Prognostic Prediction based on Temporal Dependencies

Daniel Cardoso, Cláudia Antunes
Dep. Computer Science and Engineering
Instituto Superior Técnico – Universidade de Lisboa
{daniel.cardoso,claudia.antunes}@tecnico.ulisboa.pt

Abstract. Data mining techniques have been increasingly more relevant in the healthcare domain, supporting physicians in the decision making process. The most common goal is automatic diagnosis, where those techniques have been used with success. When dealing with prognosis, however, the results are much less successful. In this paper, we argue that the reason for this dissimilarity is the inability of the techniques used to consider the inherent temporal dependencies present in the data. Moreover, we propose an alternative approach where we turn the problem of prognosis into a diagnosis problem. We do this by using the temporal sequence values of every feature, to estimate their future values and performing diagnosis on those. We end this paper by discussing the results obtained by estimating the values through the use of both tabular classifiers and Hidden Markov Models.

Keywords: prognosis; classification; temporal dependencies;

1 Introduction

The role of data analysis in healthcare has gained more attention, as available mining techniques have achieved higher levels of maturity. In particular, classification methods become to play a decisive role when applied to clinical trials, by providing high quality external evidence to support evidence-based medicine [1]. The rigorous metrics available to evaluate the confidence about the collected evidence on those trials, allied to the variety of techniques suited to different kinds of data, revealed to be fundamental to keep expertise up-to-date and available worldwide.

Despite the success of those techniques, they are mostly appropriate to analyze tabular data, described by a set of independent variables. Actually, we can see this kind of data as a static snapshot of the status of some entity, which is completely suited to represent patient records collected during their diagnosing process. On the other hand, prognosis may be seen as the prediction of an outcome in a future instant, considering all available data collected along time. In this manner, we may think of prognosis as the task of predicting an outcome, given a set of time-ordered snapshots. While in a single snapshot, methods may assume some level of independency among variables, this assumption is clearly unlikely in a set of snapshots, where the same variable is measured along different instants of time.
Actually, and despite this dependency among snapshots, a large number of classification-based approaches have been proposed for prognosis (see [2], [3], [4], for example). In our opinion, the results achieved through them have been impaired due to the dependency among the different values for the same variable along time.

In this paper, we argue that the simple prediction of the prognosis outcome by traditional classification methods, given a set of snapshots, can be significantly improved by exploring the temporal relations, or evolution verified in each variable that compose the snapshots. In order to validate our claim, we formalize the problem addressed, and present an approach to take those dependencies into account in the process of outcome prediction. We also perform a comparative analysis between two techniques used to estimate the future values of some features.

After the formalization of the prognosis problem, we review a set of case studies on several different diseases, with the most well-known classification techniques (section 2). In section 3 we describe our approach, and propose two distinct implementations of it, followed by a description of some experiments that compare the accuracy of both traditional classifiers and our approach using two different techniques for the estimation phase (section 4). The paper concludes with a discussion of the improvements achieved, the issues constraining those improvements and proposing some guidelines for the next steps (section 5).

2 Prognostic Prediction

In the medical context, diagnosis is the use of patient’s data, demographic and clinical, in order to understand and classify the current health condition of a patient [5].

From a formal point of view, and in the computer-based context, let $A$ be a set of attributes and $C$ a set of possible classes. Given an instance $\vec{x}_i$ described by a set of $m$ attributes from $A$, say $\vec{x}_i = (x_{i1}, \ldots, x_{im})$, the goal is to discover the most probable value $y_i$, which corresponds to its class or status, with $y_i \in C$, as in (1).

$\vec{x}_i = (x_{i1}, \ldots, x_{im}) \rightarrow y_i \quad (1)$

In a classification context, this is done in two steps: first by producing a classification model $M$, based on a set of known pairs $(x_i, y_i)$, and second, by applying the discovered model to each instance to classify.

On the other hand, prognosis is the foreseeing or prediction of the risk or probability of a certain health event happening, in the future, using the clinical and non-clinical data. It is the medical prediction of how the pair patient disease is going to evolve in a specified period of time.

To do this prognosis, a physician will use data that relates the patient to a certain part of the population, i.e. demographic data, as well as the patient’s and patient’s family clinical history. This means that the evolution of the patient is important in the prediction of his next state. Simply putting, if a patient is showing improvement in a certain factor that is responsible for some disease, it is more probable that his prognosis related to that disease is better than if it the patient had the same value but that factor was deteriorating.
As previously stated, in the process of making a prognosis a physician uses the medical history of a patient. This includes the different states a patient has been in the form of various clinical analyses he had done in different points over time. The need to use this sequential information shows the utmost importance that time has when predicting someone’s survivability, risk of recurrence.

Considering all of this, then the prognosis task can be formalized as follows:

Let a patient be represented by a sequence of pairs, \((\tilde{x}_i^1, y_i^1) \ldots (\tilde{x}_i^n, y_i^n)\), then the goal is to predict his value for \(y_{i+1}\) (2). Note that the different values for \(y_i\) may be observable (available) or non-observable at time instant \(t\) for instance \(i\).

\[
(\tilde{x}_i^1, y_i^1) \ldots (\tilde{x}_i^n, y_i^n) \rightarrow y_{i+1}^n \tag{2}
\]

### 2.1 Prognosis as a Classification task

According to some authors [6], the development, validation and impact assessment of both diagnosis and prognosis can be mutatis mutandis applied. Indeed, they only differ on the amount of time until the outcome assessment, the present for diagnosis, and a future point in time for prognosis.

A survey of the works on computer-aided multivariate prognosis shows the veracity of the claim above, with almost works being based on tabular classification techniques, from decision trees to artificial neural networks, but also Bayes classifiers, Support Vector Machines and regression [7, 8].

From the different studies, it is clear that there is no best technique to perform overall prognosis and that the result of a technique depends highly on the data being used [2]. In other words, there is no general solution that can be used in more than one dataset maintaining their performance.

Additionally, none of the identified works contemplate temporal information, neither use the evolution of the patients. Moreover, it seems that the improvements are not made incrementally, with just a few being based on previous works.

### 2.2 Prediction in Time Series

Despite prognosis in healthcare has been made mainly through tabular real-valued classification tools [6], the advances on the analysis of time series are undeniable. Time series represent ordered measurements at regular temporal intervals [8], which may be uni or multivariated, representing a single or multiple co-occurring variables.

Prediction of time series is a research field with a long history, with stock markets and signal processing the most paradigmatic cases [9]. Along the time, medical time series were also been studied, with ECG and EEC the most addressed. The techniques used vary from regression models like ARMA to recurrent neural networks [10], trying to foresee the next outcome. More recently, hidden Markov models become to be applied with considerable success on omic data, and are currently be adjusted to contemplate multivariate series [11]. Along with these approaches, dynamic Bayesian networks have been proposed, but training algorithms are just being defined [12].
3 Prognosis based on Temporal Dependencies

The traditional classification approach has been applied to prognosis with modest success, as seen above. In all described cases, no evolution on single variables was explored, and actually, the different time instances of their values were addressed separately, ignoring any possible hidden structure. On the other hand, the analysis of time series has mainly been applied to predict the next outcome of a single variable.

In order to address this situation, we propose to transform the prognosis into a diagnosis task, by estimating the values of the variables that constitute the snapshot in the future point in time.

Formally, let $A$ be a set of attributes, $C$ a set of possible classes and $n$ the number of observations. Let the $k^{th}$ observation, described by $m$ variables from $A$, be the pair given by $(\bar{x}_{i,k}, y_{i,k}) = (x_{i,1,k}, \ldots, x_{i,m,k}, y_{i,k})$ that says that at observation $k$ the instance is described by $x^i$ (the observable values) and classified as $y^i \in C$ (the predicted value).

Given an instance described by an ordered set of $n$ observations, the goal is to predict the $(n+1)^{th}$ observation, as in (3).

$$(\bar{x}^1_i, y^1_i), \ldots, (\bar{x}^n_i, y^n_i) \rightarrow (\bar{x}^{n+1}_i, y^{n+1}_i) \quad (3)$$

The difference to the definition (2) is the need to predict the entire $(n+1)^{th}$ observation, not only the predicted value $y^{n+1}_i$. Indeed, if there is a model $M_D$, that from observable values is able to determine the predicted value, it is enough to estimate the observable values in the $n+1^{th}$ observation, and from them to predict the predicted value. This model $M_D$ is just a simple diagnosis model as in (1).

According to this formulation, a prognosis model, $M_p$, is then the composition of several models: one estimation model $M_{E_k}$ per each observable variable $X_k$ and a diagnosis model $M_D$ able to predict the class given an observation, as in (4), where $n$ corresponds to the number of available observations and $m$ the number of variables for describing each observation.

$$M_p((\bar{x}^1_i, y^1_i), \ldots, (\bar{x}^n_i, y^n_i)) = M_D(M_{E_1}(\bar{x}^1_i, \ldots, \bar{x}^n_i), \ldots, M_{E_m}(\bar{x}^1_i, \ldots, \bar{x}^n_i)) \quad (4)$$

By transforming the prognosis problem into a diagnosis task, the challenge becomes to be able to estimate the observation in the time point to predict, which translates into the definition of the estimation models per each observable variable.

As stated above, the art of prognosis is based on the analysis of the evolution of the different variables along time. Therefore, estimation models should be able to recognize verified evolution trends in the estimation of future values.

In this manner, we propose that an estimation model for a single variable $X_k$, say $M_{E_k}$ should be a function from a sequence of the observed values to a value of $X_k$.

In particular, we propose two different approaches: the univariate-based and the multivariate-based estimations.
A univariate-based model for variable $X_k$ (UvE) is a function from a sequence of $n$ values of $X_k$ to its next value, $x_k^{n+1}$, as in (5), where $\text{Dom}_{X_k}$ represents the domain of variable $X_k$. These models only explore the individual values of a variable, ignoring any influence from other variables.

$$M_{UvE_k}: [\text{Dom}_{X_k}]^n \rightarrow \text{Dom}_{X_k}$$

(5)

On the counterpart, a multivariate-based model for variable $X_k$ (MvE) is a function from a sequence of $n$ vectors of $m$ variables, including $X_k$, to its next value, $x_k^{n+1}$ (6).

$$M_{MvE_k}: [\text{Dom}_{X_1} \times ... \times \text{Dom}_{X_m}]^n \rightarrow \text{Dom}_{X_k}$$

$$MUvE_k(\bar{x}^1 ... \bar{x}^n) = x_k^{n+1}$$

(6)

By receiving a sequence of multi-values, recorded along $n$ observations, multivariate estimator is able to contemplate the interdependencies among the different values, and having more informed inputs, is expected to output better estimations.

3.1 Tabular-based estimations

In this section we described the methodology used when finding the estimations using decision trees. We trained one model per variable, using tenfold cross-validation and then we classified the sequences using the models trained in the last step. The final classifications were saved as the estimations to be used later with the diagnostic model.

3.2 HMM-based estimations

The aim of this section is to describe how HMM were used in the estimation process of the future values of the features, both in the UvE and MvE models.

An HMM is characterized by a five-tuple $\{X, O, A, B, \pi\}$ where: $X$ is a finite number of states; $O$ is a finite number of observations; $A$ the transition probabilities; $B$ the observation probabilities; $\pi$ the initial state distribution and $\lambda = \{A, B, \pi\}$.

The HMMs we used had one state per time step used, so if we had a sequence with data from 7 time instances our hmm would have 7 states. All the probabilities distributions, $\lambda$, would then be initialized randomly and normalized so that the probability distribution equals to 1.

We would then train one HMM per class, using the Baum-Welch algorithm, which is used to adjust $\lambda$ to maximize the likelihood of the training set. The training set was composed by a subset of the data that had the specific class.

The prediction phase was done by concatenating all the possible classes to the observed sequence and applying the forward algorithm with that sequence and the matching class HMM. The forward algorithm calculates the likelihood that the hmm generated the sequence. The sequence with the highest likelihood was chosen and so the concatenated class was the estimation.
4 Experimental Results

The Hepatitis dataset was made available as part of the ECML/PKDD 2005 Discovery Challenge (available in http://lisp.vse.cz/challenge/CURRENT/), it contains information about 771 patients, and more than 2 million examinations between 1982 and 2001. Based on the work of [13] the data was reduced to the most significant exams. In the end 17 variables were used: gender, age, birthdate, birth decade, 11 of the most significant exams (GOT, GPT, ZTT, TTT, T-BIL, D-BIL, I-BIL, ALB, CHE, T-CHO and TP) and the results from the active biopsies at the time of the exams (type, activity and fibrosis). Fibrosis is the objective class and it is described by integer values between 0 and 4, where a value of 0 means that there is no fibrosis and a value of 4 that the fibrosis is severe. The subset used contains 488 patients and a distribution the following distribution of classes: 0 has 2.05%, 1 has 45.9%, 2 has 21.35%, 3 has 15.19% and finally 4 has 15.40%.

4.1 Diagnosis Models

As a baseline for comparison with the proposed approaches we used two models: BaselineSingleObservation and BaselineMultipleObservations. BaselineSingleObservation is a diagnostic model where a single observation in time is used to perform the prognosis. In other words, the state of a patient at instant $n$ is used to predict his class at instant $n+1$. On the other hand, BaselineMultipleObservations instead of using a single observation, multiple observations are used. All the past information is used here to predict the class at instant $n+1$.

A collection of techniques were used with these models, with both achieving similar results: the precision ranged between 40% and 55%, depending on the technique and number of time points used, as seen in Fig. 1.

![Fig. 1. Precision of BaselineSingleObs (left) and BaselineMultipleObs (right) models with different classifiers and number of observations](image-url)
4.2 Estimation Models

Before assessing the results of our prognosis approach, we evaluate the impact of the number of observations used on the quality of the estimations made through the two estimation models proposed.

Since Hepatitis is described by nominal variables, we measured the number of correct predictions.

Fig. 2 and Fig. 3 show the results with univariate and multivariate estimation models, respectively. Both estimation models were applied using a different number of observations, Hidden Markov Models and Decision Trees (J48).

When using Decision Trees both models reach similar levels of accuracy, with quite good results for the majority of the Hepatitis variables (above 80%). It is interesting that there is a slight trend to increase the accuracy as the number of observations get higher.

Using HMM the univariate model performs as well as the J48, but the multivariate model achieves considerably worse results. This might happen because of the noise introduced by the other variables which might not be dependent on the one that is being estimated. With this technique we can also notice the inverse relation between the number of observations and the precision of the estimation, with higher number of observations ending up with worse estimations.

Despite our expectations, it seems that there is no improvement on using multivariate-based estimation.

![Fig. 2](image1.png)  ![Fig. 3](image2.png)

**Fig. 2.** Impact of the number of observations on the precision of estimation models for each variable with univariate-based estimation

**Fig. 3.** Impact of the number of observations on the precision of estimation models for each variable with multivariate-based estimation
4.3 Prognosis Results

The overall precision of our prognosis approaches using different techniques can be seen in Fig. 4. Even though the precision using the estimations achieved by using HMM are somewhat lower, on average, than the ones using J48, improvements can be seen in both techniques compared to the baseline models (see Fig. 1).

In this dataset the improvements round about 20% when using J48 estimations, and with HMM while smaller the improvement can still be clearly seen on the univariate approach. The worse performance by the multivariate approach compared to the multiple observations baseline was caused by the error introduced by the estimations used.

Fig. 5 shows the impact of the number of observations on the overall prognosis precision using both, UvE and MvE estimation models, and a variety of techniques. It is interesting to notice that the increase of the number of observations does not improve the final precision of the UvE model on either of the techniques and actually worsens the performance of the MvE model with the HMM estimations, showing that the further values from the long past do not help in estimating future values.

It is also noticeable that in both models decision trees through the C4.5 algorithm and the Random Forest Ensemble performed better than the remaining techniques.
5 Discussion

Currently, medical practice is helped by a variety of computer-aided tools, dedicated to help physicians taking the most appropriate decisions. However, despite the importance of prognosis, it did not deserved dedicated tools, and in the majority of situations, it has been addressed as a simple diagnosis problem, without exploring the temporality involved.

In order to mimic physicians practice, computer-aided prognosis should take into attention patients’ evolution, considering the different observations made along time. In this paper, we formalize both diagnosis and prognosis problems, making clear the differences between them, and propose a method to transform the prognosis into a diagnosis task, based on the composition of classification over the estimation of observation values. As described above, what distinguishes this approach, from what is found in the literature, is the use of temporal dependencies of the data in order to estimate the future values of every feature and with those values perform a diagnostic in the future.

From the experimental comparison of the different approaches using different estimation techniques on the hepatitis dataset we can see a general improvement when using the temporal informed methods proposed. We can also see that the UvE estimation process achieved similar results with HMM and J48, but the final prognosis had better results with the J48. We attribute this difference to the fact that small precision differences in very significant features may result in a noteworthy overall final precision. In the MvE estimation model the performance of HMM was disappointing with results much lower than J48, as stated, this may be caused by the noise introduced by the large number of features that might be independent of the feature in question.

This differences in the performance of the estimation models, need to be studied, and other techniques (like Dynamic Bayesian networks) should be explored to enrich the estimation process. In either cases, the temporality of this kind of data should be considered as a core aspect of the prognosis.

6 Acknowledgments

This work is partially supported by FCT Fundação para a Ciência e a Técnologia, under research project D2PM (PTDC/EIA-EIA/110074/2009).

7 References


