Applied Soft Computing 12 (2012) 2693–2706

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

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

Multi-instance genetic programming for predicting student performance in web based educational environments

Amelia Zafra*, Sebastián Ventura

Department of Computer Science and Numerical Analysis, University of Cordoba, Spain

ARTICLE INFO

Article history:
Received 3 July 2010
Received in revised form 3 February 2012
Accepted 13 March 2012
Available online 25 April 2012

Keywords:
Educational data mining
Multiple instance learning
Genetic programming
Classification

ABSTRACT

A considerable amount of e-learning content is available via virtual learning environments. These platforms keep track of learners' activities including the content viewed, assignments submission, time spent and quiz results, which all provide us with a unique opportunity to apply data mining methods. This paper presents an approach based on grammar guided genetic programming, G3P-MI, which classifies students in order to predict their final grade based on features extracted from logged data in a web based education system. Our proposal works with multiple instance learning, a relatively new learning framework that can eliminate the great number of missing values that appear when the problem is represented by traditional supervised learning. Experimental results are carried out on data sets with information about several courses and demonstrate that G3P-MI successfully achieves better accuracy and yields trade-off between such contradictory metrics as sensitivity and specificity compared to the most popular techniques of multiple instance learning. This method could be quite useful for early identification of students at risk, especially in very large classes, and allows the instructor to provide information about the most relevant activities to help students have a better chance to pass a course.

1. Introduction

The widespread accessibility of the World Wide Web and the increase of easy tools to browse the resources on the Web have made that web based education systems extremely are popular and the means of choice for both distance education and as a complement for face to face education. This has led to the development and use of a number of sophisticated web-based learning and course management tools around the world. These systems called virtual learning environments (VLEs) include among other features, course content delivery features, quiz modules, assignment submission components, a grade reporting system and logbooks.

Today the important challenge facing higher education is to reach a stage that facilitates more efficient, effective and accurate educational processes for universities. Data mining is considered the most appropriate technology for giving additional insight to the lecturer, student, manager, and other educational staff and acts as an active automated assistant in helping them to make better decisions about their educational activities.

Accurately predicting student performance is useful in many different contexts in universities. For example, identifying exceptional students for scholarships is an essential part of the admission process in undergraduate and postgraduate institutions and identifying weak students who are likely to fail, is also important for allocating limited tutoring resources. In this study we apply data mining methods in order to identify by means of the student's work and the use of the platform if he/she has a higher or lower probability of passing the course. The idea is to discover if students that pass or fail both use online resources in a different way. If this is so, we identify how different types of problems influence students' achievement determining which activities are more relevant for passing a course so that we can help instructors to develop more effectively and efficiently homework. With this information, we could guide the learners' activities and intelligently recommend on-line activities or resources that would support and improve learning. Thus students could follow the learning process better and teachers could appraise on-line course structure effectiveness. Nowadays, there has been a growing interest in solving similar questions analyzing valuable information to detect possible errors, shortcomings and improvements in student performance and discovering how the student's motivation affects the way he or she interacts with the software [1–3]. All previous studies use traditional supervised learning to represent the problem. However, such representation generates instances with many missing values because the information about the problem is incomplete. Each course has different types and numbers of activities and each student carries out the number of activities he/she find the most interesting, dedicating more or less time to resolve them.

* Corresponding author. Tel.: +34 957212031; fax: +34 957218630.
E-mail addresses: azafra@uco.es (A. Zafra), sventura@uco.es (S. Ventura).

1568-4946/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2012.03.054

In order to overcome these shortcomings, we present a grammar guided genetic programming algorithm, G3P-MI, to solve the problem using a multiple instance learning (MIL) representation. This learning framework is considered to be an extension of supervised traditional learning that allows us to eliminate the missing values that appear in traditional supervised learning offering a more flexible representation that adapts itself to the information available. The most representative paradigms in MIL are compared to our proposal and experimental results show that G3P-MI is more effective for obtaining a more accurate model as well as for finding a trade-off between contradictory measurements like sensitivity and specificity. Moreover, it adds comprehensibility to the knowledge discovered, allowing interesting relationships to be obtained between activities, resources and student achievement.

The paper is organized as follows. Section 2 introduces multi-instance learning and previous works about educational data mining. Section 3 presents the problem of classifying students' performance from a multi-instance perspective. Section 4 presents the G3P-MI algorithm and Section 5 reports on experiment results which compare our proposal to the most representative multiple instance learning paradigms. Finally, Section 6 summarizes the main contributions of this paper and suggests some future research directions.

2. Background

2.1. Multiple instance learning

Multiple instance learning (MIL) introduced by Dietterich et al. [4] consists of generating a classifier that will correctly classify unseen patterns. The main characteristic of this learning is that the patterns are bags of instances where each bag can contain different numbers of instances. There is information about the bags because a bag receives a special label, but the labels of instances are unknown. Although the actual learning process is quite similar to traditional supervised learning, the two approaches differ in the class labels provided from which they learn. In a traditional machine learning setting, an object m is represented by a feature vector v, which is associated with a label $f(m)$. However, in the multiple instance setting, each object m may have i various instances denoted $m_1, m_2, \ldots, m_i$. Each of these variants will be represented by a (usually) distinct feature vector $V(m_i)$. A complete training example is therefore written as $(\{V(m_1), V(m_2), \ldots, V(m_i)\}, f(m))$. In this case, the $f(m)$ represents the information in the examples, but there is no information about each individual instance. The goal of learning is to find a good approximation in function $f(m_i)$, $\hat{f}(m_i)$, analyzing a set of training examples and labeled by $f(m_i)$. To obtain this function Dietterich et al. define a hypothesis that assumes that if the result observed is *positive*, then at least one of the variant instances must have produced that positive result. Furthermore, if the result observed is *negative*, then none of the variant instances could have produced a positive result. This can be modeled by introducing a second function $g(V(m_{i,j}))$ that takes a single variant instance and produces a result. The externally observed result, $f(m_i)$, can then be defined as follows:

$$f(m_i) = \begin{cases} 1, & \text{if } \exists j | g(V(m_{i,j})) = 1 \\ 0, & \text{otherwise} \end{cases}$$

The distinctive point of this learning, where each example or pattern (called bag) can be represented by a different number of instances, allows us to set out different relationships between instances in a bag and the label of that bag. This fact introduces a priori more complexity into the learning process since the number of instances that are actually positive inside a positive pattern is unknown. However, it also provides us with greater flexibility with respect to classical representation, showing a more appropriate form of representation in a great number of applications, which improves the efficiency and effectiveness obtained in traditional learning. Thus, over the past few years, many applications have been formulated as MIL problems. These include text categorization [5], content-based image retrieval [6,7] and image annotation [8,9], drug activity prediction [10,11], web index page recommendation [12], video concept detection [13,14], semantic video retrieval [15] and pedestrian detection [16]. In all cases MIL provides a more natural form of representation that manages to improve the results obtained by traditional supervised learning.

In order to solve these problems, many MIL methods have been proposed. The first MIL method is APR [4], which represents the target concept by an axis-parallel rectangle (or hyper-rectangle) in the feature space. The rectangle includes at least one instance from each positive bag but excludes all instances from the negative bags. Maron and Lozano-Pérez [10] proposed a measure called Diverse Density (DD), which essentially measures how many different positive bags have instances near a point in the feature space and how far the negative instances are from that point. A DD-based method tries to find the point with the highest DD value as the target concept. EM-DD [17] combines expectation-maximization (EM) with the DD formulation by using EM to search for the most likely concept. Several other methods try to modify standard single instance learning methods for MIL by introducing constraints derived from MIL formulation, such as multi-instance lazy learning algorithms which extend k nearest-neighbor algorithms (kNN) [18], multi-instance tree learners which adapt classic methods [19], multi-instance rule inducers which adapt the RIPPER algorithm [20], multi-instance Bayesian approach [21], multi-instance neural networks which extend standard neural networks [22], multi-instance kernel methods which adapt classic support vector machines [5,9] and multi-instance ensembles which show the use of ensembles in this learning [11]. Finally, it is worth mentioning that recently MIL has also attracted the attention of unsupervised learning in proposals on clustering [23,24].

2.2. Data mining and e-learning

Data mining is the process of extracting useful knowledge and information from a data collection. Nowadays, data mining has been used in many application domains such as the biomedical industry, retail and marketing, telecommunications, web Mining, computer auditing, financial industry, medicine and so on.

An established research community has emerged over the last decade from the synergistic fields of education. As a newer subfield of data mining, educational data mining encompasses its own unique range of research questions and approaches. There exists a wide range of ideas as to how e-learning experiences can be improved by the utilization of appropriate data mining techniques. The hidden patterns, associations, and anomalies that are discovered by data mining techniques in educational data can improve decision making processes in higher educational systems. This improvement can bring such advantages as maximizing educational system efficiency, decreasing student's drop-out rates, increasing student's promotion rates, increasing student's retention rates, increasing student's transition rates, increasing the educational improvement ratio, increasing the student's success, increasing the student's learning outcome, and reducing the cost of system processes. Recently a number of studies have been undertaken to investigate the prospect of using data mining for e-learning decision making, as well as for identifying irregular