On the need of new approaches for the novel problem of long-term prediction over multi-dimensional data

Rui Henriques and Cláudia Antunes

Abstract Mining evolving behavior over multi-dimensional structures is increasingly critical for planning tasks. On one hand, well-studied techniques to mine temporal structures are hardly applicable to multi-dimensional data. This is a result of the arbitrary-high temporal sparsity of these structures and of their attribute-multiplicity. On the other hand, multi-label classification over denormalized data do not consider temporal dependencies among attributes.
This work reviews the problem of long-term classification over multi-dimensional structures to solve planning tasks. For this purpose, firstly, it presents an essential formalization and evaluation method for this novel problem. Finally, it extensively overviews potential relevant contributions from different research streams.

1 Introduction

New planning opportunities are increasingly triggered by the growing amount, completeness and precision of temporal data. The integration of data in multi-dimensional structures have been enabled through the world-wide adoption of data warehouses. For this setting, the study of long-term prediction in evolving contexts can increasingly provide additional value [6][34]. Applications may range from clinical prevention to several planning tasks in retail, educational, commercial, financial and social security domains [26][6]. An example may be the long-term planning of hospital resources based on underlying healthcare needs (for instance, seen as the need of a patient get a specific treatment within upcoming years).

The mining of temporal dynamics using multistep-ahead classifiers has been mainly applied to temporal and sequential structures [48][10]. In practice, this body of knowledge is hardly applicable to multi-dimensional data structures. Although mappings between these structures exist, the resulting temporal event-sparsity and

Department of Computer Science and Engineering, Instituto Superior Técnico, UL, Portugal
Contact: rmch@ist.utl.pt, claudia.antunes@ist.utl.pt

attribute-multiplicity claim for new research. Additionally, although multi-label classifiers can be adopted by denormalizing multi-dimensional into tabular structures, they fail to deal with temporal dependencies. Further challenges of long-term prediction in evolving contexts include the ability to deal with different time scales [1][7], advanced temporal rules [4] and knowledge-based constraints [3].

When considering, for example, an hospital planning task, multi-dimensional structures are centered in health-records (the fact) and grouped based on a dimension, usually the patient. Health records may track a multiplicity of measures related to diagnostic, prescription or treatment dimensions. Additional challenges arise from the arbitrary sparse nature of measure recordings. Thus, this structure does not allow the application of existing well-studied predictors.

The document is structured as follows. Section 2 discusses the novelty of this problem and introduces a case to illustrate its significance. Section 3 formulates the problem of long-term prediction over multi-dimensional structures. Section 4 discusses the key aspects for its correct assessment. Finally, an overview of the relevant work in long-term prediction is synthesized and existing contributions plugged as potential principles to address the problem.

2 Why a New Long-term Prediction Formulation?

Consider a training dataset following a simple multi-dimensional structure, a star scheme centered in one fact that track events unevenly spread across time. In healthcare, a health record can flexibly capture measures related to laboratory results, prescriptions, treatments and diagnostics. Health records can be used to learn a multi-step-ahead learner that classifies a measure across multiple periods for planning tasks as the hospitalization needs of a patient in the upcoming months. Lower-level planning tasks and personalized tasks (as the prediction of evolving physiological states) can be additionally consider. Despite the relevance of this problem, to the best of our knowledge there is not yet a dedicated research stream for its solution.

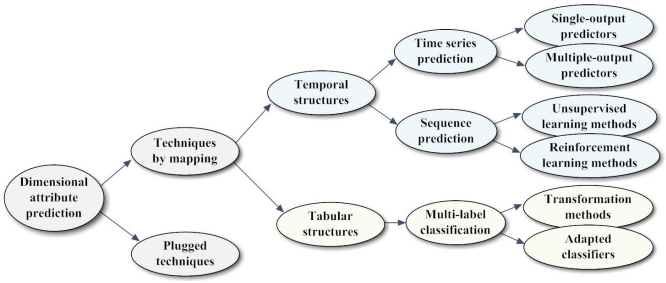

Fig. 1: Contributions from existing research streams

Fig.1 gathers the existing research streams that can provide important contributions to the development of novel approaches for long-term prediction. These contributions may be applied *as-is* or through mappings into tabular and temporal structures, as illustrated in Fig.2. Next subsections synthesize their limitations.

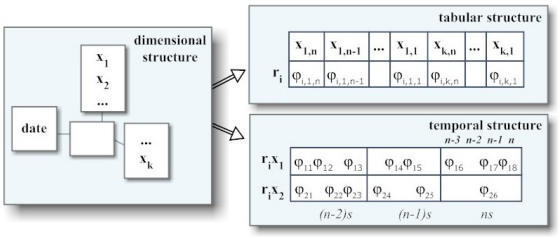

Fig. 2: Structural mappings for an adapted multistep-ahead classification

2.1 Limitations of multi-label classifiers

A simple procedure to deal with prediction over multi-dimensional structures is to denormalize these structures into a plain tabular structure and to apply a multi-label classifier. A first set of challenges appear as multi-label classifiers were developed with a different purpose – categorization and multi-parameter diagnosis. *First*, multi-label classifiers are not prone to label ordinal attributes, but to label disjoint binary attributes. Labeling of nominal attributes requires the mutual exclusive labeling of binary attributes. *Second*, attributes are considered to be independent. Temporal dependencies are not captured in the learning process.

Additionally, two core and severe problems arise when adopting tabular structures. *First*, when capturing each event occurrence as an attribute, the size of the table may grow dramatically, which can significantly reduce the efficiency of the learning process and the accuracy of the classifier. *Second*, its viability strongly depends on the ability to capture the temporal dependencies both among domain attributes and across the periods under prediction.

2.2 Limitations of sequence and time series predictors

Given a training dataset with series composed by $n+h$ observations, *multi-step-ahead prediction over univariate series* is the task of learning a model to predict the h next observations of a n-length series, where $h>1$ is the *prediction horizon*.

These series can either be sequences, as a genetic arrangement, or time series, as a physiological signal. Research streams on these areas are important if we want

to work with one measure-of-interest. To deal with the multiple measures, one can treat these measures as a multivariate sparse temporal structure. Let us assume that a mapping between a dimensional and a temporal structure is possible. In this way, the problem would be the prediction of multiple periods based on two types of input: sparse multivariate time sequences derived from fact measures, and static attributes derived from dimensions. With this formulation, three challenges arise.

The *first challenge* is to adapt long-term predictors to deal with arbitrary-high sparse time sequences. The rate of events' occurrence per fact may vary significantly across time. This happens within and among patients in healthcare systems, but may also occur with banking applications, genetic mutations or almost every dynamic daily-life system. Structural sparsity results from the alignment of events across time points. Sequential predictors can solve this problem as they are focused on causalities [40]. However, since they do not account for temporal distances among events, they are non-expressive candidates.

The *second challenge* is to enrich long-term prediction in order to deal with multivariate time sequences. For instance, determining if a patient is hospitalized across multiple periods is conditionally dependent on the patient clinical history composed by multiple attributes related to prescriptions, samplings, treatments and diagnosis.

The *third challenge* is to perform long-term prediction in evolving contexts. The relevance of understanding evolutionary behavior in planning problems through predictive rules is discussed in [26][22]. Predictors can either scan large or local periods of a fact measure depending if the measure is considered to be stationary or non-stationary. In the first case, cyclic behavior is key [26]. In the second case, understanding of evolving and calendric behavior is critical and rarely considered.

2.3 Illustration

Challenges for long-term prediction in healthcare are synthesized in Table 1.

Healthcare data challenges	**Requirement**
Health records may define multiple measures of interest;	Strategies to deal with multivariate structures
The number of health records can be significantly high and its flexible nature may hamper the learning;	Background knowledge guidance to avoid efficiency and domain-noise problems
Health records are irregularly collected due to an uneven schedule of visits or measurements made;	Methods to deal with missing values and event sparsity
Health record sampling grid varies both among and within patients;	Efficient structural operations for record alignment and time partitioning
Different measures can be recorded at distinct time scales;	Calendric-mining and aggregation techniques to deal with the different sampling rate of health records
Evolving patterns, as the progress of a disorder or a reaction to a prescription, are spread across many non-relevant records;	Convolutional memory techniques and pattern-based learning ability to detect evolving health trends

Table 1: Requirements for long-term prediction over dimensional healthcare data

3 Problem formulation

This section formalizes the target problem. For this task, a formal review of the underlying concepts is introduced.

3.1 Underlying definitions

Def. 1 Consider a dataset of training instances, $D = \{x_1, ...x_n\}$, of the form $x_i = (a, \mathbf{y})$, where $a = \{a_1 \in A_1, .., a_m \in A_m\}$ is a set of input attributes and $\mathbf{y} = (y_1, .., y_h) \in Y^h$ is a vector of either numeric or categorical symbols, where $h > 1$ is the horizon of prediction. Given a training dataset, the task of *long-term prediction* is to learn a mapping model $M : A \rightarrow Y^h$ that accurately and efficiently predicts y based on a particular x, i.e. y=$M(x)$.

Given a training dataset D, the task of *long-term prediction over tabular data* is to learn a model $M : A \rightarrow Y^h$, where the domain is a set of alphabets, $A = \{\Sigma_1, .., \Sigma_m\}$.

Although this problem is similar to a multi-label classification problem, its definition explicitly considers conditional temporal-dependency among $\mathbf{y}$ symbols and includes ordinal attributes.

Def. 2 Let Σ be an alphabet of symbols σ, and $\tau \in \mathbb{R}$ be the sample time interval of a series, $\{\theta_i \mid \theta_i = \tau_0 + i\tau; i \in \mathbb{N}\}$. A *sequence* s is a vector of symbols

s=$(\varphi_1, .., \varphi_n)$, with φ_i=$[\varphi_{i,1}, .., \varphi_{i,d}] \in (\mathbb{R}|\Sigma_i)^d$.

A *time series* t with regard to θ_i, is given by

t=$\{(\varphi_i, \theta_i)|\varphi_i$=$[\varphi_{i,1}, .., \varphi_{i,d}] \in (\mathbb{R}|\Sigma_i)^d, i$=$1, .., n\} \in \mathbb{T}^{n,d}$.

A *time sequence* w is a multi-set of events

w=$\{(\varphi_i, \theta_j)|\varphi_i$=$[\varphi_{i,1}, .., \varphi_{i,d}] \in (\mathbb{R}|\Sigma_i)^d; i$=$1, .., n; j \in \mathbb{N}\}$.

s, t and w are univariate if $d = 1$ and *multivariate* if $d > 1$.

The domains of sequences is $\mathbb{S}^{n,d}$, time series is $\mathbb{T}^{n,d}$, and time sequences is $\mathbb{W}^{n,d}$, where n is the length and d the multivariate order.

Exemplifying, a univariate time series capturing monthly hospitalizations can be $\mathbf{y}$=$\{(0, \tau_1), (3\text{-}5, \tau_2), (>5, \tau_3), (2, \tau_4)\}$, with $y \in \mathbb{T}^{4,1}$. A multivariate time sequence capturing two physiological measures from blood tests can be $a = \{([2\ 21], \tau_2), ([3\ 19], \tau_3), ([2\ 20], \tau_5)\}$, with $a \in \mathbb{W}^{6,2}$.

Given a training dataset D, the well-studied task of *time series long-term prediction* problem is to learn a mapping model $M : A \rightarrow Y^h$, where $A = \mathbb{T}^{m,1}$, and A and Y values are either numeric or share the same alphabet Σ.

Given a training dataset D, the task of *long-term prediction over multivariate sparse temporal structures* is to define the map model $M : A \rightarrow Y^h$, where $A = \mathbb{W}^{m,d}$.

Def. 3 Given a training dataset D with tuples in the form of (a=$\{a_1 \in A_1, .., a_m \in A_m\}, \mathbf{y} \in Y^h$), the task of *long-term prediction over multi-dimensional data* is to construct a mapping model $M : \{A_1, .., A_m\} \rightarrow Y^h$, where a attributes are either a symbol or a time sequence of l-length and multivariate d-order ($X_i = \Sigma \mid \mathbb{W}^{l,d}$), $\mathbf{y}$ is a vector of h symbols, and $h > 1$.

Exemplifying, an instance, $(\{x_1, x_2, x_3\}, y)$, can represent a patient, where x_1 is his age, x_2 and x_3 are two multivariate time sequences capturing measures from blood and urine tests, and $\mathbf{y}$ is his number of hospitalizations across different periods. The goal is to learn a model, based on a training dataset, to predict multi-period hospitalizations $\mathbf{y}$ for a patient based on health-related data x.

3.2 Long-term prediction over dimensional data

In order to understand how to derive x from a multi-dimensional dataset, some concepts are formalized below.

A *multi-dimensional data structure*, $\{(\cup_i \{Dim_i\}) \cup Fact\}$, is defined by a set of dimensions, Dim_i, and one fact, $Fact$. Each dimension contains a primary key, a set of attributes a, and no foreign keys. The fact contains one foreign key for each dimension and a set of measures $(b_1, .., b_d)$.

Two special dimensions, the *time* and *select* dimensions, need to be identified. The select dimension, the patient in the healthcare example, is used to group the multiple fact occurrences across time in n instances $(x_1, .., x_n)$ according to the primary keys in this dimension. The number of instances, n, is given by the number of these primary keys.

Given a multi-dimensional dataset D, its mapping in a set of instances of the form $(a_1, .., a_m)$ follows a three-stage process. *First*, using D select-dimension, the set of all fact occurrences is grouped in n instances $(x_1, .., x_n)$. *Second*, the set of fact occurrences for each instance is mapped into a multivariate time sequence, b=$\{(b_i, \theta_j) \mid \varphi_i$=$[b_{i,1}, .., b_{i,d}]; i$=$1, .., n; j \in \mathbb{N}\}$, where the order d is the number of fact measures. *Third*, the attributes from dimensions are captured as one-valued attributes, a_{Dim_i}=$(a_{i,1}, .., a_{i,|Dim_i|})$. After this three-step process the instances follow the form $(a_1, .., a_m)$, where a_i attribute is either derived from a dimension or a multivariate time sequence of l-length derived from a fact (a_i=$b \in \mathbb{W}^{l,d}$).

Illustrating, consider the dimensional dataset $\{Dim_{patient}$= $\{id_2, A_{1,1}, A_{1,2}\}, Dim_{lab}$ =$\{id_3, A_{2,1}\}, Dim_{time}, Fact_{hr}$=$\{fk_1, fk_2, fk_3, B_1, B_2\}\}$. An example of a retrieved patient tuple is $(a_{1,1}, a_{1,2}, a_{2,1}, \{([b_{1,1}\ b_{1,2}], \theta_1), ([b_{2,1}\ b_{2,2}], \theta_4), ([b_{3,1}\ b_{3,2}], \theta_5)\})$.

In real-world planning tasks, the training dataset may not be temporally compliant with the instance under prediction. Two strategies can be used in these cases: allowance of temporal shifts to the training tuples and project temporal behavior from unsupervised learning (transiting from a pure supervised into an hybrid solution). For instance, if we consider health records between 2005 and 2011, and we want to predict the hospitalizations for a patient until 2014, the model can either rely on a 3-year temporal shift and on the projection of cyclic and calendric patterns.

Finally, the domain and properties of the adopted datasets (either multi-dimensional, relational or series-based) should be made available. Variables should not only include sensitivity to temporal shifts, but additionally the degree of sparsity, noise sensitivity, completeness, length, degree of stationarity, presence of static features, discretization constraints, and, in the case of temporal structures, allowance for itemsets, multivariate order and alphabet amplitude.

4 Evaluation

Long-term prediction requires different metrics than those used in traditional single-label classification. This section presents the set of metrics adopted in the literature, and proposes a roadmap to evaluate long-term predictors.

Predictor's *efficiency* is measured in terms of memory and time cost for both the training and testing stages. The *accuracy* of a predictive model is the probability that the predictor correctly labels multiple time points, $P(\hat{y} = y)$. This probability is usually calculated over a train dataset using a 10-fold cross-validation scheme. If not, disclosure of the adopted sampling test technique (e.g. holdout, random subsampling, bootstrap) needs to be present. Accuracy can be employed using similarity or loss functions applied along the horizon of prediction. Next sections review ways to translate horizon-axis plots of accuracy into a single metric.

4.1 Predictor's accuracy

First, we visit metrics both from time series prediction and multi-label classification, required if someone wants to establish comparisons with these works. Second, we introduce key metrics to cover different accuracy perspectives for this problem.

Multistep-ahead prediction:

The simplest way of understanding the accuracy of a multistep-ahead predictor is to use the mean absolute error (MAE), the simple mean squared error (MSE), the mean relative absolute error (RAR) or the average normalized mean squared error (NMSE), the ratio between the MSE and the time series variance. The normalized root mean squared error (NRMSE) either uses the series amplitude (when the attribute under prediction is numeric) or the number of labels (when the target attribute

is ordinal) to normalize the error. The accuracy is sometimes assessed through the symmetric mean absolute percentage of error (SMAPE) [8]. The average SMAPE over all time series under test is referred as SMAPE*. Other less frequent metrics, as the average minus log predictive density (mLPD) or relative root mean squared-error (RRMSE), are only desired for very specific types of datasets and, therefore, are not consider.

In fact, every similarity function can be used to compute a normalized distance error. A detailed survey of similarity-measures is done in [21]. Euclidean-distance, similarly to SMAPE, is simple and competitive. Dynamic Time Warping treats misalignments, which is important when dealing with long horizons of prediction. Longest Common Subsequence deals with gap constraints. Pattern-based functions consider shifting and scaling in both temporal and amplitude axis. These similarity functions have the advantage of smoothing error accumulation, but the clear drawback of the computed accuracy to not be easily comparable with literature results.

$$NMSE(y,\hat{y}) = \frac{\frac{1}{h}\Sigma_{i=1}^{h}(y_i-\hat{y}_i)^2}{\frac{1}{h-1}\Sigma_{i=1}^{h}(y_i-\bar{y})^2}$$

$$NRMSE(y,\hat{y}) = \frac{\sqrt{\frac{1}{h}\Sigma_{i=1}^{h}(y_i-\hat{y}_i)^2}}{y_{max}-y_{min}} \in [0,1]$$

$$SMAPE(y,\hat{y}) = \frac{1}{h}\Sigma_{i=1}^{h}\frac{|y_i-\hat{y}_i|}{(y_i+\hat{y}_i)/2} \in [0,1]$$

To compute the predictor's accuracy, the multiple correlation coefficient R^2 is adopted. Both the average and the harmonic mean (minimizing the problems of the simple mean) are here proposed. A threshold for a set of testing instances below 0.9 is consider non-acceptable in many domains.

$$Acc_i(y,\hat{y}) = 1-(\text{NRMSE}(y,\hat{y}) \vee \text{SMAPE}(y,\hat{y}))$$

$$Accuracy = \frac{1}{n}\Sigma_{j=1}^{n}Acc_i(y^j,\hat{y}^j) \vee n\left(\Sigma_{j=1}^{n}\frac{1}{Acc_i(y^j,\hat{y}^j)}\right)^{-1}$$

In sequence learning, additional accuracy metrics consider functions applied to subsequences. The simplest case is of boolean functions that verify the correct labeling of contiguous points. The variance of functions applied to subsets of contiguous periods is key if the performance of the predictor deteriorates heavily across the horizon of prediction. This metric, here referred as error accumulation, avoid the need of a visual comparison of accuracy across the horizon.

Multi-label classification:

Multi-label classification metrics are relevant to compare results when the class under multi-period prediction is nominal. Beyond the common intersection operator used to compute accuracy, additional functions can be adopted to differentiated costs for false positives and true negatives or to allow for XOR differences.

$$Accuracy = \frac{1}{n}\Sigma_{j=1}^{n}\frac{|y^j \cap \hat{y}^j|}{|y^j \cup \hat{y}^j|} \text{ [53]}$$

$$HammingLoss = \frac{1}{n}\Sigma_{j=1}^{n}\frac{|y^j \ XOR \ \hat{y}^j|}{h} \text{ [53]}$$

Target accuracy metrics:

When the class for prediction is numeric or ordinal, the accuracy of the long-term predictor should follow one of the loss functions adopted in multistep-ahead prediction. Preferably, *NRMSE* and *SMAPE* if the goal is to compare with literature results. A similarity function that treats misalignments should be complementary applied for further understanding of the predictor's performance.

If the class for prediction is nominal, the accuracy should follow the adapted multi-label accuracy metric defined below, and be potentially complemented with other loss functions to deal with temporal labeling misalignments.

$$Acc_i(y^j,\hat{y}^j)=\frac{1}{h}\Sigma_{i=1}^{h} \mid y_i^j \cap \hat{y}_i^j \mid \vee\ 1-\text{LossF}(y^j,\hat{y}^j)$$

$$Accuracy=\frac{1}{m}\Sigma_{j=1}^{m} Acc_i \vee n(\Sigma_{j=1}^{n}\frac{1}{Acc_i(y^j,\hat{y}^j)})^{-1}$$

Accuracy may not suffice to evaluate long-term predictors. Specificity, sensitivity and precision can be evaluated recurring to a 3-dimensional decision matrix, where, for instance, a mean metric can be applied to eliminate the temporal dimension.

In non-balanced datasets, as the target healthcare datasets, most of the considered instances are in a non-relevant category. For instance, critical patients are just a small subset of all instances. A system tuned to maximize accuracy can appear to perform well by simply deeming all instances non-relevant to all queries. In the given example a predictor that outputs zero hospitalizations for every patient may achieve a high accuracy rate. A deep understanding can be made by studying recall, fraction of correctly predicted instances that are relevant, and precision, the fraction of relevant instances that are correctly predicted. F-measure trades-off precision versus recall in a single metric. By default, $\alpha = 1/2$, the balanced F-measure, equally weights precision and recall.

To redefine these metrics, a boolean criteria T is required to decide whether an instance is of interest. For example, relevant patients have average yearly hospitalizations above 2. Table 2 presents the confusion matrix for the target predictors, from which a complementary set of metrics were retrieved.

	Relevant	Non-relevant
Positive	tp=$\Sigma_{j=1}^{n} T(y)\wedge Acc(y,\hat{y})\geq\beta$	fp=$\Sigma_{i=1}^{n}$(1-$T(y)$)$\wedge Acc(y,\hat{y})<\beta$
Negative	fn=$\Sigma_{j=1}^{n} T(y)\wedge Acc(y,\hat{y})<\beta$	tn=$\Sigma_{i=1}^{n}$(1-$T(y)$)$\wedge Acc(y,\hat{y})\geq\beta$

Table 2: Confusion matrix for long-term predictors

$$Precision=\frac{tp}{tp+fp}=\frac{\Sigma_{j=1}^{n}(T(y^j)\wedge Acc(y^j,\hat{y}^j)\geq\beta)}{\Sigma_{j=1}^{n}(T(y^j)\wedge Acc(y^j,\hat{y}^j)\geq\beta)\vee(1-T(y^j)\wedge Acc(y^j,\hat{y}^j)<\beta)}$$

$$Recall=\frac{tp}{tp+fn}=\frac{\Sigma_{j=1}^{n}(T(y^j)\wedge Acc(y^j,\hat{y}^j)\geq\beta)}{\Sigma_{j=1}^{n}T(y^j)}$$

$$F_{Measure}=\frac{1}{\alpha\frac{1}{Precision}+(1-\alpha)\frac{1}{Recall}}, \text{ where } \alpha\in[0,1]$$

$$RoundAccuracy = \frac{tp+fp}{tp+tn+fp+fn} = \frac{1}{n}\Sigma_{j=1}^{n}{}^{(Acc(y^j,\hat{y}^j)\geq\beta)}$$

4.2 Other relevant metrics

Predictor's *error accumulation*, the propagation of past prediction errors into future predictions, can be expressed by a bias-variance for squared loss functions [17].

Predictor *utility* defines the interestingness of long-term predictors based on usefulness, novelty and understandability metrics. Usefulness concerns the probability of an arbitrary instance to have their unlabeled multi-points classified according to a well-defined behavior. Novelty measures the contribution of a predictive model to increase the knowledge of the domain. Finally, understandability refers to the ability of retrieving knowledge from the learner. This work does not consider utility due to domain-driven multiplicity of usefulness and novelty criteria [1] and as consequence of the increasingly available methods to achieve high understandability [43].

Finally, *smoothness* metrics [17] evaluate the ability of the predictor avoid overfitting when noise fluctuations are present.

5 Related Research

Work on active research streams, illustrated in Fig.5, have presenting important results to constrain the solution space.

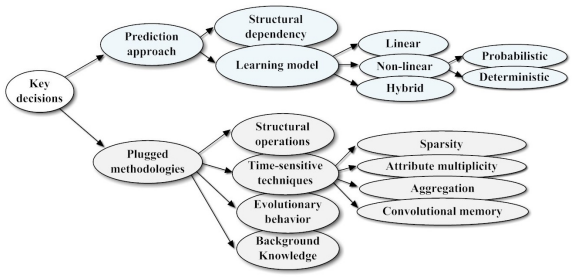

Fig. 3: Key areas for the definition of principles across different settings

5.1 Prediction approach

A. Structural dependency. A decision axis is whether to consider or not dependencies among the periods under prediction. Conventional approaches follow a

multiple-input single-output mapping. In *iterated* methods [11][8], a h-step-ahead prediction problem is tackled by iterating, h times, the one-step-ahead predictor. Taking estimated values as inputs has an evident negative impact in error propagation [46]. *Direct* methods perform the h-step-ahead prediction by learning h models, each returning a direct forecast. Although not prone to error accumulation [46], they require higher functional complexity to model the stochastic dependencies between two non-similar series. Additionally, the fact that the n models are learned independently, prevents this approach from considering underlying dependencies among the predicted variables that may result in a biased learning [10]. In literature, successful *hybrids* that combine both approaches exist [47].

Multiple-Input Multiple-Output (MIMO) methods learn one model that preserves the stochastic dependencies for a reduced bias, although reduces the flexibility and variability of single-output approaches that may result in a new bias [10][8]. To avoid this, intermediate configurations can be set by decomposing the original task into $k = h/s$ tasks, each output with size s, with $s \in \{1,...,n\}$. This approach, *Multiple-Input Several Multiple-Outputs* (MISMO), trades off the property of preserving the stochastic dependency among future values with a greater flexibility of the predictor [52].

B. Learning model. Independently of the structural dependency choice, several learning paradigms exist. All of them, either implicitly or explicitly, model the multivariate conditional distribution $P(Y|A)$.

Learners can either follow linear or non-linear predictive models. Linear models include simple, logistic or Poisson regression, as integrated recurrent autoregressions and feed-forward moving average mappings [34].

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) Markov models (HMM) [35], variable-memory Markov models [5], conditional random fields [30], and stochastic grammars [16]. Deterministic predictors include recurrent, time-delay and associate neural networks [25][31], multiple adaptive regression splines [36], regression and model trees [13][44], support vector machines (SVM) [15], and genetic solvers [18].

C. Plugged Methodologies. Significant performance improvements are triggered by plugged temporal methodologies that predictors may adopt [43].

C1. Structural operations. Suitable dataset representations, similarity-measures and time-partitioning strategies are required for a quick and flexible learning. Criteria for temporal partitioning include clustering, user-defined granularities, fuzzy characterization, split-based sequential-trees, domain-driven ontologies and symbolic interleaving [42][40][1].

C2. Time-sensitive techniques. Strategies to enhance the performance of long-term predictors for healthcare planning tasks are required to answer the introduced requirements. Techniques to deal with *data sparsity* have been proposed, for instance, in [38][27]. The goal is to avoid the exponential growth of the target data struc-

tures and to correctly interpret empty time points. Time windows and feature-based descriptions have been proposed to deal with *temporal granularity*. [23] and [2] provide initial principles for hierarchical temporal zooming operations and calendars.

Techniques to deal with *data attributes multiplicity* have been addressed in multivariate response prediction research streams. The task is to predict a matrix of responses based for multivariate time series [14]. Due to the complexity of this task, existing solutions are linear vector auto-regressions [32]. Although multivariate responses are useful to assist the prediction of class-of-interest, content-temporal dependencies among the input attributes are not considered.

Finally, covariance functions to deal with *memory sampling*, following either a parametric or non-parametric approach for the selective retaining of decisive events have been proposed in [51]. Strategies, as binary or exponentially decaying weighted of an input function, set a trade-off between *depth* (how far memory goes) and *resolution* (degree of data preservation).

C3. Evolutionary behavior. The understanding of evolving behavior to balance the smoothing and overfitting problems of long-term predictors is still a youth research stream. Prediction rules, which specify a causal and temporal correlation between time points, have been used to assist prediction [43]. In [22], emerging or evolutionary patterns, patterns whose support increases significantly over time, are adopted.

C4. Background knowledge. Finally, background knowledge is increasingly claimed as a requirement for long-term prediction, as it guides the definition of time windows [42]; provides methods to bridge different time scales, to treat monitoring holes and to remove domain-specific noise [6]; defines criteria to prune the explosion of multiple-equivalent patterns [3]; and fosters the ability to incrementally improve results by refining the way domain-knowledge is represented [3]. A hierarchy of flexible content constraints, and of taxonomical and relational time relaxations is given in [1]. Further modeling of domain-driven temporal dynamics is required [2].

5.2 Related research streams

Time series long-term prediction, sequence learning and multi-label classification are the research streams with major relevant contributions.

Long-term prediction. Although traditionally applied over time series, it can be extended to deal with time sequences.

In [17], a comparative study on the performance of iterated, direct and hybrid single-output approaches in terms of their error accumulation, smoothness of prediction, and learning difficulty is done. Selected literature have been provided different methods to define *s*-variable in MISMO approaches (as cross-validation for different values or as a function of the current query point). Experimental studies [52] show that the choice of *s* strongly varies according to the case, with s=1 (Direct method) and s=n (MIMO method) being good performers in less than 20% of

the cases. Improvements have been achieved in case of a large horizon h by adopting time series operators as the total or partial autocorrelation in multiple-output approaches [52]. A comparison of five multi-step-ahead predictors is done in [8].

Linear AIRMA models have been applied to deal with non-stationarity by defining a separated model learning for each suitable temporal window assumed to be stationary. Alternatively, in [45], clustering is combined with linear function approximation. In [17], an hybrid HMM-regression is evaluated using different regression orders and predicting windows sizes. Regression trees [13] and model trees [44] are adapted decision trees, where each leaf stores a linear predictor.

Evaluation of three multiple-output neural network predictors (simple feed-forward, modular feed-forward and Elman) is done in [9]. In [39] temporal convolution machines use Gaussian distributions to learn a class of multimodesl distributions over temporal data using three recurrent neural network variants. In [12], Bayesian learning is applied to deal with noisy and non-stationary series.

In [10], multiple-output approaches are extended with query-based criteria grounded on local learning. In [29], least-squares support-vector-machines (LS-SVM) are adopted with a local criteria for input selection, mutual information, to estimate dependencies according to Shanon entropy principle. In [46], k-nearest neighbors selection and noise estimation are additional criteria applied to select parameters to guide ARIMA and neural networks with encouraging results.

Sequence Learning. Sequence learning methods are adopted when the mining goal is sequence prediction, sequence recognition or sequential decision making [48]. The sequence recognition problem can be formulated as a prediction problem, $\hat{\mathbf{y}} = \mathbf{y}$ where $\hat{\mathbf{y}} = M(a_1, .., a_m)$. In the field of sequential decision making, sequence elements represent system states and the goal is to compose actions, Z, to reach a specific state $P(Z_{A \rightarrow Y}|AY)$ or to satisfy a goal $P(Z_{G=true}|A)$ [48]. Only contributions to the sequence prediction problem will be considered. Although sequence prediction only considers the causal ordering of elements, it provides important principles to consider in the solution space.

Unsupervised and reinforcement learning techniques from machine learning have been applied to sequence prediction, although still not scalable for large data volumes. We will briefly cover these contributions as they define important principles to solve the introduced problem. Additionally, learning techniques as expectation maximization, gradient descendant, policy iteration, hierarchical structuring or grammar training can be transversally applied to different implementations [30].

First, *unsupervised learning* are required in long-term prediction to avoid a biased learning towards smoothing or overfitting, and to deal with temporally non-compliant instances. Motifs, calendric rules, episodes, containers and partially-ordered tones [1][43][40] may be patterns of interest to assist prediction. Different approaches for their use within predictors exits. In [37], patterns are translated into boolean features to guide SVMs and logistic regressions.

Second, *reinforcement learning* [50] with two major types of predictors: *inductive logic predictors* that learn symbolic knowledge from sequences in the form of expressive rules [33], and *evolutionary computing predictors* that use heuristic-

search over probabilistic models of pattern likelihood [41]. Both methods are applied with temporal-difference methods [50]. These techniques are the preference when one is not interested in a specific temporal horizon, but rather in predicting the occurrence of a certain symbol or pattern. In [20], sequence-generating rule models are defined to constrain which symbol can appear. In [19], time series are discretized into feature vectors to train trees by varying parameters as the width of the sliding window, from which rules are retrieved and combined with logical operators.

A large spectrum of implementations are, in fact, hybrid predictors. Examples include the use of symbolic rules and evolutionary computation applied to neural networks [49]. Although formal rule-based languages obtained by induction can be used for long-term prediction, these methods have not been extensively applied due to inference complexity [43].

Multi-label Classification. In [53] an overview of simple and hierarchical multi-label classifiers is done. Multi-label learning provide basic principles to deal with the long-term classification of nominal classes. Five methods that transform the multi-label classification problem either into single-label classification or regression problems are introduced. A set of classifiers and predictors are adapted for multi-label data. Examples include a revised C4.5 with an adapted entropy calculation [53], a kNN lazy learner that includes label-ranking probabilities [28], an extended AdaBoost and a novel probabilistic generative model [24].

6 Discussion

This work formalizes the problem of long-term prediction over multi-dimensional structures. It discusses the novelty and relevance of the problem in real-world applications. Accuracy, error propagation, noise sensitivity and complementary metrics to deal with non-balanced datasets were pointed as critical and defined.

Limitations and potential contributions are detailed from the three related research streams – multistep-ahead prediction, sequence learning and multi-label classification. Attribute multiplicity, conditional-dependency, and occurrence-sparsity are key challenges to solve the target problem. Empirical contributions, in the form of principles assessing one or more of these challenges, are the required next steps to promote an efficient learning of accurate predictors.

Acknowledgment

This work is supported by *Fundação para a Ciência e Tecnologia* under the 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. Bacchus, F., Kabanza, F.: Using temporal logics to express search control knowledge for planning. A.I. **116**, 123–191 (2000)
5. Begleiter, R., El-Yaniv, R., Yona, G.: On prediction using variable order markov models. J. Artif. Int. Res. **22**, 385–421 (2004)
6. Bellazzi, R., Ferrazzi, F., Sacchi, L.: Predictive data mining in clinical medicine: a focus on selected methods and applications. Wiley Interdisc, DM and KD **1**(5), 416–430 (2011)
7. Bellazzi, R., Zupan, B.: Predictive data mining in clinical medicine: Current issues and guidelines. IJ Medical Information **77**(2), 81–97 (2008)
8. Ben Taieb, S., Sorjamaa, A., Bontempi, G.: Multiple-output modeling for multi-step-ahead time series forecasting. Neurocomput. **73**, 1950–1957 (2010)
9. Bengio, S., Fessant, F., Collobert, D.: Use of modular architectures for time series prediction. Neural Proc. Lett. **3**, 101–106 (1996)
10. 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)
11. Bontempi, G., Birattari, M., Bersini, H.: Lazy learning for iterated time-series prediction. In: IW on A.B.B.T. for Nonlinear Modeling, pp. 62–68. Katholieke U.L., Leuven, Belgium (1998)
12. Brahim-Belhouari, S., Bermak, A.: Gaussian process for nonstationary time series prediction. Computational Statistics and Data Analysis **47**(4), 705 – 712 (2004)
13. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Chapman & Hall, New York (1984)
14. Brown, P.J., Vannucci, M., Fearn, T.: Multivariate bayesian variable selection and prediction. Journal of the Royal Statistical Society **60**(3), 627–641 (1998)
15. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2**, 121–167 (1998)
16. 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)
17. Cheng, H., Tan, P.N., Gao, J., Scripps, J.: Multistep-ahead time series prediction. In: Advances in Knowl. Disc. and DM, *LNCS*, vol. 3918, pp. 765–774. Springer Berlin, Heidelberg (2006)
18. 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)
19. Cotofrei, P., Stoffel, K.: First-order logic based formalism for temporal data mining. In: Foundations of Data Mining and knowledge Discovery, *Studies in Computational Intelligence*, vol. 6, pp. 185–210. Springer Berlin, Heidelberg (2005)
20. Dietterich, T.G., Michalski, R.S.: Discovering patterns in sequences of events. Artif. Intell. **25**, 187–232 (1985)
21. 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)
22. Dong, G., Li, J.: Efficient mining of emerging patterns: discovering trends and differences. In: 5th ACM SIGKDD, KDD, pp. 43–52. ACM, NY, USA (1999)
23. Fang, Y., Koreisha, S.G.: Updating arma predictions for temporal aggregates. Journal of Forecasting **23**(4), 275–296 (2004)
24. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Proc. of the 2nd European Conf. on Comput. Learning Theory, pp. 23–37. Springer-Verlag, London, UK (1995)

25. Guimarães, G.: The induction of temporal grammatical rules from multivariate time series. In: 5th ICGI, pp. 127–140. Springer-Verlag, London, UK (2000)
26. Henriques, R., Antunes, C.: An integrated approach for healthcare planning over dimensional data using long-term prediction. In: 1st Proc. in Healthcare Information Systems. Springer-Verlag, Beijing, China (2012)
27. Hsu, C.N., Chung, H.H., Huang, H.S.: Mining skewed and sparse transaction data for personalized shopping recommendation. Machine Learning **57**, 35–59 (2004)
28. The IEEE Computational Intelligence Society: A k-Nearest Neighbor Based Algorithm for Multi-label Classification, vol. 2 (2005)
29. Ji, Hao, Reyhani, Lendasse: Direct and recursive prediction of time series using mutual information selection. In: IWANN, vol. 3512, pp. 1010–1017. Springer (2005)
30. 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)
31. Kleinfeld, D., Sompolinsky, H.: Associative neural network model for the generation of temporal patterns: Theory and application to central pattern generators. Biophysical J. **54**(6), 1039–1051 (1988)
32. Koch, I., Naito, K.: Prediction of multivariate responses with a selected number of principal components. Comput. Statistical Data Analysis **54**, 1791–1807 (2010)
33. Lavrac, N., Dzeroski, S.: Inductive Logic Programming: Techniques and Applications. Ellis Horwood, New York, NY, USA (1994)
34. Laxman, S., Sastry, P.S.: A survey of temporal data mining. Sadhana-academy Proc. in Eng. Sciences **31**, 173–198 (2006)
35. Laxman, S., Sastry, P.S., Unnikrishnan, K.P.: Discovering frequent episodes and learning hidden markov models: A formal connection. IEEE Trans. on KDE **17**, 1505–1517 (2005)
36. Lee, T., Chiu, C., Chou, Y., Lu, C.: Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. CS & DA **50**(4), 1113–1130 (2006)
37. 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)
38. Liu, J., Yuan, L., Ye, J.: An efficient algorithm for a class of fused lasso problems. In: Proc. of the 16th ACM SIGKDD, KDD, pp. 323–332. ACM, NY, USA (2010)
39. Lockett, A.J., Miikkulainen, R.: Temporal convolution machines for sequence learning. Tech. Rep. AI-09-04, University of Texas at Austin (2009)
40. Mannila, H., Toivonen, H., Inkeri Verkamo, A.: Discovery of frequent episodes in event sequences. IJ of DMKD **1**, 259–289 (1997). DOI http://dx.doi.org/10.1023/A:1009748302351
41. Meeden, L.A.: An incremental approach to developing intelligent neural network controllers for robots. IEEE Trans. on Sys. Man and Cyber. **26**(3), 474–485 (1996)
42. Moerchen, F.: Tutorial cidm-t temporal pattern mining in symbolic time point and time interval data. In: CIDM. IEEE, Nashville, USA (2009)
43. Mörchen, F.: Time series knowledge mining. W. in Dissertationen. G&W (2006)
44. Quinlan, J.R.: Learning with continuous Classes. In: Aust. J.C. on A.I., pp. 343–348 (1992)
45. Sfetsos, A., Siriopoulos, C.: Time series forecasting with a hybrid clustering scheme and pattern recognition. IEEE Trans. on Systems Man and Cybernetics, Part A **34**(3), 399–405 (2004)
46. Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y., Lendasse, A.: Methodology for long-term prediction of time series. Neurocomput. **70**, 2861–2869 (2007)
47. Sorjamaa, A., Lendasse, A.: Time series prediction using dirrec strategy. In: ESANN, pp. 143–148 (2006)
48. Sun, R., Giles, C.L.: Sequence learning: From recognition and prediction to sequential decision making. IEEE Intelligent Systems **16**, 67–70 (2001)
49. Sun, R., Peterson, T.: Autonomous learning of sequential tasks: experiments and analyses. IEEE Transactions on Neural Networks **9**(6), 1217–1234 (1998)
50. Sutton, R., Barto, A.: Reinforcement learning: an introduction. AC & ML. MIT Press (1998)
51. Sutton, R.S.: Learning to predict by the methods of temporal differences. ML **3**, 9–44 (1988)
52. Taieb, S.B., Bontempi, G., Sorjamaa, A., Lendasse, A.: Long-term prediction of time series by combining direct and mimo strategies. In: IJCNN, pp. 1559–1566. IEEE Press, USA (2009)
53. Tsoumakas, G., Katakis, I.: Multi Label Classification: An Overview. IJ of Data Warehouse and Mining **3**(3), 1–13 (2007)