Clinical Pathway Analysis Using Graph-Based Approach and Markov Models

Haytham Elghazel[1], Véronique Deslandres[1], Kassem Kallel[2], Alain Dussauchoy[1]
Université de Lyon, Lyon, F-69003, France ; université Lyon 1, EA4125, LIESP, Villeurbanne, F-69622, France
[1] {elghazel, deslandres, dussauchoy}@bat710.univ-lyon1.fr
[2] kassem.kallel@gmail.com

Abstract

Cluster analysis is one of the most important aspects in the data mining process for discovering groups and identifying interesting distributions or patterns over the considered data sets. A new method for sequences clustering and prediction is presented in this paper, which is based on a hybrid model that uses our b-coloring based clustering approach as well as Markov chain models. The paper focuses on clinical pathway analysis but the method applies to every kind of sequences, and a generic decision support framework has been developed for managers and experts. The interesting result is that the clusters obtained have a twofold representation. Firstly, there is a set of dominant sequences which reflects the properties of the cluster and also guarantees that clusters are well separated within the partition. On the other hand, the behavior of each cluster is governed by a finite-state Markov chain model which allows probabilistic prediction. These models can be used for predicting possible paths for a new patient, and for helping medical professionals to eventually react to exceptions during the clinical process.

1. Context and Motivations

In French hospitals, the Diagnosis Related Groups (DRG) system was introduced in the eighties with the new medical information system provision called PMSI. DRG classification is a national decision and provides a certain unity in economic (cost similarity) and medical terms. Hence PMSI data can also be used for the assessment of both public and private hospital performances. Each hospital stay leads to one standard uniform discharge summary (SDS). The SDS contains data related to the nature of treatments, medical exams and diagnosis as well as patient's personal information. The SDS is then identified to correspond to one patient's group called the *Diagnosis Related Group* (*DRG*). This operation is performed using a supervised approach according to a decision tree.

The PMSI system, which involves both economic and medical data, offers a good starting point for designing a suitable strategic information system according to healthcare quality and cost control. Due to the prevalence of information technology in medical care, data collected from PMSI can be used for discovering useful patterns and especially for clinical pathway. These patterns can be analyzed automatically or by medical professionals in order to develop better strategies to improve the quality of medical treatments.

Indeed, the numerous medical activities, and therapeutic interventions performed by medical professionals and collected from PMSI may be considered as a medical histories. These records are valuable sources for anticipating future trends of clinical pathways, in order to facilitate resources and activities planning and management within French health centers.

This paper reports the work in adopting the Markov Models and a recently proposed *b-coloring based clustering approach* [1] for discovering a typology of clinical pathways in the French medical information system.

The key idea of this paper is to formulate a model-based framework for clustering a variety of clinical pathways. Each clinical pathway is represented as a sequence of stays and has therefore a different length. For our concern, each hospital stay is characterized by two qualitative items: the *Diagnosis Related Group* and the *Principal Medical Diagnosis*. The population of clinical pathway will be divided into k clusters using the *b-coloring based clustering method* in [1]. The later enables to build a fine partition of the data set even if the number of clusters k is not specified in advance. The behavior of each cluster is then governed by a finite-state Markov chain model which provides a probabilistic generative model for pathways peculiar to that group. These models can be used for predicting possible paths for new arrival of patients, and can also support medical people when faced with exceptions.

1-4244-1476-8/07/$25.00 ©2007 IEEE.

Section 2 introduces the theory of sequential data clustering. Section 3 is devoted to describe a novel framework for clinical pathways, and explain how the Markov Models was adopted to represent clinical pathways behavior. Some experiments using real PMSI data set are presented in Section 4. We summarize our contribution and give future research in Section 5.

2. Related Work

Clustering of sequences is the division of a collection of time series (or sequences) into groups of similar objects [2]. In clustering, some details are disregarded for data simplification purposes. Clustering can be viewed as a data modeling technique that provides an attractive mechanism to automatically find the hidden structure of large data sets. Sequence clustering is therefore related to many disciplines and plays an important role in a broad range of applications. Specific applications are: clustering users based on their web navigation patterns [3], clustering patients based on Red Blood Cell Cytograms [4], and clustering biological sequences [4]. The current paper deals with the problem of clustering patients based on their clinical pathways.

There are a variety of methods for clustering sequences. The most widespread are model-based sequence clustering methods [4,5]. Among them are the Learning Mixture Markov models that constitute the most popular tool for clustering time series. Given the number k of clusters, Cadez *et al.* [4] propose a unifying probabilistic framework for clustering individuals into k clusters, when the available data measurements are not multivariate vectors of fixed dimensionality. They provide a general Expectation-Maximization algorithm (EM) [6] for clustering such data set and demonstrate its usefulness on three applications. The key idea in [4] is the fact that each cluster in the partition is depicted by a Markov chain model and the EM approach is used to generate parameter estimates (the initial state probability vector and the transition matrix for each Markov cluster) so that cluster models are constructed in a straightforward and consistent manner.

Another probabilistic model-based approach for clustering sequences is proposed in [5] using a mixture of Hidden Markov Models (HMM). A real advantage of this method is that the behavior of each cluster of the partition is governed by a HMM. HMM is a powerful stochastic method for modeling sequential or time-series data, and has been successfully used in many tasks such as speech recognition, DNA sequence analysis and information extraction from text data. In HMM, each state is characterized by two probability distributions: the transition distribution over states and the emission distribution over the output symbols. In such a model, a random state generates a sequence of output symbols as follows: at each step, the state emits an output symbol according to its emission distribution; it goes to a next state according to the transition distribution. Since the activity of the state is observed indirectly, through the sequence of output symbols, and the sequence of states is not directly observable, the states are said to be hidden [7].

While probabilistic clustering approaches have some important features: the mixture model can be used to assign sequences to clusters (*online property)*, and the given clusters are easily interpretable. Nevertheless, the induction algorithm suffers from some weaknesses:

- The number of clusters k in both approaches must be set beforehand;
- For the Hidden Markov Model based approach, the number of hidden states must be chosen before the model is fit;
- Both approaches are based on an initialization step which gives the input partitioning to the training process (EM). This initialization step is generally a random operation and the algorithm converges to a local maximum.

The other broad class in sequence clustering uses similarity measures to compare sequences. In most applications involving determination of similarity between pairs of sequences, the path lengths may be different making difficult to embed the time series in a metric space and/or calculate distances between corresponding elements of the sequences. This brings us to the second aspect of sequence matching, namely, sequence alignment. Two of the well-known similarity functions between a pair of sequences are: *Longest Common Subsequence* (LCS) sequences [8], and *Dynamic Time Warping* (DTW) [9]. They are systematic and efficient methods based on dynamic programming that identifies which correspondence among feature vectors of two sequences is best when scoring the similarity between them.

When two sequences S_1 and S_2 consist of symbolic or discrete data, we have to consider both sequences as a strings and the similarity is determined by the application of *Longest Common Subsequence* algorithm [8]. The problem is to find the longest string c such that for S_1 and S_2, the characters of c appear as a scattered subsequence of S_1 and S_2 (*i.e.* in both sequences and in the same order). The similarity between S_1 and S_2 is given by the length of c. A particularly important application is in finding a consensus among DNA sequences.

On the other hand, given two time series of continuous values S_1 and S_2, DTW [9] finds the warping of the time dimension in S_1 that minimizes the difference between the two series. It was used extensively in speech recognition, a domain in which the time series are notoriously complex and noisy.

The proximity (similarity) measure between sequences is then computed, a variety of based-similarity techniques for grouping sequences can be found in literature [10]. Intuitively, sequences within a stable cluster are more similar to each other than they are to a sequence from a different cluster.

Hence, given a number of clusters k, the similarity-based clustering approaches try to discover a partition such that the sequences within the same cluster are similar to each other (*intracluster cohesion*) while sequences from different clusters are dissimilar (*intercluster separation*). Nevertheless, clusters of the partition are not easily interpretable. Indeed, most of these methods fail to give a clear idea about the *relationships* between sequences of clusters. On the other hand, the methods cannot usually provide at least one representative member for each cluster (except centers of mass for each cluster). Consequently, it is not easy how the methods allow incremental clustering, ie classifying a new sequence in the partition, and how the methods facilitate prediction-making. It seems that in many applications that involve classification decision-making, clusters of the partition have to be described in a compact form for data abstraction purposes.

In order to alleviate the problems of both probabilistic model-based as well as similarity-based approaches, Oates *et al.* [7] have proposed to use a hybrid sequences clustering algorithm that uses dynamic time warping and hidden Markov model induction. In this approach, DTW and HMM methods complement each other: DTW produces a rough initial clustering and the HMM dynamically redeploys the sequences that do not belong to suitable clusters. The downside is that the HMM may transfer some good sequences along with the bad ones.

3. A Clinical Pathway Analysis framework

The current paper continues the challenge to mix both kinds of sequence clustering approaches. We propose a new framework for clustering time series (clinical pathways here) that exploit the real advantages of both probabilistic as well as similarity-based clustering algorithms.

Our approach formulates sequence clustering problem as a kind of graph partitioning problem in a weighted linkage graph $G = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V} = \{v_1,v_2,...,v_n\}$ is the vertex set which correspond to the sequences set $\boldsymbol{S} = \{S_1,S_2,...,S_n\}$ (clinical pathways), and $\boldsymbol{E} = \boldsymbol{V} \times \boldsymbol{V}$ is the edge set which correspond to higher dissimilarities than given threshold θ and are weighted by their dissimilarities. The graph G is traditionally represented with the corresponding weighted dissimilarity matrix, which is the $n \times n$ symmetric matrix $D=\{d_{i,j} | S_i,S_j \in \boldsymbol{S}\}$.

3.1. Problem and data description

The treated clinical pathways are formulated as a sequence of past stays characterized each one by a 2-dimensionnal vector (DRG;DP i.e. : *Diagnosis Related Group* and *Principal Medical Diagnosis*). As proposed by our medical experts, this information is considered in order to give a good idea on the clinical stay: the DRG represents both economic and medical aspect of the stay, although DP gives the best information about the medical state of the patient. The following Table shows an example of such data.

Table 1: An example of patients characterized by clinical pathways

$\mathbf{S_1}$	DRG series	DRG_{12}	DRG_7	DRG_7	DRG_{13}
	DP series	DP_1	DP_6	DP_5	DP_3
$\mathbf{S_2}$	DRG series	DRG_8	DRG_{11}	DRG_{12}	
	DP series	DP_2	DP_{33}	DP_1	
$\mathbf{S_3}$	DRG series	DRG_1	DRG_1	DRG_7	DRG_{23}
	DP series	DP_{21}	DP_{21}	DP_6	DP_2

3.2. Dissimilarity level

The dissimilarity measure D between sequences S_i and S_j is given by the *Euclidian distance* as follow:

$$d_{i,j} = D(S_i,S_j) = \sqrt{\sum_{k=1}^{2} \left(g_k\left(S_{i,k},S_{j,k}\right)\right)^2} \qquad \textbf{(1)}$$

where g_k ($k \in \{1,2\}$) is the *modified edit distance* between both attribute-series $S_{i,k}$ and $S_{j,k}$ (k=1 for DRG and k=2 for DP) corresponding respectively to the sequences S_i and S_j. The *edit distance* is related to the LCS problem. Our interest in LCS come from the LCS is a special case of the Dynamic Time Warping (DTW) algorithm. Therefore LCS inherits all the DTW features.

The *original edit distance* given in eq.(2) suffers from some limitations due to the different lengths of sequences and cannot be meaningfully used in our application. The following example illustrates this point. Consider three sequences X, Y and Z. Suppose these pathways are labels (*abab*, *cdcd*, *abefghij*). So, the distance between X and Z (which have two symbol a and b in common) should be smaller than that between X and Y (without any symbol in common). However, the *original edit distance* (eq.(2)) has a value 8 between X and Y (4+4-2*0=8) and also between X and Z (4+8-2*2=8).

$$g_k\left(S_{i,k},S_{j,k}\right) = \left|S_{i,k}\right| + \left|S_{j,k}\right| - 2 * \left|LCS\left(S_{i,k},S_{j,k}\right)\right| \qquad \textbf{(2)}$$

where $\left|S_{i,k}\right|$ is the length of attribute-series $S_{i,k}$ and $\left|LCS\left(S_{i,k},S_{j,k}\right)\right|$ is the length of the *Longest Common Subsequence* between both attribute-series $S_{i,k}$ and $S_{j,k}$.

This limitation motivates the definition of a *modified edit distance* that correctly captures the similarity between the sequences. The dissimilarity

between two sequences is defined as the ratio of the *original edit distance* to the sum of the length of both sequences (*cf.* eq.(3)). The minimum and maximum values of this distance measure are 0.0 and 1.0, respectively.

$$g_k(S_{i,k}, S_{j,k}) = \frac{|S_{i,k}| + |S_{j,k}| - 2 * |LCS(S_{i,k}, S_{j,k})|}{|S_{i,k}| + |S_{j,k}|} \quad (3)$$

3.2. Clinical Pathways Clustering

The sequences to be clustered $\boldsymbol{S} = \{S_1, S_2, \ldots, S_n\}$ are now depicted by a *non-complete edge-weighted graph* $G=(\boldsymbol{V},\boldsymbol{E})$. In order to divide the vertex set $\boldsymbol{V}$ into a partition $\boldsymbol{P} = \{C_1, C_2, \ldots, C_k\}$ where for $\forall C_i, C_j \in \boldsymbol{P}$, $C_i \cap C_j = \varnothing$ *for* $i \neq j$ (when the number of clusters k is not pre-defined), our new clustering approach [1], based on *graph b-coloring* is applied to the graph G. A *graph b-coloring* is the assignment of colors (clusters) to the vertices of the graph such that: (i) no two adjacent vertices have the same color (*proper coloring*), (ii) for each color there exists at least one *dominating vertex* which is adjacent to all the other colors.

The *graph-based clustering algorithm* [1] performs on two steps: 1) initializing the colors of vertices of G with maximum number of colors, and 2) removing colors without any dominating vertex using a *greedy procedure*.

An advantage of this method is that it offers a real representation of each cluster by a set of *dominant sequences* which reflects the properties of the cluster and also guarantees that the cluster has a distinct separation from other clusters of the partition.

As an illustration, let suppose $\{S_1, S_2, S_3, S_4, S_5, S_6\}$ the clinical pathways data set related to the weighted dissimilarity Matrix D in Table 2. Figure 1 shows the *superior threshold graph* θ =0.15 for Table 2. The edges are labeled with the corresponding dissimilarities. Therefore, using the *b-coloring* of $G_{>0.15}$ (*cf.* Fig.2), the sequences set is divided into three clusters, namely: $\{\boldsymbol{S_1}, \boldsymbol{S_3}\}$, $\{\boldsymbol{S_2}, S_6\}$ and $\{\boldsymbol{S_4}, S_5\}$. The vertices with the same color (shape) are grouped into the same cluster and the nodes with bold letter are the dominating vertices.

Table 2. A dissimilarity Matrix

$\boldsymbol{S_i}$	S_1	S_2	S_3	S_4	S_5	S_6
S_1	0					
S_2	0.20	0				
S_3	0.10	0.20	0			
S_4	0.20	0.20	0.25	0		
S_5	0.10	0.20	0.10	0.05	0	
S_6	0.40	0.075	0.15	0.15	0.15	0

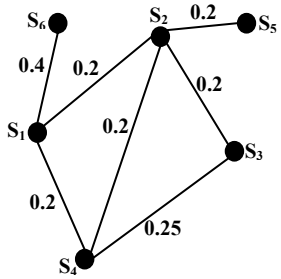

Fig. 1. ***Superior threshold graph*** $G_{>0.15}$ **(**θ **=0.15)**

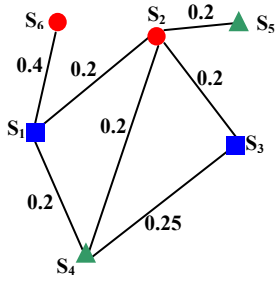

Fig. 2. ***b-coloring*** **of graph** $G_{>0.15}$ **: three classes are identified**

The clustering algorithm is iterative and performs multiple runs, each of them increasing the value of the dissimilarity threshold θ (selected from the dissimilarity matrix D). Once all threshold values passed, the algorithm provides the optimal partitioning (corresponding to one threshold value θ_o) which maximizes *Dunn's generalized index* ($Dunn_G$) [11]. $Dunn_G$ is designed to offer a compromise between the *intercluster separation* and the *intracluster cohesion*. So, it is the more appropriated to partition data set in *compact* and *well-separated* clusters. As an illustration, successive threshold graphs are constructed for each threshold value θ selected from the dissimilarity matrix in Table 1, and our approach is used to give the *b-coloring* partition of each corresponding threshold graph. The value of the *Dunn's generalized index* is computed for the each partition obtained. We conclude that the partition θ=0.15 has the maximal $Dunn_G$ among other ones with different θ.

A second part of our framework gives a model-representation for the k clusters of the partition. Indeed, each cluster c, $1 \leq c \leq k$, is represented by a finite-state Markov chain model which provides a probabilistic generative model for pathways from that group. Each *Markov cluster* c has a data-generating model with parameters Φ_c (the initial state probabilities $\pi_c(s_i)$ and m *x* m transition matrix $a_c(s_i,s_j)$ for that cluster, where $1 \leq s_i, s_j \leq m$ denote discrete states as the different pairs (*DRG;DP*) in the sequences set).

In other words, $\pi_c(s_i)$ is the probability of starting in state s_i in all the sequences grouped within the cluster c as follows:

$$\pi_c(s_i) = \frac{\text{\# sequences starting in state } s_i \text{ within } c}{\text{\# sequences within } c} \quad \textbf{(4)}$$

$a_c(s_i,s_j)$ is the probability of going from state s_i to state s_j in all the sequences associated with the cluster c.

$$a_c(s_i, s_j) = \frac{\text{\# transitions from state } s_i \text{ to state } s_j \text{ within } c}{\text{\# transitions from state } s_i \text{ within } c} \quad \textbf{(5)}$$

The population of clinical pathways is now divided into k Markov clusters. Representing the clinical pathway cluster using Markov chains models makes medical practitioners easier to know the clinical process transitions and to accumulate knowledge from the clinical processes. The discovered models can then be used to:

- Fulfill the *online property* of the clustering. Based on the Markov definition, the probability of a given sequence $S_i = (s_{i,1}, s_{i,2}, \dots, s_{i,T_i})$ for an individual i, conditioned on assuming that S_i is a member of cluster c (*i.e.* c_i=c), is then computed, in order to assign it to one of the existing clusters, as follow:

$$P(S_i|c_i = c, \Phi_c) = \pi_c(s_{i,1}) \prod_{j=1}^{T_i-1} a_c(s_{i,j}, s_{i,j+1}) \quad \textbf{(6)}$$

- Predict possible paths for new admitted patients, using the probabilistic transition matrix, in order to help medical professionals to react to exceptions during the clinical process.

4. Experiments

The effectiveness of our framework is assessed based on two real medical data sets selected from the PMSI information system of *Rhone-Alpes* region. These pathways are obtained during a one-year tracking procedure of patients by using anonymous numbers. They consist of respectively 406 and 2050 heterogeneous instances of clinical pathways.

The cluster quality from *our framework* is compared to the results from the mixture Markov models framework introduced earlier [4]. The key idea is: considering the number of clusters k returned from our *b-coloring based-clustering approach*, the Expectation-Maximization algorithm is performed for clustering the set of clinical pathways. In addition to producing an estimate of k, the EM training process requires an initial partitioning of the sequence as an input clusters. In this evaluation step, the clusters are initialized using clinical pathways chosen at random from the data set.

For an interesting assess of the results gained with both clustering frameworks, our evaluation will be based on two quality indices:

Prediction performance index (*PP*): used to examine the predictability of the evolution of all the sequences S_i within a data set X. The key idea is to select the S_i sequences separately and to:

- Eliminate the final state s_{i,T_i} from the sequence S_i.
- Classify the new truncated sequence in one of the k current clusters using the *online property* given by the formula in *eq.*(6). The chosen cluster is denoted by c_i.
- Predict which state z can appear next in the sequence by using the transition matrix of the selected cluster c_i. This state will be compared with the original eliminated state s_{i,T_i} as follow:

$$PP_X = \frac{\sum_{i \in |X|} \omega_i}{|X|} \quad \textbf{(7)}$$

$$where \begin{cases} \omega_i = 1, \ if \ s_{i,T_i} = \underset{1 \le z \le m}{argmax}\{a_{c_i}(s_{i,T_i-1}, z)\} \\ \omega_i = 0, \qquad otherwise \end{cases}$$

The effects on *prediction performance* are then examined for both data sets. The results were obtained using a 5-fold cross-validation. In particular, the data set (clinical pathways) is divided into 5 sub-samples. 80% serve to a training sample and the remaining 20% are considered as a test sample. This process is repeated five times (each of which uses one sub-sample as a test sample) and averaged the results. The training sample is used to generate a partition (Markov clusters) of the data using both clustering frameworks (ours and the EM framework). The sample test is used to evaluate the *prediction performance*. For a good assessment, the prediction performance is also computed on the training sample.

Intracluster homogeneity index (*IH*): such index is fundamental in the cluster validation problem. Considered as a probability scheme, the *intracluster homogeneity* is used to reflect the compactness of the discovered clusters. The greater this value, the more cohesive are clusters of partition. For a partition $\boldsymbol{P} = \{C_1, C_2, \dots, C_k\}$ of $\boldsymbol{S} = \{S_1, S_2, \dots, S_n\}$, this function is defined as the average *intracluster homogeneity* of all the k clusters of $\boldsymbol{P}$ as follows:

$$IH = \frac{\sum_{c=1}^{k} IH_c}{k} \quad \textbf{(8)}$$

$$where, \quad IH_c = \sum_{S_i \in c} \delta_i \quad \textbf{(9)}$$

$where$:

$$\begin{cases} \delta_i = 0 \ if \ P(c_i = c|S_i, \Phi) < 0{,}5 \\ \delta_i = 1 \ if \ P(c_i = c|S_i, \Phi) \ge 0{,}5 \end{cases}$$

$c_i \ is \ the \ clutser \ of \ S_i$

$\Phi = \{\Phi_1, \Phi_2, \dots, \Phi_k\}$ *represents the parmaters for all the clutsers*

The *intracluster homogeneity* of a cluster c (IH_c) is based on the the *conditional probability* of sequence S_i to be assigned to cluster c called $P(c_i = c|S_i, \Phi)$ defined as:

$$P(c_i = c|S_i, \Phi) = \frac{P(S_i|c_i=c,\Phi_c)P(c)}{\sum_{u=1}^{k} P(S_i|c_i=c,\Phi_u)P(u)} \quad (10)$$

$$where, P(c) = \frac{\#\ sequences\ in\ the\ cluster\ c}{n}$$

In the following tables 3 and 4, the experimental results show that our framework gives better performances than the mixture-based models framework. The usefulness of our framework is confirmed as a decision-aid system. Hence, it allows us to directly address the two problems mentioned above to (1) build a fine partition of a sequential data set (heterogeneous and with different lengths) with more *cohesive* and *easily interpretable* clusters, and (2) ensure prediction-making of sequences.

Table 3. Performances on the first data set (406 clinical pathways)

Clustering Framework	*HI*	*PP (training)*	*PP (test)*
Our framework	0.94	76,4%	58,5%
Mixture Markov framework	0.43	61,2%	45,1%

Table 4. Performances on the second data set (2050 clinical pathways)

Clustering Framework	*IH*	*PP (training)*	*PP (test)*
Our framework	0.98	86,48%	68,4%
Mixture Markov framework	0.73	75,8%	49,3%

In order to assess the performance of our clustering framework, the stability of the resulting clusters is also examined. These clusters are considered as an input to the training EM process. Ultimately, the process converges to a final clustering of the data and a generative model (the Markov chain models) for each cluster of the partition. In our case, we have observed that the EM algorithm converges after only one iteration which proves the high stability of the clusters of our partition. In order to better assess the stability of these clusters, the percentage of dominant sequences in the resulting partition is computed. It is found to be 78,44% (respectively 85,56%) for a partition of 406 (respectively 2050) clinical pathways. Since a dominant sequence guarantees that their cluster has a distinct *separation* from all other clusters of the partition, the high number of dominant sequences in such a partition increases the stability of the clusters.

5. Conclusion

In this paper, a new framework for clinical pathways analysis was presented, which is based on a hybrid model that uses a recently proposed *b-coloring based clustering approach* and *Markov chain models*. The underlying clustering method result is that each cluster is described by dominant members (this ensures the global stability of the partition), as well as a finite-state Markov chain model which allows probabilistic prediction for pathways from that group.

Using a number of quality indices that give a good idea of the *intracluster homogeneity* as well as the *prediction performance* of our models, the usefulness of the framework is confirmed as a decision-aid system, and the results are better than those given by the mixture-based models. The validation stage of the *clinical pathway typology* will be completed with the participation of several specialists from the medical domain.

6. References

[1] Elghazel, H., V. Deslandres, M-S. Hacid, A. Dussauchoy and H. Kheddouci, "A new clustering approach for symbolic data and its validation: Application to the healthcare data", *In F. Esposito et al., editors, ISMIS2006 (Springer Verlag LNAI 4208)*, pp.473–482, 2006.

[2] Antunes, C. and A. Oliveira, "Temporal data mining: an overview" *In KDD Workshop on Temporal Data Mining*, pp.1–13, 2001.

[3] Cadez, I. V., D. Heckerman, C. Meek, P. Smyth and S. White, "Visualization of navigation patterns on a Web site using model-based clustering", *In Knowledge Discovery and Data Mining*, pp.280-284, 2000.

[4] Cadez I. V., S. Gaffney, and P. Smyth, "A general probabilistic framework for clustering individuals and objects". *In Knowledge Discovery and Data Mining*, pp.140–149, 2000.

[5] Smyth, P., "Clustering sequences with hidden markov models", *In Advances in Neural Information Processing*, Vol. 9. 1997.

[6] Dempster, A. P., N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM-Algorithm", *Journal of the Royal Statistical Society*, Series B, Vol. 39, pp.1-38, 1977.

[7] Oates T., L. Firoiu, and P. Cohen, "Clustering time series with hidden Markov models and dynamic time warping", *In Proceedings of the IJCAI-99 Workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning*, pp.17-21, 1999.

[8] Paterson M. and V. Dancik, "Longest Common Subsequences". *Mathematical Foundations of Computer Science*, Vol. 841, pp.127-142, 1994.

[9] Kruskall, J. B., and M. Liberman, "The symmetric time warping problem: From continuous to discrete", *In Kruskal J.B., Sankoff D., editors. Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. Stanford: CSLI Publications*, pp.125–161, 1999.

[10] Jain, A.K., M.N. Murty and P.J. Flynn, "Data Clustering: A Review", *ACM Computing Surveys*, Vol. 31, pp.264-323, 1999.

[11] Kalyani, M. and M. Sushmita, "Clustering and its validation in a symbolic framework", *Pattern Recognition Letters*, 24(14), pp.2367-2376, 2003.