Journal of Biomedical Informatics 48 (2014) 94–105

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Learning Bayesian networks for clinical time series analysis

Maarten van der Heijden [a,b,*], Marina Velikova [a], Peter J.F. Lucas [a,c]

[a] Institute for Computing and Information Sciences, Radboud University Nijmegen, The Netherlands
[b] Department of Primary and Community Care, Radboud University Nijmegen Medical Centre, The Netherlands
[c] Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands

ARTICLE INFO

Article history:
Received 28 March 2013
Accepted 10 December 2013
Available online 18 December 2013

Keywords:
Chronic disease management
Bayesian networks
Machine learning
Temporal modelling
Clinical time series
Chronic obstructive pulmonary disease

ABSTRACT

Introduction: Autonomous chronic disease management requires models that are able to interpret time series data from patients. However, construction of such models by means of machine learning requires the availability of costly health-care data, often resulting in small samples. We analysed data from chronic obstructive pulmonary disease (COPD) patients with the goal of constructing a model to predict the occurrence of exacerbation events, i.e., episodes of decreased pulmonary health status.
Methods: Data from 10 COPD patients, gathered with our home monitoring system, were used for temporal Bayesian network learning, combined with bootstrapping methods for data analysis of small data samples. For comparison a temporal variant of augmented naive Bayes models and a temporal nodes Bayesian network (TNBN) were constructed. The performances of the methods were first tested with synthetic data. Subsequently, different COPD models were compared to each other using an external validation data set.
Results: The model learning methods are capable of finding good predictive models for our COPD data. Model averaging over models based on bootstrap replications is able to find a good balance between true and false positive rates on predicting COPD exacerbation events. Temporal naive Bayes offers an alternative that trades some performance for a reduction in computation time and easier interpretation.

1. Introduction

Clinical data often takes the form of time series, which, when interpreting all the variables concerned in their mutual context, offer a description of the progression of a disease over time. While insight into the evolution of a disease is an important aspect of the management of any disease, whether acute or chronic, for patients with a chronic disease the evolution is of even more importance, as often the disease will not disappear. In this context it is in particular important to be able to detect when the disease becomes worse, i.e., to detect and possibly prevent *exacerbations*. For any chronic disease it is therefore of clinical interest to study the interaction between different variables, ranging from signs and symptoms to environmental factors, in terms of both static and temporal relationships. If we can capture this knowledge in a model, predictions regarding health status made by means of such models can be used to assist in chronic disease management, for example by advising on therapy adjustment. Furthermore, disease models are important for epidemiological purposes, for example for survival analysis, as well as for cost-effectiveness analysis and policy planning.

A complication that arises when analysing clinical time series is that it is often hard to obtain sufficient data, for example because the event of interest is relatively low frequency or because taking measurements is costly, time consuming or inconvenient for the patient. In addition, the reality of gathering clinical data is that observations are made at irregular time intervals and the data will contain missing values. Patients will sometimes forget to provide data, or omit some evidence for unknown reasons. Also measuring devices may sometimes fail or readings may not be recorded. These observations pose a challenging research question, which we seek to answer in this paper, namely, whether we can learn useful predictive models from clinical data with the combined characteristics of missing values and limited availability.

The research methods we propose draw their inspiration from various existing methods, which have proven to be successful in machine learning applications. Yet the combination of these methods has not been applied to clinical time series analysis. Temporal variants of Bayesian networks are our main tools to reason about causal and temporal processes in a probabilistic manner. In particular we use dynamic Bayesian networks (DBNs), where 'dynamic' should be interpreted as modelling the temporal dynamics of the process. They provide interpretable and versatile models to

* Corresponding author at: Institute for Computing and Information Sciences, Radboud University Nijmegen, The Netherlands.
E-mail addresses: m.vanderheijden@cs.ru.nl (M. van der Heijden), peterl@cs.ru.nl (P.J.F. Lucas).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.12.007

describe time series data and can be used to classify states and make predictions about future states. To learn network structures in the presence of missing observations we make use of the structural expectation–maximisation (EM) algorithm [1], which iteratively completes the data and performs a search for the best network structure to explain the data. Here we employ a variant of structural EM, with a different approach to filling in values for the missing data. Finally, to tackle the problem of data sparsity, we consider bootstrap methods originally developed in statistics [2]. In the context of Bayesian networks these methods have been applied to the analysis of micro-array data and we extend them here to the learning of temporal models from clinical time series.

To provide a concrete clinical context for this research, we focus on chronic obstructive pulmonary disease (COPD) as an application area. This disease has many characteristics that are typical for any chronic disease, although symptoms and signs will be mostly different from those other diseases. COPD is a major chronic disease in terms of morbidity and mortality; it affects the respiratory system, decreasing lung capacity and obstructing airways, thus interfering with normal breathing. An important aspect of COPD which is particularly relevant in the present context is the progressive nature of the disease. Specifically episodes of acute deterioration have a profound impact on patient well-being and on health-care costs [3]. These exacerbations are mainly caused by airway infections resulting in symptom worsening [4]. Important to note is also that patients with frequent exacerbations usually have faster disease progression, which makes exacerbation prevention a particularly relevant goal. Additionally, a faster treatment response to exacerbations appears to lead to better recovery [5].

The main contributions of the paper are as follows:

- We formulate an algorithm to learn temporal probabilistic models from limited clinical time series with missing values. The main novelty of the algorithm lies in combining learning of dynamic Bayesian networks from clinical data using structural EM with block bootstrapping for small data samples.
- We propose a variant of our learning algorithm based on naive Bayes networks, which has the attractive properties of reduced computational complexity, thus easy construction, while offering good prediction performance.
- The proposed learning algorithms are used to build predictive models of COPD patient's health status, focussing on day to day progress of signs and symptoms that can rapidly change during exacerbation events. These predictive models are novel in the context of COPD as they can handle both the dynamic nature and uncertainty inherent in the disease progression. As such, these models can be embedded within clinical or home-based applications for chronic disease management.
- We evaluate the learning procedure on COPD synthetic and patient data and show that it is effective in terms of structure discovery of interesting variable relationships, interpretability and prediction performance of the models learned.
- The results from this research demonstrate important clinical implications not only for the prediction of COPD exacerbations but also for the clinical relevance of the methods proposed for chronic disease management applications in general.

2. Related research

2.1. Clinical time series analysis

Survival analysis is a popular clinical application of time series analysis and here the technique of Cox-regression is normally used for model construction [6]. However, Cox-regression has important limitations; in particular, it is not suitable to model independence assumptions. Our work is more closely related to the research described in [7], which argues for using DBNs for prognostic models in medicine. Variants of DBNs have also been used to model temporal dynamics of organ failure in patients in intensive care units (ICU) [8], although there no structure learning was used. Further, a Bayesian network has been developed on the basis of electronic health record data to predict the onset of COPD in asthma patients; however, temporal information was not explicitly taken into account [9]. In the specific context of long term disease management for COPD, related research has focussed on facilitating remote communication [10,11] and automatic data interpretation is still uncommon. In [12] a telehealth system is described that has been applied to COPD and contains a decision support component. However, currently the decision support is limited to rule based detection of abnormal values and trend detection. Automatic interpretation of monitoring data using machine learning while taking time and uncertainty into account, therefore, appears to be a useful contribution to the area of chronic disease management.

2.2. Modelling and machine learning techniques

Early work on using Bayesian networks for prediction includes dynamic network models [13], with, for example, a clinical application to predicting sleep apnea [14]. In our work, we used and extended techniques for learning Bayesian networks from data with missing values [1] and learning from small samples using bootstrapping [15]. These methods are used extensively in bioinformatics [16], but application to the domain of clinical time series analysis constitutes a new and interesting challenge. The bootstrap methods used for small data samples are related to what is known as *bagging* in the machine learning literature [17], but are usually applied to learning decision trees instead of Bayesian networks. Our augmented temporal naive Bayes model is an extension of the TAN classifier from [18] to a prediction model that takes time dependencies into account.

In [19] a method is proposed to learn DBNs with changing dependency structures. The models include hidden variables that influence the structure of the model depending on the value of a particular variable that controls the structure change. This method is of interest when sufficient data is available to learn changing dependencies. Finally, in [20] an approach is described to use steady state information in addition to time series data to learn DBNs. Steady state data from the limiting distribution of the Markov chain describing the process is used as an additional source of information.

A different approach to modelling temporal processes is used in temporal nodes Bayesian networks (TNBNs), which model events instead of dynamics [21]. TNBNs are similar to Bayesian networks, but temporal nodes take time intervals as values. The intervals represent the time since a parent event. We made a TNBN model for our COPD domain to compare to the dynamic models, which will be discussed in Sections 5.3 and 6.3.

3. The Aerial project: mobile COPD management

The methods we developed and used in this paper were needed as part of a research project, called Aerial, aimed at the detection of worsening in patients with COPD, i.e., exacerbations. Here we briefly describe the system and the design of the study to sketch the practical clinical context of the work.

3.1. A system that supports self-management

In order to facilitate self-management of COPD by patients we developed a system with the capability to gather patient-specific