Computer Methods and Programs in Biomedicine 80 Suppl. 1 (2005) S39-S49

ELSEVIER

Computer Methods and Programs in Biomedicine

www.intl.elsevierhealth.com/journals/cmpb

Knowledge discovery with classification rules in a cardiovascular dataset

Vili Podgorelec[a,*], Peter Kokol[a], Milojka Molan Stiglic[b], Marjan Heričko[a], Ivan Rozman[a]

[a]*University of Maribor - FERI, Smetanova 17, SI-2000 Maribor, Slovenia*
[b]*Maribor Teaching Hospital, Department of Pediatric Surgery, Maribor, Slovenia*

KEYWORDS
Machine learning;
Knowledge discovery;
Classification rules;
Pediatric cardiology;
Medical data mining

Summary In this paper we study an evolutionary machine learning approach to data mining and knowledge discovery based on the induction of classification rules. A method for automatic rules induction called AREX using evolutionary induction of decision trees and automatic programming is introduced. The proposed algorithm is applied to a cardiovascular dataset consisting of different groups of attributes which should possibly reveal the presence of some specific cardiovascular problems in young patients. A case study is presented that shows the use of AREX for the classification of patients and for discovering possible new medical knowledge from the dataset. The defined knowledge discovery loop comprises a medical expert's assessment of induced rules to drive the evolution of rule sets towards more appropriate solutions. The final result is the discovery of a possible new medical knowledge in the field of pediatric cardiology.

1. Introduction

Modern medicine generates huge amounts of data and there is an acute and widening gap between data collection and data comprehension. Obviously it is very difficult for a human to make use of such amount of information (i.e. hundreds of attributes, thousand of images, several channels of 24 hours of ECG or EEG signals) and to be able to find basic patterns, relations or trends in the data. Thus, data becomes less and less useful, the transformation data ⇒ information harder and harder, and the transformation data ⇒ information ⇒ knowledge almost impossible. Thus, there is a great need to find new methods for data analysis to facilitate the creation of knowledge that can be used for clinical decision making. Intelligent systems for knowledge extraction are tools that can help in achieving this goal.

1.1. Objectives and scope of the paper

This paper has two main objectives. The first is to introduce a new intelligent knowledge extraction paradigm based on evolutionary rule sets induction. We present a new hybrid classification algorithm based on genetic algorithms (GAs) and genetic programming (GP) - the AREX approach.

*Correspondence to: Dr. Vili Podgorelec. University of Maribor - FERI, Smetanova ulica 17, SI-2000 Maribor, Slovenia. Tel.: +386 2 235 5121; fax: +386 2 23 44 134.
E-mail: vili.podgorelec@uni-mb.si

0169-2607/$ - see front matter

AREX (Automatic Rules Extractor) is a general hybrid method that incorporates two original, independent algorithms along with a simple genetic algorithm that together solve the problem of automatic induction of classification rules. The first algorithm is a multi-population self-adapting genetic algorithm for the induction of decision trees. The second is a system for the evolution of programs in an arbitrary programming language, which is used to evolve classification rules. Finally, an optimal set of classification rules is determined with a simple genetic algorithm.

The second objective is to present a case study of using the developed algorithm to discover new knowledge in a problem of early and accurate identification of cardiovascular problems in pediatric patients. It is shown how AREX can be used to extract medical knowledge, and the results obtained in this manner are evaluated by a medical expert. To objectively compare the developed AREX approach with existing methods the results are compared to those obtained with other classification methods.

The paper is organized as follows. Section 2 presents a short overview of data mining and knowledge discovery, with emphasis on the evolutionary induction of decision trees; it indicates some reasons for the development of AREX. Section 3 presents the developed AREX algorithm in detail. Section 4 presents a case study of using AREX upon a cardiovascular database, where all the obtained results are evaluated and compared with the existing classification algorithms. Finally, Section 5 presents a discussion that concludes the paper.

2. Data mining and knowledge discovery

Although a great deal of time and effort is spent in building and maintaining all kinds of databases, it is nonetheless rare that the full potential of his valuable resource is realised. The principal reason for this paradox is that the majority of organisations lack the insight and/or expertise to effectively translate information into usable knowledge [1]. In light of these conditions, there exists a clear need for automated methods and tools to assist in exploiting the vast amount of available data. This requirement has led to the development of data mining technology. Data mining is an umbrella term which describes the process of uncovering patterns, associations, changes, anomalies and statistically significant structures and events in data. Traditional data analysis is assumption driven in the sense that a hypothesis is manually formed and validated (by statistical means) against the data. In contrast, data mining is discovery driven in that useful patterns are automatically extracted from the data [2]. In order to accomplish this task, data mining systems frequently utilise methods from disciplines such as artificial intelligence, machine learning and pattern recognition [3].

Data mining algorithms usually operate on data sets composed of vectors (instances) of independent variables (features, attributes). For example, a database may describe a group of people in terms of their age, sex, income and occupation. In this case, age is an example of an attribute and each instance corresponds to a distinct individual.

To discover the hidden patterns in data, it is essential to build a model consisting of independent variables that can be used to determine a dependent variable (also known as class or decision). Building such a model therefore consists of identifying the relevant independent variables (attributes) and minimising the predictive error [4]. It is also highly desirable to find the simplest possible model that fits the data, since these are typically the most meaningful and easiest to interpret. This last requirement reflects the principle of Occam's Razor which tells us to prefer the simplest model that fits the data [5].

Before we proceed, it is important to distinguish data mining from pattern recognition as these terms are sometimes confused with each other. Pattern recognition is primarily concerned with the construction of accurate classifiers. A classifier is fundamentally a mapping between a set of input variables $x_1, \ldots, x_n$ to an output variable y whose value represents the class label $\omega_1, \ldots, \omega_m$ [5]. In general, representing the knowledge embodied within the classifier structure is not a priority. Consequently, while there is no shortage of extremely accurate classifiers, some of the best are akin to a black box; that is, they give little or no insight into why they make decisions. Neural networks [5,6] exemplify these types of systems because their classification rules are embedded in their structure. Since neural network components (node activation functions, connection weights, etc.) encode complex mathematical functions, articulating the rules they represent is a difficult problem [7]. In contrast, the primary purpose of data mining is not simply classification, but to provide meaningful knowledge to the user regarding the classification process. Thus, the models produced by data mining algorithms should be in a form that lends itself to analysis by the user. Decision rule sets which linearly partition the data space into class-homogeneous regions meet this requirement. Examples of techniques that accomplish decision rule induction from data