A Hidden Markov Model-based approach to sequential data clustering

Antonello Panuccio, Manuele Bicego, and Vittorio Murino
University of Verona, Italy†
{panuccio|bicego|murino}@sci.univr.it

Objectives

Clustering of sequential or temporal data with an Hidden Markov Model (HMM)-based technique.

Main aspects:

- use of HMM to derive new proximity distances, in the likelihood sense, between sequences;
- partitional clustering algorithm which alleviates computational burden characterizing traditional hierarchical agglomerative approaches.

The method is demonstrated on real world data sequences, i.e. the EEG signals:

- temporal complexity;
- growing interest in the emerging field of Brain Computer Interfaces.

Hidden Markov Models

A discrete HMM is formally defined by the following elements :

- *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.

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

- *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.

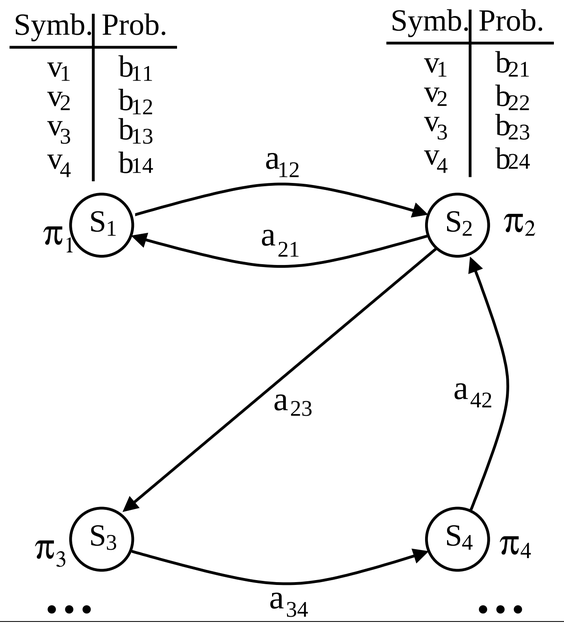
$$b_j(k) = P[v_k \text{ at time t } | q_t = S_j]$$
$$1 \leq j \leq N, 1 \leq k \leq M$$

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.

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

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



Hidden Markov AR Models

A very interesting class of HMMs that seems to be particularly suitable for EEG signals are the *autoregressive HMMs*. In this case, the observation vectors $\{y_1 ... y_T\}$ are drawn from an autoregressive process and thus B is defined as

$$b_j(y_t) = P(y_t | q_t = S_i) = N(y_t - \mathbf{F}_t \hat{\mathbf{a}}_i, \sigma_i^2) \qquad (1)$$

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;
- σ_i^2 is the estimated observation noise for the i-th state, estimated using Jazwinski method.

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.

HMM for EEG signal modeling

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 the EEG signal is based on [Penny and Roberts 1998] paper:

- train an autoregressive HMM directly on the EEG signal, rather than use an intermediate AR representation;
- a Kalman filter approach is used to preliminary segment the signal in different dynamic regimes of the signal;
- assign each HMM state with a different dynamic regime.

Initialization of training procedure

Problem: The HMM training procedure (*Baum-Welch re-estimation procedure*) is sensitive to initial parameters estimate.

Solution:

Initialization of Emission Matrix $\mathbf{B}$:

- pass a Kalman filter AR model over the data, obtaining a sequence of AR coefficients;
- coefficients corresponding to low evidence are discarded;
- clusterize the remaining with Gaussian Mixture Models;
- the center of each Gaussian cluster is then used to initialize the AR coefficients in each state of the HMM-AR model.

Initialization of Transition Matrix $\mathbf{A}$:

- The prior knowledge from the problem domain about average state duration densities is used to initialize the matrix.
- each diagonal element is set to $a_{ii} = 1 - \frac{1}{d}$ to let HMM remain in state i for d samples;
- d is computed knowing that EEG data is stationary for a period of the order of half a second.

The proposed method for sequence clustering

The proposed approach, inspired by [Smyth 95], can be depicted by the following algorithm:

1. *Train an* $m-$*states HMM for each sequence* S_i, $(1 \leq i \leq N)$ *of the dataset* D*, initializing the training as former explained. The obtained* N *HMM are identified by* λ_i, $(1 \leq i \leq N)$*;*
2. *for each model* λ_i*, evaluate its probability to generate all sequences* S_j, $\quad 1 \leq j \leq N$*, obtaining a likelihood matrix* L *where*

$$L_{ij} = P(S_j | \lambda_i), \qquad 1 \leq i, j \leq N \qquad (2)$$

3. *derive a sequences distance matrix from (2);*
4. *apply a suitable clustering algorithm to the distance matrix obtaining* K *clusters on the data set* D*.*

Remarks:

- This method exploits the measure defined by (2) which naturally expresses the similarity between two observation sequences.
- Hidden Markov Models are able to model similarity between sequences, allowing to recover the difficult task of clustering sequences to standard clustering.

Derivation of Distance

Three different differences derived from (2) were investigated:

1. *Distance "SM"*: obtained by simply symmetrizing (2):

$$L_S^{ij} = \frac{1}{2}[L_{ij} + L_{ji}] \qquad (3)$$

2. *Distance "KL"*: similar to the Kullback-Leibler information number:

$$L_{KL}^{ij} = L_{ii}\left[\ln \frac{L_{ii}}{L_{ji}}\right] + L_{ij}\left[\ln \frac{L_{ij}}{L_{jj}}\right] \qquad (4)$$

3. *Distance "BP"*: proposed in this paper:

$$L_{BP}^{ij} = \frac{1}{2}\left\{\frac{L_{ij} - L_{ii}}{L_{ii}} + \frac{L_{ji} - L_{jj}}{L_{jj}}\right\} \qquad (5)$$

Motivations for distance "BP":

- The measure (2), 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.
- This kind of measure assumes that all sequences are modelled with the same quality without considering how well sequence S_j is modelled by the HMM λ_j: this could not always be true.
- Our proposed distance also considers the modelling goodness by evaluating the relative normalized difference between the sequence and the training likelihoods.

Clustering Algorithm

About step 4 we introduce a new partitional clustering algorithm, called DPAM:

- this methods obtains a single partition of the data;
- we compare it with *Complete Link Agglomerative Hierarchical Clustering*, a standard class of algorithms that, instead of single partition, produces a sequence of clustering of decreasing number of clusters at each step;
- partitional method have advantages in application involving large data sets for which the construction of a dendogram is computationally prohibitive.

†Address for correspondence: Department of Informatics, University of Verona, Strada Le Grazie n.15, I-37134 Verona, Italy

A Hidden Markov Model-based approach to sequential data clustering

Antonello Panuccio, Manuele Bicego, Vittorio Murino
University of Verona, Italy†
{panuccio|bicego|murino}@sci.univr.it

DPAM Partitional Clustering

The standard partitional clustering schemes, as K-means, work as follows:

- at each iteration evaluate the distance between each item and each cluster descriptor
- assign the item to the nearest cluster.
- after re-assignment, the descriptor of each cluster will be reevaluated by averaging its cluster items;
- A simple variation of the method, called "Partition Around Medoid (PAM)", determines each cluster representative by choosing the point nearest to the centroid.

Problem: in our context we cannot evaluate centroid of each cluster because we only have item distances and not values.

Proposed approach: features

- a novel partitional method is proposed, 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. Our approach propose to adopt a multiple initialization procedure, where the best resulting partition is determined by a sort of Davies-Bouldin criterion.

The Algorithm

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 (3), (4), and (5), the resulting algorithm is the following:

- for t=1 to η
  - Initial cluster representatives θ_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 jth 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.: $p = \arg\min_{t=1,\ldots,\eta}\{\mathcal{DBL}^{(t)}\}$

Results

Data set

EEG data recorded by Zak Keirn at Purdue University; the dataset contains EEGs signal recorded from different subjects which were asked to perform five mental tasks:

- *baseline task*, for which the subjects were asked to relax as much as possible;
- *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;
- *letter task*, for which the subjects were instructed to mentally compose a letter to a friend without vocalizing;
- *geometric figure rotation*, for which the subjects were asked to visualize a particular 3D block figure being rotated about an axis;
- *visual counting task*, for which the subjects were asked to image a blackboard and to visualize numbers being written on the board sequentially.

Preprocessing and Segmentation

We applied the method on a segment-by-segment basis, $1\mathrm{s}$ signals sampled at $250\mathrm{Hz}$ 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).

Experimental

- 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; the 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 the following Table, with number of mental states growing from two to five.

Hierarchical Complete Link

	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%

- Accuracies are quite satisfactory. None of the method experimented can be considered the best one;
- nevertheless, measures (3) and (4) seem to be more effective.
- Therefore we applied the partitional algorithm to the same datasets setting the number of initializations $\eta = 5$ during all the experiments. Results are presented in the following Table:

Partitional DPAM Clustering

	BP	KL	SM
2 natural clusters	95.79%	96.32%	95.79%
3 natural clusters	75.44%	72.98%	65.61%
4 natural clusters	64.21%	62.04%	50.52%
5 natural clusters	57.04%	46.74%	44.80%

- in this last case the BP distance is overall slightly better than the others experimented measures.

Final comments:

- 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.

Conclusions

- We addressed the problem of unsupervised classification of sequences using an HMM approach.
- These models, very suitable in modelling sequential data, are used to characterize the similarity between sequences in different ways.
- We extend the Smyth's ideas 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.

†Address for correspondence: Department of Informatics, University of Verona, Strada Le Grazie n.15, I-37134 Verona, Italy