Article

Journal of Educational Computing Research
0(0) 1–23

DOI: 10.1177/0735633117752614
journals.sagepub.com/home/jec
SAGE

Predicting Secondary School Students' Performance Utilizing a Semi-supervised Learning Approach

Ioannis E. Livieris[1], Konstantina Drakopoulou[2], Vassilis T. Tampakas[1], Tassos A. Mikropoulos[3], and Panagiotis Pintelas[2]

Abstract

Educational data mining constitutes a recent research field which gained popularity over the last decade because of its ability to monitor students' academic performance and predict future progression. Numerous machine learning techniques and especially supervised learning algorithms have been applied to develop accurate models to predict student's characteristics which induce their behavior and performance. In this work, we examine and evaluate the effectiveness of two wrapper methods for semisupervised learning algorithms for predicting the students' performance in the final examinations. Our preliminary numerical experiments indicate that the advantage of semisupervised methods is that the classification accuracy can be significantly improved by utilizing a few labeled and many unlabeled data for developing reliable prediction models.

Keywords

educational data mining, student's evaluation system, semisupervised methods, self-training, Yet Another Two Stage Idea

[1]Department of Computer and Informatics Engineering (DISK Lab), Technological Educational Institute of Western Greece, Patras, Greece
[2]Department of Mathematics, University of Patras, Patras, Greece
[3]The Educational Approaches to Virtual Reality Technologies Lab, University of Ioannina, Ioannina, Greece

Corresponding Author:
Ioannis E. Livieris, Department of Computer and Informatics Engineering (DISK Lab), Technological Educational Institute of Western Greece, Patra GR 263-40, Greece.
Email: livieris@teiwest.gr

Introduction

Educational data mining (EDM) is an academic field of research that seeks to obtain, from the data stored in educational environments, new and useful information in order to develop and strengthen the cognitive theories of teaching and learning (Baker & Yacef, 2009). EDM supported by computer-based approaches typically comes from two wide categories: data (e.g., textual material from course information) and knowledge (e.g., useful information that could potentially have a greater impact on educational research and practice). Its great importance springs from the fact that it allows educators and researchers to extract useful conclusions from sophisticated and complicated questions. More analytically, while traditional database queries can only answer questions such as ''find the students with poor performance,'' data mining can provide answers to more abstract questions like ''find the students who will exhibit poor performance'' (Livieris, Mikropoulos, & Pintelas, 2016, p. 1). Therefore, the application of EDM is mainly concentrated on the development of intelligent methods to discover the key characteristics from students' records and use them for understanding the students' learning behavior and predict their future performance and achievements (Acharya & Sinha, 2016; Chen, Liu, Ou, & Liu, 2000; Knauf, Sakurai, Tsuruta, & Jantke, 2010; Koedinger, McLaughlin, & Heffernan, 2010; Tsai, Ouyang, & Chang, 2016).

In Greece, like in most countries, secondary education takes place after 6 years of primary education and may be followed by higher education or vocational training. Its main objectives are to engender the best possible education and to enhance a balanced and all-round development of the students' personality at a cognitive and emotional level. It comprises two main stages: Gymnasium covers the first 3 years which aims to enrich students' knowledge in all fields of learning and support the development of composite and critical thinking. The next 3 years are covered by Lyceum which further cultivates the students' personalities while at the same time prepares them for admission in higher education. Essentially, Lyceum acts like a bridge between school education and higher learning specializations that are offered by universities.

Therefore, the ability to predict the students' performance with high accuracy in many stages of the school period is considered essential not only for the educators and the educational institutes but also for the students. More comprehensively, the ''knowledge discovery'' can assist educators conduct their classes' better, identify learning difficulties, and improve their teaching methods while the students could be provided with a first evaluation of their progress and possibly enhance their performance. Furthermore, analyzing students' learning and making predictions regarding further aspects of their performance is essential for an educational system in order to provide personalized learning activities tailored to each student's special needs or even guiding them to follow technical

education. Thus, it is of major importance to closely monitor the students' performance in order to identify possible retardation and proactively intervene toward their academic enhancement through the assignment of extra learning material, small group training, and so forth.

However, the prediction of students' performance or even the identification of students who are likely to exhibit poor performance is often a difficult and challenging task. Moreover, if such identification is possible, it is usually too late to avoid students' failure. A workable solution to prevent this trend is to analyze the knowledge acquired by students' previous academic performance. The core of this research is the acquisition of knowledge from students' academic performance records. This work proposes a methodology to develop decision systems able to guide students through the enrollment process starting from this knowledge.

Semisupervised learning constitutes the appropriate technique to exploit data originated from educational institutes, since there is often a lack of labeled data, while unlabeled is vast. However, the majority proportion of these studies examines the efficiency of supervised methods, and especially classification (usually pass or fail), while semisupervised learning methodologies have been rarely and only recently applied to the educational field (Kostopoulos et al., 2015a, 2015b) for predicting students' performance in distance higher education.

In this work, we examine the effectiveness of two wrapper methods for semisupervised learning for the prediction of high school students' in the final examinations. More analytically, we evaluated the performances of self-training and Yet Another Two Stage Idea (YATSI) approaches with the most classic classification methods. Self-training and YATSI constitute two of the most efficient and frequently utilized semisupervised algorithms which have been successfully used in a variety of real-world applications (Catal & Diri, 2009; Driessens, Reutemann, Pfahringer, & Leschi, 2006; Levatic, Dzeroski, Supek, & Smuc, 2013; Roli & Marcialis 2006; Rosenberg, Hebert, & Schneiderman, 2005; Sigdel et al., 2014) providing some promising classification results. Our preliminary numerical experiments illustrate that the classification accuracy can be significantly improved, utilizing a few labeled and many unlabeled data for developing reliable prediction models.

The remainder of this article is organized as follows: The next section presents recent studies of data mining application in education. A Review of Semisupervised Machine Learning Techniques section presents some elementary machine learning definitions and a brief description of the semisupervised techniques used in our study. The Dataset section presents a description of the educational dataset and a detailed analysis of the data attributes. In the Experimental Results section, we present a series of tests in order to examine the accuracy of semisupervised learning algorithms with well-known supervised methods. The final section presents our conclusion and our proposals for future research.

Literature Review of Related Work

During the last two decades, the application of data mining techniques has gained popularity in the modern educational era, spurred by the fact that it enables all educational stakeholders to discover new, interesting, and useful knowledge about students and potentially improve some aspects of the quality of education (Livieris et al., 2016). Some excellent surveys (Baker & Yacef, 2009; Romero & Ventura 2007, 2010; Romero, Ventura, Pechenizkiy, & Baker, 2010) presented the major trends in EDM research, describing in detail the process of mining learning data to discover new insights and how those insights impact the activity or practitioners in education. A number of rewarding studies have been carried out in recent years and some of them are presented as follows.

Oladokun, Adebanjo, and Charles-Owaba (2008) investigated the prediction ability of neural networks for predicting the likely performance of a candidate being considered for admission into the university. The attributes in their study concerned a number of socio-economic, biological, environmental and academic factors. They developed another accurate prediction neural network model to early identify the students' final achievement exhibiting 74% successful classification performance.

Cortez and Silva (2008) predicted the secondary student grades of two core classes (Mathematics and Portuguese) of two secondary school students. The data were extracted from school records, as well as data provided by the students through questionnaires. They applied four classification algorithms on three data setups, with different combination of attributes, trying to find out those with more effect on the prediction. Based on their numerical experiments, the authors concluded that the students' achievement is highly correlated with their performance in the past years, and less with other academic, social, and cultural characteristic of the students and their contexts.

Thai-Nghe, Janecek, and Haddawy (2007) used machine learning techniques for predicting the performance of undergraduate and postgraduate students at two academic institutes. Along this line, Thai-Nghe, Busche, and Schmidt-Thieme (2009) presented an extensive study to deal with the class imbalance problem in order to improve the prediction results of academic performances. First, they balanced the datasets and then they used both cost-insensitive and -sensitive learning with support vector machine for the small datasets and decision tree for the larger datasets which provided satisfactory classification results.

Ramaswami and Bhaskaran (2010) presented the Chi-squared Automatic Interaction Detector prediction model which was utilized to analyze the interrelation between variables that were used to predict the performance at higher secondary school education. Their comparative study revealed that features such as medium of instruction, marks in written assignments and test, location of school, living area, and type of secondary education were the strongest

indicators. The Chi-squared Automatic Interaction Detector prediction model of student performance was constructed with seven-class predictor variables.

Ramesh, Parkav, and Rama (2013) tried to identify the factors influencing the students' performance in final examinations, based on a dataset including questionnaire data and performance details collected. Their motivation is based on identifying the essential predictive variables which affect the performance of higher secondary students, determine the best classification algorithm, and predict the grade at higher examinations. Their study showed that the type of school (coed or boys or girls) does did influence the students' performance but parent's occupation and possibly financial status play a major role. Moreover, their numerical experiments revealed that the multilayer perceptron exhibited the best classification accuracy.

In more recent works, Kostopoulos et al. (2015a, 2015b) examined the effectiveness of semisupervised methods for predicting students' performance in distance higher education. Several experiments were conducted using a variety of semisupervised learning algorithms compared with well-known supervised methods which revealed some very promising results, especially the self-training and the tri-training algorithm.

Livieris et al. (2016) presented a user-friendly decision-support software for predicting the students' performance, together with a case study concerning the final examinations in the course of ''Mathematics.'' Their proposed software is based on a hybrid predicting system incorporating a number of possible learning methods which achieves better performance than any examined single learning algorithm. Based on their preliminary results, the authors concluded that the application of data mining can gain significant insights into student progress and performance.

Elbadrawy et al. (2016) proposed a recommendation system based on personalized analysis to predict the students' performance. Their numerical experiments revealed that the proposed method based on multiple regression and improved matrix decomposition is capable of successfully forecasting student performance in a timely and accurate manner outperforming traditional methods.

A Review of Semisupervised Machine Learning Techniques

The growing research and developments in computer science and information technology contributes to the exponential generation of data in size, dimension, and complexity. Moreover, these datasets have nonlinear relationship between inputs and outcomes, hindering their analysis and modeling. Thus, machine learning algorithms have been widely used for knowledge extraction from datasets and constitute a significant role in their exploration and analysis. There are mainly three basic and commonly used types of machine learning: supervised, unsupervised, and semisupervised learning.

In supervised learning, the objective is to derive a prediction model or classification function f for predicting the true labels of unseen future data while the training dataset consists only of labeled data. If the output classes are discrete, the function f is called classifier, while, if the output classes are continuous then the function f is called regression function (Mitchell, 1997). In unsupervised learning, the main goal is to organize the unseen (unlabeled) data in order to derive interesting and regular patterns from them. Unsupervised learning can be categorized into three different settings: clustering, novelty detection, and dimensionality reduction (Zhu & Goldberg, 2009).

Semisupervised learning consists of a mixture of supervised and unsupervised learning, aiming to obtain better classification results by exploiting the lack of labeled examples by using unlabeled ones. Vast research efforts have focused on semisupervised algorithms as an alternative to traditional methods of machine learning that depict remarkable performance over labeled data but lack the ability to be applied on large amounts of unlabeled data. The general assumption in this class of algorithms is that data points in a high-density region are likely to have same classes and the decision boundary lies in low-density regions (Zhu, 2006). Therefore, this class of methods has the advantage of reducing the effort of supervision to a minimum, while still preserving competitive recognition performance. More specifically, semisupervised learning methods utilize only a small proportion of the whole amount of labeled data for accomplishing their task. This attribute known as labeled ratio R is defined by

$$R = \frac{\text{Number of labeled instances}}{\text{Number of all instances}}$$

and it is usually provided in percentage values (%). Having chosen the labeled ratio, all the available data split into two different subsets, the labeled (L) and the unlabeled (U) set. The generic representation of the examples included in each of these subsets is respectively defined as follows:

$$\begin{cases} x_L = \{\text{Feature set} \mid \text{Class}\} \\ x_U = \{\text{Feature set} \mid \text{Not known class}\} \end{cases}$$

Depending on the theory of each semisupervised method, these subsets interact in various ways with the labeling of examples in U and their incorporation in L.

In the literature, several semisupervised methods have been proposed so far, based on expectation-maximization algorithms, self-training, cotraining, active learning, tri-training, and graph-based techniques. We refer the reader to the studies by Pise and Kulkarni (2008), Triguero and Garćıa (2015), and Zhu (2006) and the references therein for an overview on semisupervised learning methods. During the last decade, many researchers have applied these methods

in many real-world applications which have stated that the classification accuracy can be significantly improved if a large number of unlabeled data are used together with a small number of labeled data (Chapelle, Scholkopf, & Zien, 2009; Kostopoulos et al., 2015a, 2015b; Levatic et al., 2013; Liu & Yuen, 2011; Sigdel et al., 2014; Triguero, Śaez, Luengo, Garćıa, & Herrera, 2014; Wang & Chen, 2013; Zhu, 2006, 2011).

In this study, we investigate the classification accuracy utilizing two wrapper-based semisupervised learning techniques—self-training and YATSI. Self-training has established as a very popular algorithm due to its simplicity and it has been successfully utilized to tackle difficult real-world problems and is often found to be more efficient and more accurate than other semisupervised algorithms (Kostopoulos et al., 2015a, 2015b; Roli & Marcialis 2006; Rosenberg et al., 2005; Sigdel et al., 2014). YATSI is a wrapper-supervised algorithm which has been recently proposed in the literature and consists of a two-stage classifier that improves its predictive accuracy by making use of the available unlabeled data, providing some promising results (Catal & Diri, 2009; Levatic et al., 2013; Sigdel et al., 2014). Subsequently, we briefly describe the characteristics of both semisupervised classification algorithms.

Self-Training

Self-training is a wrapper-based semisupervised approach which constitutes an iterative procedure of self-labeling unlabeled data. According to Ng and Cardie (2003), ''Self-training is a single-view weakly supervised algorithm'' (p. 2). Initially, an arbitrary classifier is trained with a small amount of labeled data, which have been randomly chosen from the training set. Subsequently, the training set is iteratively augmented gradually using a classifier trained on its own most confident predictions. More specifically, each classified unlabeled instance that has achieved a probability value over a defined threshold c is considered enough reliable to be added to the training set for the following training phases. Finally, these instances are added to the initial training set, increasing in this way its efficiency and robustness. Therefore, the retraining of the classifier is done using the new enlarged training set until stopping criteria are satisfied. A high-level schematic representation of the self-training procedure is presented in Figure 1.

An important reason why performance may fluctuate compared with supervised algorithms' performance is the fact that, during the training phase of the former, some of the unlabeled examples will not get labeled, since the termination of the algorithm will have been preceded (Schwenker & Trentin, 2014). However, since the success of the self-training algorithm is heavily depended on its own predictions, its weakness is that erroneous initial predictions will probably lead the classifier to generate incorrectly labeled data (Zhu & Goldberg, 2009).

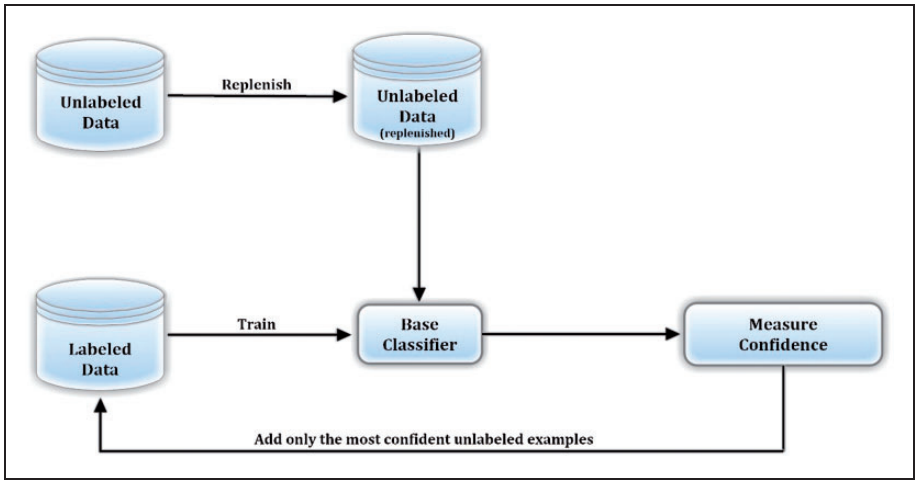

Figure 1. Self-training framework.

Yet Another Two-Stage Idea

The YATSI algorithm (Driessens et al., 2006) is similar to the self-training concept, since it uses its own predictions in the training process and it can be wrapped around any classifier. As the name implies, YATSI classifier uses both labeled and unlabeled data in a two-stage set-up (Figures 2 and 3).

In the first stage, a prediction model is generated on the available training data utilizing a supervised base classifier and subsequently the classifier is used to predict a severity level for each unlabeled instance. The output unlabeled data with predicted severity are called ''prelabeled data.''

In the second stage, these prelabeled examples are then used together with the original training data in a weighted nearest neighbor algorithm. The weights used by the nearest neighbor classifier are meant to limit the amount of trust the algorithm puts in the labels generated by the model from the first step (Driessens et al., 2006).

Dataset

In this study, we have utilized a dataset concerning the performance of 3,716 students in courses of ''Mathematics'' of the first 5 years of secondary school. The data have been collected by the *Microsoft showcase school* ''Avgoulea-Linardatou'' during the years 2007 to 2016. Table 1 presents 12 attributes which characterize each instance in the dataset and are based on several written

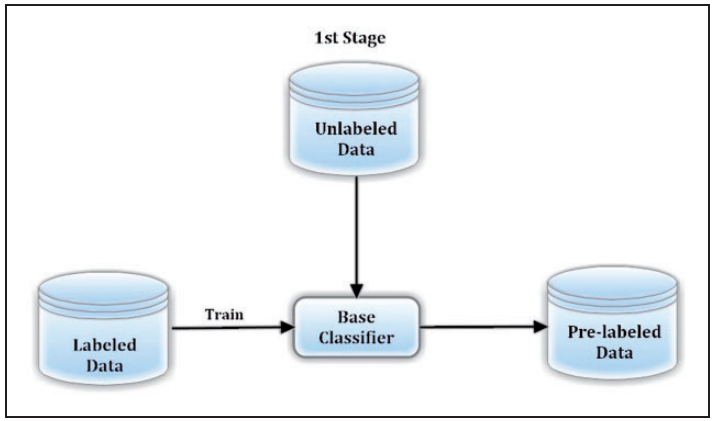

Figure 2. YATSI framework (first stage).

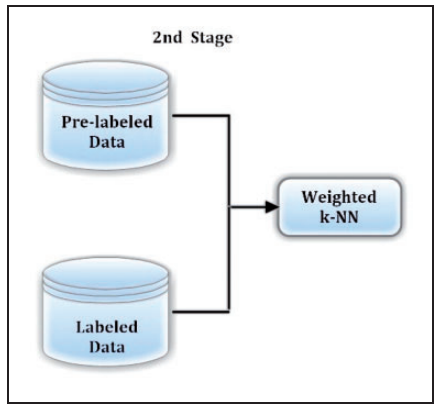

Figure 3. YATSI framework (second stage).

assignments and frequent oral questions which assess students' understanding of important mathematical concepts and topics daily.

The first attribute is related with secondary stage type of each student. More specifically, AG, BG, and CG stands for the first, second, and third class of Gymnasium, respectively, while AL and BL stands for the corresponding first and second class of Lyceum. The next 10 values are time-variant attributes and refer to the students' performance on both the academic semesters, utilizing a 20-point grading scale, where 0 is the lowest grade and 20 is the perfect score. Many

Table 1. Attributes Description.

Attribute	Values
Secondary stage type	{AG, BG, CG, AL, BL}
Oral grade of the first semester	[0, 20]
Grade of the first test of the first semester	[0, 20]
Grade of the second test of the first semester	[0, 20]
Grade of the final examination of the first semester	[0, 20]
Final grade of the first semester	[0, 20]
Oral grade of the second semester	[0, 20]
Grade of the first test of the second semester	[0, 20]
Grade of the second test of the second semester	[0, 20]
Grade of the final examination of the second semester	[0, 20]
Final grade of the second semester	[0, 20]
Grade in the final examinations	"Fail," "Good," "Very good," "Excellent"

related studies have shown that such attributes have a material impact in students' success in the examinations (Cortez & Silva, 2008; Livieris, Drakopoulou, & Pintelas, 2012; Livieris et al., 2016; Ramaswami & Bhaskaran, 2010). The assessment of students during the academic year consists of oral examination, two 15-minute prewarned tests, a 1-hour examination, and the overall semester performance of each student in the first and second semester. The 15-minute tests include multiple-choice questions and short answer problems while the 1-hour examinations include several theoretic and multiple-choice questions, as well as a variety of difficult mathematical problems requiring arithmetic skills, solving techniques, and critical analysis. The overall semester performance of each student addresses the personal engagement of the student in the lesson and his progress. Finally, the students were classified according to their grade in the final examinations (2-hour examinations) utilizing a four-level classification: 0 to 9 (*Fail*), 10 to 14 (*Good*), 15 to 17 (*Very good*), and 18 to 20 (*Excellent*).

Similar to Livieris et al. (2012, 2016), since it is of great importance for an educator to recognize weak students in the middle of the academic period, two datasets have been created based on the attributes presented in Table 1.

- DATA_{A}: It contains the attributes which concern the students' performance of the first semester.
- DATA_{AB}: It contains the attributes which concern the students' performance of the first and second semesters.

Experimental Results

In this section, we evaluate the classification performances of two semisupervised techniques—Self-training and YATSI—with that of the corresponding supervised learning algorithms. Our experiments assume limited labeled data availability, namely, we evaluate the performance of selected classifiers for three different training sizes (10%, 20%, and 30%) of the labeled data.

Table 2. Classification Performance of Supervised and Self-Training Algorithms.

	$DATA_A$			$DATA_{AB}$		
	Training size			Training size		
Classifier	10%	20%	30%	10%	20%	30%
NB	68.5%	67.9%	69.0%	66.7%	66.1%	66.9%
S-NB ($c = 0.8$)	**71.9%**	**72.1%**	71.2%	70.0%	**70.2%**	**70.1%**
S-NB ($c = 0.9$)	71.6%	71.3%	71.7%	**70.1%**	70.1%	70.1%
S-NB ($c = 0.95$)	71.6%	71.2%	**71.8%**	69.8%	70.0%	69.8%
SMO	73.9%	72.0%	75.2%	72.6%	72.4%	75.5%
S-SMO ($c = 0.8$)	**75.4%**	75.1%	**75.6%**	**77.2%**	77.3%	77.2%
S-SMO ($c = 0.9$)	75.1%	74.9%	74.9%	77.1%	**77.5%**	**77.4%**
S-SMO ($c = 0.95$)	74.2%	**75.2%**	75.4%	74.9%	74.4%	76.5%
MLP	70.4%	72.8%	73.4%	70.4%	71.6%	73.4%
S-MLP ($c = 0.8$)	**76.9%**	76.0%	76.9%	**79.4%**	78.8%	79.1%
S-MLP ($c = 0.9$)	74.8%	**76.2%**	76.1%	78.7%	**79.3%**	**79.2%**
S-MLP ($c = 0.95$)	74.8%	74.9%	**77.0%**	78.1%	77.4%	78.8%
C4.5	74.2%	75.0%	74.0%	74.7%	78.1%	76.6%
S-C4.5 ($c = 0.8$)	76.7%	76.8%	**76.9%**	**79.9%**	**79.7%**	79.8%
S-C4.5 ($c = 0.9$)	**76.8%**	**76.9%**	76.8%	78.6%	79.5%	78.0%
S-C4.5 ($c = 0.95$)	76.0%	76.6%	76.2%	78.2%	79.1%	**78.9%**
JRip	73.9%	73.9%	76.4%	71.2%	75.1%	77.2%
S-JRip ($c = 0.8$)	**77.3%**	77.0%	**77.0%**	**80.9%**	80.6%	**80.6%**
S-JRip ($c = 0.9$)	76.4%	**77.4%**	76.7%	80.1%	**80.9%**	79.7%
S-JRip ($c = 0.95$)	76.5%	76.1%	76.0%	79.7%	79.8%	79.5%

Note. NB = naive Bayes; SMO = sequential minimum optimization; MLP = multilayer perceptron; JRip = RIPPER algorithm.

The classification accuracy of all learning algorithms was evaluated utilizing the standard procedure called stratified 10-fold cross-validation and the implementation code was written in JAVA, using WEKA Machine Learning Toolkit (Hall et al., 2009). For each classifier, the largest value is illustrated in boldface to indicate the best classifier for each training size.

Performance Comparison With Self-training

We consider the following five supervised classification techniques—naive Bayes (NB; Domingos & Pazzani, 1997), sequential minimum optimization (SMO; Platt, 1999), C4.5 (Quinlan, 1993), multilayer perceptron (MLP; Rumelhart, Hinton, & Williams, 1986), RIPPER algorithm (JRip; Cohen, 1995), and corresponding Self-trained learning counterparts—Self-training naive Bayes (S-NB), self-training SMO (S-SMO), self-training C4.5 (S-C4.5), self-training MLP (S-MLP), and self-training JRipper (S-JRip). All self-trained classifiers are evaluated for three different values of parameter c, namely 0.8, 0.9, and 0.95, for minimum confidence used to select the prelabeled instances for retraining (Sigdel et al., 2014).

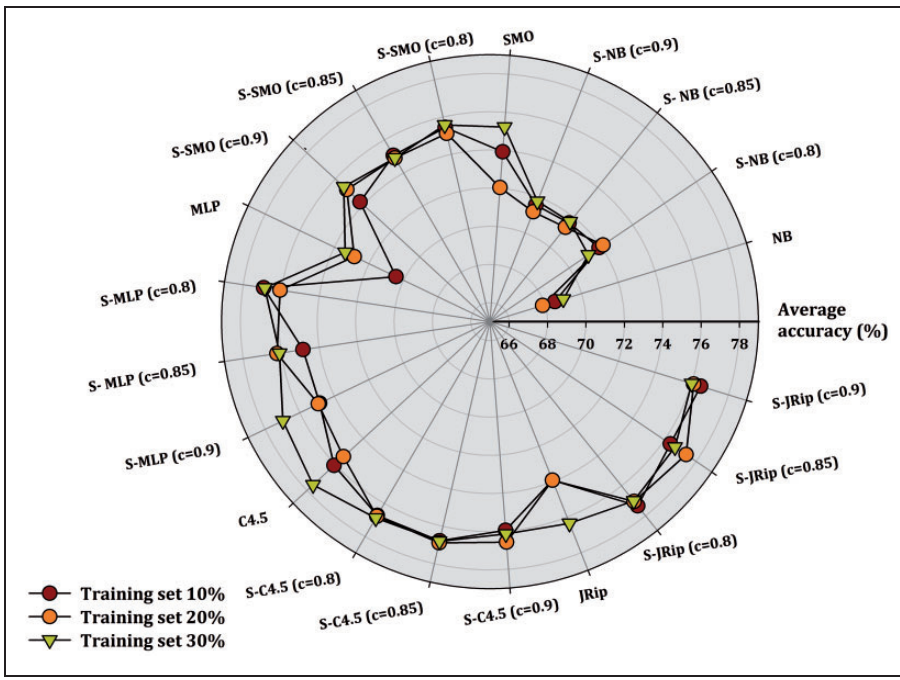

Figure 4. Comparison of accuracy of supervised and self-training algorithms for $DATA_A$.

Table 2 shows the experimental results for both datasets for the supervised classifiers and the corresponding self-training classifiers. In addition, we also provide a more representative visualization of the classification accuracy ability of each algorithm for each training set in Figures 4 and 5. In these figures, we have mapped each different ratio of labeled examples with different color and line format across a radar plot.

Clearly, all classifiers using self-training improve their accuracies with respect to supervised learning, relative to both datasets. For $DATA_A$, the classification performance of SMO and C4.5 was slightly improved from 0.2% to 3.2% with self-training, while for MLP, the accuracy with self-training is improved up to 6.9% over the accuracy of MLP alone. For $DATA_{AB}$, the performance with self-training of NB, SMO, and C4.5 is improved by 4.1% to 5.2, while for MLP and JRip, the classification performance is significantly improved up to 9.4% to 9.7%. Moreover, for most algorithms, the classification accuracy is usually improved for lower value of parameter c as well as the size of the training set increases.

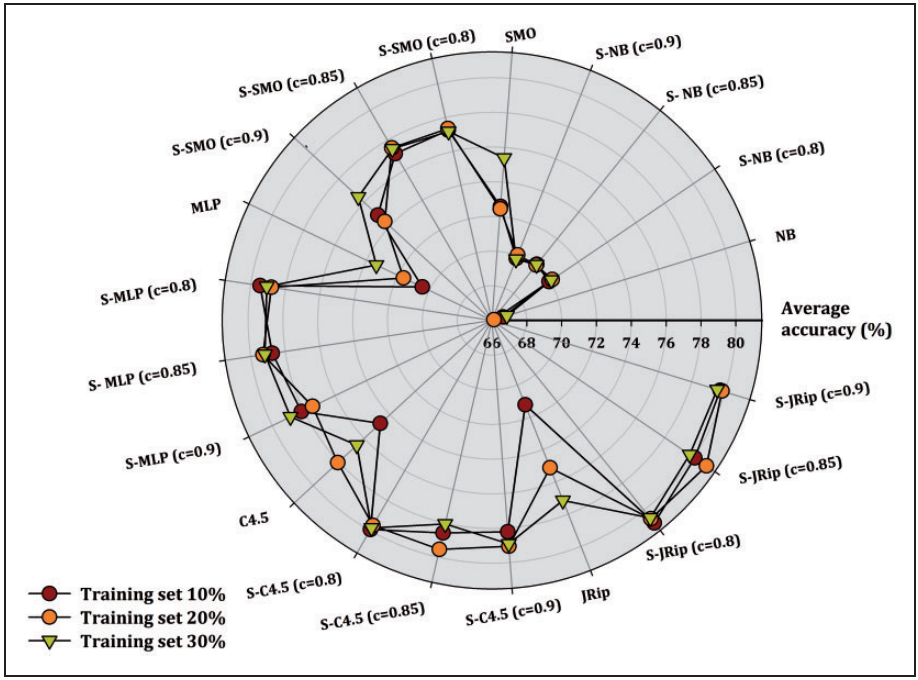

Figure 5. Comparison of accuracy of supervised and self-training algorithms for $DATA_{AB}$.

Performance Comparison With YATSI

Next, we evaluate the performance of the presented five supervised classification techniques with that of the corresponding YATSI learning counterparts—YATSI with naive Bayes (Y-NB), YATSI with SMO (Y-SMO), YATSI with C4.5 (Y-C4.5), YATSI with MLP (Y-MLP), and YATSI with JRipper (Y-JRip). For YATSI classifiers, we test k-nearest neighbors with 1, 5, and 10 neighbors, while the weighting factor for prelabeled data (F) is set to 1 (Sigdel et al., 2014).

Table 3. Classification Performance of Supervised Algorithms and YATSI.

	$DATA_A$			$DATA_{AB}$		
	Training size			Training size		
Classifier	10%	20%	30%	10%	20%	30%
NB	68.5%	67.9%	69.0%	66.7%	66.1%	66.9%
Y-NB ($k = 1$)	72.6%	**72.5%**	**74.5%**	67.6%	**69.4%**	**70.9%**
Y-NB ($k = 5$)	72.6%	71.2%	72.4%	**68.4%**	68.5%	69.7%
Y-NB ($k = 10$)	**73.0%**	71.3%	72.2%	68.1%	67.8%	69.8%
SMO	**73.9%**	72.0%	75.2%	**72.6%**	72.4%	75.5%
Y-SMO ($k = 1$)	71.8%	**75.1%**	**78.0%**	68.6%	**72.7%**	**75.5%**
Y-SMO ($k = 5$)	72.1%	73.5%	74.9%	68.8%	72.2%	73.8%
Y-SMO ($k = 10$)	72.4%	72.8%	74.1%	68.9%	71.2%	73.5%
MLP	70.4%	72.8%	73.4%	70.4%	71.6%	73.4%
Y-MLP ($k = 1$)	**73.5%**	**74.3%**	**77.9%**	69.1%	71.8%	**76.3%**
Y-MLP ($k = 5$)	**73.5%**	72.8%	75.8%	71.0%	71.6%	74.8%
Y-MLP ($k = 10$)	73.4%	72.0%	75.5%	**71.5%**	**72.0%**	74.5%
C4.5	74.2%	75.0%	74.0%	**74.7%**	78.1%	76.6%
Y-C4.5 ($k = 1$)	74.9%	**75.9%**	**78.5%**	70.8%	**80.1%**	**79.3%**
Y-C4.5 ($k = 5$)	**75.0%**	73.9%	75.1%	73.7%	79.7%	76.5%
Y-C4.5 ($k = 10$)	74.8%	73.4%	75.1%	74.0%	79.1%	77.2%
JRip	**73.9%**	73.9%	76.4%	71.2%	**75.1%**	**77.2%**
Y-JRip ($k = 1$)	73.3%	**75.4%**	**76.6%**	69.7%	74.2%	76.5%
Y-JRip ($k = 5$)	73.5%	74.1%	74.4%	**71.7%**	74.0%	74.8%
Y-JRip ($k = 10$)	73.1%	73.0%	74.1%	71.3%	73.8%	74.3%

Note. NB = naive Bayes; SMO = sequential minimum optimization; MLP = multilayer perceptron; JRip = RIPPER algorithm.

Table 3 shows the experimental results for both datasets for the supervised classifiers and the corresponding YATSI classifiers. Moreover, in Figures 6 and 7, we visualize the classification accuracy ability of each algorithm by mapping each different ratio of labeled examples with different color and line format across a radar plot for each training set. All classifiers using YATSI improve their classification performance relative to supervised learning for DATAA. Notice that NB and MLP were the most improved algorithms increasing their accuracy by 2.5% to 5.5%, while JRip was the least benefit with YATSI.

For $DATA_{AB}$, YATSI does not always improve the performance of naive Bayes when unlabeled data are used together with labeled data. More specifically, for 10% training, only NB, MLP, and JRip were benefited from YATSI method, while for 20% and 30% training, all classifiers except JRip improved their performance with YATSI. NB and MLP were the most benefited algorithms with the YATSI whose performance significantly increased by 1.5% to 4.0%. Moreover, Y-C4.5 illustrated the best classification performance regarding both datasets. It is worth noticing that the accuracy of YATSI worsens as k usually increases from 1 to 10.

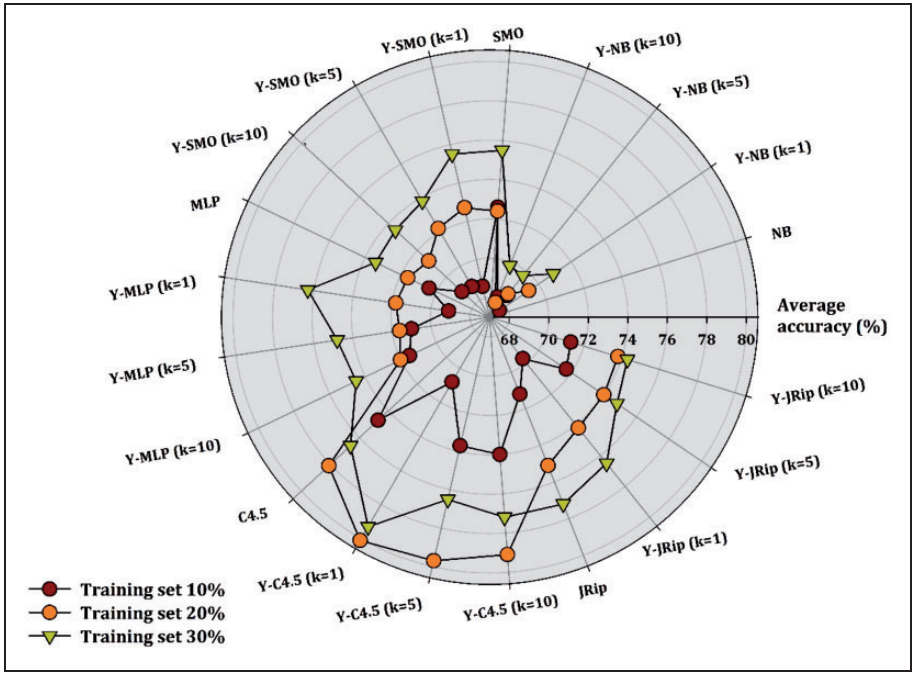

Figure 6. Comparison of accuracy of supervised and YATSI algorithms for $DATA_A$.

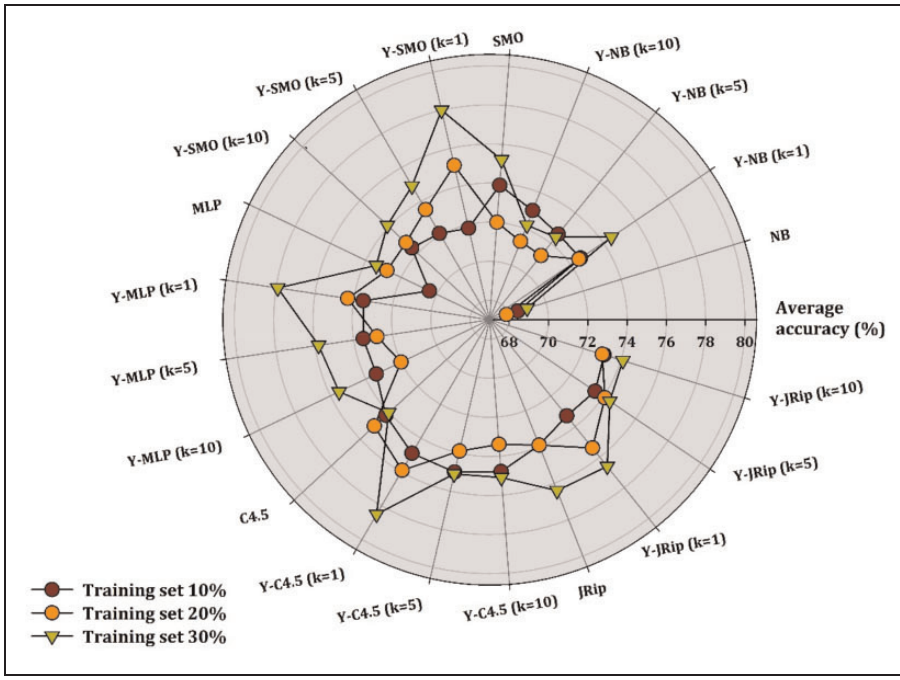

Figure 7. Comparison of accuracy of supervised and YATSI algorithms for $DATA_{AB}$.

To illustrate the classification performance of self-training and YATSI, we compare them with two of the most famous semisupervised learning algorithms, tri-training and cotraining, regarding both datasets (Table 4).

Table 4 presents the evaluation of self-training and YATSI algorithms using the five supervised classification algorithms as base learners with the corresponding tri-training and cotraining counterparts namely—tri-training with naive Bayes (Tri-NB), cotraining with naive Bayes (Co-NB), tri-training with SMO (Tri-SMO), cotraining with SMO (Co-SMO), tri-training with C4.5 (Tri-C4.5), cotraining with C4.5 (Co-C4.5), tri-training with MLP (Tri-MLP), cotraining with MLP (Co-MLP), tri-training with JRipper (Tri-JRip), and cotraining with JRipper (Co-JRip). Clearly, self-training reported the best classification accuracy utilizing 10% and 20% as labeled data ratio, relative to all base learners. Moreover, YATSI presented the highest classification results using 30% as labeled ratio, followed by self-training.

Subsequently, we evaluate the performance of the supervised learning algorithms utilizing 100% of the training set with the corresponding semisupervised algorithms (self-training and YATSI) which archived the best classification accuracy regarding both datasets.

Table 4. Classification Performance of Semisupervised Algorithms, Tri-Training, Cotraining, Self-Training, and YATSI.

Classifier	$DATA_A$ Training size			$DATA_{AB}$ Training size		
	10%	20%	30%	10%	20%	30%
Tri-NB	71.7%	71.8%	71.9%	69.9%	69.9%	70.0%
Co-NB	71.8%	71.8%	71.3%	**70.1%**	69.8%	69.8%
S-NB	71.9%	72.1%	71.8%	**70.1%**	**70.2%**	70.1%
Y-NB	**73.0%**	**72.5%**	**74.5%**	68.4%	69.4%	**70.9%**
Tri-SMO	75.2%	75.1%	75.3%	**77.2%**	77.2%	**77.4%**
Co-SMO	75.0%	75.0%	75.2%	75.5%	75.0%	75.7%
S-SMO	**75.5%**	**75.2%**	75.6%	**77.2%**	**77.5%**	**77.4%**
Y-SMO	73.9%	75.1%	**78.0%**	72.6%	72.7%	75.5%
Tri-MLP	**77.7%**	**77.6%**	77.7%	**80.4%**	**80.4%**	**80.3%**
Co-MLP	76.6%	76.6%	76.4%	77.6%	79.0%	78.5%
S-MLP	76.9%	76.2%	77.0%	79.4%	79.3%	79.2%
Y-MLP	73.5%	74.3%	**77.9%**	71.5%	72.0%	76.3%
Tri-C4.5	**77.0%**	**77.1%**	76.9%	79.8%	79.8%	79.1%
Co-C4.5	75.8%	75.8%	75.2%	77.3%	77.5%	78.5%
S-C4.5	76.8%	76.9%	76.9%	**79.9%**	79.7%	78.9%
Y-C4.5	75.0%	75.9%	**78.5%**	74.7%	**80.1%**	**79.3%**
Tri-JRip	76.3%	76.3%	75.2%	79.3%	79.3%	78.8%
Co-JRip	75.9%	75.9%	74.6%	76.7%	77.7%	77.3%
S-JRip	**77.3%**	**77.4%**	**77.0%**	**80.9%**	**80.9%**	**80.6%**
Y-JRip	73.9%	75.4%	76.6%	71.7%	75.1%	77.2%

Note. Tri-NB = tri-training with naive Bayes; Co-NB = cotraining with naive Bayes; Tri-SMO = tri-training with SMO; Co-SMO = cotraining with SMO; Tri-C4.5 = tri-training with C4.5; Co-C4.5 = cotraining with C4.5; Tri-MLP = tri-training with MLP; Co-MLP = cotraining with MLP; Tri-JRip = tri-training with JRipper; Co-JRip = cotraining with JRipper.

The interpretation of Table 5 indicates that the best self-trained and YATSI classifiers are all competitive and, in most cases, outperform the performance of the corresponding supervised ones which used 100% of the training data. YATSI method reported the highest classification performance for $DATA_A$, with Y-C4.5 reporting the best accuracy of corrected classified instances. Nevertheless, for $DATA_{AB}$, self-trained method reported the best classification results for three out of five classifiers with S-JRip illustrating the highest performance.

Table 5. Classification Performance of Supervised Learning Algorithms Utilizing 100% of the Training Set With the Best Self-Training and YATSI Classifiers.

Classifier	$DATA_A$	$DATA_{AB}$
NB	71.6%	69.9%
S-NB	72.1%	70.2%
Y-NB	**72.6%**	**70.9%**
SMO	75.8%	**77.5%**
S-SMO	75.6%	**77.5%**
Y-SMO	**78.0%**	75.5%
MLP	75.9%	78.0%
S-MLP	77.0%	**79.9%**
Y-MLP	**77.9%**	76.3%
C4.5	77.1%	79.0%
S-C4.5	76.9%	79.9%
Y-C4.5	**78.5%**	**80.1%**
JRip	75.8%	78.7%
S-JRip	**77.4%**	**80.9%**
Y-JRip	76.6%	76.5%

Note. NB = naive Bayes; SMO = sequential minimum optimization; MLP = multilayer perceptron; JRip = RIPPER algorithm.

Conclusions

In this work, we examined the effectiveness of two wrapper methods for semi-supervised learning for the prediction of high-school students' performance in the final examinations. More specifically, we evaluated the performances of self-training and YATSI approaches with the most classic supervised methods and with two of the most popular and frequently used semisupervised algorithms. The selected attributes related to written assignments, oral examinations, short tests, and examinations during the academic year are marked according to specific assessment criteria and are utilized to evaluate the final grade in examinations using semisupervised learning methods with a considerable accuracy, as reflected from the experimental results. In conclusion, we point out that semisupervised algorithms can improve the classification accuracy utilizing a few labeled and many unlabeled data for developing reliable prediction models.

Furthermore, it is worth mentioning that the students' attributes utilized in our work do not constitute a conclusive list. An extension can introduce new attributes and other criteria which were not in the current database, but are collectable by tutors and may potentially influence the performance and the quality of the prediction of student's performance, that is, students' social and cultural characteristics, more tests, and more projects. Moreover, all students' learning interactions with the educational system can be analyzed and parameters such as learning activities tried, activities correctly completed, and the time spent to complete each one could also be taken into account. However, the questions of which attributes should be utilized or which have higher impact on the predicted outcome is still under consideration by many researchers (see Baker & Yacef, 2009; Romero & Ventura, 2007, 2010; Romero et al., 2010, and the references there in). Probably, the research to answer these questions is very likely to reveal additional and crucial information about why and how some variables affect student performance.

Since the experimental results are quite encouraging, similar to Livieris et al. (2016; Livieris, Drakopoulou, Kotsilieris, Tampakas, & Pintelas, 2017), a next step could be the development of a decision-support tool based on a semisupervised learning algorithm, concerning the prediction of the students' performance in the final examinations of a school year. Through the use of a predictive tool, the educators are able to forecast students' success in a course and identify those at risk and the students with learning difficulties in order to offer customize assistance according to the students' needs. Accurate prediction of student's success is one way to reach the highest level of quality in an education system. Moreover, another direction for a future research would be to enlarge our experiments with more schools and school years and collect data from all 6 years of secondary education and apply our methodology for predicting the students' performance at Panhellenic (national)-level examinations for admission to universities.

Acknowledgments

The authors are grateful to the Microsoft showcase school ''Avgoulea-Linardatou'' for the collection of the data used in our study, for the evaluation of the tool, and for their kind and valuable comments, which essentially improved our work.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Acharya, A., & Sinha, D. (2016). An intelligent web-based system for diagnosing student learning problems using concept maps. *Journal of Educational Computing Research*, *55*, 323–345.

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review future visions. *Journal of Educational Data Mining*, *1*(1), 3–17.

Catal, C., & Diri, B. (2009). Unlabelled extra data do not always mean extra performance for semi-supervised fault prediction. *Expert Systems*, *26*(5), 458–471.

Chapelle, O., Scholkopf, B., & Zien, A. (2009). Semi-supervised learning. *IEEE Transactions on Neural Networks*, *20*(3), 542–542.

Chen, G., Liu, C., Ou, K., & Liu, B. (2000). Discovering decision knowledge from web log portfolio for managing classroom processes by applying decision tree and data cube technology. *Journal of Educational Computing Research*, *23*(3), 305–332.

Cohen, W. (1995). Fast effective rule induction. *Proceedings of the 12th International Conference on Machine Learning, Tahoe City, CA*, 115–123.

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. *Proceedings of 5th Annual Future Business Technology Conference, Porto, Portugal*, 5–12.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*, 103–130.

Driessens, K. Reutemann, P. Pfahringer, B., & Leschi, C. (2006). Using weighted nearest neighbor to benefit from unlabeled data. *Proceedings of the 10th Pacific-Asia conference on knowledge discovery and data mining* (pp. 60–69). Heidelberg, Germany: Springer.

Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G., & Rangwala, H. (2016). Predicting student performance using personalized analytics. *Computer*, *49*(4), 61–69.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: An update. *SIGKDD Explorations Newsletters*, *11*, 10–18.

Knauf, R., Sakurai, Y., Tsuruta, S., & Jantke, K. (2010). Modeling didactic knowledge by storyboarding. *Journal of Educational Computing Research*, *42*(4), 355–383.

Koedinger, K., McLaughlin, E., & Heffernan, N. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational Computing Research*, *43*(4), 489–510.

Kostopoulos, G. Kotsiantis, S., & Pintelas, P. (2015a). Estimating student dropout in distance higher education using semi-supervised techniques. *Proceedings of the 19th panhellenic conference on informatics* (pp. 38–43). New York, NY: ACM.

Kostopoulos, G. Kotsiantis, S., & Pintelas, P. (2015b). Predicting student performance in distance higher education using semi-supervised techniques. *Proceedings of the 5th international conference on model and data engineering* (pp. 259–270). New York, NY: Springer.

Levatic, J., Dzeroski, S., Supek, F., & Smuc, T. (2013). Semi-supervised learning for quantitative structure-activity modeling. *Informatica*, *37*(2), 173.

Liu, C., & Yuen, P. (2011). A boosted co-training algorithm for human action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *21*(9), 1203–1213.

Livieris, I. Drakopoulou, K. Kotsilieris, T. Tampakas, V., & Pintelas, P. (2017). DSS-PSP—A decision support tool for evaluating students' performance. *International conference on engineering applications of neural networks* (pp. 63–74). Cham, Switzerland: Springer.

Livieris, I. Drakopoulou, K., & Pintelas, P. (2012). Predicting students' performance using artificial neural networks. *Proceedings of Information and Communication Technologies in Education*, 321–328.

Livieris, I., Mikropoulos, T., & Pintelas, P. (2016). A decision support system for predicting students' performance. *Themes in Science and Technology Education*, *9*, 43–57.

Mitchell, T. (1997). *Machine learning*. New York, NY: McGraw Hill.

Ng, V., & Cardie, C. (2003). Weakly supervised natural language learning without redundant views. *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology, 1*, 94–101. Stroudsburg, PA: Association for Computational Linguistics.

Oladokun, V., Adebanjo, A., & Charles-Owaba, O. (2008). Predicting students' academic performance using artificial neural network: A case study of an engineering course. *The Pacific Journal of Science and Technology*, *9*(1), 72–79.

Pise, N., & Kulkarni, P. (2008). A survey of semi-supervised learning methods. *Proceedings of the 2008 International Conference on Computational Intelligence and Security, 2*, 30–34. Washington, DC: IEEE Computer Society.

Platt, J. (1999). Using sparseness and analytic QP to speed training of support vector machines. In Kearns, M., Solla, S. & Cohn, D. (Eds.), *Advances in neural information processing systems* (pp. 557–563). Cambridge, MA: MIT Press.

Quinlan, J. (1993). *C4.5: Programs for machine learning*. San Francisco, CA: Morgan Kaufmann.

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. *International Journal of Computer Science Issues*, *7*(1), 135–146.

Ramesh, V., Parkav, P., & Rama, K. (2013). Predicting student performance: A statistical and data mining. *International Journal of Computer Applications*, *63*(8), 35–39.

Roli, F., & Marcialis, G. (2006). Semi-supervised PCA-based face recognition using self-training. *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, 560–568. Heidelberg, Germany: Springer.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems With Applications*, *33*, 135–146.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE on Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, *40*(6), 601–618.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (2010). Handbook of educational data mining. In *Data mining and Knowledge series*. Boca Raton, FL: Chapman & Hall/CRC press.

Rosenberg, C. Hebert, M., & Schneiderman, H. (2005). *Semi-supervised self-training of object detection models*. Paper published in Workshop on Application of Computer Vision, Breckenridge, CO.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In Rumelhart, D. & McClelland, J. (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 318–362). Cambridge, MA: MIT Press.

Schwenker, F., & Trentin, E. (2014). Pattern classification and clustering: A review of partially supervised learning approaches. *Pattern Recognition Letters*, *37*, 4–14.

Sigdel, M. Dinç, I. Dinç, S. Sigdel, M. Pusey, M., & Aygun, R. (2014). Evaluation of semi-supervised learning for classification of protein crystallization imagery. *Conference on Southeastcon 2014, IEEE, Lexington, KY*, 1–6. Washington, DC: IEEE.

Thai-Nghe, N. Busche, A., & Schmidt-Thieme, L. (2009). Improving academic performance prediction by dealing with class imbalance. *9th International Conference on Intelligent Systems Design and Applications (ISDA'09)*, 878–883.Washington, DC: IEEE.

Thai-Nghe, N. Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. *Proceeding of 37th IEEE Frontiers in Education Conference* (pp. 7–12). Milwaukee, WI: IEEE.

Triguero, I., & Garćıa, S. (2015). Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study. *Knowledge and Information Systems*, *42*(2), 245–284.

Triguero, I., Śaez, J., Luengo, J., Garćıa, S., & Herrera, F. (2014). On the characterization of noise filters for self-training semi-supervised in nearest neighbor classification. *Neurocomputing*, *132*, 30–41.

Tsai, Y., Ouyang, C., & Chang, Y. (2016). Identifying engineering students' English sentence reading comprehension errors. *Journal of Educational Computing Research*, *54*(1), 62–84.

Wang, Y., & Chen, S. (2013). Safety-aware semi-supervised classification. *IEEE Transactions on Neural Networks and Learning Systems*, *24*(11), 1763–1772.

Zhu, X. (2006). *Semi-supervised learning literature survey*. Madison, WI: University of Wisconsin–Madison (Technical Report 1530).

Zhu, X. (2011). Semi-supervised learning. In Sammut, C. & Webb, G. (Eds.), *Encyclopedia of machine learning* (pp. 892–897). Berlin, Germany: Springer.

Zhu, X., & Goldberg, A. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *3*(1), 1–130.

Author Biographies

Ioannis E. Livieris received his BSc, MSc, and PhD degrees in Mathematics from the University of Patras, Greece in 2006, 2008, and 2012, respectively. He is currently a lecturer in Technological Educational Institute of Western Greece. His research interests include numerical optimization, neural networks, data mining, and machine learning algorithms and its applications.

Konstantina Drakopoulou received her BSc and MSc in 2012 and 2014, respectively. Currently, she is a PhD student and an educator in the Microsoft showcase

school ''Avgoulea-Linardatou.'' Her research is performed as a member of the Educational Software Development Laboratory.

Vassilis T. Tampakas received his diploma from the Computer Engineering and Informatics Department, Patras University and his PhD on distributed computing and algorithms. He worked as a senior researcher in the Computer Technology Institute. In 1995, he joined Technological Educational Institute of Patras as a faculty member. He is currently a professor in the Department of Computer and Informatics Engineering, Technological Educational Institute of Western Greece. He is also the director of DISK Lab (Distributed/Parallel & Knowledge Management Systems Laboratory) at the same Institution. His research interests include distributed systems and algorithms, mobile networks, parallel and distributed information retrieval, data mining, and machine learning techniques or algorithms. He has extensively published in major international journals and conferences.

Tassos A. Mikropoulos is a professor of the Department of Primary Education and the director of the ''Educational Approaches to Virtual Reality Technologies laboratory'' at the University of Ioannina, Greece (www.earthlab.uoi.gr). Tassos Mikropoulos is the elected chair of the Hellenic Association of ICT in Education. His research interests are on educational technology, virtual reality, and educational neuroscience. His work has been published in many referred journals and conferences with more than 1,000 citations. He is a member of the editorial board and reviewer for many international journals. Professor Mikropoulos has been project director, principal investigator, and consultant in numerous research and development and educational projects.

Panagiotis Pintelas is a professor of Computer Science with the Informatics Division of Department of Mathematics at Patras University, Greece. His research interests include software engineering, AI and ICT in education, machine learning, and data mining. He was involved in or directed several dozens of National and European research and development projects.