Similarity-Based Clustering of Sequences using Hidden Markov Models

Manuele Bicego[1], Vittorio Murino[1], and Mário A.T. Figueiredo[2]

[1] Dipartimento di Informatica, Università di Verona
Ca' Vignal 2, Strada Le Grazie 15, 37134 Verona, Italy
bicego,murino@sci.univr.it
[2] Instituto de Telecomunicações, Instituto Superior Técnico
1049-001 Lisboa, Portugal
mtf@lx.it.pt

Abstract. Hidden Markov models constitute a widely employed tool for sequential data modelling; nevertheless, their use in the clustering context has been poorly investigated. In this paper a novel scheme for HMM-based sequential data clustering is proposed, inspired on the similarity-based paradigm recently introduced in the supervised learning context. With this approach, a new representation space is built, in which each object is described by the vector of its similarities with respect to a predeterminate set of other objects. These similarities are determined using hidden Markov models. Clustering is then performed in such a space. By way of this, the difficult problem of clustering of sequences is thus transposed to a more manageable format, the clustering of points (vectors of features). Experimental evaluation on synthetic and real data shows that the proposed approach largely outperforms standard HMM clustering schemes.

1 Introduction

Unsupervised classification (or clustering) of data [1] is undoubtedly an interesting and challenging research area: it could be defined as the organization of a collection of patterns into groups, based on similarity. It is well known that data clustering is inherently a more difficult task if compared to supervised classification, in which classes are already identified, so that a system can be adequately trained. This intrinsic difficulty worsens if sequential data are considered: the structure of the underlying process is often difficult to infer, and typically different length sequences have to be dealt with. Clustering of sequences has assumed an increasing importance in recent years, due to its wide applicability in emergent contexts like data mining and DNA genome modelling and analysis.

[2] M. Figueiredo's work partially supported by the (Portuguese) Foundation for Science and Technology (FCT) under grant POSI/SRI/33143/1999.

Sequential data clustering methods could be generally classified into three categories: *proximity-based* methods, *feature-based* methods and *model-based* methods. In the *proximity-based* approaches, the main effort of the clustering process is in devising similarity or distance measures between sequences. With such measures, any standard distance-based method (as agglomerative clustering) can be applied. *Feature-based* methods extract a set of features from each individual data sequence that captures temporal information. The problem of sequence clustering is thus reduced to a more addressable point (vector of features) clustering. Finally, *model-based* approaches assume an analytical model for each cluster, and the aim of clustering is to find a set of such models that best fit the data. Examples of models that can be employed include time series models, spectral models, and finite state automata, as *hidden Markov models* (HMM) [2]. HMMs are a widely used tool for sequence modelling, whose importance has rapidly grown in the last decade. In the context of sequence clustering, HMMs have not been extensively used, and only a few papers can be found in the literature: the corresponding state of the art is presented in Section 2. The proposed approaches mainly fall into the first (proximity-based) and in the third (model-based) categories. In this paper, an alternative HMM clustering scheme is proposed, classifiable as belonging to the *feature-based* class, that extends the similarity-based paradigm [3–8]. This paradigm, which has been introduced recently for supervised classification purposes, differs from typical pattern recognition approaches where objects are represented by sets (vectors) of features. In the similarity-based paradigm, objects are described using pairwise (dis)similarities, i.e., distances from other objects in the data set. The state of the art of the similarity-based paradigm is reviewed in Section 2.

In this paper, we propose to extend this paradigm to the problem of clustering sequences, using a new feature space, where each sequence is characterized by its similarity to all other sequences. The problem is to find a suitable metric for measuring (dis)similarities between sequences, and, as shown in [9, 10], HMMs are a suitable tool for that purpose. In that space, clustering is then performed using some standard techniques: the difficult task of sequence clustering is thus transposed to a more manageable format, that of clustering points (vectors of features). Experimental evaluation on synthetic and real data shows that this approach largely outperforms standard HMM clustering schemes.

The rest of the paper is organized as follows: Section 2 summarizes the state of the art in HMM-based clustering of sequences and reviews the similarity-based paradigm. Section 3 reviews the fundamentals of hidden Markov models, while Section 4 details the proposed strategy. Experimental results are reported in Section 5. Finally, Section 6 is devoted to presenting conclusions and future work directions.

2 State of the art

2.1 HMM-Based Sequence Clustering

HMMs have not been extensively employed for clustering sequences, with only a few papers exploring this direction. More specifically, early approaches related to speech recognition were presented in [11–13]. All these methods belong to the proximity-based clustering class. HMMs were employed to compute similarities between sequences, using different approaches (see for example [10, 14]), and standard pairwise distance matrix-based approaches (as agglomerative hierarchical) were then used to obtain clustering. This strategy, which is considered the standard method for HMM-based clustering of sequences, is better detailed in the Section 3.1.

The first approach not directly linked to speech was presented by Smyth [9] (see also the more general and more recent [15]). This approach consists in two steps: first, it devises a pairwise distance between observed sequences, by computing a symmetrized similarity. This similarity is obtained by training 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 an LL matrix which is then used to cluster the sequences in K groups, using a hierarchical algorithm. In the second step, one HMM is trained for each cluster; the resulting K models are then merged into a "composite" global HMM, where each HMM is used to design a disjoint part of this "composite" model. This initial estimate is then refined using the standard Baum-Welch procedure. As a result, a global HMM modelling all the data is obtained. The number of clusters is selected using a cross-validation method. With respect to the above mentioned taxonomy, this approach can be classified as belonging to both the proximity-based class (a pairwise distance is derived to initialize the model) and the model-based class (a model for clustering data is finally obtained).

An example of an HMM-based method for sequence clustering is the one proposed in [16], where HMMs are used as cluster prototypes. The clustering is obtained by employing the *rival penalized competitive learning* (RPCL) algorithm [17] (a method originally developed for point clustering) together with a state merging strategy, aimed at finding smaller HMMs.

A relevant contribution to the model-based HMM clustering methodology was made by Li and Biswas [18–22]). Basically, in their approach [18], the clustering problem is addressed by focusing on the model selection issue, *i.e.* the search for the HMM topology best representing data, and the clustering structure issue, *i.e.* finding the most likely number of clusters. In [19], the former issue is addressed using the *Bayesian information criterion* [23], and extending to the continuous case the *Bayesian model merging* approach [24]. 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 [20], the same problems are addressed in terms of Bayesian model selection, using BIC [23], and the *Cheesman-Stutz* (CS) ap-

proximation [25]. A more comprehensive version of this paper has appeared in [22], where the method is also tested on real world ecological data. These clustering methodologies have been applied to specific domains, as physiology, ecology and social science, where the dynamic model structure is not readily available. Obtained results have been published in [21].

2.2 Similarity-based classification

The literature on similarity-based classification is not vast. Jain and Zongker [3] have obtained a dissimilarity measure, for a handwritten digit recognition problem, based on deformable templates; a multidimensional scaling approach was then used to project this dissimilarity space onto a low-dimensional space, where a 1-nearest-neighbor (1-NN) classifier was employed to classify new objects. In [4], Graepel *et al* investigate the problem of learning a classifier based on data represented in terms of their pairwise proximities, using an approach based on Vapnik's structural risk minimization [26]. Jacobs and Weinshall [5] have studied distance-based classification with non-metric distance functions (*i.e.*, that do not verify the triangle inequality). Duin and Pekalska are very active authors in this area[3] having recently produced several papers [6–8]. Motivation and basic features of similarity-based methods were first described in [6]; it was shown, by experiments in two real applications, that a Bayesian classifier (the RLNC - regularized linear normal density-based classifier) in the dissimilarity space outperforms the nearest neighbor rule. These aspects were more thoroughly investigated in [8], where other classifiers in the dissimilarity space were studied, namely on digit recognition and bioinformatics problems. Finally, in [7], a generalized kernel approach was introduced, dealing with classification aspects of the dissimilarity kernels.

3 Hidden Markov Models

A discrete-time hidden Markov model $\boldsymbol{\lambda}$ can be viewed as a Markov model whose states are not directly observed: instead, each state is characterized by a probability distribution function, modelling the observations corresponding to that state. More formally, an HMM is defined by the following entities [2]:

- $S = \{S_1, S_2, \cdots, S_N\}$ the finite set of possible (hidden) states;
- the transition matrix $\mathbf{A} = \{a_{ij},\ 1 \leq j \leq N\}$ representing the probability of moving from state S_i to state S_j,

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

  with $a_{ij} \geq 0$, $\sum_{j=1}^{N} a_{ij} = 1$, and where q_t denotes the state occupied by the model at time t.

[3] See `http://www.ph.tn.tudelft.nl/Research/neural/index.html`

– the emission matrix $\mathbf{B} = \{b(o|S_j)\}$, indicating the probability of emission of symbol $o \in V$ when system state is S_j; V can be a discrete alphabet or a continuous set (e.g. $V = I\!R$), in which case $b(o|S_j)$ is a probability density function.
– $\boldsymbol{\pi} = \{\pi_i\}$, the initial state probability distribution,

$$\pi_i = P[q_1 = S_i], \quad 1 \leq i \leq N$$

with $\pi_i \geq 0$ and $\sum_{i=1}^{N} \pi_i = 1$.

For convenience, we represent an HMM by a triplet $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$.

Learning the HMM parameters, given a set of observed sequences $\{\mathbf{O}_i\}$, is usually performed using the well-known Baum-Welch algorithm [2], which is able to determine the parameters maximizing the likelihood $P(\{\mathbf{O}_i\}|\boldsymbol{\lambda})$. One of the steps of the Baum-Welch algorithm is an evaluation step, where it is required to compute $P(\mathbf{O}|\boldsymbol{\lambda})$, given a model $\boldsymbol{\lambda}$ and a sequence $\mathbf{O}$; this can be computed using the *forward-backward procedure* [2].

3.1 Standard HMM-based clustering of sequences

The standard proximity-based method for clustering sequences using HMMs can be sumarized by the following algorithm. Consider a given a set of N sequences $\{\mathbf{O}_1 ... \mathbf{O}_N\}$ to be clustered; the algorithm performs the following steps:

1. Train one HMM $\boldsymbol{\lambda}_i$ for each sequence $\mathbf{O}_i$.
2. Compute the distance matrix $D = \{D(\mathbf{O}_i, \mathbf{O}_j)\}$, representing a similarity measure between sequences or between models; this is typically obtained from the forward probability $P(\mathbf{O}_j|\boldsymbol{\lambda}_i)$, or by devising a measure of distances between models. In the past, few authors have proposed approaches to computing these distances: early approaches were based on the Euclidean distance of the discrete observation probability, others on entropy, or on co-emission probability of two models, or, very recently, on the Bayes probability of error (see [14] and the references therein).
3. Use a pairwise distance-matrix-based method (*e.g.*, an agglomerative method) to perform the clustering.

4 Proposed Strategy

The idea at the basis of the proposed approach is conceptually simple: to build a new representation space, using the similarity values between sequences obtained via the HMMs, and to perform the clustering in that space. Similarity values allow discrimination, since this quantity is high for similar objects/sequences, i.e., belonging to the same group, and low for objects of different clusters. Therefore, we can interpret the similarity measure $\mathcal{D}(\mathbf{O}, \mathbf{O}_i)$ between a sequence $\mathbf{O}$ and another "reference" sequence $\mathbf{O}_i$ as a "feature" of the sequence $\mathbf{O}$. This fact

suggests the construction of a feature vector for $\mathbf{O}$ by taking the similarities between $\mathbf{O}$ and a set of reference sequences $\mathcal{R} = \{\mathbf{O}_k\}$, so that $\mathbf{O}$ is characterized by a *pattern* (*i.e.*, a set of features) $\{\mathcal{D}(\mathbf{O}, \mathbf{O}_k),\ \mathbf{O}_k \in \mathcal{R}\}$.

More formally, given a set of sequences $\mathcal{T} = \{\mathbf{O}^1 ... \mathbf{O}^N\}$ to be clustered, the proposed approach can be briefly described as follows:

- let $\mathcal{R} = \{\mathbf{P}_1, ..., \mathbf{P}_R\}$ be a set of R "reference" or "representative" objects; these objects may belong to the set of sequences ($\mathcal{R} \subseteq \mathcal{T}$) or may be otherwise defined. In a basic case it could be $\mathcal{R} = \mathcal{T}$.
- train one HMM $\boldsymbol{\lambda}_r$ for each sequence $\mathbf{P}_r \in \mathcal{R}$;
- represent each sequence $\mathbf{O}_i$ of the data set by the set of similarities $\mathcal{D}_{\mathcal{R}}(\mathbf{O}_i)$ to the elements of the representative set $\mathcal{R}$, computed with the HMMs $\boldsymbol{\lambda}_1 ... \boldsymbol{\lambda}_R$ as:

$$\mathcal{D}_{\mathcal{R}}(\mathbf{O}_i) = \begin{bmatrix} \mathcal{D}(\mathbf{O}_i, \mathbf{P}_1) \\ \mathcal{D}(\mathbf{O}_i, \mathbf{P}_2) \\ \vdots \\ \mathcal{D}(\mathbf{O}_i, \mathbf{P}_R) \end{bmatrix} = \frac{1}{T_i} \begin{bmatrix} \log P(\mathbf{O}_i|\boldsymbol{\lambda}_1) \\ \log P(\mathbf{O}_i|\boldsymbol{\lambda}_2) \\ \vdots \\ \log P(\mathbf{O}_i|\boldsymbol{\lambda}_R) \end{bmatrix} \quad (1)$$

  where T_i is the length of the sequence $\mathbf{O}_i$.
- perform clustering in $\mathbb{R}^{|\mathcal{R}|}$, where $|\mathcal{R}|$ denotes the cardinality of $\mathcal{R}$, using any general technique (not necessarily hierarchical) appropriate for clustering points in an Euclidean space.

In the simplest case, the representative set $\mathcal{R}$ is the whole data set $\mathcal{T}$, resulting in a similarity space of dimensionality N. Even if computationally heavy for large data sets, it is interesting to analyze the discriminative power of such a space.

5 Experimental results

In this section, the proposed technique is compared with the standard HMM clustering scheme presented in Section 3. Once the likelihood similarity matrix is obtained, clustering (step 3) is performed by using three algorithms:

- two variants of the agglomerative hierarchical clustering techniques: the *complete link scheme*, and the *Ward scheme* [1].
- a non parametric, pairwise distance-based clustering technique, called *clustering by friends* [27]: this technique produces a partition of the data using only the similarity matrix. The partition is obtained by iteratively applying a two-step transformation to the proximity matrix. The first step of the transformation represents each point by its relation to all other data points, and the second step re-estimates the pairwise distances using a proximity measure on these representations. Using these transformations, the algorithm partitions the data into two clusters. To partition the data into more than two clusters, the method has to be applied several times, recursively.

Regarding the proposed approach, after obtaining the similarity representation with $\mathcal{R} = \mathcal{T}$ (*i.e.* by using all sequences as representatives), we have used three clustering algorithms:

– again the hierarchical agglomerative complete link and Ward methods, where distance is the Euclidean metrics in the similarity space: this is performed to compare the two representations with the same algorithms;
– standard K-means algorithm [1].

Clustering accuracies were measured on synthetic and real data. Regarding the synthetic case, we consider a 3-class problem, where sequences were generated from the three HMMs defined in Fig. 1. The data set is composed of 30 sequences

$$\mathbf{A} = \begin{array}{|c|c|c|}\hline 1/3 & 1/3 & 1/3 \\ \hline 1/3 & 1/3 & 1/3 \\ \hline 1/3 & 1/3 & 1/3 \\ \hline \end{array} \quad \boldsymbol{\pi} = \begin{array}{|c|}\hline 1/3 \\ \hline 1/3 \\ \hline 1/3 \\ \hline \end{array} \quad \mathbf{B} = \begin{array}{|c|c|}\hline \mu_1 = 1 & \sigma_1^2 = 0.6 \\ \hline \mu_2 = 3 & \sigma_2^2 = 0.6 \\ \hline \mu_3 = 5 & \sigma_3^2 = 0.6 \\ \hline \end{array}$$
(a)

$$\mathbf{A} = \begin{array}{|c|c|c|}\hline 1/3 & 1/3 & 1/3 \\ \hline 1/3 & 1/3 & 1/3 \\ \hline 1/3 & 1/3 & 1/3 \\ \hline \end{array} \quad \boldsymbol{\pi} = \begin{array}{|c|}\hline 1/3 \\ \hline 1/3 \\ \hline 1/3 \\ \hline \end{array} \quad \mathbf{B} = \begin{array}{|c|c|}\hline \mu_1 = 1 & \sigma_1^2 = 0.5 \\ \hline \mu_2 = 3 & \sigma_2^2 = 0.5 \\ \hline \mu_3 = 5 & \sigma_3^2 = 0.5 \\ \hline \end{array}$$
(b)

$$\mathbf{A} = \begin{array}{|c|c|c|}\hline 1/3 & 1/3 & 1/3 \\ \hline 1/3 & 1/3 & 1/3 \\ \hline 1/3 & 1/3 & 1/3 \\ \hline \end{array} \quad \boldsymbol{\pi} = \begin{array}{|c|}\hline 1/3 \\ \hline 1/3 \\ \hline 1/3 \\ \hline \end{array} \quad \mathbf{B} = \begin{array}{|c|c|}\hline \mu_1 = 1 & \sigma_1^2 = 0.4 \\ \hline \mu_2 = 3 & \sigma_2^2 = 0.4 \\ \hline \mu_3 = 5 & \sigma_3^2 = 0.4 \\ \hline \end{array}$$
(c)

Fig. 1. Generative HMMs for synthetic data testing: $\mathbf{A}$ is the transition matrix, $\boldsymbol{\pi}$ is the initial state probability, and $\mathbf{B}$ contains the parameters of the emission density (Gaussians with the indicated means and variances).

(of length 400) from each of the three classes; the dimensionality of the similarity vectors is thus $N = 90$. Notice that this clustering task is not easy, as the three HMMs are very similar to each other, only differing slightly in the variances of the emission densities. The accuracy of clustering can be quantitatively assessed, by computing the number of errors: a clustering error occurs if a sequence is assigned to a cluster in which the majority of the sequences are from another class. Results are presented in Table 1, averaged over 10 repetitions. From this table it is possible to notice that the proposed methodology largely outperforms standard clustering approaches: the best performing algorithm is the partitional k-means on the similarity space, which produces an almost perfect clustering. In order to have a better insight into the discriminative power of the proposed feature space, we also computed the supervised classification results on this synthetic example. Decisions were taken using the standard *maximum likelihood* (ML) approach, where an unknown sequence is assigned to the class whose model shows the highest likelihood. Note that this classification scheme does not make use of the similarity space introduced in this paper, and represents the supervised

Table 1. Clustering results on synthetic experiments.

Standard classification	
ML classification	94.78%
Standard clustering	
Aggl. complete link	64.89%
Aggl. Ward	71.33%
Clus. by Friends	70.11%
Clustering on similarity space $\mathcal{S}_T$	
Aggl. complete link	95.44%
Aggl. Ward	97.89%
k-means	98.33%

counterpart of the standard clustering approach proposed in Section 3.1. The classification error is computed using the standard *leave one out* (LOO) scheme [28]. It is important to note that clustering results in the similarity space are better than the classification results, confirming the high discrimination ability of the similarity space.

The real data experiment regards 2D shape recognition, where shapes were modelled as proposed in [29]; briefly, object contours are described using curvature, and these curvature sequences are modelled using HMMs with Gaussian mixtures as emission probabilities. The object database used is the one from Sebastian *et al.* [30], and is shown in Fig. 2. In this case, only the number of clusters is known. The clustering algorithms try to group the shapes into different clusters, based on their similarity. Results, averaged over 10 repetitions, are presented in Table 2. From these tables it is evident that the proposed representation permits greater discrimination, resulting in a increasing of the clustering accuracies. Also in this case, the ML classification accuracy was computed, using

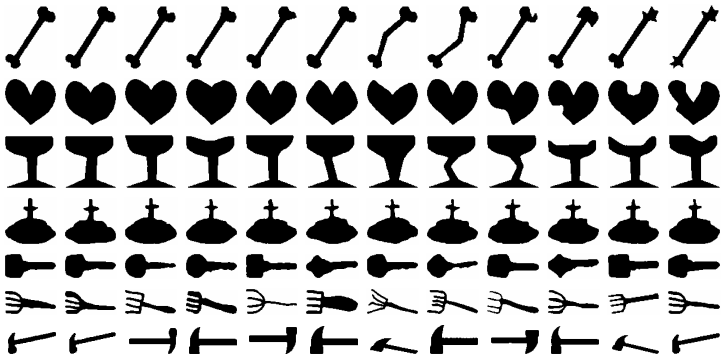

Fig. 2. Objects set used for testing.

Table 2. Clustering results on real experiments.

Standard classification	
ML classification	81.55%
Standard clustering	
Aggl. complete link	78.69%
Aggl. Ward	22.86%
Clus. by Friends	70.0%
Clustering on the similarity space $\mathcal{S}_\mathcal{T}$	
Aggl. complete link	63.10%
Aggl. Ward	77.62%
k-means	88.21%

the LOO scheme. From table 2 it is possible to note that the clustering results are better than the classification performances, confirming the high discriminative potentiality of the proposed similarity space.

6 Conclusions

In this paper, a scheme for sequence clustering, based on hidden Markov modelling and the similarity-based paradigm, was proposed. The approach builds features in which each sequence is represented by the vector of its similarities to a predefined set of reference sequences. A standard point clustering method is then performed on those representations. As a consequence, the difficult process of clustering sequences is cast into a simpler problem of clustering points, for which well established techniques have been proposed. Experimental evaluation on synthetic and real problems has shown that the proposed approach largely outperforms the standard HMM-based clustering approaches.

The main drawback of this approach is the high dimensionality of the resulting feature space, which is equal to the cardinality of the data set. This is obviously a problem, and represents a central topic for future investigation. We have previously addressed this issue in the context of similarity-based supervised learning [31]. In this unsupervised context, one idea could be to use some linear reduction techniques, in order to reduce the dimensionality of the space. Another idea is to directly address the problem of adequately choosing the representatives: this problem could be casted in the context of feature selection for unsupervised where the prototypes to be chosen are the features to be selected.

References

1. Jain, A., Dubes, R.: Algorithms for clustering data. Prentice Hall (1988)
2. Rabiner, L.: A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc. of IEEE **77** (1989) 257–286

3. Jain, A., Zongker, D.: Representation and recognition of handwritten digits using deformable templates. IEEE Trans. Pattern Analysis and Machine Intelligence **19** (1997) 1386–1391
4. Graepel, T., Herbrich, R., Bollmann-Sdorra, P., Obermayer, K.: Classification on pairwise proximity data. In M. Kearns, S. Solla, D.C., ed.: Advances in Neural Information Processing. Volume 11., MIT Press (1999)
5. Jacobs, D., Weinshall, D.: Classification with nonmetric distances: Image retrieval and class representation. IEEE Trans. Pattern Analysis and Machine Intelligence **22** (2000) 583–600
6. Pekalska, E., Duin, R.: Automatic pattern recognition by similarity representations. Electronics Letters **37** (2001) 159–160
7. Pekalska, E., Paclik, P., Duin, R.: A generalized kernel approach to dissimilarity-based classification. Journal of Machine Learning Research **2** (2002) 175–211
8. Pekalska, E., Duin, R.: Dissimilarity representations allow for building good classifiers. Pattern Recognition Letters **23** (2002) 943–956
9. Smyth, P.: Clustering sequences with hidden Markov models. In Mozer, M., Jordan, M., Petsche, T., eds.: Advances in Neural Information Processing. Volume 9., MIT Press (1997)
10. Panuccio, A., Bicego, M., Murino, V.: A Hidden Markov Model-based approach to sequential data clustering. In Caelli, T., Amin, A., Duin, R., Kamel, M., de Ridder, D., eds.: Structural, Syntactic and Statistical Pattern Recognition. LNCS 2396, Springer (2002) 734–742
11. Rabiner, L., Lee, C., Juang, B., Wilpon, J.: HMM clustering for connected word recognition. In: Proc. of IEEE ICASSP. (1989) 405–408
12. Lee, K.: Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition. IEEE Transactions on Acoustics, Speech and Signal Processing **38** (1990) 599–609
13. Kosaka, T., Matsunaga, S., Kuraoka, M.: Speaker-independent phone modeling based on speaker-dependent hmm's composition and clustering. In: Int. Proc. on Acoustics, Speech, and Signal Processing. Volume 1. (1995) 441–444
14. Bahlmann, C., Burkhardt, H.: Measuring hmm similarity with the bayes probability of error and its application to online handwriting recognition. In: Proc. Int. Conf. Document Analysis and Recognition. (2001) 406–411
15. Cadez, I., Gaffney, S., Smyth, P.: A general probabilistic framework for clustering individuals. In: Proc. of ACM SIGKDD 2000. (2000)
16. Law, M., Kwok, J.: Rival penalized competitive learning for model-based sequence. In: Proc. Int. Conf. Pattern Recognition. Volume 2. (2000) 195–198
17. Xu, L., Krzyzak, A., Oja, E.: Rival penalized competitive learning for clustering analysis, RBF nets, and curve detection. IEEE Trans. on Neural Networks **4** (1993) 636–648
18. Li, C.: A Bayesian Approach to Temporal Data Clustering using Hidden Markov Model Methodology. PhD thesis, Vanderbilt University (2000)
19. Li, C., Biswas, G.: Clustering sequence data using hidden Markov model representation. In: Proc. of SPIE'99 Conf. on Data Mining and Knowledge Discovery: Theory, Tools, and Technology. (1999) 14–21
20. Li, C., Biswas, G.: A bayesian approach to temporal data clustering using hidden Markov models. In: Proc. Int. Conf. on Machine Learning. (2000) 543–550
21. Li, C., Biswas, G.: Applying the Hidden Markov Model methodology for unsupervised learning of temporal data. Int. Journal of Knowledge-based Intelligent Engineering Systems **6** (2002) 152–160

22. Li, C., Biswas, G., Dale, M., Dale, P.: Matryoshka: A HMM based temporal data clustering methodology for modeling system dynamics. Intelligent Data Analysis Journal **in press** (2002)
23. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics **6** (1978) 461–464
24. Stolcke, A., Omohundro, S.: Hidden Markov Model induction by Bayesian model merging. In Hanson, S., Cowan, J., Giles, C., eds.: Advances in Neural Information Processing Systems. Volume 5., Morgan Kaufmann, San Mateo, CA (1993) 11–18
25. Cheeseman, P., Stutz, J.: Bayesian classification (autoclass): Theory and results. In: Advances in Knowledge discovery and data mining. (1996) 153–180
26. Vapnik, V.: Statistical Learning Theory. John Wiley, New York (1998)
27. Dubnov, S., El-Yaniv, R., Gdalyahu, Y., Schneidman, E., Tishby, N., Yona, G.: A new nonparametric pairwise clustering algorithm based on iterative estimation of distance profiles. Machine Learning **47** (2002) 35–61
28. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Academic Press (1999)
29. Bicego, M., Murino, V.: Investigating Hidden Markov Models' capabilities in 2D shape classification. Submitted for publication (2002)
30. Sebastian, T., Klein, P., Kimia, B.: Recognition of shapes by editing Shock Graphs. In: Proc. Int Conf. Computer Vision. (2001) 755–762
31. Bicego, M., Murino, V., Figueiredo, M.: Similarity-based classification of sequences using hidden Markov models (2002) Submitted for publication.