Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 100 (2016) 731 – 738

Procedia Computer Science

Conference on ENTERprise Information Systems / International Conference on Project MANagement / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2016, October 5-7, 2016

Clinical Data Analysis: An opportunity to compare machine learning methods

A. Salcedo-Bernal[a*], M.P. Villamil-Giraldo[a], A.D.Moreno-Barbosa[a]

[a]*Systems and Computing Engineering Department, School of Engineering, Universidad de los Andes, Colombia.*

Abstract

In the literature there are multiple machine learning techniques that have been used successfully in clinical data analysis. However, there is little information about the parameter configurations, the required data transformations to prepare the data used to train and evaluate the models and the impact of these decisions in the accuracy of the predictive model. This research tackles these issues, using the clinical data of MIMICII to build features from physiological measure patterns to predict the decease of patients inside the hospital in the next 24 hours, building predictive models based on Logistic Regression, Neural Networks, Decision Trees and Nearest Neighbors. In particular, we use data associated to physiological measures of 3220 patients, where 2385 left the hospital alive and 835 passed in the hospital. The results show that the chosen strategy for building features from physiological data gives good results with Neural Networks and Logistic Regression with radial kernel models and the parameter configuration plays a fundamental role in the models performance.

Peer-review under responsibility of the organizing committee of CENTERIS 2016

Keywords: Machine Learning; Comparison;MIMIC; Clinical Data Analysis; Logistic Regression; Neural Network; Decision Tree

* Corresponding author. Tel.: *+573118671663*
E-mail address: a.salcedo49@uniandes.edu.co

Peer-review under responsibility of the organizing committee of CENTERIS 2016
doi:10.1016/j.procs.2016.09.218

1. Introduction

Traditional technologies, such as physiological monitors, in medical Intense Care Units produce big amount of data of patient's state through his stay. This data often is only used to alert the medical staff of some situation that is occurring to some patients in the moment, but using this data together with stream mining techniques could produce prediction systems that help medical staff to anticipate these situations.

Physiological measures are a valuable input for generating these kinds of analyzes. However two issues arise when dealing with this information: (1) Information is taken from patients at non-uniform intervals, making it difficult to reconstruct the patients' original information since there are some periods with no information, (2) There are no direct comparisons between these techniques because different researches use distinct data sets and variables to generate and evaluate their models. In this way, it is difficult to identify the best practices to be used in different kind of clinical analysis. In particular, data set characteristics enable or not the effective use of this kind of techniques.

Many options are available to adequately prepare the data for analysis, for example: Using the average, minimal, maximum, standard deviation of the observed metrics to describe a patient in a given timespan. Each one of these design choices has a direct impact on the predictive accuracy of the analyses.

Multiple machine learning techniques have been proposed for clinical data analysis[12] , most of them aim to predict patient's state in certain periods of time that can vary from very short time prediction, like respiratory issues, heart diseases or the need of intense care; to med-long term predictions such as future diseases or life expectancy. Although the proposed predictive models are useful, there is little information about the parameter configurations, the required data transformations to prepare the dataset used to train and evaluate the models, and the impact of these decisions in the accuracy of the predictive model. Moreover, the lack of this information makes it difficult to compare the results across the proposals, making it difficult to generalize these results in other kinds of research and to identify the good practices to be used according to the data characteristics.

In order to focus on the aforementioned issues, this paper explores the use of four classification Machine Learning techniques: Artificial Neural Networks, Logistic Regression, Decision Trees and Nearest Neighbors, using medical records to predict death in the next 24 hours among patients using the MIMICII database. We present a detailed description about the preparation of the data, parameter configuration of each of the techniques used, and a comparison of the predictive accuracy of the models produced by these methods.

This paper is structured as follows: Methodology based on CRISP-DM is presented in section 2. Section 3 presents results obtained using the four classification machine learning techniques. Finally, section 4 shows conclusions and future work related to this research.

2. Methodology

This work uses the CRISP-DM[13] methodology: Data understanding and preparation, modeling and evaluation.

2.1. Data Understanding

This research uses the clinical data of MIMICII database (not the data in wave form), in particular the data associated to physiological measures of medical intensive care unit patients. This data is obtained manually by the "caregivers" in each care unit, where the regular procedure consist in writing down the measures showed by the physiological monitors and then type it in the hospital system.

To process this kind of information has certain challenges because its collection method is susceptible to human errors. In particular, we observed that there are measures that are out of range (for example heart rates below 20 bps) and also is observed that there is no periodicity in the measures. For example, the Heart Rate of a patient could be measured each hour during certain amount of time, then, isn't measured during the next time period and finally it's measured every two hours. So each patient has a unique periodicity for his physiological measures. This represents a problem for traditional analysis techniques for data streams since these techniques process the data as a time series.
Measures contained in the database vary in name and scale depending on the care unit where were captured. In this way, according to the types of analysis, it is important to identify the way to collect this data.