S_j|\lambda_i), \qquad 1 \leq i, j \leq N \tag{6}$$

3. We apply a suitable clustering algorithm to the matrix L obtaining K clusters on the data set D.

This method aims to exploits the measure defined by (6) which naturally expresses the similarity between two observation sequences. Through the use of Hidden Markov Models, that are able to describe a sequence with a simple scalar number, we could transform the difficult task of clustering sequences in the easier one of clustering points.

About step 3 we can apply several clustering algorithms but first of all we need to "symmetrize" the matrix L because the result of step 2 is not really a distance matrix. Thus we define

$$L_S^{ij} = \frac{1}{2}\left[L_{ij} + L_{ji}\right] \tag{7}$$

Another kind of HMM based measure that we applied, which remind the Kullback-Leibler information number, defines the distance L_{KL} between two HMM λ_i and λ_j, and its symmetrized version L_{KLS}, as

$$L_{KL}^{ij} = L_{ii}\left[\ln \frac{L_{ii}}{L_{ji}}\right] + L_{ij}\left[\ln \frac{L_{ij}}{L_{jj}}\right], \quad L_{KLS}^{ij} = \frac{1}{2}\left[L_{KL}^{ij} + L_{KL}^{ji}\right] \tag{8}$$

Finally, we introduced another measure, called BP metric, defined as

$$L_{BP}^{ij} = \frac{1}{2}\left\{\frac{L_{ij} - L_{ii}}{L_{ii}} + \frac{L_{ji} - L_{jj}}{L_{jj}}\right\} \tag{9}$$

motivated by the following considerations: the measure (6), defines a similarity measure between two sequences S_i and S_j as the likelihood of the sequence S_i with respect to the model λ_j, trained on S_j, without really taking into account the sequence S_j. In other words this kind of measure assumes that all sequences are modeled with the same quality without considering how well sequence S_j is modeled by the HMM λ_j: this could not always be true. Our proposed distance also considers the modeling goodness by evaluating the relative normalized difference between the sequence and the training likelihoods. About step 3 we investigated two clustering algorithms [21], namely

- Complete Link Agglomerative Hierarchical Clustering: this class of algorithms produces a sequence of clustering of decreasing number of clusters at each step. The clustering produced at each step results from the previous one by merging two clusters into one.
- Partitional Clustering: this methods obtains a single partition of the data instead of a clustering structure, such as a dendogram produced by hierarchical technique. Partitional method have advantages in application involving large data sets for which the construction of a dendogram is computationally prohibitive. In this context we developed an ad hoc partitional method described in the next section and henceforth called "DPAM".

4.1 DPAM Partitional Clustering Algorithm

The proposed algorithm shares the ideas of the well known k-means techniques. This method finds the optimal partition by evaluating at each iteration the distance between each item and each cluster descriptor, and assigning it to the nearest class. At each step, the descriptor of each cluster will be reevaluated by averaging its cluster items. A simple variation of the method, partition around medoid (PAM) [24], determines each cluster representative by choosing the point nearest to the centroid. In our context we cannot evaluate centroid of each cluster because we only have item distances and not values.

To address this problem a novel algorithm is proposed. This method is able to determine cluster descriptors in a PAM paradigm, using item distances instead of their values. Moreover, the choice of the initial descriptors could affect algorithm performances. To overcome this problem we have adopted a multiple initialization procedure, where the best resulting partition is determined by a sort of Davies-Bouldin criterion [21].

Fixed η as the number of tested initializations, N the number of sequences, k the number of clusters and L the proximity matrix characterized by previously defined distances (7), (8), (9), the resulting algorithm is the following:

```
– for t=1 to $\eta$
    • Initial cluster representatives $\theta_j$ are randomly chosen ($j = 1, \ldots, k$,
      $\theta_j \in \{1, \ldots, N\}$).
    • Repeat:
        ∗ Partition evaluation step:
          Compute the cluster which each sequence $S_i$, $i = 1, \ldots, N$
          belongs to; $S_i$ lies in the $j$ cluster for which the distance
          $L(S_i, \theta_j)$, $i = 1, \ldots, N$, $j = 1, \ldots k$ is minimum.
        ∗ Parameters upgrade:
            · Compute the sum of the distance of each element of cluster $C_j$ from each other element of the $j$th cluster
            · Determine the index of the element in $C_j$ for which this sum is minimal
            · Use that index as new descriptor for cluster $C_j$
```

- Until the representatives θ_j values between two successive iterations don't change.
- $\mathcal{R}_t = \{C_1, C_2, \ldots, C_k\}$
- Compute the Davies–Bouldin–like index defined as:

$$\mathcal{DBL}^{(t)} = \frac{1}{k}\sum_{r=1}^{k} \max_{s \neq r} \left\{ \frac{S_c^L(C_r, \theta_r) + S_c^L(C_s, \theta_s)}{L(\theta_r, \theta_s)} \right\}$$

where S_c is an intra–cluster measure defined by:

$$S_c^L(C_r, \theta_r) = \frac{\sum_{i \in C_r} L(i, \theta_r)}{|C_r|}$$

- endfor t
- *Final solution:* The best clustering $\mathcal{R}_p$ has the minimum Davies–Bouldin–like index, viz.: $p = \arg\min_{t=1,\ldots,\eta}\{\mathcal{DBL}^{(t)}\}$

5 Experiments

In order to validate the exposed modeling technique we worked primarily on EEG data recorded by Zak Keirn at Purdue University [25]. The dataset contains EEGs signal recorded from different subjects which were asked to perform five mental tasks: a *baseline* task, for which the subjects were asked to relax as much as possible; the *math task*, for which the subjects were given nontrivial multiplications problems, such as 27*36, and were asked to solve them without vocalizing or making any other physical movements; the *letter task*, for which the subjects were instructed to mentally compose a letter to a friend without vocalizing; the *geometric figure rotation*, for which the subjects were asked to visualize a particular 3D block figure being rotated about an axis; and a *visual counting task*, for which the subjects were asked to image a blackboard and to visualize numbers being written on the board sequentially. We applied the method on a segment-by-segment basis, 1s signals sampled at 250Hz and drawn from a dataset of cardinality varying from 190 (two mental states) to 473 sequences (five mental states) where we removed segments biased by signal spikes arising human artifact (e.g. ocular blinks).

The proposed HMM clustering algorithm has been first applied to two mental states: *baseline* and *math task*, then we extend trials to all available data. Accuracies are computed by comparing the clustering results with real segment labels, percentage is merely the ratio of correct assigned label with respect to the total number of segments. First we applied the hierarchical complete link technique, varying the proximity measure: results are shown in Table 1(a), with number of mental states growing from two to five.

We note that accuracies are quite satisfactory. None of the method experimented can be considered the best one, nevertheless, measures (7) and (8) seem to be more effective. Therefore we applied the partitional algorithm to the same

Table 1. Results for (a) Hierarchical Complete Link and (b) Partitional DPAM Clustering varying the distances defined in (9) BP, (8) KL and (7) SM

	BP	KL	SM
2 natural clusters	97.37%	97.89%	97.37%
3 natural clusters	71.23%	79.30%	81.40%
4 natural clusters	62.63%	57.36%	65.81%
5 natural clusters	46.74%	54.10%	49.69%

(a)

BP	KL	SM
95.79%	96.32%	95.79%
75.44%	72.98%	65.61%
64.21%	62.04%	50.52%
57.04%	46.74%	44.80%

(b)

datasets setting the number of initializations $\eta = 5$ during all the experiments. Results are presented in Table 1(b): in this last case the BP distance is overall slightly better than the others experimented measures. A final comparison of partitional and agglomerative hierarchical algorithms underlines that there are no remarkable differences between the proposed approaches. Clearly, partitional approaches alleviates computational burden, thus they should be preferred when dealing with complex signals clustering (e.g. EEG). The comparison of clustering and classification results (obtained in earlier works) shown that the latter are just slightly better. This strengthen the quality of the proposed method, considering that unsupervised classification is inherently a more difficult task.

6 Conclusions

In this paper we addressed the problem of unsupervised classification of sequences using an HMM approach. These models, very suitable in modeling sequential data, are used to characterize the similarity between sequences in different ways. We extend the ideas exposed in [8] by defining a new metric in likelihood sense between data sequences and by applying to these distance matrices two clustering algorithms: the traditional hierarchical agglomerative method and a novel partitional technique. Partitional algorithms are generally less computational demanding than hierarchical, but could not be applied in this context without some proper adaptations, proposed in this paper. Finally we tested our approach on real data, using complex temporal signals, the EEG, that are increasing in importance due to recent interest in Brain Computer Interface. Results shown that the proposed method is able to infer the natural partitions with patterns characterizing a complex and noisy signal like the EEG ones.

References

1. Rabiner, L. R.: A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. of IEEE **77(2)** (1989) 257–286. 734, 736

2. Hu, J., Brown, M. K., Turin, W.: HMM based on-line handwriting recognition. IEEE Trans. Pattern Analysis and Machine Intelligence, **18(10)** (1996) 1039–1045. 734
3. Hughey, R., Krogh, A.: Hidden Markov Model for sequence analysis: extension and analysis of the basic method. Comp. Appl. in the Biosciences **12** (1996) 95–107. 734
4. Eickeler, S., Kosmala, A., Rigoll, G.: Hidden Markov Model based online gesture recognition. Proc. Int. Conf. on Pattern Recognition (ICPR) (1998) 1755–1757. 734
5. Jebara, T., Pentland, A.: Action Reaction Learning: Automatic Visual Analysis and Synthesis of interactive behavior. In 1st Intl. Conf. on Computer Vision Systems (ICVS'99) (1999). 734
6. Rabiner, L. R., Lee, C. H., Juang, B. H., Wilpon, J. G.: HMM Clustering for Connected Word Recognition. Proceedings of IEEE ICASSP (1989) 405–408. 734
7. Lee, K. F.: Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition. IEEE Transactions on Acoustics, Speech and Signal Processing **38(4)** (1990) 599–609. 734
8. Smyth, P.: Clustering sequences with HMM, in Advances in Neural Information Processing (M. Mozer, M. Jordan, and T. Petsche, eds.) MIT Press **9** (1997). 734, 735, 738, 741
9. Li, C., Biswas, G.: Clustering Sequence Data using Hidden Markov Model Representation, SPIE'99 Conference on Data Mining and Knowledge Discovery: Theory, Tools, and Technology, (1999) 14–21. 735
10. Li, C., Biswas, G.: A Bayesian Approach to Temporal Data Clustering using Hidden Markov Models. Intl. Conference on Machine Learning (2000) 543–550. 735
11. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics, **6(2)** (1978) 461–464. 735
12. Stolcke, A., Omohundro, S.: Hidden Markov Model Induction by Bayesian Model Merging. Hanson, S. J., Cowan, J. D., Giles, C. L. eds. Advances in Neural Information Processing Systems **5** (1993) 11–18. 735
13. Cheeseman, P., Stutz, J.: Bayesian Classification (autoclass): Theory and Results. Advances in Knowledge discovery and data mining, (1996) 153–180. 735
14. Law, M. H., Kwok, J. T.: Rival penalized competitive learning for model-based sequence Proceedings Intl Conf. on Pattern Recognition (ICPR) **2** (2000) 195–198. 735
15. Penny, W. D., Roberts, S. J., Curran, E., Stokes, M.: EEG-based communication: a PR approach. IEEE Trans. Rehabilitation Engineering **8(2)** (2000) 214–215. 735
16. Juang, B. H., Levinson, S. E., Sondhi, M. M.: Maximum likelihood estimation for multivariate mixture observations of Markov Chain. IEEE Trans. Informat. Theory **32(2)** (1986) 307–309. 736
17. Juang, B. H., Rabiner, L. R.: Mixture autoregressive hidden Markov models for speech signals. IEEE Trans. Acoust. Speech Signal Proc. **33(6)** (1985) 1404–1413. 736
18. Penny, W. D., Roberts, S. J.: Dynamic models for nonstationary signal segmentation. Computers and Biomedical Research **32(6)** (1998) 483–502. 737
19. Kalman, R. E.: A New Approach to Linear Filtering and Prediction Problems. Transaction of the ASME - Journal of Basic Engineering (1960) 35–45. 737
20. Jazwinski, A.: Adaptive Filtering. Automatica **5** (1969) 475–485. 737
21. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Academic Press (1999). 737, 738, 739

22. Anderson, C. W., Stolz, E. A., Shamsunder, S.: Multivariate autoregressive models for classification of spontaneous electroencephalogram during mental tasks. IEEE Transactions on Biomedical Engineering, **45(3)** (1998) 277–286. 737
23. Nunez, P. L.: Neocortical Dynamics and Human EEG Rhythms. Oxford University Press, (1995). 737
24. Kaufman, L., Rousseuw, P.: Findings groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons – New York (1990). 739
25. Keirn, Z.: Alternative modes of communication between man and machine. Master's thesis. Purdue University (1988). 740