Knowledge-Based Systems 23 (2010) 529–535

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education

S. Kotsiantis, K. Patriarcheas, M. Xenos *

Hellenic Open University, School of Sciences and Technology, Computer Science, Greece

ARTICLE INFO

Article history:
Received 11 September 2009
Received in revised form 17 February 2010
Accepted 19 March 2010
Available online 23 March 2010

Keywords:
Educational data mining
Online learning algorithms
Classifiers
Voting methods

ABSTRACT

The ability to predict a student's performance could be useful in a great number of different ways associated with university-level distance learning. Students' marks in a few written assignments can constitute the training set for a supervised machine learning algorithm. Along with the explosive increase of data and information, incremental learning ability has become more and more important for machine learning approaches. The online algorithms try to forget irrelevant information instead of synthesizing all available information (as opposed to classic batch learning algorithms). Nowadays, combining classifiers is proposed as a new direction for the improvement of the classification accuracy. However, most ensemble algorithms operate in batch mode. Therefore a better proposal is an online ensemble of classifiers that combines an incremental version of Naive Bayes, the 1-NN and the WINNOW algorithms using the voting methodology. Among other significant conclusions it was found that the proposed algorithm is the most appropriate to be used for the construction of a software support tool.

1. Introduction

Distance education is an educational method whose main characteristic, setting it apart from other educational methods, is that the student is being taught and instructed without the physical presence of a tutor in a teaching classroom, based on a special, tutorially designed learning material and on his/her communication with the tutor [1]. In distance education the student often feels isolated consequently the communication between the student and the teacher as well with the other students is an important parameter for the success of a distance education program [2,3]. The student may become disheartened by difficulties encountered in the learning process and which may lead to a slowdown and/or quitting his/her studies. Student encouragement and support are a primary concern of a good, modern programme of distance study. Consequently, for a web-based open and dynamic learning environment, personalized support for learners becomes more important [4]. Previous research studies [5,6] have shown that students come to distance education courses with variable expectations of the levels of service and support they will receive from their tutors. It is obvious, the tutors in a distance education course have a particular role and it is important to be able to recognize and locate students with high probability of poor performance in order to take precautions and be better prepared to face such cases. Consequently, given that the distance education it addresses to adults with special educational needs and incongruity (age, professional and family obligations, etc.), there is a growing interest in the factors predicting the student's performance particularly in distance education environments [7–12]. A very promising arena to attain this objective is the use of data mining and machine learning algorithms [13].

Supervised learning algorithms are presented with instances, which have already been pre-classified in some way. That is, each instance has a label, which identifies the class to which it belongs and so this set of instances is sub-divided into classes. Supervised machine learning explores algorithms that reason from the externally supplied instances to produce general hypotheses, which will make predictions about future instances.

To induce a hypothesis from a given dataset, a learning system needs to make assumptions about the hypothesis to be learned. These assumptions are called biases. A learning system without any assumptions cannot generate a useful hypothesis since the number of hypotheses that are consistent with the dataset is usually enormous. Since every learning algorithm uses some biases, it behaves well in some domains where its biases are appropriate while it performs poorly in other domains [14]. Therefore, combining classifiers is proposed as a new direction for the improvement of the classification accuracy.

However, most ensemble algorithms operate in batch mode, i.e., they repeatedly read and process the entire training set. Basically, they require at least one pass through the training set for every

* Corresponding author. Address: Software Quality Laboratory, School of Sciences and Technology, Hellenic Open University, 12–15 Tsamadou str, Patras, Zip code: 26222, Greece. Tel.: +30 2610 367405; fax: +30 2610 367520.
E-mail address: xenos@eap.gr (M. Xenos).

0950-7051/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2010.03.010

base model to be included in the ensemble. The base model learning algorithms themselves may require several passes through the training set to create each base model. In situations where data is being generated continuously as in an educational environment, storing data for batch learning is impractical, which makes using these ensemble-learning algorithms impossible.

Incremental learning ability is very important to machine learning approaches designed for solving real-world problems due to two reasons. Firstly, it is almost impossible to collect all helpful training examples before the trained system is put into use. Therefore when new examples are fed, the learning approach should have the ability of doing some revisions on the trained system so that unlearned knowledge encoded in those new examples can be incorporated. Secondly, modifying a trained system may be cheaper in time cost than building a new system from scratch, which is useful especially in real-time applications.

In the present work an ensemble that combines an incremental version of Naive Bayes, the 1-NN and the WINNOW algorithms using the voting methodology is proposed. This paper uses the proposed ensemble in order to predict the students' performance in a distance learning system.

The application of the proposed technique in predicting students' performance proved to be useful for identifying poor performers and it can enable tutors to take precautionary measures at an earlier stage, even from the beginning of an academic year, in order to provide additional help to the groups at risk. The probability of more accurate diagnosis of students' performance is increased as new curriculum data has entered during the academic year, offering the tutors more effective results.

This paper is organised as follows: Section 2 introduces some basic themes about Educational data mining for predicting student performance and in online learning algorithms and incremental ensemble classifiers, while Section 3 discusses the proposed ensemble method. The dataset used for experiments is described in Section 4. Experiment results and comparisons of the proposed combining method with other learning algorithms are presented in Section 5. Finally, Section 6 describes summary and further research topics.

2. Background

At this point it is advisable to present some basic themes about Educational data mining for predicting student performance and in online learning algorithms and incremental ensemble classifiers.

2.1. Educational data mining for predicting student performance

To implement real intelligence or adaptivity, the models for tutoring systems should be learnt from data. However, the data sets are so small that machine learning methods cannot apply directly. Hämäläinen and Vinni [15] tackled this problem, and gives general for creating accurate classifiers for educational data. They describe experiment to predict course success with more 80% accuracy.

Minaei-Bidgoli et al. [16] present an approach to classifying students in order to predict their final grade based on features extracted from logged data in an education web-based system and they demonstrates a genetic algorithm (GA) to successfully improve the accuracy of combined classifier performance, about 10–12% when comparing to non-GA classifier.

Some of the most useful data mining tasks and methods are classification, clustering, visualization, association rule and statistics mining [17]. These methods uncover new, interesting and useful knowledge based on students' usage data [18]. Some of the main e-learning problems or subjects to which data mining techniques have been applied [19] are dealing with the assessment of student's learning performance, provide course adaptation and learning recommendations based on the students' learning behaviour, dealing with the evaluation of learning material and educational web-based courses, provide feedback to both teachers and students of e-learning courses, and detection of atypical student's learning behaviour.

As for the classification (one of the most useful educational data mining tasks in e-learning), there are different educational objectives for using classification, such as: to group students who are hint-driven or failure-driven and find common misconceptions that students possess [20], to predict/classify students when using intelligent tutoring systems [15], etc. Some other examples are: predicting a student's academic success (to classify as low, medium and high risk classes) using different data mining methods [21]; using neural network models from Moodle logs [22]. Finally, Lykourentzou et al. [23] use feed-forward neural networks, support vector machines and probabilistic ensemble simplified fuzzy ARTMAP for the prediction of student dropout.

2.2. Online learning algorithms and incremental ensemble classifiers

When comparing online and batch algorithms, it is worthwhile to keep in mind the different types of setting where they may be applied. In a batch setting, an algorithm has a fixed collection of examples in hand, and uses them to construct a hypothesis, which is used thereafter for classification without further modification. In an online setting, the algorithm continually modifies its hypothesis as it is being used; it repeatedly receives a pattern, predicts its classification, finds out the correct classification, and possibly updates its hypothesis accordingly.

The on-line learning task is to acquire a set of concept descriptions from labelled training data distributed over time. This type of learning is important for many applications, such as computer security, intelligent user interfaces, and market-basket analysis. For instance, customer preferences change as new products and services become available. Algorithms for coping with concept drift must converge quickly and accurately to new target concepts, while being efficient in time and space.

Desirable characteristics for incremental learning systems in environments with changing contexts are:

- the ability to detect a context change without the presentation of explicit information about the context change to the system.
- the ability to quickly recover from a context change and adjust the hypotheses to fit the new context.
- the capability to make use of previous experience in situations where old contexts reappear.

Online learning algorithms process each training instance once "on arrival" without the need for storage and reprocessing, and maintain a current hypothesis that reflects all the training instances seen so far. Such algorithms are also useful with very large datasets, for which the multiple passes required by most batch algorithms are prohibitively expensive.

Numerous surveys have been conducted to study the incremental ensemble classifiers for a number of applications [24–31].

Researchers as Swere et al. [32] have developed online algorithms for learning traditional machine learning models such as decision trees. Given an existing decision tree and a new example, this algorithm adds the example to the example sets at the appropriate non-terminal and leaf nodes and then confirms that all the attributes at the non-terminal nodes and the class at the leaf node are still the best.

Batch neural network learning is often performed by making multiple passes (known in the literature as epochs) through the