Death and Morbidity Prediction Using Data Mining in Perforated Peptic Ulcers

Hugo Peixoto, Algoritmi Research Center, University of Minho, Braga, Portugal
https://orcid.org/0000-0003-3957-2121
Lara Silva, University of Minho, Braga, Portugal
Soraia Pereira, University of Minho, Braga, Portugal
Tiago Jesus, University of Minho, Braga, Portugal
Vitor Neves Lopes, Centro Hospitalar do Tâmega e Sousa, Guilhufe, Portugal
António Abela, Algoritmi Research Center, University of Minho, Braga, Portugal

ABSTRACT

Peptic ulcers are not the most common complication in gastrointestinal mucosa, but these defects stand out as being the complication with the highest mortality rate. Several scoring systems based on clinical and biochemical parameters, such as the Boey and PULP scoring systems have been developed to predict the probability of mortality. In this study, a data mining process is performed in the medical data available, in order to evaluate how the scoring systems perform when trying to predict mortality and patients’ state complication. Furthermore, the presented paper studies the two scoring systems presented to define which one outperforms the other. On one hand PULP scoring allows a better mortality prediction achieving, above a 90% accuracy. One the other hand, regarding complications, the Boey system achieves better results leading to a better prediction when it comes to predicting patients’ state complication.

KEYWORDS
Boey, CRISP-DM, Data Mining, Health Complications, Knowledge Extraction, PULP, Scoring Systems

INTRODUCTION

Peptic ulcers are defined as defects in the gastrointestinal mucosa that extend through the muscularis mucosae (Sandler et al., 2002). Complications of peptic ulcer disease include bleeding, perforation, penetration, and gastric outlet obstruction. Its incidence, defined as the rate of new (or newly diagnosed) cases of the disease, varies geographically — in developed countries, hemorrhage is the most common cause (up to 73%), followed by perforation (9%) and obstruction (3%). Although not being the most common complication, the perforations stand out as being the complication with the highest mortality rate (Wang et al., 2010). Different numbers present in countries categorized as developing countries, as a review from Nigeria demonstrates, with obstruction being the most common cause of complication (56%), followed by perforation (30%) and bleeding (10%) (Irabor, 2005). Pieces of evidence implicate that the primary causes of complicated peptic ulcer disease are H. pylori infection, a common bacteria, and nonsteroidal inflammatory drugs. This work is based on a dataset that evaluates several clinical and biochemical parameters in patients with the diagnosis of peptic ulcer perforation, in order to classify them in two scoring systems – Peptic Ulcer Perforation

DOI: 10.4018/IJRQEH.2020010104

Copyright © 2020, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
(PULP) (Lohsiriwat et al., 2009) and Boey (Møller et al., 2012) – previously established to determine a patient’s prognosis with this pathology.

The PULP (Peptic Ulcer Perforation Score), is a score attributed through 11 different variables that predict 30-day mortality in patients operated on with PPU (Peptic Perforated Ulcers). The 11 variables have a different weight on the total score, which allows the quantification of the perforation. On the other hand, the Boey is attributed by three variables, with the same weight in the final result, that allows assessing the prognosis of patients with PPU that are submitted to surgery, helping in the assessment of mortality and morbidity (Agarwal et al., 2016).

The main objective of this work is to establish a relation between the scoring systems (PULP and Boey) in patients with perforated peptic ulcer and the outcome (mortality and morbidity) (Sanchez-Delgado et al., 2011). The first goal of the work is the attempt to understand which one of the scores predict better the parameters in the study (mortality and complications). At the same time, it is intended to understand which of the parameters in which are based on the two scores are more related to the two parameters in order to predict mortality and complications.

In order to achieve all the goals proposed, it was necessary to follow a detailed process that allows you to analyze and try to discover patterns in large amounts of data. This process is entitled “Data Mining” and involves methods of machine learning, statistics and database systems. To easily understand the method used, first will be explored, superficially, the concept of Data Mining and the main features of all the process. Then, we will explore, briefly, some of the work already done in this area. Only after that, the Data Mining process will effectively start, passing through all the five phases that constitute the process: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. In the end, it will be done an extensive analysis of the results and drawn some conclusions relevant to the propose.

Methodologies, Materials and Methods - Data Mining

Data mining is the process of extracting knowledge that matters from massive amounts of data. The idea is to build computer programs that sift through databases automatically, seeking patterns. If found, these regularities will likely generalize to make accurate predictions or descriptions on future data. It is an interdisciplinary field with contributions from many areas, such as statistics, machine learning, information retrieval, pattern recognition and bioinformatics. The main techniques for data mining include classification and prediction, clustering, outlier detection, association rules, sequence analysis, time series analysis and text mining, and some new techniques such as social network analysis and sentiment analysis. It should also be noted that the process described can be divided into two major categories, different from each other, the supervised and the unsupervised methods. The major difference between the two is the final goal sought with the application of the method. While the supervised search the achievement of value, requiring a target attribute, the unsupervised search for relations between the data, such as affinities and intrinsic structures (Neto et al., 2017; Zhao, 2012; Pereira et al., 2017).

Six Data Mining techniques were used to induce the data mining models: JRip, OneR, NB, DT, J48 and K-Means. JRip is one of the most basic and most popular algorithms. It implements a propositional rule learner, repeated incremental pruning to produce error reduction (RIPPER). It as a building state (formed by a growing phase and a pruning phase) and an optimization stage (Cohen, 1995; Kohavi, 1995). In its turn, OneR (short for One Rule) is a class for building a OneR classifier: it is a simple classification algorithm that uses a one-level decision tree. In other words, it uses the minimum-error attribute for prediction, discretizing numeric attributes (Najm, 2011). Then, the NB stands for Naive Bayes; it is a class for a Naive Bayes classifier that uses estimator classes. Numeric estimator precision values are chosen based on an analysis of the training data. The NB classifier learns very quickly and is easy to apply (Parsania et al., 2014; Wu et al., 2008).

Another model used was DT, that stands for Decision Table, and it is a class for building and using a simple decision table majority classifier. Decision tables are very easy to understand, and
One Moment in Time: A Patient's Story
www.igi-global.com/article/one-moment-time/76681?camid=4v1a