2014 47th Hawaii International Conference on System Science

Learning Predictive Models from Integrated Healthcare Data: Extending Pattern-based and Generative Models to Capture Temporal and Cross-Attribute Dependencies

Rui Henriques
Dep. Comp. Science Eng. *and* KDBIO, Inesc-ID
Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portual
rmch@tecnico.ulisboa.pt

Cláudia Antunes
Dep. Computer Science and Engineering
Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portual
claudia.antunes@tecnico.ulisboa.pt

***Abstract*—Modeling the dependencies among multiple temporal attributes derived from integrated healthcare databases represents an unprecedented opportunity to support medical and administrative decisions. However, existing predictive models are not yet able to successfully anticipate health conditions based on multiple (sparse) time sequences derived from repositories of health-records. To tackle this problem, we propose new predictive models able to learn from an expressive temporal structure, a time-enriched itemset sequence, which captures both temporal and cross-attribute dependencies. Revised pattern-based models and hidden Markov models are proposed to address the properties of the target integrative temporal structures. The conducted experiments hold evidence for the utility and accuracy of the proposed predictive models to anticipate health conditions, such as the need for surgeries.**

***Index Terms*—integrated healthcare data, sparse temporal data, cross-attribute dependencies, pattern-based prediction, Markov-based prediction.**

I. Introduction

The increasing integration and availability of healthcare data is triggering new opportunities to support medical and administrative decisions. In this paper, we tackle the problem of defining predictive models able to deal with multiple sparse time sequences. Each time sequence is derived from a monitored attribute of interest across a time period. Attributes provide alternative healthcare views related with diagnoses, treatments, prescriptions or lab records. This problem differs from multivariate time series analysis since the domain of the target attributes varies and their records are not temporally aligned.

The existing predictive models for temporal data are not able to consider multiple time sequences per instance (patient) and, therefore, cannot capture their structural interdependencies. Alternatively, the existing integrative learning proposals are a simple composition of separated predictive models learned for each time sequence followed by a voting stage [1]. Similarly, feature-based classifiers rely on features that are independently collected from each time sequence. These learning settings prevent existing predictive models from considering structural dependencies among alternative health-related aspects, which are critical to support prognostics and planning tasks from integrated healthcare data.

In this paper we aim to define and study the behavior of alternative classifiers that are able to adequately model the dependencies among multiple (sparse) time sequences from integrated healthcare data. For this purpose we, first, propose a data mapping step that combines the multiple attributes into one single temporal structure. Second, we propose two predictive models to capture these dependencies. The first model, P2MID, relies on classification rules based on time-enriched sequential patterns to capture integrated healthcare profiles. The second model, M2ID, relies on customized hidden Markov models to be sensitive to the varying length and sparsity degree of patient health-records. These directions are compared and evaluated against baseline classifiers for different medical tasks: prediction of hospitalization and surgery needs, and prediction of future healthcare conditions and procedures (in accordance with ICD-9-CM and CPT standards [2], [3]).

The paper is structured as follows. In *Section II*, the target task is motivated and contributions from existing research covered. In *Section III*, we describe the proposed solutions. Finally, results and critical implications are synthesized in *Section IV*.

II. Background

The challenges of defining predictive models from large repositories of health-records have been largely synthesized [4]–[7]. In particular, supervised learning from integrated healthcare data is challenged by two major aspects. First, the length and the sampling occurrence grid of healthcare events vary both among patients and across different time periods within each patient, which leads to the need to deal with *arbitrary levels of sparsity*. Second, an integrated analysis of *multiple healthcare attributes* is required for more accurate decisions. Exemplifying, determining if a patient needs to be hospitalized is not only dependent on the past hospitalizations but conditionally dependent on the patient clinical history composed by evaluations, treatments, prescriptions and diagnosis. In this section we cover the contributions and limitations of existing work for each one of these two challenges.

Def. 2.1: Let Σ be an alphabet of symbols, and t be a timestamp. A *time sequence* $w \in \mathbb{W}$ is an ordered multi-set

978-1-4799-2504-9/14 $31.00 © 2014 IEEE
DOI 10.1109/HICSS.2014.322

of *events* $\{(\sigma_i, t_i) \mid \sigma_i \in \Sigma, t_{i+1} \geq t_i, i = 1..n\}$ with $n \in \mathbb{N}$ occurrences, where $\mathbb{W}$ is the set of all time sequences.

Supervised learning from temporal structures has been mainly centered on time series, a specific time sequence where events are temporally equally distant, thus, not allowing for co-occurrences and sparsity. The less researched learning methods over time sequences have been driven by two major tasks: *prediction* [8], estimation of future time points of a single time sequence based on training time sequences, and *sequence classification* [9], labeling of an unlabeled time sequence from labeled time sequences. In this context, generative and pattern-based methods have been proposed. Generative methods include formal languages, dynamic Bayesian networks, Markov chains, and time-sensitive neural networks (NNs) such as recurrent or time-delay NNs [9]–[12]. Pattern-based methods for prediction include supervised rule discovery to predict or constrain upcoming events [13]–[15]. Pattern-based methods for sequence classification commonly rely on the extraction and weighting of discriminative prototype features for each class [16]. Common features include sequential patterns and wavelets, among others [17], [18].

A shortcoming of the majority of these methods is that sparsity is roughly treated by ignoring the temporal distance among events. Only ordering relations among events are considered. An alternative strategy to avoid sparsity that also results in a significant loss of information is to convert time sequences into feature vectors by adopting an aggregation criterion as the counting of events across sequent periods [1]. Although the supervised inference of temporal rules can solve this problem, this option is not yet scalable [14].

Additionally, these learning methods are not able to consider multiple time sequences per instance. Note that existing contributions from multivariate time series analysis and multivariate responses prediction are not applicable for the target integrated settings since both the domain and length across the target multiple time sequences per instance varies.

Def. 2.2: Given a training dataset $D = \{x_1, .., x_m\}$ with tuples in the form of $x_i = \{w_1 \in \mathbb{W}^1, ..., w_p \in \mathbb{W}^p, y \in Y\}$, the target task of *classification from integrated temporal data* is to learn a mapping model $M : \mathbb{W}^1, ..., \mathbb{W}^p \rightarrow Y$, where Y is the set of classes.

The target task of classification from multiple time sequences is formalized in *Def.2.2*. In literature, three strategies are adopted to answer this task. One direction has been to learn one model separately for each time sequence [19]. A model for disease anticipation is addressed in [20] relying on administrative records tracking drug prescriptions, hospitalizations, and daily hospital activities. The drawback of these solutions is the loss of critical integrated views that does not show up when each attribute is analyzed separately.

A second option is to extract features from multiple time sequences by relying on clustering methods that rely on edit-distance metrics based on insert-delete-replace operations [21]. The drawback here resides on the complexity of defining effective distance metrics suitable for different patients and attributes.

A final option is to perform a preprocessing stage that incorporates multiple attributes at several time points as done in [22]. However, this solution is only practical when there is background knowledge to select specific events of interest from each attribute.

III. Solution

To deal with the identified three challenges we propose a solution that relies on a data mapping stage followed by the learning of predictive models able to deal with the specificities of the mapped data.

First, in *Section III.A*, record-centered healthcare databases are mapped into a single temporal structure per patient that is both able to preserve the temporal distance among events and to provide an integrated view of the multiple time sequences per patient. Second, in *Sections III.B* and *III.C*, existing predictive models are adapted to be able to effectively and efficiently learn medical conditions using the proposed temporal structure. In particular, we want this learning to be shaped by relevant interdependencies across the selected healthcare attributes. We propose a deterministic classifier, where these interdependencies are captured through temporal patterns, and generative classifiers under a Markov assumption, where interdependencies are captured by the learned transition and emission probabilities of the underlying lattices.

A. Data Mapping

For the proposed mapping, we assume that the input healthcare databases can be mapped into a multi-dimensional scheme centered on health-records to organize a wide variety of health-related aspects [23]. A health-record can, thus, be seen as a central fact table that maintains a high multiplicity of measures related with multiple dimensions, such as the calendar date, patient identity, payer, provider, prescription and location. Since only a small subset of all measures is captured per health-record, the fact table commonly defines the type and value of the monitored measures to guarantee the compactness of the database. Fig.1 provides an illustrative record-centered database.

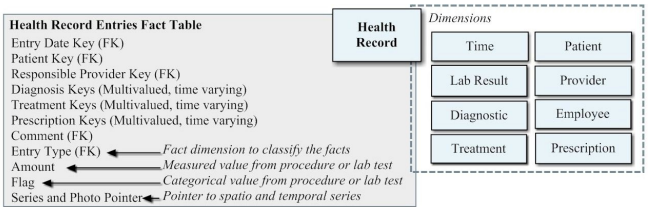

Fig. 1: Health record-centric multi-dimensional structure

Under this input data scheme, the retrieval of multiple time sequences can be done in three simple steps without loss of information. First, the *aggregation* dimension, commonly the *patient* dimension, is used to split the fact occurrences into a set of instances.

Second, a denormalization procedure is applied over the fact entries *<measure type, measure value, dimension identifiers>* to group patient records per measure.

Finally, for each instance, the *calendar-date* dimension is used to compose one time sequence per measure of interest.

Having described the steps required to retrieve multiple time sequences from an integrated healthcare database, next we propose a mapping that expressively combines the multiple time sequences into one single temporal structure that still preserves cross-attribute dependencies. This temporal structure, a *time-enriched itemset sequence*, maps the occurring events from the multiple time sequences to sets of itemsets according to a specific time granularity.

Def. 3.1: Let an item be an element from an alphabet Σ. An *itemset I* is an ordered set of items. An *event*, *e*, is a tuple (I,t), where $e.I$ is an itemset and $e.t$ a time point.

Def. 3.2: An *itemset sequence*, *s*, is an ordered set of itemsets $s{=}<e_1.I,...,e_n.I>$, from events with time points respecting $\forall_{i\in\mathbb{N}\wedge i<n}\, e_i.t < e_{i+1}.t$.

Def. 3.3: Let a time partitioning be defined by a set of φ_i contiguous intervals, $\forall_{i\in\mathbb{N}}[i\times\delta,(i+1)\times\delta[$, where δ is the considered time granularity. Given a time granularity δ, a *time-enriched itemset sequence* is an ordered set of itemsets $<\Phi_1.I,...,\Phi_n.I>$, where Φ_i is the set of events occurring within φ_i partition and $\Phi_i.I$ is the union of all itemsets occurring within φ_i, $\cup_k e_k.I$.

Contrasting with simple itemset sequences, time-enriched itemset sequences allows for the explicit representation of empty itemsets corresponding to time partitions without occurrences of health-records. Exemplifying, the illustrative set of events, $<e_1{=}(a,t_1),\ e_2{=}(d,t_1),\ e_3{=}(b,t_5),\ e_4{=}(c,t_6)>$, can be mapped as a simple itemset sequence, $<\{a,d\},\{b\},\{c\}>{=}(ad)bc$, or as a time-enriched sequence, $<\{a,d\},\varnothing,\{b,c\}>{=}(ad)\varnothing(bc)$, when considering δ=2.

We propose a four-step methodology for the composition of a single time-enriched itemset sequence per instance from multiple time sequences.

First, numeric time sequences are discretized into ordinal time sequences using lengthy alphabets.

Second, the dimensionality, $|\Sigma|$, of each time sequence w_i is balanced according to a homogeneity criterion. For the adopted databases in this work, we consider the following homogeneity criterion:

$max(|\Sigma_1|,..,|\Sigma_p|) \leq \Delta min(|\Sigma_1|,..,|\Sigma_p|)$, with $\Delta = 3$.

This balancing can be easily done by either aggregating closed symbols if the alphabet is ordinal, or by applying a hierarchical clustering method to group symbols according to the observed *Y* classes if the alphabet is nominal.

Potential conflicts between the domains of the input time sequences (items from different attributes sharing the same symbol) are treated through a simple redefinition of symbols.

Third, a time granularity is adopted to partition the timeline. For each time sequence, the values from event occurrences are grouped per time partition. Symbol repetitions are removed. In this way, each time sequence is mapped as a time-enriched itemset sequence.

Finally, the itemsets for each time-enriched itemset sequence are merged as a single itemset per partition, which leads to the target single temporal structure.

To illustrate this methodology, consider a dataset where each patient x_i has a domain with three time sequences w_i derived from an integrated healthcare database, corresponding to the monitored prescriptions, diagnosis and treatments. Since these time sequences are categorical, there is only the need to balance their dimensionality by revising their domains (or by adopting balanced categorizations such as the ICD-9-CM and CPT standards [2], [3]), and to aggregate their occurrences according to a time partitioning. Exemplifying, consider a patient with treatments={(Radiology,Feb), (SurgeryUrinary,Apr)}, conditions={(CancerB,Jan), (RenalFail,Feb), (Seizures,May)}, and prescriptions={(AnalgesicDrug,Jan)}. Given a month granularity, δ=1, the resulting time-enriched itemset sequence would be <{CancerB,AnalgesicDrug}, {RenalFail}, $\varnothing$, {SurgeryUrinary}, {Seizures}>.

The proposed mapping delivers a single temporal structure that expressively addresses the three target challenges: structural sparsity, attribute multiplicity and specific properties from healthcare databases. A simplified view of this mapping is provided in Fig.2.

B. Pattern-based Predictive Models

Under the previous mapping, the existing methods for the analysis of itemset sequences can be extended to be time-sensitive, and their output used to guide and shape the target predictive models.

The most common task in this context is the discovery of sequential patterns to mine frequent precedences and co-occurrences. Sequential patterns discovered over the target temporal structure are able to include items from different time sequences, and, therefore, to explicitly model interdependencies across the multiple health attributes of interest.

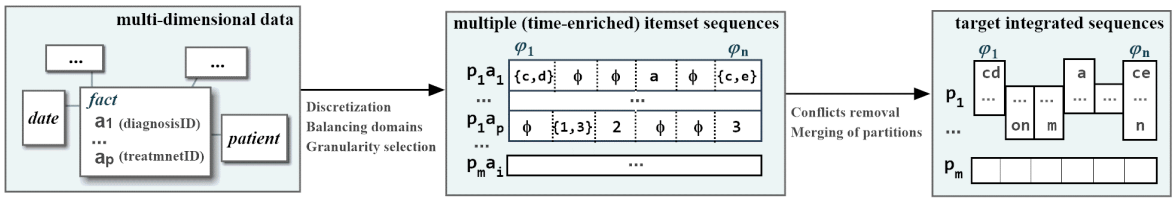

Fig. 2: Mapping integrated healthcared databases as a single temporal structure per patient

Different strategies have been proposed for the use of temporal patterns for classification, including simple ranking scores to identify the patterns more able to discriminate each class, probabilistic induction [16], alternative learners [24], and optimization methods based on confusion matrices [25]. Despite their utility to mine itemset sequences, these methods suffer from two problems. First, they are only prepared to capture frequent precedences and co-occurrences and, therefore, are not able to consider temporal distances between events, which is a critical requirement for the definition of predictive models. Second, they have been developed in the scope of genomic studies and multivariate time series analysis, and, consequently, the argued levels of performance no longer remain valid for the target sparse healthcare settings, where relevant sequential patterns are observed for a small number of instances that share a similar healthcare profile.

For these reasons we propose a new pattern-based classifier, referred as P2MID (Pattern-based Predictive Models from Integrated Data), which is a variant of existing contributions. The behavior of the P2MID classifier can be described according to its training and testing stages. In the *training stage* a discriminative model is defined in three steps.

First, a set of time-enriched sequential patterns is generated for each medical condition. Below, we introduce the revised temporal notion of a sequential pattern.

Def. 3.4: Let an itemset sequence a=$<a_1,..,a_n>$ be a subsequence of b=$<b_1,..,b_m>$ if $\exists_{1\leq i_1<..<i_n\leq m}\, a_1\subseteq b_{i_1},..,a_n\subseteq b_{i_n}$. Given a set of itemset sequences S and a minimum support threshold θ, a *sequential pattern* is a sequence $s \in S$ that is contained in at least θ sequences.

Def. 3.5: A *time-enriched sequential pattern* is a sequential pattern observed for specific time interval $[\varphi_i, \varphi_f]$. A time-enriched sequential pattern is subset of another, $a \subseteq b$, if it is both a subsequence and its time interval is contained in a time range ($\varphi_{a_i}\leq\varphi_{b_i} \wedge \varphi_{a_f}\geq\varphi_{b_f}$).

Considering the illustrative set, $\{(ac)\varnothing db, (ac)d\varnothing\varnothing, (ac)\varnothing\varnothing(bd)\}$, and a minimum support θ=2, $(ac)d$ is a simple sequential pattern, while $\{(ac)d\}[\varphi_i$=0, φ_f=2] or $\{db\}[\varphi_i$=3, φ_f=4] are illustrative time-enriched sequential patterns for a granularity δ=1.

P2MID computes these temporal patterns by defining multiple temporal aggregations ($\delta \in \{1,2,..\}$) followed by the discovery of co-occurrences for coarser-grained aggregations under a penalization factor to benefit the discovery sequential patterns that occur within small time ranges.

Second, the confidence of each pattern in relation to a particular class is evaluated to compose a new type of rules of the form $s \Rightarrow y$, where s is the temporal pattern and y the class.

Third, and similarly to CMAR [26], these rules are inserted in a tree structure if: i) the χ^2 test over the rule is above a specified α-significance level, and if ii) the tree does not contain a rule with higher priority. Understandably, since CMAR is not able to deal with sequential patterns, but only with frequent itemsets, we propose a new priority criterion. A rule $R_1 : s_1 \Rightarrow y$ is said to have priority over $R_2 : s_2 \Rightarrow y$ if $s_1 \subseteq s_2$ or if:

$$conf(R_1)>conf(R_2) \vee (conf(R_1)=conf(R_2) \wedge sup(R_1)>sup(R_2)) \vee (conf(R_1)=conf(R_2) \wedge sup(R_1)=sup(R_2) \wedge |s_1|<|s_2|)$$

Finally, the tree is pruned based on the computed priorities. This tree defines the discriminative pattern-based model, which is a simple ordered set of tuples (pattern s, class y, weight β).

In the *testing stage* of P2MID, this discriminative model is used to classify a specific patient by identifying the closest temporal patterns and relying on their classes and matching score. The strength of each group of conditions is calculated by computing the weighted-χ^2 across all the rules $s \Rightarrow y$ that satisfy a *matching* criterion between the temporal pattern s and the testing instance.

$$\begin{aligned} &weighted\text{-}\chi^2(y) = \Sigma_{match(s_i \Rightarrow y)}(\chi^2(s_i) \times \chi^2(s_i))/MCS, \\ &\text{with } MCS=(min(sup(s), sup(y))\text{-}sup(s)sup(y)/N)^2 \times N \times e, \\ &\text{where } N = \mid match(s_i \Rightarrow y) \mid \text{ and } e = \frac{1}{sup(y)^2} + \frac{1}{sup(s_i)N} - \\ &\qquad sup(y) + \frac{1}{N} - sup(s_i)sup(y) + \frac{1}{(N-sup(s_i))(N-sup(y))} \end{aligned}$$

Matching occurs if the time-enriched sequential pattern is observed for the testing instance within the specified time frame or if the pattern is observed for the testing instance with the specified duration but for different time partitions (time shift condition). For this latter case, the rule score is divided by the difference of partitions to penalize the temporal shift.

Finally, the strongest condition, $y \in Y$, is output as the estimated class for a deterministic result, or, alternatively, the computed strength for each class is delivered as their probabilistic value.

C. Generative Predictive Models

Unlike pattern-based models, classifiers relying on generative models provide a rather different behavior. Instead of modeling the local properties of time sequences using temporal patterns, they test the fit of the overall time sequence against the learned lattices. From the wide-range of generative learners, we selected hidden Markov models (HMMs) due to their expressive power, compactness, easily parameterized behavior and propensity to deal with sequential data.

Def. 3.6: A first-order *HMM* is a stochastic finite automaton, where a set of hidden states are connected according to a probability transition matrix, T, and have observable emissions described by a probability matrix, E. The (T, E) pair defines the HMM *architecture*.

However, existing Markov-based models are only prepared to deal with sequences of fixed multivariate order [9]. To be able to learn HMMs from (time-enriched) itemset sequences, we propose a new classifier, referred as M2ID (Markov-based Models from Integrated Data), that relies on a simple data mapping applied over extended HMM architectures.

The mapping step transforms a (time-enriched) sequence of itemsets into a univariate sequence by relying on an additional symbol to represent the delimiters of each time partition. Illustrating, the time-enriched itemset sequence, $(ac)\varnothing\varnothing d\varnothing(ad)$,

is mapped into a univariate sequence \$ab\$\$\$d\$\$ad\$, where \$ is the symbol that delimits co-occurrences.

By explicitly capturing time partitions, the generative models became sensitive to the underlying data sparsity since the parsing of delimiters will shape the transition and emission probabilities.

Under this mapping, learning generative models from time-enriched itemset sequences is simply a matter of adjusting the underlying architectures.

Fully inter-connected architectures, the default case, are only able to adequately deliver effective models when the number of delimiters is significantly small in comparison with the average number of items per instance. This behavioral problem is related with the heightened convergence of emission probabilities towards the delimiter symbol. Understandably, more expedite architectures need to be adopted for datasets with high sparsity or large number of time partitions.

To guarantee an accurate generative modeling of integrated healthcare data for different settings, we propose an extension of an alternative Markov-based architecture: Left-to-Right Architecture (LRA) [27]. LRAs define one unidirected path with *main* states intertwined by *insertion* and *deletion* states. Insertion states can be used both to skip rare events and events that do not significantly discriminate a particular class. Deletion states in conjunction with insertion states allow the generative model to adequately cope with the varying rates of sparsity observed within and across patients.

To turn these architectures able to model time-enriched itemset sequences, we propose the selection of specific states (*aligning* states) along the main path to only emit the delimiters of time partitions. Additionally, we forbid any other state to emit these delimiters by initializing their emission probabilities as zero.

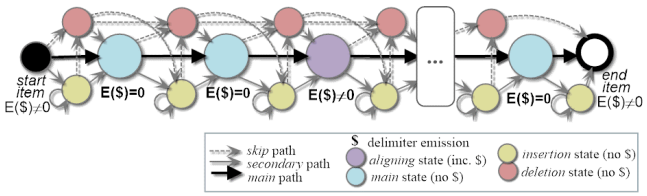

Fig. 3: Extended LRA for mining itemset sequences

In this way, we are forcing the predictive models to create time points of alignment and, therefore, to use the main states between such special aligning states as the transition-emission lattices responsible to capture the dependencies across the multiple health-related attributes of interest. Fig.3 illustrates the extended LRA architecture.

M2ID adopts Viterbi, an efficient and robust algorithm, to both learn the generative models and to classify each testing instance based on its likelihood of being generated by the learned lattices per class.

D. Illustrative Case

Consider the illustrative task of predicting the need of a specific treatment $T1$ for the upcoming quarter from healthcare data monitored along two years. Consider the presence of 15 clinical procedures ($T1$-15), 10 major health conditions ($D1$-10), 20 lab-test assessments ($L1$-20) and 15 categories of prescriptions ($P1$-15). Let us assume that, under a selected month granularity, the learned P2MID predictor has the following top 5 rules: $\{(D5L8P3)P3\}[20,24]\Rightarrow T1$ (confidence c=97% and priority score β=37), $\{T1\}[12,24]\Rightarrow\neg T1$ (c=91%,β=35), $\{(D4L8P2)P2\}[16,22]\Rightarrow T1$ (c=89%,β=34), $\{L6P7\}[18,24]\Rightarrow\neg T1$ (c=78%,β=31) and $\{P4(D4D5)\}[14,20]\Rightarrow T1$ (c=88%,β=28). Complementary, let us consider two Markov-based lattices, HMM_{T1} and $HMM_{\neg T1}$, following an extended LRA architecture learned from a balanced class setting, where HMM_T have weighter transitions towards insert states and the emissions from the latter hidden states along the main path have superior convergence towards the $\{D4, P3, D5, L8, P2\}$ set of events.

Understandably, a patient under prediction with the following time-enriched itemset sequence, $\{(T2P8)(L2P8)(D5P3)P3\varnothing(L8D4P2P3)(P2P3)\}$[18,24], is prone to be classified as a candidate for $T1$ treatment under both pattern-based and generative approaches. Pattern-based matching criteria is enough expressive to consider temporal misalignments and item gaps. Complementary, this testing sequence is more likely to be generated by the HMM_{T1} lattice than the $HMM_{\neg T1}$ lattice.

IV. Results

To evaluate the proposed solutions, the healthcare heritage prize database[1] was adopted. This database integrates healthcare aspects, such as detailed claims, hospitalizations and monthly number of laboratory tests and prescribed drugs, over 150,000 patients across multiple providers and specialties. The original relational database was mapped into a multi-dimensional database and pre-processed according to the mapping methodology proposed in section 3.1 to derive time-enriched time sequences with varying temporal granularities (month, quarter and semester). We collected a random population of 20,000 patients for this assessment. For the month granularity, each patient has an average number of 4 items per itemset (σ=2) and 36 itemsets per sequence. For the quarter granularity, patients have an average number of 12 items per itemset (σ=3) and 12 itemsets.

We selected three distinct medical tasks from data monitored along three years: anticipation of surgery, prescription and hospitalization needs.

The codification of the M2ID method relies on HMMs and learning settings adapted from the HMM-WEKA extension (implemented according to [9], [11] sources). Both the mapping and the target classifiers, P2MID and M2ID, were implemented in Java (JVM version 1.6.0-24). The following experiments were computed using an Intel Core i5 2.80GHz with 6GB of RAM.

A. Observations

The assessment of the prediction performance for each medical condition is centered on the accuracy levels from a 10-

[1] http://www.heritagehealthprize.com/c/hhp/data
(under a granted permission for this publication)

fold cross-validation scheme. A sample reduction method was applied over the original population to balance the number of instances per class. To test if there are statistically significant differences among the levels of accuracy of the different learners, we preserved the 10-folds across experiments and performed a paired two-sample two-tailed *t*-test using a *t*-Student distribution with 9 degrees of freedom.

In order to understand the improvements in performance from capturing both temporal dependencies and dependencies between multiple attributes of interest, we included the results observed from a traditional classification setting (baseline classifier). For this approach, we mapped the values for the last 4 occurring events per attribute of interest as simple features, and, then, applied different classifiers (C4.5, kNN, Naive Bayes, NN, and SVM) from Weka [28] and selected the best result. Missing values were introduced for instances that showed less than four occurrences for one or more of the monitored health attributes.

Three different medical conditions were target in our experiments: *i)* surgery anticipation for the upcoming quarter, *ii)* prediction of the average number of drug prescriptions for the upcoming month (by grouping prescription levels into four classes $\{0,1,2-5,6+\}$), and *iii)* Boolean prediction of hospitalization needs for the upcoming year. Upcoming predictions are accomplished by removing from training data the last temporal partition. In particular, for the last task, we additionally removed from data the records related with hospitalizations, i.e., these predictive models were only learned from attributes related with claims, diagnoses, lab analysis and drug prescriptions.

Fig.4 illustrates the observed accuracy levels for the different medical tasks. The horizontal line marks the default accuracy from a random learner based on the number of classes per task, $\frac{1}{|Y|}$. Two major observations can be synthesized.

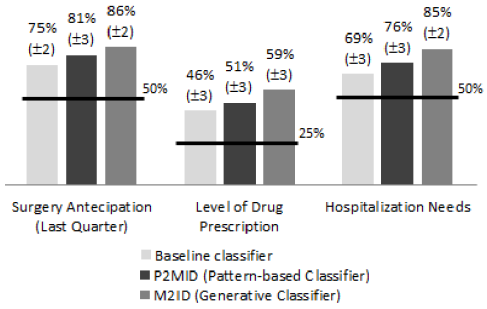

Fig. 4: Classification *accuracy* for different tasks

First, the proposed classifiers perform better than traditional classifiers across the selected prediction tasks. This improvement is statistical significant (at α=1%) and motivates the importance of modeling temporal and cross-attribute dependencies.

Second, generative models seem to be generally the most suitable choice. However, differences in performance between M2ID and M2ID were not found to be statistically significant (at α=1%) across all tasks. For instance, no statistical significance difference was found for the surgery prediction task, which is potentially related with the fact that its anticipation is well modeled by specific compact sets of health-records with heightened discriminative power.

To test the behavior of the target predictive models for varying levels of sparsity, we varied the time granularity from a month scale up to a semester scale ($\delta \in \{1,3,6\}$). This strongly impacts the number and average length of itemsets from the mapped time-enriched itemset sequences. Fig.5 compares the accuracy of the target classifiers with different time granularities for the anticipation of hospitalization needs.

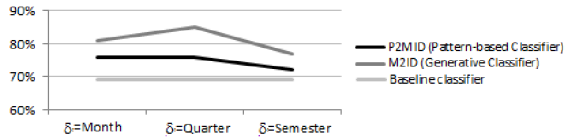

Fig. 5: *Accuracy* for varying time scales

The adoption of coarse-grained time partitions (semester granularity) deteriorates the performance of both P2MID and M2ID methods. For P2MID, the decrease on the number of partitions leads to the loss of significant precedences as they are captured as co-occurrences. M2ID behavior becomes less centered on sequential analysis through temporal alignments (less number of states to dedicatedly emit the delimiter symbol) and more focused on learning the interdependencies among a higher number of events (items), which is not the original purpose of predictive models under a Markov assumption.

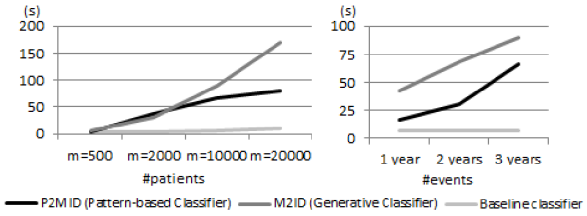

Fig. 6: *Efficiency* for varying size of datasets

The efficiency of the proposed predictive models was assessed for a varying length of patients and varying number of events (by varying the monitored time period). Fig.6 gathers the results from this analysis. The efficiency of the proposed predictive models is significantly worse than traditional classifiers. This is explained by two factors. First, traditional classifiers only learn from a very small subset of the overall events (the last four records for each one of the target attributes). Second, the training time of P2MID is expensive since the (time-enriched) sequential pattern mining task is performed for very low levels of support. M2ID performance is penalized by the length and complexity of the extended left-to-right architectures.

B. Why P2MID and M2ID methods address the target challenges?

To support the previously described decisions made for both the P2MID and M2ID methods, Fig.7 collects the results from selecting different strategies.

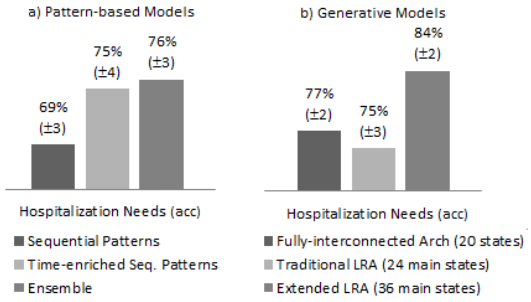

Fig. 7: Impact of P2MID and M2ID decisions

In Fig.7a, the impact of adopting simple sequential patterns versus time-enriched sequential patterns is evaluated. The difference in performance is statistically significant (at α=2%). Two main reasons explain the observed differences in performance.

First, the proposed discriminative models tend to score preferentially patterns occurring near the time period under prediction. Also, the allowance of temporal shifts under a penalization factor during the testing stage offers a time-dependent informative context for classification. Contrasting, simple sequential patterns cannot offer temporal guarantees, and, therefore, the influence of both recent and old events to discriminate the class under prediction is not clearly differentiated.

Second, the time partitioning strategy allows to deal with arbitrary high levels sparsity by choosing adequate granularity levels with impact on the degree of precedences vs. co-occurrences.

The integration of both simple and time-enriched sequential patterns was observed to slightly increase the overall levels of accuracy. This can be explained by the inclusion of precedences (modeled by simple sequential patterns) with larger time frames.

In Fig.7b, the impact of adopting alternative HMM architectures is evaluated. The adoption of fully-interconnected and left-to-right architectures as-is have a significantly worse performance when compared with the extended left-to-right architectures. This is explained by two factors.

First, since in the original architectures there are not dedicated states to emit delimiters, there is not an explicitly way of temporally aligning time-enriched itemset sequences of varying lengths. Although the inclusion of delimiters turns the learning sensitive to varying levels of sparsity (varying number of events per time partition), they are associated with convergence problems on the emission probabilities for the original architectures since delimiters occur with a significant larger frequency than normal items.

Second, the selection of specific main states in a left-to-right architecture to dedicatedly emit delimiters, creates an adequate generative setting sensitive to the underlying sparsity (given by the weight of transition probabilities towards deletion or insertion states) and attribute interdependencies (given by the most probable emissions along the main path).

C. Implications

The collected results show that the definition of predictive models prepared to deal with integrated temporal structures, such as time-enriched itemset sequences, is a promising direction to deal with the temporal and cross-attribute dependencies of integrated healthcare data.

The observed behaviors for the two proposed predictive models present contrasting properties that should be known before selecting the learner. The decision can vary depending on multiple factors, such as the considered time granularity, the selected temporal attributes of interest and the medical condition under prediction.

On one hand, pattern-based models are able to discriminate specific integrated profiles of interest. A time-enriched sequential pattern is able to combine items derived from different attributes (such as a particular diagnosis, type of prescribed drug or applied procedure) within a time frame to discriminate a medical condition (e.g. need for surgery). Pattern-based methods are particularly prone to predict medical conditions from small sets of events. In particular, specific records, such as pregnancy diagnoses, are able to robustly discriminate conditions. Under this setting, the inclusion of a matching criterion that allows temporal shifts under a penalized factor is critical to guarantee that a reasonable number of sequential patterns are selected for each testing instance. Pattern-based models are the choice for data contexts with a large number of uninformative health-records.

One the other hand, generative methods relying on extended HMM architectures offer a more smoothed behavior as they consider all the observed events to shape the transition-emission probabilities (training stage) or to compute the generation likelihood of a particular instance (testing stage). However, contrasting with pattern-based models, the accuracy of generative models degrades in healthcare scenarios where only a small subset of events per patient effectively discriminate a medical condition. Additionally, the covered architectures are not able to prefer more recent events from all the events, which commonly have a more heightened influence to discriminate the conditions under prediction.

V. Discussion

This work addresses the problem of learning predictive models to support medical decisions from integrated healthcare databases, with an incidence on both the temporal and inter-attribute dependencies. These challenges are motivated, and the contributions and limitations of existing research to answer them are synthesized.

To deal with the varying temporal sparsity across patients and the potential high multiplicity of attributes of interest,

we propose a new methodology centered in two major steps. First, a data mapping is proposed to combine multiple time sequences derived from multi-dimensional healthcare databases into a single temporal structure, a time-enriched itemset sequence. This structure offers an integrated view of each patient by preserving conditional dependencies across attributes and by allowing arbitrary time scales.

Second, two predictive models, one deterministic (P2MID) and one generative (M2ID), are proposed and compared under this mapping. P2MID relies on a discriminative model based on a time-enriched notion of sequential patterns to deliver time distance guarantees between events, to penalize temporal misalignments and to favor recent events. M2ID relies on a modified left-to-right HMM architecture able to discard non-discriminative events and to deal with patients with varying number of events. The conducted experiments hold evidence for the accuracy and utility of the proposed predictive models. This observation supports the need for developing new classifiers that are able to model the underlying temporal and cross-attribute dependencies of integrated healthcare data from time-enriched itemset sequences.

ACKNOWLEDGMENTS

This work was supported by *Fundação para a Ciência e Tecnologia* (FCT) under the research project D2PM, PTDC/EIA-EIA/110074/2009, and the PhD grant SFRH/BD/75924/2011.

REFERENCES

[1] V. Tseng and C.-H. Lee, "Effective temporal data classification by integrating sequential pattern mining and probabilistic induction," *Expert Sys.App.*, vol. 36, no. 5, pp. 9524–9532, 2009.

[2] G. Escobar, J. Greene, P. Scheirer, M. Gardner, D. Draper, and P. Kipnis, "Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases," *Medical Care*, vol. 46, no. 3, pp. 232–239, 2008.

[3] M. Abraham, J. Ahlman, A. Boudreau, J. Connelly, and D. Evans, *CPT 2011: Standard Edition*, ser. CPT / Current Procedural Terminology. American Medical Association Press, 2010.

[4] R. Bellazzi, F. Ferrazzi, and L. Sacchi, "Predictive data mining in clinical medicine: a focus on selected methods and applications," *Data Min. Know. Disc.*, vol. 1, no. 5, pp. 416–430, 2011.

[5] R. Henriques, S. Pina, and C. Antunes, "Temporal mining of integrated healthcare data: Methods, revealings and implications," in *SDM IW on Data Mining for Medicine and Healthcare*. SIAM, 2013, pp. 52–60.

[6] G. Norén, J. Hopstadius, Bate, Star, and I. Edwards, "Temporal pattern discovery in longitudinal electronic patient records," *Data Min. Knowl. Discov.*, vol. 20, no. 3, pp. 361–387, May 2010.

[7] R. Henriques and C. Antunes, "An integrated approach for healthcare planning over multi-dimensional data using long-term prediction," in *Health Information Science*, ser. LNCS. Springer Berlin/Heidelberg, 2012, vol. 7231, pp. 36–48.

[8] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Mach. Learn.*, vol. 3, no. 1, pp. 9–44, Aug. 1988.

[9] C. Bishop, *Pattern Recognition and Machine Learning*, ser. Information Science and Statistics. Springer, 2006.

[10] E. Wan, "Temporal backpropagation for fir neural networks," in *IJC on Neural Networks*, 1990, pp. 575–580 vol.1.

[11] K. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*. University of California, Berkeley, 2002.

[12] G. Guimarães, "The induction of temporal grammatical rules from multivariate time series," in *Int. Colloq. on Grammatical Inference: Alg. and App.* London, UK: Springer-Verlag, 2000, pp. 127–140.

[13] H. Lu, J. Han, and L. Feng, "Stock movement prediction and n-dimensional inter-transaction association rules," 1998.

[14] F. Mörchen, *Time series knowledge mining*, ser. Wissenschaft in Dissertationen. Görich & Weiershäuser, 2006.

[15] T. G. Dietterich and R. S. Michalski, "Discovering patterns in sequences of events," *Artif. Intell.*, vol. 25, pp. 187–232, February 1985.

[16] A. Nanopoulos, R. Alcock, and Y. Manolopoulos, "Information processing and technology," in *Feature-based classification of time-series data*. Commack, NY, USA: Nova Science Publishers, 2001, pp. 49–61.

[17] T. Oates, "Identifying distinctive subsequences in multivariate time series by clustering," in *KDD*. New York, NY, USA: ACM, 1999, pp. 322–326.

[18] D. Roverso, "Multivariate temporal classification by windowed wavelet decomposition and recurrent neural networks," in *ANS Int. Topical Meeting on NPICHMI*, 2000.

[19] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Norwell, MA, USA: Kluwer Academic Publishers, 1998.

[20] K. Choi, S. Chung, H. Rhee, and Y. Suh, "Classification and sequential pattern analysis for improving managerial efficiency and providing better medical service in public healthcare centers," *Healthc Inform Res.*, vol. 16, no. 2, pp. 67–76, 2010.

[21] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

[22] R. A. Baxter, G. J. Williams, and H. He, "Feature selection for temporal health records," in *PAKDD*. London, UK, UK: Springer-Verlag, 2001, pp. 198–209.

[23] R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*, 2nd ed. USA: John Wiley & Sons, Inc., 2002.

[24] N. Lesh, M. J. Zaki, and M. Ogihara, "Mining features for sequence classification," in *KDD*. New York, NY, USA: ACM, 1999, pp. 342–346.

[25] T. P. Exarchos, M. G. Tsipouras, C. Papaloukas, and D. I. Fotiadis, "A two-stage methodology for sequence classification based on sequential pattern mining and optimization," *Data Knowl. Eng.*, vol. 66, no. 3, pp. 467–487, 2008.

[26] W. Li, J. Han, and J. Pei, "Cmar: Accurate and efficient classification based on multiple class-association rules," in *ICDM*. IEEE CS, 2001, pp. 369–376.

[27] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*, 2nd ed., ser. Adaptive Computation and Machine Learning. MIT Press, 2001.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.