Expert Systems with Applications 40 (2013) 1054–1060

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Data mining applied to the cognitive rehabilitation of patients with acquired brain injury

A. Marcano-Cedeño [a,*], Paloma Chausa [a], Alejandro García [b], César Cáceres [a], Josep M. Tormos [b], Enrique J. Gómez [a]

[a] Bioengineering and Telemedicine Center, ETSI de Telecomunicación – UPM, 28040 Madrid, Spain
[b] Institut Guttmann Neurorehabilitation Hospital, 08916 Badalona, Spain

ARTICLE INFO

Keywords:
Acquired brain injury
Cognitive rehabilitation
Data mining
Decision tree
Multilayer perceptron
General regression neural network

ABSTRACT

Acquired brain injury (ABI) is one of the leading causes of death and disability in the world and is associated with high health care costs as a result of the acute treatment and long term rehabilitation involved. Different algorithms and methods have been proposed to predict the effectiveness of rehabilitation programs. In general, research has focused on predicting the overall improvement of patients with ABI. The purpose of this study is the novel application of data mining (DM) techniques to predict the outcomes of cognitive rehabilitation in patients with ABI. We generate three predictive models that allow us to obtain new knowledge to evaluate and improve the effectiveness of the cognitive rehabilitation process. Decision tree (DT), multilayer perceptron (MLP) and general regression neural network (GRNN) have been used to construct the prediction models. 10-fold cross validation was carried out in order to test the algorithms, using the Institut Guttmann Neurorehabilitation Hospital (IG) patients database. Performance of the models was tested through specificity, sensitivity and accuracy analysis and confusion matrix analysis. The experimental results obtained by DT are clearly superior with a prediction average accuracy of 90.38%, while MLP and GRRN obtained a 78.7% and 75.96%, respectively. This study allows to increase the knowledge about the contributing factors of an ABI patient recovery and to estimate treatment efficacy in individual patients.

1. Introduction

Acquired brain injury (ABI) is one of the leading causes of death and disability in the world. In Europe, brain injuries from traumatic and non-traumatic causes are responsible for more years of disability than any other cause (The Lancet Neurology, 2010). Because most of these patients are young people, their remaining functional limitations and psychosocial problems contribute significantly to health care related costs and loss of productivity.

After sustaining an ABI, patients have impairments consisting of not only physical, but also cognitive, social, and behavioral limitations. The most frequently occurring cognitive sequelae after an ABI pertain to mental process slowness, attention deficits, memory impairments, and executive problems. The injury dramatically changes the life of patients and their families (Pérez et al., 2010). The rapid growth on ABI case numbers and the importance of cognitive functions in daily activities, both demand efficient programs of cognitive rehabilitation.

Recovery from ABI can be facilitated with cognitive rehabilitation. Cognitive rehabilitation aims to compensate, or restore when possible, lost brain functions, improving the quality of life of the patients (Fundaci Institut Guttmann, 2008; Sohlberg & Mateer, 2001).

One of the problems of the rehabilitation process is its time length, that in many cases is inadequate for a complete and effective rehabilitation. To improve and expand the cognitive rehabilitation process, automated systems for cognitive rehabilitation of patients with ABI have been recently introduced (Solana, Cáceres, Gómez, Ferrer-Celma, & Ferre-Bergada, 2011; Tormos, García-Molina, García-Rudolph, & Roig, 2009). These systems generate large amounts of data. The analysis of these data, using data mining techniques, allows us to obtain new knowledge to evaluate and improve the effectiveness of the rehabilitation process. Also using information analysis and data mining techniques, we can create predictive models and decision support systems for the treatment of patients with ABI.

The data used in this study were obtained from the PREVIRNEC© platform. PREVIRNEC© is a cognitive tele-rehabilitation platform, developed over a web-based architecture based on web technologies and it's conceived as a tool to enhance cognitive rehabilitation, strengthening the relationship between the neuropsychologist and the patient, extending the treatment duration and frequency, allowing personalization of treatment and monitoring the performance of rehabilitation tasks. PPREVIRNEC© has been developed

* Corresponding author. Tel.: +34 915495700x33; fax: +34 913366828.
E-mail address: amarcano@gbt.tfo.upm.es (A. Marcano-Cedeño).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.08.034

during the past six years by the Universitat Rovira i Virgili and Technical University of Madrid (Spain), together with the Institut Guttmann Neurorehabilitation Hospital, IG (Spain) neuropsychology and research departments (Solana et al., 2011). This platform has been included in the hospital clinical protocols since 2005 and at the moment of this analysis PREVIRNEC© database stores 1120 patients, with a total of 183047 rehabilitation tasks executions.

Different statistical methodologies and predictive data mining methods have been applied to predict clinical outcomes of rehabilitation of patients with ABI (Rughani et al., 2010; Ji, Smith, Huynh, & Najarian, 2009; Pang et al., 2007; Segal et al., 2006; Brown et al., 2005; Rovlias & Kotsou, 2004; Andrews et al., 2002). Most of these studies are focused in determining survival, predicting disability or the recovery of patients, and looking for the factors that are better at predicting the patient's condition after suffering an ABI.

The purpose of this study is the novel application of data mining to predict the outcomes of the cognitive rehabilitation of patients with ABI. Three algorithms were used in this study: decision tree (DT), multilayer perceptron (MLP) and a general regression neural network (GRNN). PREVIRNEC© database and IG's Electronic Health Records (EHR) (Institut Guttmann Neurorehabilitation Hospital, 1997) has been used to test the algorithms. For assessing the algorithm's accuracy of prediction, we used the most common performance measures: specificity, sensitivity, accuracy and confusion matrix. The results obtained were validated using the 10-fold cross-validation method.

The remainder of this paper is organized as follows. Section 2 presents a brief introduction to data mining, the algorithms used in this research and a detailed description of the database. Section 3 shows the experimental results obtained. Section 4 presents a discussion of these results. Finally, Section 5 describes the summarized conclusions.

2. Materials and methods

2.1. Review of data mining techniques

2.1.1. Knowledge discovery in databases and Data mining

Today there is still some confusion about the terms *Knowledge Discovery in Databases* (KDD) and *Data Mining (DM)*. Often these two terms are used interchangeably. The term KDD is used to denote the overall process of turning low-level data into high-level knowledge. KDD is defined as: *the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data* (Fayyad et al., 1996). On the other hand, data mining is commonly defined as the extraction of patterns or models from observed data. Therefore DM is at the core of the knowledge discovery process, although this step usually takes only a small part (estimated at 15–25%) of the overall KDD effort (Brachman & Anand, 1996). The overall KDD process contemplates five phases: data preparation, data preprocessing, data mining, evaluation and interpretation, and implementation (Fayyad et al., 1996).

Data mining classification technique is split in two phases: the first one is the construction of a classification model that consists in training the algorithm using a data set in order to build the predictive classification model. The second phase is the evaluation of the model classification efficiency, using a testing data set (Yen, Chen, & Chen, 2011; Köksa, Batmaz, & Caner, 2011; Seng, & Chen, 2010) .

2.1.2. Decision trees (J48)

Decision tree (DT) provides powerful techniques for classification and prediction, that are widely used in data mining. The most commonly used DT algorithms include Quinlans ID3, C4.5, C5 (Quinlan, 1993), Breimans classification and regression tree (CART) (Breiman, Friedman, Olshen, & Stone, 1984) and Chi-squared Automatic Interaction Detector (CHAID) (Hartigan, 1975). As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of improving the prediction accuracy. In doing so, it uses mathematical algorithms to identify a variable and corresponding threshold for the variable that splits the input observation into two or more subgroups (Yeh, Hou, & Chang, 2012). The most commonly mathematical algorithm used for splitting includes Entropy based information gain (used in ID3, C4.5, C5), Gini index (used in CART), and chi-squared test (used in CHAID). This step is repeated at each leaf node until the complete tree is constructed. The objective of the splitting algorithm is to find a variable-threshold pair that maximizes the homogeneity (order) of the resulting two or more subgroups of samples (Delen, Fuller, McCann, & Ray, 2009).

Based on the favorable prediction results we have obtained from the preliminary runs, in this study we use the J48 algorithm as our decision tree method. The J48 algorithm is an implementation of the C4.5 algorithm (Witten & Frank, 2005) included in the WEKA software platform (weka, 2011). In order to tune the J48 algorithm to optimize its performance, we varied the confidence factor (default value of confidence factor is 0.5), which is a value that is used by the algorithm to prune developed trees (pruning of a DT is conducted to avoid over-fitting the model on the records used for modelling). A lower confidence factor results in more pruning (Witten & Frank, 2005), and a minimum number of objects for a leaf of 2.

2.1.3. Multilayer perceptron (MLP)

Multilayer perceptron (MLP) are the most commonly used feedforward neural networks due to their fast operation, ease of implementation, and smaller training set requirements (Haykin, 1994). The MLP consists of three sequential layers: input, hidden and output layers. The hidden layer processes and transmits the input information to the output layer. A MLP model with insufficient or excessive number of neurons in the hidden layer may cause problems of bad generalization and over-fitting. There is no analytical method for determining the number of neurons in the hidden layer, so it is usually chose empirically (Haykin, 1994; Marcano-Cedeño, Quintanilla-Domínguez, & Andina, 2011).

Each neuron j in the hidden layer sums its input signals x_i impinging onto it after multiplying them by their respective connection weights w_{ji}. The output of each neuron is described as follows:

$$y_i = f\left(\sum w_{ji} x_i\right) \tag{1}$$

where f is an activation function using the weighted summations of the inputs. An activation function can be a simple threshold, sigmoidal, or hyperbolic tangent function (Marcano-Cedeño et al., 2011; Güler, Gökçil, & Gülbandilar, 2009; Orhan, Hekim, & Ozer 2011). In this study, a sigmoidal transfer function was used as the activation function.

The sum of squared differences between the desired and actual values of the output neurons E is defined as follows (Marcano-Cedeño et al., 2011; Güler et al., 2009):

$$E = \frac{1}{2}\sum (y_{dj} - y_j)^2 \tag{2}$$

where y_{dj} is the desired value of output neuron j and y_j is the actual output of that neuron. Each w_{ji} weight is adjusted to minimize the value E depending on the training algorithm adopted. In this context, the backpropagation method (BP) is widely used as a primary part of an artificial neural network model. However, since the BP has some constraints such as slow convergence (Haykin, 1994; Marcano-Cedeño et al., 2011; Güler et al., 2009) or not being