An integrated approach for healthcare planning over multi-dimensional data using long-term prediction

Rui Henriques[1] and Cláudia Antunes[2]

[1] D2PM, IST–UTL, Portugal, rmch@ist.utl.pt,
[2] D2PM, IST–UTL, Portugal, claudia.antunes@ist.utl.pt

Abstract. The mining of temporal aspects over multi-dimensional data is increasingly critical for healthcare planning tasks. A healthcare planning task is, in essence, a classification problem over health-related attributes across temporal horizons. The increasingly integration of healthcare data through multi-dimensional structures triggers new opportunities for an adequate long-term planning of resources within and among clinical, pharmaceutical, laboratorial, insurance and e-health providers. However, the flexible nature and random occurrence of health records claim for the ability to deal with both structural attribute-multiplicity and arbitrarily-high temporal sparsity. For this purpose, two solutions using different structural mappings are proposed: an adapted multi-label classifier over denormalized tabular data and an adapted multiple time-point classifier over multivariate sparse time sequences. This work motivates the problem of long-term prediction in healthcare, and places key requirements and principles for its accurate and efficient solution.

1 Introduction

New planning opportunities are increasingly triggered by the growing amount, quality and integration of healthcare data through multi-dimensional structures. Research in classification over healthcare domains has been focused on early diagnosis, series description and treatment selection [5]. The mining of temporal dynamics have been mainly applied over physiological signals, with few additional methods over sequential genomic and proteomic structures.

Dealing with temporal aspects in broader contexts represents an unprecedented opportunity for healthcare planning. An example is the task of planning hospital resources by predicting if a patient will need a specific treatment within upcoming periods. For this task, a multi-dimensional structure centered in health records is commonly adopted [25]. Although there are mappings to temporal and tabular structures for the ready-application of classifiers, as illustrated in Fig.1, the resultant temporal event-sparsity and attribute-multiplicity trigger the need for a new understanding and formulation. Additionally, challenges of long-term prediction in the healthcare include the ability to deal with different time scales [1][5], advanced temporal rules [33] and knowledge-based constraints for an accurate and efficient long-term learning with minimum domain-specific noise [3].

The document is structured as follows. Section 2 introduces the current challenges of healthcare planning tasks. In sections 3 and 4, the problem of long-term prediction over multi-dimensional structures is motivated and formulated. Section 5 places key requirements for its accurate and efficient solution. Finally, an overview of the relevant work in long-term prediction is synthesized.

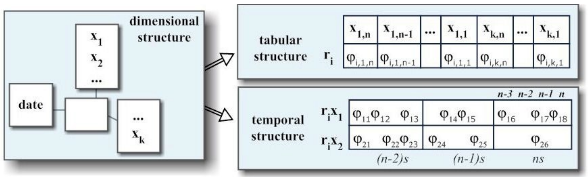


Fig.1: Structural mappings to apply long-term predictors

2 Healthcare planning challenges

Healthcare planning tasks include improvement of care pathways, detection of inefficiencies, resources allocation, health trends retrieval, and study of drug reactions and treatment effects. These applications have the potential to improve health, increase patient satisfaction and reduce costs. Similarly to predictive medicine, planning models can personalize guidelines to the characteristics of a single patient [5]. However, the assessment of his current state does not suffice. Evolving behavior across temporal horizons needs to be present.

2.1 Integrated healthcare data

In the last decade, new patient-centric data sources emerged. Countries as United Kingdom and Netherlands, already track patients' movements across health providers, payors and suppliers. The changing landscape has been shaped by: *i)* consumer-pushed demand through direct-access to risk and diagnosis information outside of the hospital setting, *ii)* new requirements for drug and treatment development, *iii)* the use of expert-systems to support medical decisions for quality compliance, and *iv)* remote home monitoring.

The stakeholders of healthcare planning are: *i)* data generators as hospitals, clinics, payors, pharmacies, e-prescription companies, laboratories and diagnostic-services providers; *ii)* data collectors as e-record vendors; and *iii)* data analyzers as pharmacos, application vendors and the research community.

Datasets are increasingly less fragmented, with appearing both cross-country and cross-player offerings, as provided by Cegedim and IMS. Datasets are derived from claims (Ingenix, D2Hawkeye, CMS), e-health records (McKesson, GE, PracticeFusion), imported health records (GoogleHealth, HealthVault), content aggregators (Walters Kluer, Reed Elsevier, Thomson), patient communities (Alere, Pharos, SilverLink, WebMD, HealthBoards), consumer reports (Anthem, vimo, hospitalcompare), online worksite healthcare (iTrax, webConsult), and physician portals (Medstory, Sermo, Doctors.net.uk).

This work uses health records as the mean to organize the wide variety of episodes into a single and compact fact [25]. In order to deal with record data flexibility, which may include laboratory results, prescriptions, treatments, diagnostics, free-text and complex structures as time series, an health record defines what the fact represents and the type of its fields. Amounts are mined as ordinal symbols, free-text is ignored, and complex data is converted into categorical sets of symbols. In Fig.2, an illustrative health record is presented.

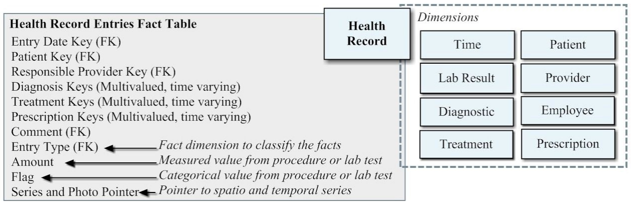

Health Record Entries Fact Table
Entry Date Key (FK)
Patient Key (FK)
Responsible Provider Key (FK)
Diagnosis Keys (Multivalued, time varying)
Treatment Keys (Multivalued, time varying)
Prescription Keys (Multivalued, time varying)
Comment (FK)
Entry Type (FK)
Amount
Flag
Series and Photo Pointer
Fact dimension to classify the facts
Measured value from procedure or lab test
Categorical value from procedure or lab test
Pointer to spatio and temporal series
Health Record
Dimensions
Time
Patient
Lab Result
Provider
Diagnostic
Employee
Treatment
Prescription

Fig.2: Health record-centered multi-dimensional structure

2.2 Emerging challenges

The increasing integration and volume of healthcare data trigger new challenges in terms of learning efficiency, attribute multiplicity and occurrence sparsity. These challenges are synthesized in Table 1.

Predictor requirement	**Healthcare data properties**
Methods to deal with missing values and event sparsity	Health records are irregularly collected due to an uneven schedule of visits or measurements made;
Strategies to deal with multivariate structures	Health records can flexibly define a high-multiplicity of attributes (e.g. wearables can produce measures for more than 20 attributes);
Efficient structural operations for record alignment and temporal partitioning	Health records' sampling grid varies both within and across patients;
Calendric-pattern discovery and aggregation techniques to deal with the different sampling rate of health records	Physiological measurements may be continuously generated, while administrative records as time-stamped prescriptions or hospitalizations exist at a coarser scale;
Convolutional memory techniques and pattern-based learning ability to detect evolving health trends	Evolutionary patterns, as the slow progress of a disorder or a reaction to a prescription or treatment, are spread across many potentially non-relevant health records;
Background knowledge guidance to avoid efficiency and domain-noise problems	The number of health records can be significantly high and its flexible nature hampers the learning;

Table 1: Critical requirements of healthcare planning

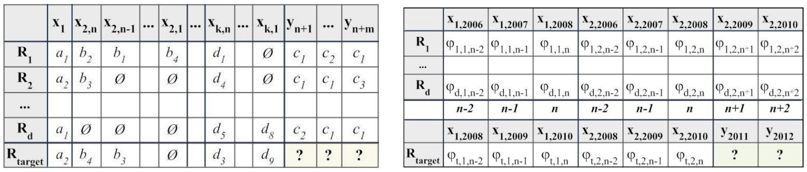
3 A need for a novel long-term prediction formulation

Given a training dataset of series composed by $n{+}h$ observations of the form (x, y), where $x{=}\{\varphi_1, ..., \varphi_n\}$ and $y{=}\{\varphi_{n+1}, ..., \varphi_{n+h}\}$, the task of traditional *long-term prediction* is to learn a predictive model to label the h next observations, where $h > 1$ is the *prediction horizon*.

This definition enters the scope of long-term prediction over series of elements. Since this work targets multi-dimensional structures characterized by multivariate and non-equally distant observations, there is the need to understand existing approaches and to incrementally extend this definition.

3.1 Limitations of long-term prediction over tabular structures

A simple way to deal with long-term prediction over multi-dimensional structures is to denormalize them into plain tabular structures. The long-term prediction task over tabular data can be target by an adapted multi-label classifier, as illustrated in Fig.3a. The goal of multi-label classification is to learn a model from an input dataset to predict a set of attributes whose class label is unknown.



	x_1	$x_{2,n}$	$x_{2,n-1}$	...	$x_{2,1}$	...	$x_{k,n}$	...	$x_{k,1}$	y_{n+1}	...	y_{n+m}
R_1	a_1	b_2	b_1		b_4		d_1		Ø	c_1	c_2	c_1
R_2	a_2	b_3	Ø		Ø		d_4		Ø	c_1	c_1	c_3
...												
R_d	a_1	Ø	Ø		Ø		d_5		d_8	c_2	c_1	c_1
R_{target}	a_2	b_4	b_3		Ø		d_3		d_9	?	?	?

(a) Denormalization

	$x_{1,2006}$	$x_{1,2007}$	$x_{1,2008}$	$x_{2,2006}$	$x_{2,2007}$	$x_{2,2008}$	$x_{2,2009}$	$x_{2,2010}$
R_1	$\varphi_{1,1,n-2}$	$\varphi_{1,1,n-1}$	$\varphi_{1,1,n}$	$\varphi_{1,2,n-2}$	$\varphi_{1,2,n-1}$	$\varphi_{1,2,n}$	$\varphi_{1,2,n+1}$	$\varphi_{1,2,n+2}$
...								
R_d	$\varphi_{d,1,n-2}$	$\varphi_{d,1,n-1}$	$\varphi_{d,1,n}$	$\varphi_{d,2,n-2}$	$\varphi_{d,2,n-1}$	$\varphi_{d,2,n}$	$\varphi_{d,2,n+1}$	$\varphi_{d,2,n+2}$
	n-2	*n-1*	*n*	*n-2*	*n-1*	*n*	*n+1*	*n+2*
	$x_{1,2008}$	$x_{1,2009}$	$x_{1,2010}$	$x_{2,2008}$	$x_{2,2009}$	$x_{2,2010}$	y_{2011}	y_{2012}
R_{target}	$\varphi_{t,1,n-2}$	$\varphi_{t,1,n-1}$	$\varphi_{t,1,n}$	$\varphi_{t,2,n-2}$	$\varphi_{t,2,n-1}$	$\varphi_{t,2,n}$	?	?

(b) Temporal shifting

Fig.3: Long-term prediction over tabular data

This option has several challenges. *First*, its viability strongly depends on the ability to represent the predicting horizon as attributes, and on the temporal compliance with the dataset instances. *Second*, by capturing each health record as a set of attributes, the size of the table may grow dramatically, which can significantly reduce the efficiency of the learning process and the accuracy of the classification model. *Third*, multi-label classifiers are neither prone to deal with ordinal attributes nor to capture the temporal dependencies among them. To solve these challenges, adaptations to multi-label classifiers are required to consider temporal dynamics and to constrain the learning over large datasets.

3.2 Limitations of long-term prediction over temporal structures

Let us assume that a mapping between a multi-dimensional and a temporal structure is possible, as illustrated in Fig.4. Let time sequences be the target temporal structure, as they do not put constraints on the arriving distribution of events. With this formulation, three challenges arise. The *first* challenge is of adapting predictors to deal with arbitrary-high sparse time sequences. When mapping multi-dimensional data to time sequences, the rate of health records' occurrence per patient and across patients may vary significantly. Structural sparsity results from the alignment of records across time points. Additionally, a mapping between a multi-dimensional dataset into non-temporal sequences would only consider events' precedences (Fig.4). Thus, existing learners are not ready for time-sensitive prediction over time sequences.

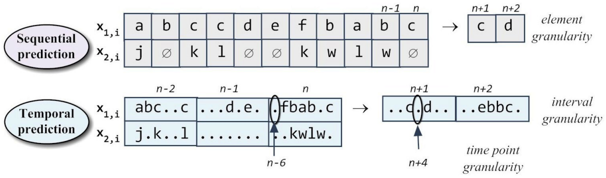


Fig.4: Sequential and temporal prediction over time sequences

The *second* challenge is of dealing with multivariate time sequences. Each arriving health record can be seen as a vector of optional attributes. When considering the task of predicting long-term hospitalizations, the attribute under prediction can be the first vector position, while the remainder vector positions correspond to optional prescriptions, symptoms, exams and diagnostics. These adjunct attributes may influence the attribute under prediction (determining if either a patient is hospitalized or not) and, therefore, are not conditionally independent from the attribute under prediction. At a finer granular level, vector positions can be physiological measures. Research on multivariate responses has been focusing on projecting multivariate attribute in the form of a matrix of responses, but assuming independence among them [26][11].

The *third* challenge is of performing long-term prediction in evolving contexts. Its relevance in planning problems is discussed in [19]. Different problems can be identified depending on the adopted predictor. Predictors can either scan local or large partitions depending, respectively, whether the time sequence is considered non-stationary or stationary. When considering local partitions nothing but cyclic behavior can be mined. When large partitions are considered and a lazy learner is adopted, over-fitting and pattern-negligence can arise. Alternatively, by adopting a non-lazy approach, the learner has to collapse all the predictive ability within a model. In this case, smoothing is an undesirable but possible result in order to avoid model complexity and over-fit propensity. The understanding of evolutionary patterns, beyond cyclic and calendric patterns, may be required to complement the learning [33]. An evolutionary pattern, as an evolving disease, can either be a subsequence whose occurrence is not cyclic but dependent on a time function or a cyclic subsequence whose arrangement progressively changes. The mining of evolutionary aspects can increase prediction accuracy, particularly for time-points near the horizon of prediction.

4 Problem formulation

Conventional predictors define a multiple-input single-output mapping. In *iterated* methods [6], a h-step-ahead prediction problem is tackled by iterating h times the one-step-ahead predictor. Taking estimated values as inputs, instead of actual observations, has negative impact in error propagation [37][32]. *Direct* methods learn h models, each returning a direct forecast. Although not prone to the accumulation of prediction errors [37], they require higher complexity to model the stochastic dependencies between non-similar series. Additionally, the fact that the n models are learned independently:

$$P(y|x) = P(\{\varphi_{n+1}, ..., \varphi_{n+h}\} \mid x) = \Pi_{i=1}^{h} P(\varphi_{n+i} \mid x),$$

prevents this approach from considering underlying dependencies among the predicted variables that may result in a biased learning [9][32].

Multiple-Input Multiple-Output (MIMO) methods learn one multiple-output predictive model. This favors efficiency and preserves the stochastic dependencies for a reduced bias, even though it reduces the flexibility of single-output approaches that may result in a new bias [9][6]. To avoid this, *Multiple-Input Several Multiple-Outputs* (MISMO) uses intermediate configurations to decompose the original task into $k = h/s$ prediction tasks, where h is the prediction horizon and s is the size of horizon's partitions. This approach trades off the property of preserving the stochastic dependency among future values with a greater flexibility of the predictor [41].

Def. 1: Single-output approaches either predict $\hat{\varphi}_i$: *i)* directly: $f_{i\in\{1,..,h\}}$ $(\varphi_n, .., \varphi_{n-d})$, where h is the prediction horizon, d is a subset of total observations (embedding dimension), and f is the stochastic predictor; or *ii)* iteratively as $f_{i\in\{1,..,h\}}$=$\{f_{i=1}(\varphi_n, ..., \varphi_{n-d}), f_{i\in\{2,...,d-1\}}(\hat{\varphi}_{n+i-1}, ..., \hat{\varphi}_{n+1}, \varphi_n, ..., \varphi_{n-d+i}),$ $f_{i\in\{d,...,h\}}(\hat{\varphi}_{n+i-1}, ..., \hat{\varphi}_{n+i-d})\}$. *Multiple-output* approaches predict the h=ks time-points within k steps $\{\varphi_{n+ps}, ..., \varphi_{n+(p-1)s+1}\}$=$f_p(\varphi_n, \varphi_{n-1}, ..., \varphi_{n-d+1})$, with $p \in \{1, ..., k\}$, where s, the predictor's variance, constrains the perseveration of stochastic properties of the series, null if s=1 and maximal if s=n.

4.1 Tabular formulation

Def.2 Consider a training dataset consisting of a set of m instances of the form $(x_1, ..., x_n, y_1, ...y_h)$, such that $(y_1, .., y_h) \in Y$ is either a numeric or a categorical vector (Y=$\mathbb{R}^h|\mathit{\Sigma}^h$); and each x_i takes on values from a domain $X_i = \cup_k\{(\mathit{\Sigma}_i|\mathbb{R}, k)\}$, where k∈$\mathbb{N}$ defines the event occurrence's order. The task of **long-term prediction over tabular data** is to construct either a single-output or multiple-output mapping model $M : \{X_1, ..., X_n\} \rightarrow Y$ for the multi-period classification of new tuples.

Using a multi-dimensional dataset, of we want to predict the number of hospitalizations for a patient j over a period i, y_i^j, we need to perform three steps. *First*, to use the patient dimension to select the health records per instance, x^j, based on patient primary key. *Second*, to use the time dimension for its ordering. *Finally*, to denormalize the health record measures in n attributes. For instance, using blood pressure measures, an example of patient j attributes is x^j=$\{x^j_{highbp}, x^j_{lowbp}\}$, with x^j_{highbp}=$\{(10, t_1), (9, t_2), (11, t_3), (\emptyset, t_4), .., (\emptyset, t_{max})\}$, and x^j_{lowbp}=$\{(7, t_1), (5, t_2), (7, t_3), (\emptyset, t_4), .., (\emptyset, t_{max})\}$.

4.2 Time sequence formulation

Considering a *sampling interval* $\tau \in \mathbb{R}$ and the *alphabet* $\mathit{\Sigma}$:

Def.3 A **time series** with regard to a series of equally-distant time points, $\mathbb{I} = \{\theta_1, ..., \theta_m\}$ with $\{\theta_i = \tau_0 + i\tau; i \in \mathbb{R}\}$ of length $m \in \mathbb{N}$, is z=$\{(\theta_i, \varphi_i) \mid \varphi_i$=$[\varphi_{i,1}, ..., \varphi_{i,d}]^T \in (\mathit{\Sigma}|\mathbb{R})^d, i$=$1, ..., n\}$. A **time sequence** is a multi-set of events, z=$\{(\varphi_i, \theta_j) \mid \varphi_i$=$[\varphi_{i,1}, ..., \varphi_{i,d}]^T \in (\mathit{\Sigma}|\mathbb{R})^d; i$=$1, ..., n; j \in \mathbb{N}^+\}$. z is univariate if d=1 and **multivariate** if d>1.

Considering the illustrative sparse time sequence $x = \{([10\ 7], \theta_1)$, $([\emptyset\ \emptyset], \theta_2)$, $([\emptyset\ \emptyset], \theta_3), ([9\ 5], \theta_4), ([13\ 7], \theta_5)$ composed of d=2 multivariate observations across n=5 time points. Long-term prediction consists of using a dataset of similar time sequences to predict a class of interest across h=$p \times s$ periods.

Def.4 Given a training dataset consisting of m instances of the form $(x, y_1, ..,$ $y_h)$, where x=$\{(\varphi_1, \theta_1), .., (\varphi_n, \theta_n)\} \in X$ is a time sequence of n-length and $y \in Y$ is a univariate time series a set of h-length, the task of **longterm prediction over multivariate and sparse temporal structures** is to construct either a single-output model or multiple-output model $M : X \rightarrow Y$, where h>1 and $\varphi_i = [\varphi_{i,1}, .., \varphi_{i,d}]^T$ with d>1.

The process of mapping a multi-dimensional dataset into a multivariate time sequence differs from the previous on the third step. The time dimension is now used to distribute the events' occurrence according to a timeline, instead of a simple ordering. The definition of aggregating functions as the average, sum or count, can be used for events' composition in coarser-granular time scales.

4.3 Problem generalization

Depending on the availability of training tuples compliant with the temporal horizon of prediction and on the allowance of temporal shifts, we may benefit to transit from a pure supervised solution into an hybrid one. Fig.3b illustrates a case where a 2-year shift is required. If significant noise results from this action, additional semi-supervised and unsupervised principles are required [33][1].

4.4 Evaluation

The evaluation of long-term predictors requires different metrics than those used in traditional classification. The *accuracy* of a predictive model is the probability that the predictor correctly labels multiple time points, $P(\hat{y}{=}y)$.

When the class for prediction is ordinal, the accuracy of the long-term predictor should be based on a similarity function. The average normalized root mean squared error (NRMSE) and the symmetric mean absolute percentage of error (SMAPE) have been employed in the literature. If the class for prediction is nominal, the similarity function should be replaced by the intersection operator.

$$NRMSE(y,\hat{y}){=}\tfrac{1}{h}\tfrac{\Sigma_{i=1}^{h}\sqrt{(y_i-\hat{y}_i)^2}}{y_{max}-y_{min}} \quad SMAPE(y,\hat{y}){=}\tfrac{1}{h}\Sigma_{i=1}^{h}\tfrac{|y_i-\hat{y}_i|}{(y_i+\hat{y}_i)/2} \text{ [6]}$$

$$Accuracy_{ord}{=}\tfrac{1}{m}\Sigma_{j=1}^{m}1-(\mathrm{NRMSE}(y^j,\hat{y}^j) \mid \mathrm{SMAPE}(y^j,\hat{y}^j))$$

$$Accuracy_{nom}{=}\tfrac{1}{m}\Sigma_{j=1}^{m}(\tfrac{1}{h}\Sigma_{i=1}^{h} \mid y_i^j \cap \hat{y}_i^j \mid)$$

Predictor's *efficiency* should be measured in terms of memory consumed and time elapsed for both the training (model learning) and prediction stages.

Finally, complementary metrics to understand the predictor's *error accumulation* [14] and *smoothness* [14], when noise fluctuations are present, should be adopted for a deeper understanding of the predictor's behavior.

5 Solution Space

Key variables, illustrated in Fig.5, must be considered to solve the introduced requirements of long-term prediction in healthcare planning.

For instance, the *target data* variable is dependent on the health records' representation, degree of sparsity, noise sensitivity, completeness, length, degree of content-stationarity, presence of static features, multivariate order, patterns presence, attributes' alphabet amplitude, and sensitivity to temporal shifts.

5.1 Adopted learning approach

Several implementations for both single-output and multiple-output approaches exist. All of them, implicitly or explicitly, work around the multivariate con-

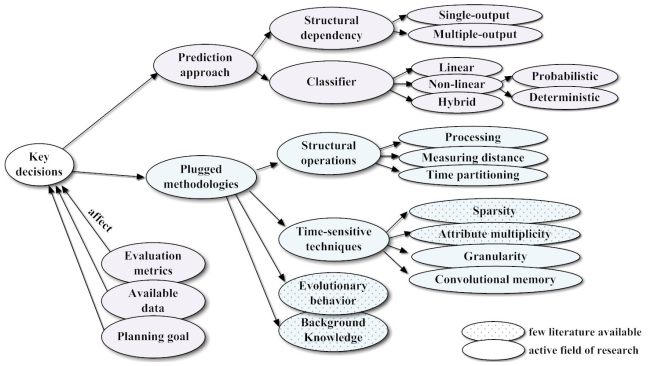


Fig.5: Key research areas for long-term prediction

ditional distribution $P(Y|X)$. Learners can either follow linear or non-linear predictive models. *Linear models* can either follow a simple, logistic or Poisson regression, as auto-regressions and feed-forward moving average mappings [28].

Non-linear long-term predictors can either define probabilistic or deterministic models. Most are adaptations of traditional classifiers using temporal sliding windows. Probabilistic predictors include (hidden) conditional random fields [24]; hidden and variable-memory Markov models (HMM) [4]; and stochastic grammars [13]. Deterministic predictors include support vector machines (SVM) [12]; recurrent, time-delay and associate neural networks [21][8]; multiple adaptive regression splines [28]; regression and model trees [35]; multiple lazy learning alternatives [9]; and genetic solvers [15].

5.2 Plugged methodologies

Despite the relevance of the learning approach choice, the significant performance improvements is triggered by the temporal criteria that predictors may adopt [33]. Since the learning of long-term predictors are NP-hard [31], the understanding of efficient structural operations, time-based strategies, temporal rules and knowledge-based constraints is key.

Structural operations. Suitable data representations [18], similarity-measures [18] and time-partitioning strategies are required for a quick and flexible learning. Criteria for temporal partitioning through clustering, user-defined granularities, fuzzy characterization, split-based sequential-trees, episodes, domain-driven ontologies and symbolic interleaving can be consulted in [34][31][1].

Time-sensitive techniques. Strategies to enhance the performance of long-term predictors for healthcare planning tasks are synthesized in Table 2.

Evolutionary behavior. Mining of evolutionary behavior, discussed in section 3, is required to avoid smoothing and overfitting problems. An understandable case is prediction rules, which specify a causal and temporal correlation between two time points or patterns. In [19], emerging or evolutionary patterns are defined as patterns whose support increases significantly over time. Although pattern-based classifiers seem a suitable choice, other approaches should not be excluded.

Requirement	**Research contributions**
data sparsity	Time windows can be adopted to create ordering or temporal partitions [34]. This can be used to constraint the exponential growth of denormalized tables and to selected the most recent health records' occurrences in order to avoid missing values. In time sequences, they can be used for flexible learning approaches, but additional techniques are required to deal with missing values, as proposed in [30][23];
temporal granularity	Different levels of granularity defined using time windows and feature-based descriptions hold the promise of minimizing problems of sparsity and efficiency, even though the data loss can degrade predictors' accuracy. To manage this trade-off, the study of improvements using hierarchical zooming operations [20] and calendars [2] is key;
data attributes multiplicity	Dependencies among attributes have been captured either through the use of additional attributes or by defining weights on how each attribute (e.g. prescription) influences a different attribute (e.g. symptom) across the time horizon [33]. In time sequences, learning strategies can be adopted to deal with sparse multivariate vectors that may constrain the vector under prediction [26]. One option is to project each vector or attribute to the target horizon [11], and to derive from them the vector under prediction;
memory sampling	Covariance functions, following either a parametric or non-parametric approach, are key for the selective forgetting of unimportant events and retaining of decisive events [21][8]. In both tabular and temporal structures, these functions can assign weights to each attribute or time point to be used by long-term predictors. Binary or exponentially decaying weighted average of an input function can be used to set a trade-off between *depth* (how far back memory goes) and *resolution* (the degree to which information about individual time-points is preserved).

Table 2: Long-term prediction principles and hypotheses

Examples may include the late combination of temporal rules within a predictive model or, alternatively, their initial retrieval to assist its learning [34].

Background knowledge. Finally, background knowledge is increasingly claimed as a requirement for long-term prediction as it guides the definition of time windows [34]; provides methods to bridge different time scales, to treat monitoring holes and to remove domain-specific noise [5]; defines criteria to prune the explosion of multiple-equivalent patterns [3]; and fosters the ability to adapt and incrementally improve results by refining the way domain-knowledge is represented [3]. A hierarchy of flexible sequential constraints, and of relaxations ranging from conservative to distance-based approximations is introduced in [1]. Further research on domain-driven time modeling is required [2].

6 Related research

Time series long-term prediction and *sequence learning* are the research streams with major relevant contributions for the introduced problem.

Long-term prediction. A comparative study on the performance of iterated and direct single-output approaches in terms of error accumulation, smoothness, and learning difficulty is presented in [38]. In literature, *hybrid* solutions that combine both approaches exist [38]. Experimental results in [32] show that the robustness and error reduction obtained using direct and hybrid forecasts do not justify the price paid in terms of increased sampling variance. Values for the MISMO variance parameter can be derived from query-point functions. Experimental studies [41] show that the s best-value strongly varies according dataset, with s=1 (Direct method) and s=n (MIMO method) being good performers in less than 20% of the cases. For large horizons h, improvements in multiple-output approaches have been achieved by adopting operators as the partial autocorrelation [41]. A comparison of five multi-step-ahead prediction methods, two single-output and three multiple-output predictors is done in [6].

Potential linear, probabilistic and deterministic classifiers were discussed. In [14] an hybrid HMM-regression is evaluated using different regression orders and time-windows sizes. Evaluation of three multiple-output neural predictors, simple feed-forward, modular feed-forward and Elman, is done in [7]. In [10], Bayesian learning is applied to recurrent neural networks to deal with noisy and non-stationary time series. In [9], multiple-output approaches, as least-squares SVM, are extended with query-based criteria grounded on local learning.

Sequence Learning. Sequence learning methods are adopted when the mining goal is either sequence prediction, sequence recognition or sequential decision making [39]. Sequence recognition can be formulated as a prediction problem.

Learning techniques as expectation maximization, gradient descendant, policy iteration, hierarchical structuring or grammar training can be transversally applied to different implementations [24]. Additionally, unsupervised and reinforcement learning techniques from machine learning have been applied to sequence prediction, even though are still not scalable for large data volumes.

First, *unsupervised learning* rely on motifs, calendric rules, episodes, containers and partially-ordered tones [1][33][31] to assist prediction. In [29], patterns are translated into boolean features to guide SVM and logistic regressions.

Second, *reinforcement learning* have been applied in inductive-logic predictors (that learn symbolic knowledge from sequences in the form of expressive rules) [27] and in evolutionary computing predictors (that use heuristic-search over probabilistic models of pattern likelihood) [40]. In [17], sequence-generating rules constrain which symbol can appear. In [16], series are used to train trees, from which rules are retrieved and combined with logical operators.

Finally, a large spectrum of implementations are hybrid. An example is the use of symbolic rules and evolutionary computation applied to neural networks [40]. The introduced reinforcement learning techniques are usually preferred when one is not interested in a specific temporal horizon, but rather in predicting the occurrence of a certain symbol or pattern.

Healthcare planning. A good survey covering temporal classification advances in healthcare can be found in [5]. In [36], simple planning problems were address relying on administrative health records including drug prescriptions, hospitalizations, outpatient visits, and daily hospital activities. State-based characterization using Markov models were, for instance, applied to predict the risk of stroke in sickle cell anemia patients [5]. Temporal abstractions have been used for multiple time-point classification of physiological signals. In [22], a collaborative approach is designed to mine biomedical multivariate time series to understand vector evolution. The mining of evolving health aspects for planning tasks have, however, received few attention. In [5], its combination with large-scale genomics and proteomics is pointed as a decisive step to characterize disease progression.

7 Conclusion

This work addresses healthcare planning problems using long-term prediction over multi-dimensional structures centered in health records. It introduced two formulations based on structural mappings into tabular structures and multivariate sparse temporal structures. The combination of unsupervised and reinforcement learning techniques should be present when the training and testing tuples are not temporally compliant and sensitive to temporal shifts. Evaluation metrics for the target problem are proposed.

A set of requirements were introduced to deal with attribute multiplicity and temporal-sparsity of health records. Contributions were identified. Literature is either focused on long-term prediction over single-attributes or on causal learning, not answering the introduced challenges. Empirical contributions in the form of principles that satisfy one or more of these requirements are the expected next steps to promote an efficient learning of accurate long-term predictors with minimum domain-specific noise.

Acknowledgment: this work is partially supported by *Fundação para a Ciência e Tecnologia* under the research project D2PM, PTDC/EIA-EIA/110074/2009, and the PhD grant SFRH/BD/75924/2011.

References

1. Antunes, C.: Pattern Mining over Nominal Event Sequences using Constraint Relaxations. Ph.D. thesis, Instituto Superior Tecnico (2005)
2. Antunes, C.: Temporal pattern mining using a time ontology. In: EPIA. pp. 23–34. Associação Portuguesa para a Inteligência Artificial (2007)
3. Antunes, C.: An ontology-based framework for mining patterns in the presence of background knowledge. In: ICAI. pp. 163–168. PTP, Beijing, China (2008)
4. Begleiter, R., El-Yaniv, R., Yona, G.: On prediction using variable order markov models. J. Artif. Int. Res. 22, 385–421 (2004)
5. Bellazzi, R., Ferrazzi, F., Sacchi, L.: Predictive data mining in clinical medicine: a focus on selected methods and applications. Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery 1(5), 416–430 (2011)
6. Ben Taieb, S., Sorjamaa, A., Bontempi, G.: Multiple-output modeling for multi-step-ahead time series forecasting. Neurocomput. 73, 1950–1957 (2010)
7. Bengio, S., Fessant, F., Collobert, D.: Use of modular architectures for time series prediction. Neural Process. Lett. 3, 101–106 (1996)
8. Berthold, M., Hand, D.J. (eds.): Intelligent Data Analysis: An Introduction. Springer-Verlag New York, Inc., Secaucus, NJ, USA (1999)
9. Bontempi, G., Ben Taieb, S.: Conditionally dependent strategies for multiple-step-ahead prediction in local learning. Int. J. of Forecasting 27(2004), 689–699 (2011)
10. Brahim-Belhouari, S., Bermak, A.: Gaussian process for nonstationary time series prediction. Computational Statistics and Data Analysis 47(4), 705 – 712 (2004)
11. Brown, P.J., Vannucci, M., Fearn, T.: Multivariate bayesian variable selection and prediction. Journal of the Royal Statistical Society 60(3), 627–641 (1998)
12. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. 2, 121–167 (1998)
13. Carrasco, R.C., Oncina, J.: Learning stochastic regular grammars by means of a state merging method. In: ICGI. LNCS, vol. 862, pp. 139–152. Springer (1994)
14. Cheng, H., Tan, P.N., Gao, J., Scripps, J.: Multistep-ahead time series prediction. In: Advances in Knowl. Disc. and Data Mining, LNCS, vol. 3918, pp. 765–774. Springer Berlin, Heidelberg (2006)
15. Cortez, P., Rocha, M., Neves, J.: A Meta-Genetic Algorithm for Time Series Forecasting. In: Proc. of AIFTSA'01, EPIA'01. pp. 21–31. Porto, Portugal (2001)
16. Cotofrei, P., Neuchâtel, U.: Rule extraction from time series databases using classification trees. In: Proc. of the 20th IASTED. pp. 327–332. ACTA Press (2002)

17. Dietterich, T.G., Michalski, R.S.: Discovering patterns in sequences of events. Artif. Intell. 25, 187–232 (1985)
18. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E.J.: Querying and mining of time series data: experimental comparison of representations and distance measures. Proceedings of the VLDB Endowment 1(2), 1542–1552 (2008)
19. Dong, G., Li, J.: Efficient mining of emerging patterns: discovering trends and differences. In: 5th ACM SIGKDD. pp. 43–52. KDD, ACM, NY, USA (1999)
20. Fang, Y., Koreisha, S.G.: Updating arma predictions for temporal aggregates. Journal of Forecasting 23(4), 275–296 (2004)
21. Guimarães, G.: The induction of temporal grammatical rules from multivariate time series. In: 5th ICGI. pp. 127–140. Springer-Verlag, London, UK (2000)
22. Guyet, T., Garbay, C., Dojat, M.: Knowledge construction from time series data using a collaborative exploration system. J. of Biomedical Inf. 40, 672–687 (2007)
23. Hsu, C.N., Chung, H.H., Huang, H.S.: Mining skewed and sparse transaction data for personalized shopping recommendation. Mach. Learn. 57, 35–59 (2004)
24. Kersting, K., Raedt, L.D., Gutmann, B., Karwath, A., Landwehr, N.: Relational sequence learning. In: Probab. ILP. LNCS, vol. 4911, pp. 28–55. Springer (2008)
25. Kimball, R., Ross, M.: The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. John Wiley & Sons, Inc., NY, USA, 2nd edn. (2002)
26. Koch, I., Naito, K.: Prediction of multivariate responses with a selected number of principal components. Comput. Stat. Data Anal. 54, 1791–1807 (2010)
27. Lavrac, N., Dzeroski, S.: Inductive Logic Programming: Techniques and Applications. Ellis Horwood, New York, NY, USA (1994)
28. Laxman, S., Sastry, P.S.: A survey of temporal data mining. Sadhana-academy Proceedings in Engineering Sciences 31, 173–198 (2006)
29. Lesh, N., Zaki, M.J., Ogihara, M.: Mining features for sequence classification. In: Proc. of the 5th ACM SIGKDD. pp. 342–346. ACM, NY, USA (1999)
30. Liu, J., Yuan, L., Ye, J.: An efficient algorithm for a class of fused lasso problems. In: Proc. of the 16th ACM SIGKDD. pp. 323–332. KDD, ACM, NY, USA (2010)
31. Mannila, H., Toivonen, H., Inkeri Verkamo, A.: Discovery of frequent episodes in event sequences. Data Min. Knowl. Discov. 1, 259–289 (1997)
32. Marcellino, M., Stock, J.H., Watson, M.W.: A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. Journal of Econometrics 135(1-2), 499–526 (2006)
33. Mörchen, F.: Time series knowledge mining. W. in Dissertationen, G&W (2006)
34. Mörchen, F.: Tutorial cidm-t temporal pattern mining in symbolic time point and time interval data. In: CIDM. IEEE (2009)
35. Quinlan, J.R.: Learning with continuous Classes. In: 5th Australian Joint Conf. on Artificial Intelligence. pp. 343–348 (1992)
36. Silberschatz, A., Tuzhilin, A.: What makes patterns interesting in knowledge discovery systems. IEEE Trans. on Knowl. and Data Eng. 8, 970–974 (1996)
37. Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y., Lendasse, A.: Methodology for long-term prediction of time series. Neurocomput. 70, 2861–2869 (2007)
38. Sorjamaa, A., Lendasse, A.: Time series prediction using dirrec strategy. In: ESANN. pp. 143–148 (2006)
39. Sun, R., Giles, C.L.: Sequence learning: From recognition and prediction to sequential decision making. IEEE Intelligent Systems 16, 67–70 (2001)
40. Sun, R., Peterson, T.: Autonomous learning of sequential tasks: experiments and analyses. IEEE Transactions on Neural Networks 9(6), 1217–1234 (1998)
41. Taieb, S.B., Bontempi, G., Sorjamaa, A., Lendasse, A.: Long-term prediction of time series by combining direct and mimo strategies. In: Proc. of the 2009 IJCNN. pp. 1559–1566. IEEE Press, Piscataway, USA (2009)