Exploration of Temporal Patterns in Classification Problems

Daniel Sousa Veloso de Oliveira Cardoso

Thesis to obtain the Master of Science Degree in
Information Systems and Computer Engineering

Supervisor: Prof. Cláudia Martins Antunes

Examination Committee
Chairperson: Prof. José Manuel da Costa Alves Marques
Supervisor: Prof. Cláudia Martins Antunes
Member of the Committee: Prof. Sara Alexandra Cordeiro Madeira

November 2014
Dedicated to my family and friends.
Acknowledgments

Firstly I would like to thank my advisor Cláudia Antunes for her relentless support during this whole journey. I would never have been able to finish this dissertation without her excellent advice, patience and expertise in the field.

I would also like to thank my friends and colleagues, Antonio Barreto, Daniel Serrano and Fabio Carballo, for providing me with a great atmosphere, where despite the work we were trying to achieve they always managed to keep the spirits up and help each other moving forward. I would also like to thank Andreia Silva for all her help with the datasets and the extremely helpful advices.

I would like to thank my parents and my brother for understanding that I did not help as much with house work during times of stress and for always keeping me in check with deadlines by asking how everything was going every 10 minutes.

Also, I’d like to recognise that this work would not have been possible without the financial support by FCT Fundação para a Ciência e a Tecnologia, under research project D2PM (PTDC/EIA-EIA/110074/2009).
Resumo

O uso de técnicas de data mining no campo da saúde tem vindo a ganhar relevância nos últimos anos, sendo aplicadas com os mais variados objectivos. O mais comum é o processo de diagnóstico automático. Neste processo, técnicas de data mining têm conseguido resultados interessantes e com sucesso. No entanto, no que toca a prognóstico a mesma qualidade de resultados não está a ser obtida. Nós argumentamos que isto acontece graças à incapacidade, das técnicas, em capturar as dependências temporais que estão inerentemente presentes nos dados. Especificamente, a evolução temporal de um paciente não está a ser tida em conta aquando da realização do prognóstico. Nesta dissertação, propomos uma nova abordagem, independente do domínio, para atacar este problema. Apresentamos os resultados obtidos em dois datasets diferentes que mostram uma melhoria na precisão do prognóstico.

Palavras-chave: Prognóstico, Classificação, Dependencias Temporais
Abstract

The use of data mining techniques in healthcare has been noticing an increased relevance over the last few years, being applied with a variety of objectives, with the most common one being the automatic diagnostic process. In this process, data mining techniques have achieved interesting and successful results. However, when it comes to prognosis the same quality of results is not being achieved. We argue that this happens due to the inability of the used techniques to capture the inherent temporal dependencies present on the data. Specifically, the temporal evolution of a patient is not being taken into account when performing prognosis. In this dissertation, we propose a different approach, independent of the domain, to address this issue. We present our results on two different datasets that show an improvement in the overall precision of the prognosis.

Keywords: Prognosis, Classification, Temporal Dependencies
Contents

Acknowledgments .................................................. v
Resumo .......................................................... vii
Abstract ........................................................ ix
List of Tables ...................................................... xiii
List of Figures ...................................................... xvi
Nomenclature ...................................................... 1
Glossary ........................................................... 1

1 Introduction 1

2 Background 3

  2.1 Medical Diagnosis versus Prognosis ...................... 3
  2.2 Data Mining Techniques .................................. 4
    2.2.1 Decision Tree ........................................ 4
    2.2.2 Artificial Neural Networks & Support Vector Machines 5
    2.2.3 Bayesian Classifiers ................................ 6
    2.2.4 Regression Analysis ................................. 8
  2.3 Validation Techniques ..................................... 8
  2.4 Related Work .............................................. 9
    2.4.1 Alzheimer ............................................ 10
    2.4.2 Cancer .............................................. 11
    2.4.3 Diabetes ............................................. 14
    2.4.4 Venous Thromboembolism ............................ 16
    2.4.5 HIV/AIDS ........................................... 17
    2.4.6 Kidney Failure ..................................... 18
    2.4.7 Critical Analysis ................................... 19
    2.4.8 Challenges in using classification for prognosis .... 19

3 Approach 21

  3.1 Estimation Algorithm .................................... 23
  3.2 Parameterizations ....................................... 24
  3.3 Example .................................................. 25
# 4 Validation and Experimental Results

4.1 Dataset Description ................................................................. 27
   4.1.1 ALS Dataset ............................................................... 27
   4.1.2 Discret ALS Dataset ....................................................... 27
   4.1.3 Hepatitis Dataset ......................................................... 28

4.2 Experimental Results ............................................................. 29
   4.2.1 Diagnosis Model ........................................................... 30
   4.2.2 Regression Techniques ................................................... 31
   4.2.3 Decision Tree ............................................................. 33
   4.2.4 HMM ......................................................................... 36
   4.2.5 In Depth Metric Analysis ............................................... 39
   4.2.6 Discussion ................................................................. 44

# 5 Conclusions .............................................................................. 49

Bibliography ................................................................................. 54
List of Tables

2.1 Notations in a binary contingency table. Color coding indicates correct (green) and incorrect (pink) rates or counts in the contingency table. ........................................ 9

3.1 Patient data. .......................................................... 25
3.2 Data used on the univariate estimation approach. ...................... 25
3.3 Data used on the multivariate estimation approach. .................... 25

4.1 Discretization for both Blood Pressure features. ....................... 28
4.2 Discretization for the Pulse feature. .................................. 28
4.3 Discretization for the Respiratory Rate features. ....................... 29
4.4 Discretization for Percentage of Normal feature. ..................... 29
4.5 Discretization for the Temperature feature. ........................... 30
List of Figures

2.1 Example Decision Tree to predict if some students will play football. 4
2.2 Artificial Neural Network structure. 5
2.3 Example of 2D SVM optimal hyperplane. 6
2.4 Example of a simple Bayes Network with conditional probability tables. 7
2.5 AUCs for the neural network and logistic regression (LR). 12
4.1 Class distribution of the ALS dataset. 28
4.2 Discretization for the Weight feature. 30
4.3 Class distribution of the Hepatitis dataset. 31
4.4 BaselineSingleObs accuracy (several classifiers and number of observations). 31
4.5 BaselineMultipleObs accuracy (several classifiers and number of observations). 32
4.6 Impact of the number of observations on the accuracy of the linear regression estimation models for each variable, in the ALS dataset. 32
4.7 Impact of the number of observations on the accuracy of the logistic regression estimation models for each variable, in the hepatitis dataset. 33
4.8 Execution time of feature estimation in the ALS dataset using Linear Regression. 33
4.9 Execution time of feature estimation in the hepatitis dataset using Logistic Regression. 34
4.10 Accuracy of different models. 34
4.11 Impact of the number of observation on prognosis models. 35
4.12 Impact of the number of observations on the accuracy of the decision tree estimation models for each variable using both univariate and multivariate models on the discrete ALS dataset. 35
4.13 Impact of the number of observations on the accuracy of the decision tree estimation models for each variable using both univariate and multivariate models on the Hepatitis dataset. 36
4.14 Execution time of feature estimation in the hepatitis dataset using Decision Trees. 36
4.15 Execution time of feature estimation in the hepatitis dataset using Decision Trees. 37
4.16 Accuracy of different models using the decision trees estimations. 37
4.17 Impact of the number of observation on prognosis models using the decision trees estimations. 38
4.18 Impact of the number of observations on the accuracy of the HMM estimation models for each variable using both univariate and multivariate models in the ALS dataset.

4.19 Impact of the number of observations on the accuracy of the HMM estimation models for each variable using both univariate and multivariate models in the Hepatitis dataset.

4.20 Execution time of feature estimation in the ALS dataset using HMMs.

4.21 Execution time of feature estimation in the hepatitis dataset using HMMs.

4.22 Accuracy of different models using the HMM estimations.

4.23 Impact of the number of observation on prognosis models using the HMM estimations.

4.24 Different metrics for the overall prognosis using linear regression on the ALS dataset.

4.25 Different metrics for the overall prognosis per class using linear regression on the ALS dataset for the Univariate Approach.

4.26 Different metrics for the overall prognosis per class using linear regression on the ALS dataset for the Multivariate Approach.

4.27 Different metrics for the overall prognosis using HMM on the ALS dataset.

4.28 Different metrics for the overall prognosis per class using HMM on the ALS dataset for the Univariate Approach.

4.29 Different metrics for the overall prognosis per class using HMM on the ALS dataset for the Multivariate Approach.

4.30 Different metrics for the overall prognosis using J48 on the Hepatitis dataset.

4.31 Different metrics for the overall prognosis per class using J48 for estimation on the Hepatitis dataset for the Univariate Approach.

4.32 Different metrics for the overall prognosis per class using J48 for estimation on the Hepatitis dataset for the Multivariate Approach.
Chapter 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 [Sackett et al., 1996]. 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 [Endo et al., 2008], [Paradise et al., 2009], [Zhou et al., 2011], 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 dissertation, 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.

In particular, we propose a new formulation for the prognosis problem where a prognosis model is the composition of several models: one estimation model per each observable variable and a diagnosis
model able to predict the class given an observation. In this manner we can achieve a generalized approach independent of the domain.

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 (chapter 2). In chapter 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 (chapter 4). The dissertation concludes with a discussion of the improvements achieved, the issues constraining those improvements and proposing some guidelines for the next steps (chapter 5).
Chapter 2

Background

Data Mining is the process of gathering knowledge from raw data. It is different from information retrieval because in that case what is retrieved is information that is present explicitly in the data and in the data mining case it discovers implicit patterns using analytical tools. There are two types of data mining, descriptive and predictive. The former, like the name says, describes characteristics and relations of the existing data and the latter use the existing data to predict some future value.

Data mining has been applied in a collection of fields like Customer Relationship Management (CRM), finance, social networks and health care.

One area that is becoming increasingly important is health, with the amount of data available and even the increase of the digitalization, to take full advantage of all this data, data mining tools need to be used. Data mining can help physicians to identify the most effective treatments, find adverse drug reactions, fraud detection, performing diagnostics and prognostics.

2.1 Medical Diagnosis versus Prognosis

Diagnosis is the use of patients’ data, demographic and clinical, in order to understand and classify the current health condition of a patient.

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 analysis 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.

2.2 Data Mining Techniques

Different techniques have been used to perform all of those predictions, as we will show in the following chapter. We will start by describing the most common classification techniques used for prognosis, following with the cases where they have been applied to perform prognosis in different diseases.

2.2.1 Decision Tree

Decision trees are one of the most common classification techniques. They are a supervised learning technique that, based on the data features and a metric, that can be the Gini index, information gain, and Chi-squared test, tries to find the feature that best splits the data into more homogeneous sets in terms of the target variable.

By the end of the algorithm we have a tree where in each interior node there is one of the features and an edge per value of that feature. In the leaf nodes of this tree structure the class label is represented. An example of a decision tree is represented in Figure 2.1, where the outcome is whether some students will play football outside or not. If the outlook of the weather is overcast then the students will play, if it is sunny we need to look into the humidity, and if it is rainy then the wind is the deciding factor.

![Decision Tree Diagram](image)

**Figure 2.1:** Example Decision Tree to predict if some students will play football.

The most common algorithms to build decision trees are Quinlan’s ID3 [Quinlan, 1986], C4.5 [Quinlan, 1993] that came improve on ID3, and Breiman & et al.'s Classification And Regression Trees (CART) [Breiman et al., 1984].

In the current work on prognosis, as seen in 2.4, the use of C5.0 is also found. C5.0 is an extension of C4.5 that, among several issues, presents a considerable performance optimization.
Decision Trees result in a very easily understandable, like Figure 2.1, it is easy to see that some initial variable divides the data into two categories and then other variables split the resulting child groups. This information is very useful to the researcher who is trying to understand the underlying nature of the data being analyzed.

2.2.2 Artificial Neural Networks & Support Vector Machines

Artificial Neural Networks are computational models that approximate the functioning of the brain, in the sense that they are highly complex and non-linear. These networks are composed by a group of interconnected nodes, also called neurons, and are used for classification. They have an input layer with nodes that correspond to data features, a various number of hidden layers and an output layer where the outcome is represented as seen in Figure 2.2.

Contrarily to decision trees, neural networks do not present an easily-understandable model. A neural network is more of a “black box” that delivers results without an explanation of how the results were derived. Thus, it is difficult or impossible to explain how decisions were made based on the output of the network.

![Figure 2.2: Artificial Neural Network structure.](image)

SVMs are another supervised machine learning technique, where a hyperplane is found that correctly separates the spatial representation of the data into the various classes. For example, if the data is 2 dimensional, the hyperplane is a line that correctly divides the data and has the largest margin between itself and a data point, as seen in Figure 2.3.

SVMs have been used with high accuracy and can with the right kernel can have good results even
if the data is not linearly separable in the base feature space. They are memory intensive and require a lot of tuning and configuration.

![Figure 2.3: Example of 2D SVM optimal hyperplane.](image)

### 2.2.3 Bayesian Classifiers

Bayesian classifiers are probabilistic classifiers that get their name by making use of the Bayes Rule of Inference.

Naïve Bayes Classifier calculates the probability of a certain outcome class by considering that all features are independent, in other words by seeing how their value alone influences the outcome class.

If the Naïve Bayes conditional independence assumption actually holds, a Naïve Bayes classifier will converge quicker than discriminative models like logistic regression, so less data would be necessary.

**Bayesian Networks**

Bayesian networks are probabilistic graphical models, represented as directed acyclic graphs, where nodes represent random variables and edges the probabilistic dependency between them. These dependencies between variables are found using the theory of information.

An example can be seen in Figure 2.4, where there are three variables, "Grass Wet","Sprinkler" and "Rain". Each of this variables can be true (T) or false (False) and their relation is as follows: the grass
can be wet either because it rained or because the sprinkler is on, and the rain also effects whether the
sprinkler is used or not, for example if it rains the sprinkler is usually not on.

Figure 2.4: Example of a simple Bayes Network with conditional probability tables.

HMMs or Hidden Markov Models can be viewed as a specific case of the more general dynamic
graphical models, where particular dependencies are assumed. Thus, HMMs and their variants can be
interpreted as examples of DBNs. An HMM is a stochastic finite automaton, where each state generates
(emits) an observation. An HMM is described by a quintuple, \(N, M, A, B, \pi\) where this symbols mean:

\(N\) = number of states in the model

\(M\) = number of distinct observation symbols per state (observation symbols correspond to the physical
output of the system being modelled)

\(T\) = length of observation sequence

\(O\) = observation sequence, i.e., \(O_1, O_2, \ldots, O_T\)

\(Q\) = state sequence \(q_1, q_2, \ldots, q_T\) in the Markov model

\(A = a_{ij}\) transition matrix, where \(a_{ij}\) represents the transition probability from state \(i\) to state \(j\)

\(B = b_j(O_t)\) observation emission matrix, where \(b_j(O_t)\) represent the probability of observing \(O_t\) at state
\(j\)

\(\pi = \pi_i\) the prior probability, where \(\pi_i\) represent the probability of being in state \(i\) at the beginning of the
experiment, i.e., at time \(t = 1\)

\(\lambda = (A, B, \pi)\) the overall HMM model.

As mentioned above the HMM is characterized by \(N, M, A, B\) and \(\pi\). The \(a_{ij}, b_i(O_t),\) and \(\pi_i\) have the
properties:

\[\Sigma_j a_{ij} = 1, \Sigma_i b_i(O_t) = 1, \Sigma_j \pi_i = 1\] and \(a_{ij}, b_i(O_t),\) and \(\pi_i >= 0\) for all \(i, j, t.\)
2.2.4 Regression Analysis

Regression analysis is the use of a statistical analysis method used to measure the relation between variables. In other words, it helps to understand how a dependent variable varies with changes in one of the independent variables.

Linear Regression is an example of regression analysis where a linear function is used to model the data. When the outcome variable, the dependent variable is binary or categorical, linear regression cannot be applied. In those cases it is used logistic regression.

However, linear regression is appropriate only if the data can be modeled by a straight line function, which is often not the case.

Logistic Regression is a generalization of linear regression that, as just mentioned, is used to predict binary or categorical dependent variables. In this regression instead of predicting the estimate value of an event it predicts the probability of it occurring.

Regressions have comprehensible probabilistic interpretation and you can easily update your model to take in new data, unlike decision trees or SVMs. You can use this if you expect to receive more training data in the future that you want to be able to quickly incorporate into your model.

Another example of regression analysis that is also used in the healthcare domain is called Cox Proportional Hazard Models, which are a type of survival models, where the time to the occurrence of an event is related with one or more covariates that may be responsible. They show the influence of variables in the time to an event occurrence.

In medical studies Cox Proportional hazard models are the most common method used for survival outcomes.

It is an extension of the logistic model to the survival setting. Similar to conditional logistic regression with conditioning only at time of events. In the logistic method we use a linear predictor while in the COX mode a hazard function is used. The hazard function dictates the risk of the outcome during the follow up time.

\[
\lambda(t|X) = \lambda(t)e^{\beta X}
\] (2.1)

Where \(\lambda(t)\) is the hazard at time \(t\), and is usually estimated at the mean values of the predictors and \(\beta X\) is the linear predictor, \(\beta_1 \times x_1 + \beta_2 \times x_2 + \ldots + \beta_p \times x_p\).

The linear predictor is usually centered at the mean value of the predictors, and \(e^{\beta X}\) then indicates the hazard ratio compared to the average risk profile.

2.3 Validation Techniques

In Table 2.1 we can see notation used. The table is composed by the positive and negative predictions, \(+P\) and \(-P\) respectively, and the positive and negative real values, \(+R\) and \(-R\) also respectively. Then TP means the True Positives, TN the number of True Negatives and similarly FP and FN the number of False Positives and False Negatives Respectively, respectively. The sum by rows result in PP and
PN which are the number of predicted positives and negatives while the sum by columns results in RP and RN, the number of Real Positives and Real Negatives, respectively. The sum of all the real and predicted values gives the size of the population, Pop.

<table>
<thead>
<tr>
<th></th>
<th>+R</th>
<th>-R</th>
</tr>
</thead>
<tbody>
<tr>
<td>+P</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>-P</td>
<td>FN</td>
<td>TN</td>
</tr>
<tr>
<td></td>
<td>RP</td>
<td>RN</td>
</tr>
</tbody>
</table>

Table 2.1: Notations in a binary contingency table. Color coding indicates correct (green) and incorrect (pink) rates or counts in the contingency table.

In this dissertation the usual evaluation metrics, like **accuracy**, **precision**, **f-measure**, **sensitivity** and **specificity**, will be used and are described as follows.

**Accuracy** is the ratio of correct classifications over all the cases,

\[
Accuracy = \frac{TP + TN}{Pop}
\]  

(2.2)

**Precision**, also called positive predictive value, is the degree to which several measurements provide answers very close to each other. It is an indicator of the scatter in the data. The lesser the scatter, higher the accuracy.

\[
Precision = \frac{TP}{PP}
\]  

(2.3)

**Sensitivity**, also called **true positive rate** or **recall**, is the ability of the model to identify positive cases, in other words this metric shows the overall percentage of correctly identified classifications.

\[
Sensitivity = \frac{TP}{TP + FN}
\]  

(2.4)

Because only measuring the ability to identify the positive cases is useless (a system that always classified something as positive would have a sensitivity of 1), we also use **specificity**. Similarly, **specificity** measures the ability of the system to identify the negative cases.

\[
Specificity = \frac{TN}{FP + TN}
\]  

(2.5)

**F-measure**, also called **F1 Score**, is a measure of a test’s performance and can be considered the weighted average of the precision and recall.

\[
F\text{-measure} = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity}
\]  

(2.6)

### 2.4 Related Work

In this section it will be overviewed the work that has been done in the area of automatic prognostic and diagnostic. Diagnostic because, even though this thesis will be about prognosis, [Hendriksen et al.,
2013] states that the development, validation and impact assessment of both cases can be mutatis
mutandis applied.

The prediction classification can be a diagnostic or a prognostic depending only on the amount of
time until the outcome assessment. Being the options between the outcome assessment the present or
the future, the former is a diagnostic and the latter a prognosis.

In the field of diagnosis the techniques used revolve around the same as in prognosis. Mainly it
uses decision trees, artificial neural networks, association rules and Bayes classifiers as well as Support
Vector Machines [Kharya, 2012].

There are three types of prediction that can be done when talking about prognosis:

- We can try to predict the probability of developing a disease or a state of that disease, in other
  words we can perform a risk assessment or predict the disease susceptibility;

- We can predict if there will be recurrence of an event, for example if a cancer will recur after it was
  excised;

- We can predict if the patient will be alive at a certain time point, known as survivability.

We will separate the review on prognostic prediction by disease in order to allow the comparison
between the work being done in the various diseases. Showing that even though the same techniques
are used they require different preprocessing and the end results are very data dependent.

We can find work on prognostic prediction as far back as 1980 [Nash et al., 1980] where a regression
analysis is used to find the predictive power of 17 features when predicting the survival of breast cancer
patients. Also in the early 90s [Hanson et al., 1993] where logistic regression is used to predict Survival
of HIV infected patients and [Mangasarian et al., 1995] where dynamic programming is used to predict
the time to recurrence of an excised cancer.

2.4.1 Alzheimer

The Alzheimer’s disease (AD) is the most common form of dementia. It causes problems with memory,
thinking and behavior. Symptoms usually develop slowly and get worse over time, becoming severe
enough to interfere with daily tasks and eventually leading to death. In order to predict the progress of
the disease several techniques have been applied.

Alzheimer’s disease is associated with variable but shortened life expectancy, even at relatively early
stages. For that reason having a survivability expectancy might be important for the patients and their
carers to understand and plan ahead.

In [Paradise et al., 2009] they used Cox proportional hazards regression modeling for univariate and
multivariate statistics.

On the multivariate analysis in order to find the most predictive features a forward stepwise approach
was used followed by a backward stepwise linear regression in order to confirm if the results were robust.

The final model, SAM (Survival in Alzheimer’s Model) is a 4 point risk scale according to whether
a patient has or not the identified risk factors (increasing age, Constructional praxis, Gait apraxia). A
patient with two risk factors will have an 80% chance of surviving 12 months, but less than 50% chance of surviving 3.5 years.

This study has some limitations like the fact that one of the features where it was built upon, was clinically obtained, by a standardized assessment by the same doctor. Also this model's generalizability may be limited because the cohort was a convenience sample and was not recruited to be representative of the larger population of people with AD.

In [Zhou et al., 2011], Zhou et al. develop a new multi-task learning formulation based on the temporal group Lasso regularizer, in order to predict the Alzheimer's disease progression, based on the Mini Mental State Examination (MMSE) and Alzheimer's disease Assessment Scale cognitive subscale (ADAS-Cog) scores, that give the cognitive status of a patient. The multi-task regression approach captures the relation of the task, and the regularizer ensures that a small set of features is used for the regression and that a large deviation between successive time points is penalized.

2.4.2 Cancer

Cancer, known medically as a malignant neoplasm, is a class of diseases characterized by out-of-control cell growth. It becomes harmful when faulted cells grow into lumps of tissue that are called tumors. The cancer may also end up by spreading when cancerous cells move through the lymphatic system or bloodstream.

Cancer is seen as a deadly disease, as most people end up dying from the cancer or its treatment. The ones that actually survive have twice the probability of developing a second cancer than the people that were never diagnosed with cancer. [Rheingold et al., 2000]

Because of this the three types of prognosis are found in cancer research: there is the prediction of the probability of developing cancer, in other words we can perform a risk assessment or predict the cancer susceptibility, the prediction if there will be recurrence in the cancer after if it excised or the prediction if the patient will be alive at a certain time point.

In these three areas of prognosis we have the following work.

To perform the prediction of survival at 5, 10 and 15 years after the diagnostic, [Lundin et al., 1999] use artificial neural networks and logistic regression, showing that neural networks are consistent with logistic regression as is represented in 2.5.

In [Steyerberg et al., 2005], in order to decide which treatment is better for the patients' well-being, a Cox regression analysis is used to predict a score based on the regression coefficients, which classifies the patients in 3 different groups: the ones with good, intermediate or bad prognosis in terms of survivability. Using this knowledge of the degree of prognosis in addition with the short-term versus long-term benefits of each treatment, a better choice can be performed helping to improve the patient's quality of life.

To predict the overall survivability, at 1 year and 5 years mark, of patients with Acute Myeloid Leukemia, Breems et al. applied multivariate Cox regression analysis with stepwise backward selection on the patient's age at the time of the relapse, length of relapse free interval, previous stem cell
transplant and cytogenetics.

Like in [Steyerberg et al., 2005], they used the regression coefficients has a score function that is used to classify the patients [Breems et al., 2005].

The purpose of [Delen et al., 2005] is to develop predictive models and discover/explain relationships between certain independent variables and the survivability, 5 years after the diagnosis, in the context of breast cancer. Delen et al. perform a comparative study with decision trees (C5.0), MLP neural network and logistic regression. Showing that with the SEER dataset and using a 10-fold cross-validation, the decision tree performed the best out of the three with accuracy of 0.9362, closely followed by the neural network that achieved 0.9121 and the logistic regression that got 0.8920.

In the presence of microarray data, the clinical data is usually underused say Gevaert et al. that in [Gevaert et al., 2006] propose the usage of Bayesian networks to equally use both sources of data and that way get better results when performing the prognosis.

They evaluated three methods for integrating clinical and microarray data: decision integration, partial integration and full integration and used them to classify publicly available data on breast cancer patients into a poor and a good prognosis group. The partial integration method is most promising and has an independent test set area under the ROC curve of 0.845.

In the problem addressed in [Anagnostopoulos and Maglogiannis, 2006], a neural network calculates a time interval that corresponds to a possible right end-point of the patient’s disease-free survival time, in other words it predicts the time to recur (TTR) by classifying the patient into 4 classes, $TTR \leq 1$ year, $TTR \leq 3$ years, $TTR \leq 6$ years and $TTR > 6$ years. The accuracy of the neural network was measured through a stratified 10-fold cross-validation approach. Sensitivity ranged between 80.5 and
91.8%, while specificity ranged between 91.9 and 97.9%, depending on the tested fold and the partition of the predicted period.

In [Bellaachia and Guven, 2006] a comparison is made between different data mining techniques, Naïve Bayes, Neural Networks and Decision Trees when predicting survivability of breast cancer patients 5 years after the diagnose. For that comparison the Weka toolkit\(^1\) and the SEER Dataset is used, which is composed by demographic data (age, race, etc.) and clinical data (Extension of tumor, stage of cancer, etc.). After the tests the conclusion was that both, decision trees and neural networks, had better and similar performance with accuracy around 86%, though in the computational time the approaches did differ where the neural networks model took 12 times more to be built.

Because of the neural networks’ ability to consider variable relations and create non-linear predictions models they are a very used method for cancer survivability prediction, how long after surgery it is expected that the cancer will recur. Here in [Chi et al., 2007] it is shown that they can be used to predict the probability of survivability, and based on a threshold classify them as good or bad prognosis, with 2 different datasets.

In [Endo et al., 2008], Endo et al. use Logistic Regression model, Artificial Neural Network (ANN), Naive Bayes, Bayes Net, and a collection of Decision Trees (Decision trees with naïve Bayes, ID3 and J48 algorithms) to predict breast cancer survival at 5 years learning that Logistic regression has the highest accuracy along with J48. Decision trees tend to have high sensitivity. But is also shown that the best algorithm depends on the object and the dataset.

Because there is no use of fuzzy logic when performing cancer prognosis, most of the current work uses neural networks that yield difficult to understand models and that there is no use of hybridization of machine learning techniques, Muhammad Umer Khan et al. investigated a hybrid scheme based on fuzzy logic and decision trees on the SEER dataset. They performed experiments using different combinations of number of decision tree rules, types of fuzzy membership functions and inference techniques in order to predict the patient survivability. They end up by comparing the performance of each for cancer prognosis and found hybrid fuzzy decision tree classification is more robust and balanced than the independently applied crisp classification. [Khan et al., 2008]

In [Delen, 2009], Delen uses a handful of data mining techniques, decision trees, artificial neural networks and support vector machines along with the most common statistical analysis tool, logistic regression, to build a prediction model for prostate cancer survivability and comparing their performance. The results indicated that SVMs are the best predictor with a test data set accuracy of 92.85%, followed by ANNs with an accuracy of 91.07%, followed by decision trees with an accuracy of 90.00% and logistic regression with an accuracy of 89.61%.

Jong Pill Choi et al. compared the performance of an Artificial Neural Network, a Bayesian Network and a Hybrid Network used to predict breast cancer prognosis. The hybrid Network was a combination of ANN and Bayesian Network. All the techniques were used on nine variables of the SEER data that were clinically accepted. In this research the accuracy of ANN (88.8\(\%\)) both performed much better than the Bayesian Network. [Choi et al., 2009]

\(^{1}\)http://www.cs.waikato.ac.nz/ml/weka/
In [Sun et al., 2011] improve the L1-L2 norm SVM that has automatic feature selection for prognostic prediction to use regression, and developed the algorithm to utilize the information of censored data. The proposed method is compared with other seven prognostic prediction methods, namely CART, MARS, RSA, RRLC, L1-norm SVM, L2-norm SVM, Elastic Net, penalized Buckley-James, on three real world data sets. The experimental results show that the proposed method performs consistently better than the medium performance and that it is more efficient than other algorithms that achieved similar performance.

Kharya performs a review of use cases where data mining has been used to perform prognosis of cancer disease. It shows that the most common cases, while they may need to be tested on larger set of examples in order to find rules with higher level of statistical confidence, they do find statistically significant associations that can help predict a patients’ future. In this study they show examples using decision trees, neural networks, logistic regression as well as Bayesian networks. [Kharya, 2012]

In [Wang et al., 2012], the prediction of survivability on the 5 year mark after diagnose were performed using decision trees and logistic regression. Using the SEER dataset Wang et al. show that logistic regression, even though the accuracy is similar, outperforms decision trees by having a higher g-mean and by comparing the ROC curve and AUC.

In [Saxena et al., 2013] instead of using the complete Wisconsin Prognostic Breast Cancer data set, a pre-processing technique is used in order to reduce the number of features and improve the accuracy of polynomial neural network that was later used. The pre-processing technique is called principal component analysis (PCA) and it is a statistical procedure that returns a set of principal components. These principal components are less than or equal to the number of original features and they are ordered by their importance in the variability of the outcome. It is shown that the use of PCA is preferred to normalization, having the former more accurate results.

Using the SEER database Lakshmi et al. perform a comparison of a number of techniques when diagnosing and predicting 5 year survivability of patients diagnosed with breast cancer. The techniques that were compared were: C4.5, SVM, PNN, k-NN, Binary Logistic Regression as well as Multinomial Logistic Regression, Partial Least Squares Regression (PLS-DA), Partial Least Squares Linear Discriminant Analysis (PLS-LDA), k-means and Apriori Algorithm. In the end, this study [Lakshmi et al., 2013], shows that PLS-DA performs the best with lowest computation time and highest accuracy.

### 2.4.3 Diabetes

Diabetes, often referred to by doctors as diabetes mellitus, describes a group of metabolic diseases in which the person has high blood glucose (blood sugar), either because insulin production by the pancreas is inadequate, or because the body’s cells do not respond properly to insulin, or both. There are three types of diabetes: type 1 is when the body does not produce insulin, type 2 is when the body does not produce enough for normal function or the cells in the body do not react to insulin, insulin resistance and the third type affects females when pregnant. They develop high levels of blood sugar and don’t have enough insulin to transport it.
In [Lindström and Tuomilehto, 2003] a risk score to predict the incidence of diabetes was developed. The multivariate logistic regression model coefficients were used to assign each variable category a score. The Diabetes Risk Score was composed as the sum of these individual scores. In the final predictive model there were 7 features selected, Age, BMI, waist circumference, history of antihypertensive drug treatment and high blood glucose, physical activity, and daily consumption of fruits, berries, or vegetables.

The model was developed using a cohort study from 1987 and another from 1992 where the subjects received by mail a questionnaire on medical history and health behavior and an invitation to a clinical examination.

The score that was derived from the regression coefficients ranged from 0 to 20 and the value $\geq 9$ was able to predict diabetes with a sensitivity of 0.78 and 0.81, specificity of 0.77 and 0.76 in the 1987 and 1992 cohorts, respectively.

In order to improve the work of Lindström et al. the author of [Balkau et al., 2008] aims to describe sex specific lifestyle and clinical diabetes risk factors in a French population followed over 9 years in order to aid in identifying those at risk for incident diabetes. The data is composed by clinical along with biological data that was gathered every 3 years over a period of 9 years. In this study patients with already incident diabetes in the beginning were excluded as well as the patients with unknown status of diabetes at the end. The author performed a statistical analysis over the data in order to find the most predictive features. Balkau et al. used logistic model to test for interactions with sex. Parsimonious logistic regression models were selected using forwards and backwards as well best model selection criteria using all parameters; the Hosmer-Lemeshow goodness-of-fit test was the principal criteria for selection of a model.

The resulting models, clinical and clinical + biological, were able to predict the incidence of diabetes over the 9 year period. They studied the influence of gender in the model, learning that the predictive functions were different for each sex.

Because the currently available screening tools for identifying individuals at high risk of type 2 diabetes can be invasive, costly and time consuming Xie et al. developed a tool to identify individuals in the Chinese general population with high risk of developing type 2 diabetes (Xie, et al., 2010). Using data from 994 persons with type 2 diabetes and 13 129 persons with normal fasting glucose, test performed to find diabetic patients, aged 35-74 years. After a Classification and regression tree (CART) analysis, performed separately in men and women, two risk trees were obtained: one with 5 risk levels for men, and another with 8 for women. Being that women with a diabetes risk level (DRL) of 8 and men with a DRL of 5 are at the highest risk of type 2 diabetes. The CART results were compared with multivariable logistic regression model including the same predictors achieving both the same AUC of 0.71 vs. 0.73 in women and 0.65 vs. 0.69 in men, in the training and testing samples, indicating a good prediction above chance.

In [Chen et al., 2010] a risk score is built for the prediction of type 2 diabetes in a 5 year follow up study between 1999 and 2004, using demographic data, like age, sex and ethnicity, some feature that represent the history of the patient and clinical tests. The score was built using a logistic regression anal-
ysis where the features’ coefficients were rounded up and used as a score if that feature was present. It was found that this diabetes risk score was a useful non-evasive method to identify Australian adults at high risk of type 2 diabetes who might benefit from interventions to prevent or delay its onset.

2.4.4 Venous Thromboembolism

Venous Thrombosis is a blood clot that forms within a vein. A common cause of venous thrombosis is the deep vein thrombosis that can turn into a pulmonary embolism, which can be lethal. Venous thromboembolism is a disease that includes both deep vein thrombosis (DVT) and pulmonary embolism (PE).

In order to predict the outcome in a 30 day period of patients that had a pulmonary embolism, Aujesky et al. used clinical variables that were shown to be related with the death of patients with PE. These variables included demographics, comorbid conditions, physical examination findings, and laboratory and chest x-ray findings. On that data a stepwise logistic regression analysis was performed to create the prediction rules that classify within 5 levels of mortality risk. [Aujesky et al., 2005]

[Eichinger et al., 2010] Based on a cohort study of 929 patients that had a first unprovoked deep vein thrombosis, Eichinger et al. perform a Cox hazard proportional analysis to learn the relevance of, previously selected, clinical and laboratorial data in the recurrence of the thrombosis. Using those values a nomogram was created that can give risk probability of recurrence and correctly classify patients in risk categories.

The risk of recurrence in a patient that had an unprovoked thromboembolism is between 5 and 7% in the first year, that risk can be significantly reduced by the administration of oral anticoagulation therapy. On the other hand, the risk of major bleeding with ongoing oral anticoagulation therapy among venous thromboembolism patients is 0.9–3.0% per year with an estimated case fatality rate of 13%. Given that the long-term risk of fatal hemorrhage appears to balance the risk of fatal recurrent pulmonary embolism among patients with an unprovoked venous thromboembolism, clinicians are unsure if continuing oral anticoagulation therapy beyond 6 months is necessary. In [Rodger et al., 2008], Rodgers et al. used conditional logistic regression with forward variable selection, they conducted multivariable analysis with recurrent venous thromboembolism as the dependent variable in order to develop a risk score that may help clinicians decide whether to stop the anticoagulation therapy or not.

They concluded that it may be safe for women who have taken oral anticoagulants for 5–7 months after an unprovoked venous thromboembolism to discontinue therapy if they have 0 or 1 of the following signs or symptoms: hyperpigmentation, edema or redness of either leg; a D-dimer level of 250 µL or more while taking warfarin; BMI 30 kg/m2 or more; and age 65 years or more. A decision rule for mean was not able to be found.

In citeTosetto2012 another risk prediction score is develop for the same task as [Rodger et al., 2008], to help clinicians know if the anticoagulant therapy may stop, in this case after an initial period of at least 3 months.

The score (DASH, D-dimer, Age, Sex, Hormonal therapy) was developed firstly by identifying vari-
ables highly correlated with the recurrence by using COX regression. In the initial full model there were 7 features: D-dimer; age; patient sex; hormone use at time of VTE (in women); mode of initial presentation (DVT alone or DVT and PE); and previous history of cancer, not active at the time of initial event. At first, the model was reduced using backward selection of features, but because this may lead to an overly optimistic model they evaluated the degree of over-optimism both by a heuristic formula and by linear shrinkage with bootstrapping, this means that they adjust the regression coefficient based on the calculated optimism.

In the end, by multiplying the corrected coefficient by a common value and rounding to the nearest integer the score was found. The annualized recurrence risk was 3.1% for a score $\leq 1$, 6.4% for a score $= 2$ and 12.3% for a score $\geq 3$. By considering at low recurrence risk those patients with a score $\leq 1$, life-long anticoagulation might be avoided in about half of patients with unprovoked VTE.

2.4.5 HIV/AIDS

Human immunodeficiency virus/ acquired immunodeficiency syndrome (HIV/AIDS) is a disease that affects the human immune system when infected with HIV. Acquired Immunodeficiency Syndrome is the final stage of HIV infection. People at this stage of HIV disease have badly damaged immune systems, which put them at risk for opportunistic infections that may lead to death.

In terms of prognosis of HIV/AIDS, it usually refers to the likely outcome of HIV/AIDS. It may also include the duration of HIV/AIDS, chances of complications of HIV/AIDS, probable outcomes, prospects for recovery, recovery period for HIV/AIDS, survival rates, death rates, and other outcome possibilities in the overall prognosis of HIV/AIDS.

The ART Cohort Collaboration is an association between 13 cohort studies from Europe and North America, it gathers data from patients who are infected with HIV-1 and started highly active antiretroviral therapy (HAART). In [Egger et al., 2002], Egger et al. build a prognostic model to predict the development into AIDS or death and to death alone. The prognostic models were parametric survival models based on the Weibull, loglogistic, and lognormal distributions showing that the Weibull was the one that generalized best stratified by baseline CD4 cell count and transmission group (sexual contact, drug injection, etc.).

Using an adaptive fuzzy regression technique, Don et al. predicted the length of survival of AIDS patients based on their CD4, CD8 and viral load counts. A comparison was made with fuzzy neural networks getting both the techniques similar results. The accuracy of the prognosis ranged between 60 and 100% depending on what year was being predicted. [Dom et al., 2009]

With data from patients diagnosed with AIDS between 1987 and 2007 from the University Hospital of Kuala Lumpur. Abdul-Kareem et al. developed a Classification And Regression Tree (CART), based on clinical and demographic data, to predict survival of patients during that interval. The author managed to get an accuracy between 60-93% depending on the year that's being predicted. [Abdul-Kareem et al., 2010]
2.4.6 Kidney Failure

Kidney failure, also called renal failure or renal insufficiency, is a medical condition in which the kidneys fail to adequately filter waste products from the blood. There are 5 stages of kidney failure, being the first mildly diminished renal function, stage 2 and 3 need more level of care from the physician in order to deal with the dysfunction, stages 4 and 5 require the patients to endure in active treatment in order to survive. This active treatment may come in the form of dialysis or kidney transplant.

Due to the enormous amount of people in kidney transplantation waiting list Ahn et al. try to predict, in [Ahn et al., 2000], the one year survival of patients with kidney transplantation in order to make a more informed decision when choosing a patient for transplant. For that they built a Bayesian Network on 35,366 kidney transplants performed in the United States between 1987 and 1991.

For the same task, in [Petrovsky et al., 2002], are reported the results of training an ANN, that was able to correctly predict 84.95% of successful transplants and 71.7% of unsuccessful transplants.

Later, in [Shadabi et al., 2004], Shadabi et al. try to improve on Petrovsky’s work. For this Shadabi et al. used artificial neural networks instead of the more usually used statistical techniques that don’t provide enough information for complex problems. They tried to improve on Petrovsky and et al.’s work by using a radial basis function network and prediction of the outcome at the 2 years mark. The accuracy of this approach was very similar, when used on the same data set, to the one proposed in [Petrovsky et al., 2002] and despite the use of a range of pre-processing and ANN solutions for prediction of outcomes of kidney transplants, they found that the resultant accuracy of approximately 62% was probably too low to be of any clinical use.

Like the previous papers [Shadabi et al., 2004], etc. in [Osofisan et al., 2011] artificial neural networks are used, on data monitored when providing kidney dialysis treatment, to determine the features are related with patients’ life expectancy as well as detect the existence of renal failure. It provides a model that can help and support a better understanding of a patient’s evaluation results.

Kusiak et al. [Kusiak et al., 2005] used rough sets and decision trees to predict the survival time of patients undergoing kidney dialysis. Although they had a limited dataset and the lack of many important variables they show the potential for making accurate decisions for individual patients is enormous and the classification accuracy is high enough (above 75–85%) to warrant the use of additional resources and further research.

Wolfe et al. [Wolfe et al., 2008] use Cox regression analysis to calculate the LYFT score (life years from transplant), in order to develop a novel kidney allocation system based on this prediction of lifespan. The LYFT score was higher for younger patients and smaller for diabetic patients.

Li et al. [Li et al., 2010] present the development of a Bayesian belief network classifier for prediction of graft status and survival period in renal transplantation using the patient profile information prior to the transplantation. They developed two classifiers one to predict the status of the graft and another to predict its survival period. While the first one achieved a prediction accuracy of 97.8% and true positive values of 0.967 and 0.988 for the living and failed classes the second model showed only 68% accuracy.
Organ Failure

Prognosis work that is not about kidney failure can also be found. In [Oztekin et al., 2009] a study is performed where the authors used neural networks, decision trees as well as logistic regression, when trying to predict a patients’ survival after a combined heart-lung transplant. The predictive models’ performance in terms of 10-fold cross-validation accuracy rates for two multi-imputed datasets ranged from 79% to 86% for neural networks, from 78% to 86% for logistic regression, and from 71% to 79% for decision trees.

Also, survival analysis of liver transplant patients in Canada was done by Hong et al. [Hong et al., 2006] here they apply Cox proportional hazards analysis to evaluate many clinical and physical parameters’ relation to the survival of the patient. A drawback of that study is that they use a very limited set of variables.

Again in liver transplant there is this study [Ataide et al., 2012], where the Kaplan–Meier method and Cox regression are used to evaluate the relevance of the up-to-seven criteria, with 7 being the sum of the size and number of tumors for any given hepatocellular carcinoma (HCC), when predicting the survival of patients with hepatocellular carcinoma that perform liver transplant.

2.4.7 Critical Analysis

What we can see from the status of automated prognosis for the various diseases presented above, is that it is usually made without contemplating any temporal information. And, it is our opinion, that using it may considerably improve the results achieved, since it will mimic physicians’ procedures.

None of the previously presented approaches takes advantage of the evolution of a patient in order to increase its prognostic accuracy. The time, as a dimension, is being overlooked when building a prognostic model and it should be included in the process.

Another disadvantage of the work that has been done is that its results are data dependent, and even domain dependent, [Endo et al., 2008] states that there is no best technique to perform overall prognosis and that the result of a technique depends highly on the data being used. In other words there is no general solution that can be used in more than one dataset maintaining their performance.

Another problem identified in this review of prognosis work is that there is no evolution or search for improvement, with just a few of the papers being based and working on improving some earlier work. There is a worry to develop new prediction models before validating the already existing ones.

2.4.8 Challenges in using classification for prognosis

One of the major setbacks when trying to perform prognosis, is the fact that the data is, what is called, censored. This means that the value of a feature in the data is only partially known. In our case this feature is the outcome, where, for example, when predicting cancer recurrence, we know the value if the cancer has recurred while on the other case, we cannot say with certainty that it won’t recur, just the amount of time that has passed since the cancer was removed. This introduces a level of uncertainty in data that needs to be handled by data mining techniques.
Other difficulty when performing prognosis using classification is finding the correct dataset to train the model. The data should be from a cohort study, what enables better measurements of the features and helps to keep track on the outcome.

Also given the characteristics of the task at hand, difference between patients, using one predictor (or feature) is rarely descriptive enough to help. Doctors use several/a set of features about patients to be able to give a prognosis, and so also needs to happen when performing the prognosis computationally. As in medical prognosis, a multivariate approach should be used, by computer-based systems, in order to take into account the relations between features. Features, also called predictors, can be data from the patient's demographic (age, gender, etc.), clinical history, physical tests, and disease characteristics. They should be well defined and, so they could be used in real clinical situations.
Chapter 3

Approach

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 [Steyerberg et al., 2005]. From a formal point of view, and in the computer-based context, let \( A \) be a set of variables (either known as attributes) and \( C \) a set of possible classes. Given an instance \( \bar{x}_i \), described by a set of \( m \) variables from \( A \), say \( \bar{x}_i = (x_{i,1},...,x_{i,m}) \), 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 3.1.

\[
\bar{x}_i = (x_{i,1},...,x_{i,m}) \rightarrow y_i \tag{3.1}
\]

In a classification context, this is done in two steps: first by producing a classification model \( M_D \), based on a set of known pairs \( (x_i, y_i) \) – the training dataset, 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.

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

Let a patient be represented by a sequence of pairs, \( (\bar{x}_i^1, y_i^1), ..., (\bar{x}_i^n, y_i^n) \), then the goal is to predict his \( y_i^{n+1} \) value – equation 3.2. Note that the different values for \( y_i^t \) may be observable (available) or non-observable at time instant \( t \) for instance \( i \).

\[
(\bar{x}_i^1, y_i^1), ..., (\bar{x}_i^n, y_i^n) \rightarrow y_i^{n+1} \tag{3.2}
\]

In this context each \( \bar{x}_i^k \) is known as a snapshot. In other words, a snapshot is all the data that characterizes a patient in one, single, time point.

The traditional classification approach has been applied to prognosis mostly by preprocessing the data making use of domain knowledge, as seen above. In all described cases, the evolution of single variables was not explored, and actually, the different time instances of their values were addressed separately, ignoring any possible hidden structure, in the majority of approaches. On the other hand, the
analysis of time series is applied to predict the next outcome of a single variable.

By recognizing that estimation may be used to fill unseen variable outcomes, which in turn may be used to improve classifiers accuracy, as in asap classifiers [Antunes, 2010]. With this purpose, 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, $C$ be a set of possible classes and $n$ be the number of observations. Let the $t^{th}$ observation, described by $m$ variables from $A$, be the pair given by $(\bar{x}_t^i, y_t^i)$ that says that at observation $t$ the instance is described by $x_t^i$ (the observable values) and classified as $y_t^i \in C$ (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 equation 3.3.

$$(\bar{x}_1^i, y_1^i), \ldots, (\bar{x}_n^i, y_n^i) \rightarrow (\bar{x}_{n+1}^i, y_{n+1}^i) \tag{3.3}$$

The difference to the definition 3.2 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 equation 3.1.

According to this formulation, a prognosis model, $M_P$, is then the composition of several models: one estimation model $M_{Ek}$ per each observable variable $X_k$ and a diagnosis model $M_D$ able to predict the class given an observation, as in equation 3.4, where $n$ corresponds to the number of available observations and $m$ the number of variables for describing each observation.

$$M_P((\bar{x}_1^i, y_1^i) \ldots (\bar{x}_n^i, y_n^i)) = M_D(M_{E1}(\bar{x}_1^i \ldots \bar{x}_n^i) \ldots M_{Em}(\bar{x}_1^i \ldots \bar{x}_n^i)) \tag{3.4}$$

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 $M_{Ek}$ should be a function from a sequence of the observed values to an $X_k$ value. 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 equation 3.5, 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.

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

$$M_{UvE_k}(x_k^1 \ldots x_k^n) = x_k^{n+1} \tag{3.5}$$
On the counterpart, a **multivariate-based model** for variable $X_k(M \vee E)$ is a function from a sequence of $n$ vectors of $m$ variables, including $X_k$, to its next value, $x_{k}^{n+1}$ – see equation 3.6.

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

$$M_{M \vee E_k}(\bar{x}^1...\bar{x}^n) = x_{k}^{n+1} \quad (3.6)$$

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 Estimation Algorithm

From the previous formulation, the algorithm required to train the new classifier is simple, and is similar for both estimation models.

**Algorithm 1** Pseudocode for Univariate Estimation training

```plaintext
1: procedure UNIVARIATE_ESTIMATION(Dataset $D$, Function $alg_{estim}$, Function $alg_{class}$, int $\rho$)
2:  // $D$ – the training dataset with
3:  $D = \{(x_i^t, y_i^t) : \forall t : 1 \leq i \leq |D| \wedge 1 \leq t \leq n\}$
4:  // $alg_{estim}$ – the estimation algorithm
5:  // $alg_{class}$ – the training algorithm
6:  // $\rho$ – the number of observations to use
7:  $A \leftarrow \{\text{the set of attributes describing } D\}$

8:  // Training each estimation model
9:  for each variable $X_k$ in $A$ do
10:     $D_k \leftarrow \pi_{X_k}(D) = \{(x_{ik}^{n-\rho}, ..., x_{ik}^n) : \forall i \in D\}$
11:     $M_{Ek} \leftarrow alg_{estim}(D_k)$
12: end for

13:  // Estimating n+1 snapshot
14:  for each variable $X_i$ in $D$ do
15:      for each variable $X_k$ in $A$ do
16:         $x_{ik}^{n+1} \leftarrow M_{Ek}(x_{ik}^{n-\rho}, ..., x_{ik}^n)$
17:         $D^{n+1} \leftarrow D^{n+1} \cup \{(x_{i1}^{n+1}, ..., x_{im}^{n+1}, y_{i}^{n+1})\}$
18:      end for
19:  end for

20:  // Train the diagnosis model
21:  $M_D \leftarrow alg_{class}(D^{n+1})$

22:  // Output the composition of models
23:  Return $M_D \circ (M_{E1}, ..., M_{Ek})$
24: end procedure
```

Note, that the dataset has to be composed of records containing $n$ snapshots, as described before, and $\rho$ has to be less or equal to $n$. In terms of the classification training algorithm, it should be any tabular one, like a decision tree learner, an algorithm for training neural networks or naïve Bayes.
The difference between the \textit{UvE} and \textit{MvE} models is on the creation of the estimation models (line 10), in particular on the creation of the training dataset for each variable. While for the univariate model, it consists on the projection of $D$ in relation to each $X_k$, like in equation 3.7, the multivariate model, equation 3.8, uses the entire set of variables. In other words, instead of just using the $k^{th}$ variable values, the entire instances are used. In both cases, $\rho$ corresponds to the number of snapshots to keep in the dataset. Since, it is usual that the instants more significant for determining the next value are the previous ones, only the last $\rho$ snapshots are used.

\[
\pi_{X_k}(D) = \{(x_{ik}^{n-\rho}, ..., x_{ik}^n) : \forall \bar{x}_i \in D\}
\] (3.7)

\[
\pi_{X_k}(D) = \{\bar{x}_i^{n-\rho}, ..., \bar{x}_i^n) : \forall \bar{x}_i \in D\}
\] (3.8)

After training the estimation model for each variable, the diagnosis model is learnt from the estimated snapshot for instant $n + 1$ and the known class label. Then, the algorithm outputs the model resulting from the composition of the different estimators and the diagnosis model learnt from the estimated values (line 23).

\subsection{3.2 Parameterizations}

From the solution presented in the previous section there are few steps that deserve some special attention.

As it is possible to see in the pseudocode, of both approaches, a few parameters are configurable as arguments of our method. Namely the technique to be used in the estimation step, $alg_{estim}$, the technique to be used in the classification step, $alg_{class}$ and the amount of time steps to be used, $\rho$.

$\rho$ is the number of time steps to be used, this parameter allows us to easily change the amount of information to be used on each run of the method. In this work, when $\rho$ is smaller than the number of steps available the more recent steps are used. For example, if we have 10 time steps, and $\rho = 5$, the data from time steps 6 through 10 is taken into account.

$alg_{class}$ is the method that is used to create the Diagnostic model. This parameter can be any classification technique, like Naive Bayes, Decision Trees or Support Vector Machines. In this dissertation we decided to use the techniques whose output is an easily understandable model. This gave us techniques like Decision Trees, Naive Bayes and Logistic Regression as can be seen in Chapter 4.

$alg_{estim}$ is the method that is used in the estimation step of our solution. This technique is used to estimate the future value of every feature and it can be a various number of techniques. The technique can be chosen based on the characteristics and amount of the data available. For example, in this dissertation we use linear regression for the estimation when the data is numeric and logistic regression when it is nominal. We also try decision trees for the estimation and Hidden Markov Models.
This freedom in the configuration allows for a more generic and adaptive method that can be used on the most various data.

3.3 Example

Let’s assume we have clinical and laboratorial data for a set of patients with Alzheimer’s disease. The data itself could be represented like in Table 3.1, where HBP and LBP mean high and low blood pressure, respectively, and Degree of Progression is the objective class.

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Gender</th>
<th>Age</th>
<th>HBP</th>
<th>LBP</th>
<th>Degree of Progression</th>
<th>Time Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>M</td>
<td>65</td>
<td>160</td>
<td>88</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>M</td>
<td>65</td>
<td>140</td>
<td>90</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>M</td>
<td>65</td>
<td>138</td>
<td>85</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>M</td>
<td>65</td>
<td>134</td>
<td>81</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>25</td>
<td>M</td>
<td>65</td>
<td>141</td>
<td>88</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 3.1: Patient data.

As seen in Table 3.1, we have $N$ time steps of data, in this case 5, we are going to use those in order to perform prognosis. As described in the last section we have two approaches for doing that: in the first, we only use the past values of a feature to predict its future, that is, for example with the high blood pressure, we can use HBP at times 0, 1, 2, 3, 4 in order to predict it at instant 5, and the same for every other feature.

In Table 3.2 we can see the data to be used in order to predict HBP at instant 5.

<table>
<thead>
<tr>
<th>HBP_0</th>
<th>HBP_1</th>
<th>HBP_2</th>
<th>HBP_3</th>
<th>HBP_4</th>
<th>HBP_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>160</td>
<td>140</td>
<td>138</td>
<td>134</td>
<td>141</td>
<td>?</td>
</tr>
</tbody>
</table>

Table 3.2: Data used on the univariate estimation approach.

In the second option, we use every feature value to predict each one. In this case we would use the time steps 0 to 4 of every feature, as can be seen in Table 3.3 to predict HBP 5. And the same for every other feature.

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Gender</th>
<th>Age</th>
<th>HBP_0</th>
<th>LBP_0</th>
<th>...</th>
<th>HBP_4</th>
<th>LBP_4</th>
<th>HBP_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>M</td>
<td>65</td>
<td>160</td>
<td>88</td>
<td></td>
<td>141</td>
<td>88</td>
<td>?</td>
</tr>
</tbody>
</table>

Table 3.3: Data used on the multivariate estimation approach.

At the end of both approaches the result is the same, which is a complete snapshot at time step 5. Using that predicted data on a diagnostic model that was previously trained on data (like in Table 3.1). We would get the final prognosis.
Chapter 4

Validation and Experimental Results

4.1 Dataset Description

In order to validate our proposal, we used two different real datasets from the healthcare field: the ALS and the Hepatitis datasets.

4.1.1 ALS Dataset

The ALS dataset\(^1\) includes information from over 8500 ALS patients who participated in industry clinical trials. The data include demographic, family and medical history, the patient’s history in terms of ALS symptoms, clinical and some laboratorial data. From these, we used a subset composed by the patients that had demographic data, had performed Slow Vital Capacity exams, as well as measurements of their vitals, counting 13 variables: gender, age, height, percentage of normal, subject liters (trial 1, 2 and 3), blood pressure (systolic and diastolic), pulse, respiratory rate, temperature and Weight.

The dataset is mostly composed by numeric attributes that were normalized into the range [0,1] using the Feature Scaling method, equation 4.1. Where \(X'\) is the new value, \(X\) the current value, \(X_{\text{min}}\) and \(X_{\text{max}}\) the minimum and maximum value of that feature, respectively, and \(a\) and \(b\) are the new range minimum and maximum, or in other words \(a = 0\) and \(b = 1\).

\[
X' = a + \frac{(X - X_{\text{min}})(b - a)}{(X_{\text{max}} - X_{\text{min}})}
\]  

(4.1)

The outcome is a score that evaluates the state of the disease between 0 (severe) and 48 (normal), discretized into 4 classes (aggregations of 12 points). The subset contains 578 patients, with 5.9% for the 1st class, 22.3% for the 2nd, 29.1% for the 3rd and 42.7% for the 4th, as seen in Figure 4.1.

4.1.2 Discret ALS Dataset

Because some of the techniques used cannot be directly applied to numeric data a discretization was applied on the ALS dataset.

\(^1\)https://nctu.partners.org/ProACT/
Several different discretizations were applied to the ALS dataset, all with similar results. For that reason, only the results for the following discretization were described.

For this discretization some domain knowledge was used to find a discretization that makes more sense than just dividing the feature into \( n \) bins. In Tables 4.1, 4.2, 4.3, 4.4, 4.5 and Figure 4.2 the discretization used for the Blood Pressure, Pulse, Respiratory Rate, Percentage of Normal, Temperature and Weight features can be seen, respectively.

For the features where no information, that gave some insight into how to discretize said features, could be found, their values were discretized into \( n \) equally sized bins, with \( n = 6 \).

<table>
<thead>
<tr>
<th>Blood Pressure</th>
<th>Systolic</th>
<th>Diastolic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>&lt;120</td>
<td>&lt;80</td>
</tr>
<tr>
<td>Pre-Hypertension</td>
<td>120-139</td>
<td>80-89</td>
</tr>
<tr>
<td>High Blood Pressure Stage 1</td>
<td>140-159</td>
<td>90-99</td>
</tr>
<tr>
<td>High Blood Pressure Stage 2</td>
<td>169-179</td>
<td>100-109</td>
</tr>
<tr>
<td>Hypertensive Crysis</td>
<td>&gt;= 180</td>
<td>&gt;= 110</td>
</tr>
</tbody>
</table>

Table 4.1: Discretization for both Blood Pressure features.

<table>
<thead>
<tr>
<th>Pulse</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Slow (Bradycardia)</td>
<td>&lt;60</td>
</tr>
<tr>
<td>Resting</td>
<td>60-100</td>
</tr>
<tr>
<td>Fast (Tachycardia)</td>
<td>&gt;100</td>
</tr>
</tbody>
</table>

Table 4.2: Discretization for the Pulse feature.

### 4.1.3 Hepatitis Dataset

The Hepatitis dataset was made available as part of the ECML/PKDD 2005 Discovery Challenge\(^2\), it contains data about 771 patients, and more than 2 million examinations between 1982 and 2001. Based on the work of [Watanabe et al., 2003] 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, ...

\(^2\)http://lisp.vse.cz/challenge/CURRENT/
Respiratory Rate
Slow  <12
Normal  12-20
Fast  20-24
Very Fast  >24

Table 4.3: Discretization for the Respiratory Rate features.

| % of Normal | Very Low Breathing Capacity | <50 |
| Deteriorating | 50-80 |
| Normal | 80-100 |
| Athlete | >100 |

Table 4.4: Discretization for Percentage of Normal feature.

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 (no-fibrosis) and 4 (most severe). The subset contains 488 patients and the following distribution of classes: 2.05% of 0, 45.9% for 1, 21.35% for 2, 15.19% for 3 and 15.40% for 4, as seen in Figure 4.3.

4.2 Experimental Results

In this section we will present the experimental results of these case studies. We begin by describing the baseline for comparison and then present the performance of our approaches using different techniques.

This sections begins by presenting the baselines, the diagnosis models, the models that perform prognosis similarly to diagnosis.

Then, separating by technique used in the estimation phase (regression, decision tree or HMM), we present the performance of the different approaches compared to the baseline. It is important to note that, per estimation method, we separate the results by phase. This means that firstly we introduce the estimation performance and, after that, the overall prognosis performance using those same estimations.

It is also relevant to note that no preprocessing technique was applied due to the fact that it would affect both cases and therefore have no influence in the final performance of our method.

All The results shown use the accuracy metric because it is a universal metric and the F-measure metric showed no significant change to the presented results.

We will also show a more in depth analysis of some cases where all the metrics will be taken into account. We will show only some, due to the large number of data and charts generated by this analysis.

All tests were ran on an Asus U36SD with an Intel® Core™ i5 2430M/2410M Processor, clocked at 2.40 GHz, 8Gb of Ram and running 64 bit Windows 8.1 Pro. The Weka toolkit3, version 3.7.10, was used for the regression and decision tree estimations and classifications. While to perform the HMM estimations, the package HMM 4 for the programming language R was used.

3http://www.cs.waikato.ac.nz/ml/weka/
4http://cran.r-project.org/web/packages/HMM/index.html
4.2.1 Diagnosis Model

As a baseline for comparison with the proposed approaches we used two models. *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, *BaselineMultipleObservation* instead of using a single observation, uses multiple observations: all 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 accuracy ranged between 40% and 55%, depending on the dataset, technique and number of time points used, as seen in Figure 4.4 and Figure 4.5.

It is interesting to note that the accuracy is almost constant for ALS, for different techniques and number of time steps. However, it is clear that for Hepatitis those variables are determinant for reaching higher accuracy. The best results tend to be achieved using 3 time steps and J48 and Logistic Regression.

It is also interesting to note that those differences are smoother in the presence of the multiple observations.
4.2.2 Regression Techniques

In this section the results of applying regression techniques with out approaches is shown. Because of the different characteristic of the datasets (numeric versus nominal attributes), different regression techniques have been applied in the estimation phase of this work, namely linear regression for the numeric datasets and logistic regression for the nominal.

Estimation Models

As previously mentioned we begin the presentation of our results by analyzing the estimation phase performance.

The results with the univariate and multivariate estimation models for the ALS dataset (numeric) can be seen in Figure 4.6. These models were built using linear regression as previously said. Both estimation models were applied using a different number of observations, 3, 5 and 6 time steps. Because the dataset is numeric, we evaluated our estimation by the error, the distance, to the actual value at time \( t_{n+1} \). Both the univariate and the multivariate estimation model presented an average estimation error of around 0.165, with features having errors as high as 0.30 and as low as 0.01.

We can also see in Figure 4.7 the results of using Logistic Regression on the Hepatitis dataset. In both cases the average accuracy of estimation rounded the 80% range, with the multivariate model being consistently a bit worse than the model that uses a single variable.
While in the ALS case we cannot discriminate any clear difference between the two estimation models in the Hepatitis dataset, as previously stated, the multivariate approach performed a bit worse than the univariate model. Also in the ALS case we can see a slight trend of improvement in the estimations with the increase of snapshots used, while in the other case the opposite is noticeable.

In Figure 4.8 and Figure 4.9, we can see the performance analysis, in milliseconds, of the estimation phase. It is important to note that no significant difference was noticed between the univariate and multivariate estimations when using linear regression. The same cannot be said about logistic regression where the overall estimation of the features on the multivariate approach took about \((N_{steps} \times 3)\) times more than the univariate.

**Prognosis Results**

Finally we will evaluate the performance of the prognosis model, using the estimations presented in the previous section.

The overall prognosis accuracy achieved by using various techniques with our approaches, on the predictions achieved by using regression techniques, can be seen in Figure 4.10. It is observable that decision tree classifiers outperform the other techniques, with both of them, J48 and RandomForest, achieving better accuracies than the other techniques as well as the corresponding baselines. In the
Figure 4.7: Impact of the number of observations on the accuracy of the logistic regression estimation models for each variable, in the hepatitis dataset.

Figure 4.8: Execution time of feature estimation in the ALS dataset using Linear Regression.

ALS dataset, J48 and RandomForest were able to improve the results in more than 15%, with both estimation models. In the Hepatitis dataset the $UvE$ model clearly improved the final accuracy of the prognosis, achieving an improvement of about 20%. The $MvE$ model didn’t do so well only improving the final prognosis by 5%.

Figure 4.11 shows the relation between the number of observations and the final accuracy of the prognosis, using both, $UvE$ and $MvE$ estimation models, and a variety of techniques. It interesting to note that in the ALS dataset and using linear regression we can see distinctly that the number of time steps used and the overall accuracy are directly proportional. In the Hepatitis datasets the opposite relation is notable, as the number of time steps used increases the accuracy decreases or maintains. This leads us to think that, in this dataset, the furthest points in time are not as relevant to perform the prognosis as the ones closer to $t_{n+1}$. If this happens because of the nature of the disease or because of the characteristics of the data we are not certain.

4.2.3 Decision Tree

In this section, J48 was used as the estimation technique. J48 is an implementation of Quinlan’s C4.5 algorithm [Quinlan, 1993]. Because J48 cannot handle numeric classification this technique was used

on the hepatitis dataset and on the discretized version of the ALS dataset. The results were as follows.

**Estimation Models**

Again, 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.

Figure 4.12 and Figure 4.13 show the results with univariate and multivariate estimation models for the ALS and the Hepatitis dataset respectively. Both estimation models were applied using a different number of observations.

The ALS dataset estimations perform very poorly with an average of 21% and 32% on the UvE and MvE models, respectively. While the results are poor, there can be noticed a slight improvement by using the multivariate model. This result may be caused by the discretization that was used.

On the Hepatitis dataset 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.

In Figures 4.14 and 4.15, the average and total time of execution, in milliseconds, for the feature estimation phase, using J48, is presented.

It is important to note that, on the Hepatitis dataset, even though the estimation using logistic regression had similar results (Figure 4.7), when looking into to the accuracy of the estimation, it took much
Figure 4.11: Impact of the number of observation on prognosis models.

(a) ALS - linear regression estimations
(b) Hepatitis - logistic regression estimations

Figure 4.12: Impact of the number of observations on the accuracy of the decision tree estimation models for each variable using both univariate and multivariate models on the discrete ALS dataset.

longer to estimate the results (3× more in the fastest case and 800× more in the slowest, (Figure 4.9)).

On the ALS case, the time performance was very similar to the linear regression estimation, (see Figure 4.8), with no significant difference worth mentioning.

Prognosis Results

The overall prognosis accuracy achieved by using different techniques on the decision tree estimations can be seen in Figure 4.16. The improvements on the accuracy of our approach are always present when compared to the ones achieved by baseline models also shown in the same figure. In the Hepatitis dataset the improvements round about 20%, while in the ALS dataset the improvements are more modest with the UvE model improving around 5%, with most classification techniques, and achieving similar results as the baseline with the MvE model.

It is also curious to note that even though the MvE estimations were a little better, the overall prognosis using this estimations was consistently worse than its univariate counterpart.

Figure 4.17 shows the relation between the number of observations and the final accuracy of the prognosis, using both, UvE and MvE estimation models, and a variety of techniques. Again, and similarly to the regression estimations, it is interesting to note that on the hepatitis case the higher number of observations become prejudicial to the UvE model, which means that the values from the long past do
not help to estimate future values. And on the ALS case you can see the inverse relation while much less noticeable than when using linear regression to perform the estimations.

Again there is no clear difference between both estimation models, but decision trees (through C4.5 algorithm – J48 and the RandomForest ensemble) always perform better than the other models.

4.2.4 HMM

In this final section, HMMs were used in the estimation phase. Because HMMs cannot handle directly numeric classification the same discretization of the ALS dataset was used.

The HMMs we used had one state per time step, so if we had a sequence with data from 7 time instances our HMM would have 7 states. All the probability distributions, $\lambda$, would then be initialized randomly and normalized so that the probability distribution equals 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
Figure 4.15: Execution time of feature estimation in the hepatitis dataset using Decision Trees.

Figure 4.16: Accuracy of different models using the decision trees estimations.

likelihood was chosen and so the concatenated class was the estimation.

Estimation Models

Figure 4.18 shows the results with univariate and multivariate estimation models in the ALS dataset and Figure 4.19 shows the same results when applied on the Hepatitis dataset. A different number of time steps were used with each estimation model.

On the ALS case, the average estimation accuracy was of 13% and 4%, in the UvE and MvE approaches respectively. This poor performance might be caused by the discretization that was applied to the data. It can also be seen that the number of snapshots used has an inverse relation with the accuracy of the estimations, with the accuracy decreasing with the increase of the number of time steps used.

On the other hand, on the hepatitis dataset the average estimation accuracy was of 83% and 49%, in the UvE and MvE approaches respectively. This is a very similar results to when using regression of decision trees on the UvE model of this dataset. The MvE performed considerably worse than any other technique used on this dataset.

In Figure 4.20 and 4.21, we can see the performance analysis, in seconds, of the estimation phase using HMMs. As previously said the Baum-Welch algorithm was used in this step. This algorithm tries to maximize the likelihood of the training set and its result, the model's configuration, is a local maximum.
Because of this fact this algorithm is ran $X$ times, with $X$ being the number of iterations, to try to find the optimum solution. The execution times presented here represent the time taken to run 1 iteration of the algorithm, while the estimation results shown, were achieved by performing 50 iterations.

While in the ALS case, this technique presented a very bad performance. In the Hepatitis case, it achieved similar results in the estimation accuracy, in the $U_{vE}$ approach while the $M_{vE}$ performed significantly worse than any other, (see Figure 4.7 and Figure 4.13). But, in both cases, the time to make those estimations was much longer than any other technique. Even longer than the Logistic regression alternative on the hepatitis dataset, being between 8 and $200 \times$ slower than it and between 100 and $6880 \times$ slower than decision trees (Figure 4.9 and Figure 4.15).

**Prognosis Results**

In Figure 4.22 the accuracy of the various models is shown using different techniques and the HMM estimations. In the ALS case, as it was to expect the overall accuracy was the worst found so far, being close to the baseline, but in most cases a bit worse. In the Hepatitis case it is curious to note that, even though the accuracy of the estimations in the univariate model is very similar, the overall prognosis accuracy is consistently a bit worse. On the multivariate model, and because the estimations were so
Figure 4.19: Impact of the number of observations on the accuracy of the HMM estimation models for each variable using both univariate and multivariate models in the Hepatitis dataset.

(a) Average Estimation Time  
(b) Total Estimation Time

Figure 4.20: Execution time of feature estimation in the ALS dataset using HMMs.

poor, the overall performance was worse than any other multivariate model, on this dataset, and even worse than the respective baseline.

The impact of the amount of time steps, number of observations, used can be seen in Figure 4.23. Here no clear relation can be extracted, with the overall performance not following any pattern in relation with the technique used or the amount of snapshots.

4.2.5 In Depth Metric Analysis

In this section, as the title suggests, we will do a more in depth analysis of the performance of our method. Showing different metric performance per method of estimation and dataset, as well as performance per class.

ALS Dataset using Linear Regression Estimations

In Figure 4.24, a collection of metrics are shown for the overall prognosis using different approaches, with different number of snapshots and Linear Regression estimations. These values are a result of the weighted average of the results per class, based on their distribution, shown in Figures 4.25 and 4.26.

As it was previously mentioned, in the last section, the results for the f-measure and precision closely follow the presented accuracy. They both show an increment of around 15%, when comparing the uni-
When looking at the sensitivity and specificity we can see that, the former follows the same results as the various metrics already described, accuracy, f-measure and precision, on the other hand, the latter shows much higher results in the order of 80%. This shows that our system is better at deciding whether a patient is not in a current disease state than actually seeing if he is. Even though the sensitivity of our approach is not as good as desired, it still shows an improvement compared to the baseline’s sensitivity and the same can be said about specificity.

In Figure 4.25 and 4.26, the same collection of metrics can be seen, but in this case with results separated per class, for the univariate and multivariate approaches respectively.

When looking at the precision, of the univariate approach (Figure 4.25), a clear relation between the results and the class distribution can be noticed. Where the classes with higher number of cases reach precisions in the order of high 80% and the classes with less number of cases achieved a much lower precision, around 30%. This result show that our method is much better at positively classifying the cases which have a bigger number of cases in the data.

This result will then have some influence over the f-measure as well, since it is the harmonic mean between precision and sensitivity.

Sensitivity and specificity are both much more balanced, with sensitivity showing results averaging
Figure 4.23: Impact of the number of observation on prognosis models using the HMM estimations.

50% for all the classes and specificity around 80%, with the class "0-12" achieving 96% which was to be expected as that class represents the one that is less present in the data.

On the multivariate approach (Figure 4.26), the same relations, between metric performance and classes, can be seen just with a consistent small decrease in performance.

**Discrete ALS Dataset using HMM Estimations**

Figure 4.27, shows the performance of a group of metrics, namely f-measure, precision, sensitivity and specificity, of the overall prognosis when using estimations made using HMMs on the discret version of the ALS dataset.

What can be seen is that no significant difference can be distinguished between the various techniques, approaches and number of time points used. All metrics, apart from specificity that achieves an average of 70%, show an average performance of around 35%. Those results show a average of 5% decrease compared to the performance of the baseline.

Looking at Figures 4.28 and 4.29, might gives some insight into the reason for the poor performance of this approach.

In Figure 4.28, we can see the same collection of metrics but represented per class on the results of the univariate approach.

Immediately a proportionality relation between the class distribution and the precision performance can be seen. With the most common class achieving performances around 80% and the next two most common classes with results around 10%. The least common class achieves a precision of 0%.

The sensitivity, while also achieving 0% for the least common class, does not show a clear relation between class distribution and performance, with the other three classes obtaining similar results between 50% and 30%.

Because f-measure is a balanced metric calculated with the last two described metrics it shows a similar relationship between class distribution and performance, but this one less pronounced. Offset by the sensitivity results.

Specificity, as expected, shows the inverse relation between performance and class distribution with results between 96% and 55%.
This results tells us that, for the most common class, even though the cases that were classified with this class were correctly classified there are still a very large amount of false negatives. On the other hand, the least common classes, even though there is less precision on the classified cases, i.e. large amount of false positives, the amount of false negatives is lesser. This tends to show that this estimation technique was not able to handle the unbalancedness of the discrete data. Needing more cases of the least common classes to improve their performances.

For the multivariate approach (Figure 4.29), even thought the results are similar in the overall performance, in the precision metric there is less of a relation between class distribution and performance. In this case some techniques manage to not achieve 0% performance with the least common class and the second least common gets results between 45% and 30%, compared to the 10% of the univariate approach. On the other hand, the most common class decreases from an 80% precision to a 55%.

Sensitivity wise, the only difference is an small decrease in the "12-24" class, the second least common, contrasting with the increase in precision.

These changes have no clear effect on the overall prognosis, with the final result of the multivariate model being very similar to the univariate.

Hepatitis Dataset using J48 Estimations

Similarly to the sections above, a collection of metrics for the overall prognosis using different approaches, with different number of snapshots and J48 estimations, are shown in Figure 4.30. These
values are a result of the weighted average of the results per class, shown in Figures 4.31 and 4.32, based on their distribution.

As can be seen, the performance, as far as f-measure, precision and sensitivity, behave the same as the accuracy that was presented in the previous sections. That means that there is an overall improvement of around 20% when compared to the baseline and that there is an inverse relation between the number of snapshots used and the overall performance of one of those metrics.

Specificity, as in the other previous analysis, shows a consistently higher performance than any of the other metrics. Not dependent on the amount of snapshots used or the technique. Specificity shows results around 80%. This, again, shows the ability of our method to identify patients that are not in a specific disease state.

In Figure 4.31 and 4.32, the same collection of metrics can be seen, but in this case with results separated per class, for the univariate and multivariate approaches respectively.

In this case, with this dataset and using J48 as the estimation method, the relation between the class distribution and the results of any of the metrics cannot be identified. Except for the class "F0", which is by far the least present in the dataset, and achieves significant worse results than any other class with most techniques and metrics.

Looking at these results it can also be noticed that decision tree techniques tend to outperform the other two techniques, naive Bayes and logistic regression.

If we consider just those decision tree techniques, the results range between 80% and 50%, depend-
Figure 4.26: Different metrics for the overall prognosis per class using linear regression on the ALS dataset for the Multivariate Approach.

Specificity, again, shows much higher results than any of the other metrics. With averaging results in the order of 90%, only with the most common class, "F1", a bit worse (80%), and the least common, "F0", with the best results (98%).

When comparing both approaches, the univariate and multivariate, there can be seen that, in most cases, metric and technique wise, the multivariate approach lags a bit behind with close, but a bit worse, results.

4.2.6 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 dissertation, 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.

Taking into account the presented results of the techniques used, we can say that HMM were clearly the used method that performed the worse. Not only they took a lot longer to perform the estimations but, their results still palled when compared to the use of regressions or decision trees. The other two methods achieved really similar results, when dealing with nominal datasets, using decision tree is a better approach when it comes to execution time, while the overall prognosis results are fairly close.

If dealing with numerical dataset, linear regression managed to achieve an improvement over the baseline. Decision trees also did, achieve an improvement, but a much smaller one and HMMs only managed to achieve similar results to the baseline, or even a bit worse. These two last results though, are largely influenced by the discretization that was applied to the data, and maybe, with a different discretization the results would be better.

Even though this work focus mostly on the estimation step of the proposed approach, the diagnosis phase still has room for improvement which if done can help improve the final results. One example is the use of more complex techniques and ensembles in the classification, that are known to have better performance than simpler decision trees or regressions.

A curious result that counters what initially was thought when this approaches were planned is the performance of the multivariate approach. This approach was initially thought to be better than the univariate, because of its ability to find and use dependency relations between variables. That was in
Figure 4.28: Different metrics for the overall prognosis per class using HMM on the ALS dataset for the Univariate Approach.

fact not the case in the datasets used. This might be because the data has too much noise or simply the relations between the used variables does not exist. Either way the univariate approach was consistently better.
Figure 4.29: Different metrics for the overall prognosis per class using HMM on the ALS dataset for the Multivariate Approach.

Figure 4.30: Different metrics for the overall prognosis using J48 on the Hepatitis dataset.
Figure 4.31: Different metrics for the overall prognosis per class using J48 for estimation on the Hepatitis dataset for the Univariate Approach.

Figure 4.32: Different metrics for the overall prognosis per class using J48 for estimation on the Hepatitis dataset for the Multivariate Approach.
Chapter 5

Conclusions

There is a mismatch in the amount of data available in the field of healthcare and the data that is being used in order to gain knowledge. As it was discussed in this dissertation, diagnosis and prognosis is a very relevant subject in the area of healthcare and that it has been subject to some work in the past years. This work shows that no novel techniques are being introduced, being the same techniques used consistently throughout the years, and a visible lack of work improving on previous research with predicting models being developed independently.

We also showed that the problem of prognosis is being tackled in the same way of diagnosis, not using the patients’ evolution over time in order to improve the results.

In order to address this issue, we describe a novel approach that transforms prognosis into a diagnosis problem. This solution has two possible variants for the use of time on improving the results of a prognostic model. An univariate and a multivariate one where dependency relationships can be used to improve the final result. The method was then evaluated and discussed using two different datasets in order to show its’ generalizability.

From the survey presented the lack of use of temporality in the prognosis problem can clearly be identified. Our contribution was the proposal and definition of an extensible method that uses the temporality of the data to improve the prognosis result. It is extensible because a lot different techniques or methods can be used in each step of the approach accordingly to the characteristics of the problem and the data.

This extensibility also introduces a level of generality and domain independence to our method that, based on the problem being tackled, allows us to choose the various different techniques and configurations that may suit best the specific case.

From the experimental comparison of the different approaches, over two distinct datasets (with different data characteristics, either from the medical and the data points of view), it is clear an improvement trend when using the temporal informed methods proposed. The shallow differences between the results of the estimation models, need to be deeply 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.
Another possible variation to tackle the prognosis problem presented in this thesis would be to, instead of using the values that result from the estimation phase, like in the current approach, the model that represents the evolutionary trend of that feature would be used. Then the final classification would be performed on these models.
Bibliography


