A Prognosis of Junior High School Students' Performance Based on Active Learning Methods

Georgios Kostopoulos[1], Sotiris Kotsiantis[1] and Vassilios S. Verykios[2]

[1]Educational Software Development Laboratory (ESDLab)
Department of Mathematics, University of Patras, Greece
[2]Hellenic Open University, Greece
kostg@sch.gr,sotos@math.upatras.gr,verykios@eap.gr

Abstract. In recent years, there is a growing research interest in applying data mining techniques in education. Educational Data Mining has become an efficient tool for teachers and educational institutions trying to effectively analyze the academic behavior of students and predict their progress and performance. The main objective of this study is to classify junior high school students' performance in the final examinations of the "Geography" module in a set of five pre-defined classes using active learning. The exploitation of a small set of labeled examples together with a large set of unlabeled ones to build efficient classifiers is the key point of the active learning framework. To the best of our knowledge, no study exist dealing with the implementation of active learning methods for predicting students' performance. Several assessment attributes related to students' grades in homework assignments, oral assessment, short tests and semester exams constitute the dataset, while a number of experiments are carried out demonstrating the advantage of active learning compared to familiar supervised methods, such as the Naïve Bayes classifier.

Keywords: Pool-based active learning, uncertainty sampling strategy, prediction, student performance, junior high school.

1 Introduction

Over recent years, there is a growing research interest in applying data mining techniques in education. Educational Data Mining (EDM) has become an efficient tool for teachers and educational institutions exploiting data stored in databases. Using a wide variety of machine learning methods and tools, EDM is trying to effectively analyze the academic behavior of students based on several characteristics and predict their progress, performance or dropout rates [18]. Recently, Slater et al. [23] reviewed 40 frequently used tools for data mining in education. Therefore, it is essential for educational institutions to support weak students and increase retention rates targeting to improve learning effectiveness and provide high quality education.

The main objective of this study is to predict junior high school students' performance in the final examinations of the "Geography" module using active learning. The exploitation of a small set of labeled examples together with a large set of unla-

C. Frasson and G. Kostopoulos (Eds.): BFAL 2017, LNAI 10512, pp. 67–76, 2017
DOI: 10.1007/978-3-319-67615-9_6

beled ones to build efficient classifiers is the key point of the active learning framework. The students' examination grade has been classified into five pre-defined classes and is based on several quantitative assessment attributes, such as homework assignments, oral performance, short tests and semester exams that have been performed during the academic year. Moreover, we investigate the possibility to identify low performance students in a timely manner quite accurately. The accurate prediction of strengths and weaknesses of such students is beneficial for teachers, as well as for educational institutions. Students that are possible to fail in the final examinations need extra help and learning support. Well planned assignments and activities, additional learning material and supplementary lessons adapted to the different needs and knowledge levels of students may motivate them and enhance their performance. To the best of our knowledge there is no study dealing with the implementation of active learning methods in the educational field.

The rest of this paper is organized as follows: In Section 2 we present recent studies of machine learning technics for predicting students' performance in high school and especially familiar supervised methods. In Section 3 we briefly describe the active learning task. A description of the dataset is given in section 4 together with a detailed analysis of data attributes used. In section 5 we analyze the experiments carried out in this study and present their results while making a comparison to familiar supervised methods, such as the Naïve Bayes classifier. Finally, in Section 6 we conclude writing down some thoughts for future work.

2 A Recent Review of Data Mining Applications in Education

A number of very rewarding studies have been carried out in recent years, dealing with the implementation of familiar machine learning technics to evaluate high school students' performance in secondary education. Moreover, what all these studies have in common is the prediction (pass or fail) in the "Mathematics" module, confirming its importance for students, teachers and educational institutions. Some of these studies are analyzed below:

Cortez and Silva [2] parsed data originating from two secondary schools to predict students' performance (pass or fail) in "Mathematics" and "Portuguese language" modules in the final examinations at the end of academic year. Four familiar data mining methodologies, particularly Decision Trees, Random Forest, Neural Networks and Support Vector Machines (SVM), were tested in several demographic, social and school attributes showing a high predictive accuracy, especially in the case where the past school period grades were known.

Márquez-Vera et al. [13] implemented a Genetic Programming (GP) algorithm and recommended different classification technics from Weka [26] to predict high school students' failure in secondary education. More precisely, five rule induction algorithms, five decision tree algorithms and an evolutionary GP algorithm, named Interpretable Classification Rule Mining (ICRM), were used for successfully predicting students' final performance (pass or fail).

Stapel et al. [24] developed an online math learning platform in Germany with more than 100k interactive exercises grouped in series and arranged in digital books. The platform supports and guides students without teacher intervention, while teachers can be aware of students' performance through detailed reports. Data were collected during an academic year (2015) and included information related to students' activities in the platform 40 days before their first assessment attempt. Finally, an ensemble of classifiers was used to predict students' performance on specific learning objectives.

Livieris et al. [11] presented a user-friendly decision support tool for predicting high school students' performance in the final examinations of the "Mathematics" module. The tool incorporates familiar supervised algorithms from Weka and is based on several time-variant quantitative variables of students, such as written assignments, oral performance, short tests and exams. The notable results of the proposed tool show that it may be effectively used for the early identification of low performance students in high school.

In a recent study, Kostopoulos et al. [8] examined the effectiveness of semi-supervised methods to predict students' performance in distance higher education. Familiar algorithms from KEEL [25] (Self-training, Co-training, Tri-training, De-Tri-training, Democratic, RASCO and Rel-RASCO) were tested using several base classifiers. The experimental results were promising compared to familiar supervised methods, showing the predominance of the Tri-training algorithm with an accuracy measure exceeding 80% in the performance prediction (pass or fail) of students in the final examinations of a distance undergraduate course.

3 Active Learning

In many real world applications, there is often a lack of labeled data while unlabeled data can easily be obtained. Labeling unlabeled data may be a difficult and time consuming affair, as it requires a lot of human effort and experts. The need to utilize the hidden information in unlabeled data in order to build accurate predictive learning models has resulted into the development of significant machine learning approaches. Semi-supervised learning (SSL) and active learning are the two representative paradigms for learning from both labeled and unlabeled data [28].

In active or query learning the learning algorithm chooses the data from which it learns querying the labels of unlabeled examples from an oracle. These examples are usually the most informative ones. The essential aim of active learning is to minimize the number of queries posed, using a small number of labeled examples and build efficient predictive models at the lowest possible cost [3, 19]. Several scenarios have been proposed to ask queries such as pool-based sampling, stream-based sampling and membership queries synthesis [20]. In pool-based sampling, we consider that there are a small set L of labeled examples and a large set U of unlabeled ones. The best queries are selected from U, these examples are added to L and the iterative procedure is repeated until U is empty or a stopping criterion is met [16]. In stream-based or selective sampling an unlabeled example is picked for labeling by the oracle, and the active learner decides whether to query it or not, while in membership query synthesis the learner may request the label for any unlabeled example posing a query de novo.

A number of frequently used strategies [22] have been applied to evaluate the informativeness and representativeness of unlabeled examples and request their labels, such as uncertainty sampling, random sampling and query by committee (QBC) [12]. Uncertainty sampling is one of the most effective and simplest strategies querying the labels of the most uncertain to label examples. For binary classification margin sampling query is the most common strategy [21], while entropy sampling is used for classification problems with more than two class labels. Random sampling is another simple strategy in accordance with which, unlabeled examples are randomly selected from *U*. It is noteworthy that many studies have shown that random sampling often prevails over uncertainty sampling [16]. QBC queries the label of the most informative example, which is the one that a committee of classifiers disagree the most.

The pseudo-code description of a pool-based active learning scenario using uncertainty as sampling strategy is shown in Algorithm 1:

Algorithm 1. Pool-based Active Learning with Uncertainty Sampling

```
Input: labeled dataset $L$, unlabeled dataset $U$.
1. Initially, apply base learner $B$ to the training dataset $L$ to obtain classifier $C$.
2. Apply $C$ to unlabeled dataset $U$.
3. From $U$, select $m$ instances for which $C$ is most uncertain.
4. Ask the teacher for labels of the m instances.
5. Add the $m$ labeled instances to $L$.
6. Re-train on $L$ to obtain a new classifier, $C'$.
7. Repeat steps 2 to 6, until $U$ is empty or until some stopping criterion is met
8. Output a classifier that is trained on $L$.
```

4 Dataset Description

The present study is focused on the following two questions:

1. How do active learning techniques perform for predicting students' final performance in junior high school?
2. Can we predict the students that are going to fail or pass the final examinations in a good time?

The dataset used in our study was provided by a junior high school in Greece. For a time period of three years (2007-2010), data of 307 students (12-13 years) have been collected concerning the "Geography" module. Each instance in the dataset corresponds to an individual student and is characterized by the values of 15 performance attributes (Table 1). More specifically:

The assessment of students during the first semester consists of seven attributes: two 15-minute pre-warned tests, two oral examinations, several homework assignments, a 1-hour exam and the semesters' grade. The 15-minute tests (TEST_A1, TEST_A2) include short answer problems, while the 1-hour exam (EXAM_A) covers a wide range of the curricula. Several homework assignments and oral questions assess students' understanding of important concepts and topics in geography daily in the semester (HW_A, ORAL_A1 and ORAL_A2). Finally, the overall semester performance of each student corresponds to attribute GRADE_A. In the same way, the assessment of students during the second and third semester consists of five and three attributes respectively. The output nominal attribute "EXAMS" corresponds to the students' grade in the final examinations (2-hour exam) according to the following five-level classification: 0-9 (insufficient), 10-12 (poor), 13-14, (good), 15-17 (very good), 18-20 (excellent).

Table 1. Attributes Description

Attribute	Values	Description
TEST_A1	[1, 10]	1st semester's test1 grade
TEST_A2	[1, 10]	1st semester's test2 grade
EXAM_A	0-9, 10-12 ,13-14, 15-17, 18-20	1st semester's exam grade
HW_A	[0, 5]	1st semester's homework grade
ORAL_A1	[1, 10]	1st semester's oral1 grade
ORAL_A2	[1, 10]	1st semester's oral2 grade
GRADE_A	[1, 20]	1st semester's overall grade
TEST_B	[1, 10]	2nd semester's test1 grade
EXAM_B	0-9, 10-12 ,13-14, 15-17, 18-20	2nd semester's exam grade
HW_B	[0, 5]	2nd semester's homework
ORAL_B	[1, 10]	2nd semester's oral grade
GRADE_B	[1, 20]	2nd semester's overall grade
TEST_C	[1, 10]	3rd semester's test1 grade
ORAL_C	[1, 10]	3rd semester's oral grade
GRADE_C	[1, 20]	3rd semester's overall grade
EXAMS	0-9, 10-12 ,13-14, 15-17, 18-20	Grade in final examinations

In the following section we present the experiments that take place in our study and the experimental results.

5 Experiments

Initially, the dataset was partitioned into 10 folds of 276 instances using the 10-fold cross validation procedure. One fold was kept for assessing the predictive effectiveness of the model, while the rest 9 folds were used for the training process. In each fold, 21 instances of the training set formed the labeled set and the rest 255 formed the unlabeled set. A pool-based sampling scenario with the margin sampling query strategy was used. We have defined a maximum number of 30 iterations as a stopping criterion, while we selected a single example for labeling at each of the iterations. On this basis, at the end of the learning process there will be 51 labeled instances.

A set of five familiar supervised algorithms from Weka were used as base classifiers forming five respective active learners. These algorithms are:

- Bayes Net, representative of the Bayesian Networks [14]
- Multilayer Perceptrons (MLPs), representative of Neural Networks [6]
- Naïve Bayes [7], a very effective and simple classification algorithm [27]
- Random Forest (RF), a combination of tree-structured predictors [1]
- Sequential Minimal Optimization (SMO), a very effective SVM algorithm [15]

The experiments were conducted in two distinct phases of three sequential steps each time using the JCLAL tool. JCLAL is a computational tool for performing active learning methods for both researchers and programmers. JCLAL provides a friendly and high-level environment facilitating the implementation of existing active learning methodologies or the development of new ones [17]. It supports pool-based and stream-based sampling scenarios, as well as a variety of single-label and multi-label active learning scenarios, such as uncertainty and query by committee sampling. In each phase, the first step consists of the seven attributes (TEST_A1, TEST_A2, EXAM_A, ORAL_A1, ORAL_A2, HW_A, GRADE_A) referred to the assessment of a student during the first semester. The second step includes the attributes of both first and second semesters, while in the third step, all attributes are used.

Table 2. The Accuracy (%) of the Active Learners with the Margin Sampling Query Strategy

	Active Learner	**1st step**	**2nd step**	**3rd step**
Base algorithm	NB	59.84	66.73	68.99
	RF	61.56	64.08	67.04
	SMO	63.11	62.83	65.38
	MLPs	58.52	63.75	59.81
	Bayes Net	60.19	64.41	65.67

In the first phase of the experiments we evaluate the performance of the above mentioned active learners measuring the accuracy, which corresponds to the percentage of the correctly classified instances (Table 2). As Table 2 shows, the accuracy measure of the active learners ranges from 58.52% to 63.11% based on the attributes regarding the first semester's assessment. In the second step, the NB based active learner outweighs with 66.73% accuracy, while in the 3rd step the accuracy is 68.99%.

We evaluate the performance of the above active learners using the Friedman Aligned Ranks nonparametric test [5]. According to the test results the algorithms are ranking from the best performer to the lower one (Table 3). The NB based active learner prevails over the rest, followed by the SMO and RF based active learners.

Table 3. Friedman Aligned Ranks test (significance level of 0.05)

Algorithm	**Rank**
NB	5.33
RF	6.00
SMO	7.67
Bayes Net	7.67
MLPs	13.33

In the 2nd phase of experiments we make a comparison between the active learner using NB as the base classifier and the NB supervised classifier measuring the accuracy in each one of the experiments steps. The results (Table 4) indicate that active learning outweighs supervised learning in each of the three steps. It must be mentioned that supervised learning requires a large amount of labeled data to train the classifier (276 instances), while only a small amount of labeled data (51 instances) are needed for achieving better accuracy using active learning.

Table 4. Active Learning vs Supervised Learning (accuracy %)

Step	**Active Learner (NB base classifier)**	**NB (supervised)**
1st step	59.84	59.30
2nd step	66.73	61.60
3rd step	68.99	62.90

Regarding the active learner using NB as base classifier, and for each one of the three steps of the experiments we present a graph (Figure 1) illustrating the accuracy percentage rate related to the number of labeled instances during the iterative learning process and measure the area (Table 5) bounded by the accuracy curve and the x-axis (AUC). AUC is another common metric for assessing the performance of a binary classifier [10].

Table 5. Area Under the Accuracy Curve (NB)

Step	**AUC**
1st step	57.132
2nd step	64.614
3rd step	65.387

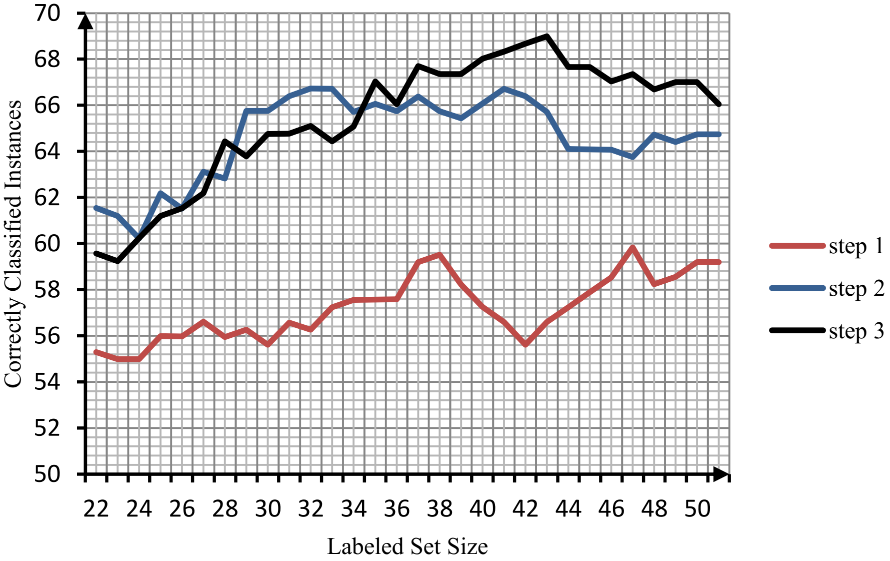

Fig. 1. Accuracy Curve of the NB based active learner

6 Conclusions

In the present study the effectiveness of active learning methodology is examined to predict the performance of junior high school students in the final examinations in the "Geography" module. Specifically, a pool-based sampling scenario was adopted making use of the margin sampling query strategy, while the dataset was based on several quantitative assessment attributes, such as homework assignments, oral performance, short tests and semester exams that have been performed during the academic year. To the best of our knowledge, no study exists, dealing with the implementation of active learning methods for the prediction of students' performance.

Several experiments were conducted measuring the accuracy of five active learners using familiar supervised algorithms as base classifiers. The experimental results indicate that it is possible to predict low performers with sufficient accuracy in a timely manner using a limited number of labeled examples as a training set. At the end of the first semester, the accuracy of the active learner using SMO as base classifier is 63.11%. At the end of the second semester the NB based active learner outweighs with 66.73% accuracy, while accuracy is 68.99% before the final examinations. Comparing the active learner using NB as base classifier and the NB supervised classifier, it is shown that active learning prevails over supervised learning in each one of the three experiment steps. It is worth noting that supervised learning requires a large amount of labeled data to train the classifier (276 instances), while only a small amount of labeled data (51 instances) were needed for achieving better accuracy using active learning.

An interesting aspect is to combine semi-supervised learning and active learning, since both methodologies aim to build effective predictive models exploiting a small labeled dataset together with a large unlabeled dataset. Studies in other domains have shown that such a combination may improve the predictive accuracy [4, 9, 29]. Further experiments and more research is needed using additional attributes (e.g. such as grades from previous years, the time of study, the number of school absences, parents education).

References

1. Breiman, L.: Random forests. Machine learning, vol. 45(1), pp. 5-32 (2001)
2. Cortez, P., & Silva, A. M. G.: Using data mining to predict secondary school student performance (2008)
3. Dasgupta, S.: Two faces of active learning. Theoretical computer science, vol. 412(19), pp. 1767-1781 (2011)
4. Hady, M. F. A., & Schwenker, F.: Combining committee-based semi-supervised learning and active learning. Journal of Computer Science and Technology, vol. 25(4), pp. 681-698 (2010)
5. Hodges, J. L., & Lehmann, E. L.: Rank methods for combination of independent experiments in analysis of variance. The Annals of Mathematical Statistics, vol. 33(2), pp. 482-497 (1962)
6. Hornik, K., Stinchcombe, M., & White, H.: Multilayer feedforward networks are universal approximators. Neural networks, vol. 2(5), pp. 359-366 (1989)
7. John, G. H., & Langley, P.: Estimating continuous distributions in Bayesian classifiers. In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, pp. 338-345. Morgan Kaufmann Publishers Inc. (1995)
8. Kostopoulos, G., Kotsiantis, S., & Pintelas, P.: Predicting Student Performance in Distance Higher Education Using Semi-supervised Techniques. In: Model and Data Engineering, pp. 259-270. Springer International Publishing (2015)
9. Leng, Y., Xu, X., & Qi, G.: Combining active learning and semi-supervised learning to construct SVM classifier. Knowledge-Based Systems, vol. 44, pp. 121-131 (2013)
10. Ling, C. X., Huang, J., & Zhang, H.: AUC: a statistically consistent and more discriminating measure than accuracy. In IJCAI, vol. 3, pp. 519-524 (2003)
11. Livieris I.E., Mikropoulos T.A., & Pintelas P.: A decision support system for predicting students' performance. Themes in Science and Technology Education, vol. 9(1), pp. 43-57 (2016)
12. Mamitsuka, N. A. H.: Query learning strategies using boosting and bagging. In Machine Learning: Proceedings of the Fifteenth International Conference (ICML'98), vol. 1. Morgan Kaufmann Pub (1998)
13. Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S.: Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. Applied intelligence, vol. 38(3), pp. 315-330 (2013)
14. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. San Francisco, CA: Morgan Kaufmann (1988)

15. Platt, J.: Sequential minimal optimization: A fast algorithm for training support vector machines (1998)
16. Ramirez-Loaiza, M. E., Sharma, M., Kumar, G., & Bilgic, M.: Active learning: an empirical study of common baselines. Data Mining and Knowledge Discovery, pp. 1-27 (2016)
17. Reyes, O., Pérez, E., del Carmen Rodriguez-Hernández, M., Fardoun, H. M., & Ventura, S.: JCLAL: a Java framework for active learning. Journal of Machine Learning Research, vol. 17(95), pp. 1-5 (2016)
18. Romero, C., & Ventura, S.: Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 3(1), pp. 12-27 (2013)
19. Settles, B., & Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In Proceedings of the conference on empirical methods in natural language processing, pp. 1070-1079. Association for Computational Linguistics (2008)
20. Settles, B.: Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 6(1), pp. 1-114 (2012)
21. Shannon, C. E.: A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review, vol. 5(1), pp. 3-55 (2001)
22. Sharma, M., & Bilgic, M.: Evidence-based uncertainty sampling for active learning. Data Mining and Knowledge Discovery, vol. 31(1), pp. 164-202 (2017)
23. Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D.: Tools for Educational Data Mining A Review. Journal of Educational and Behavioral Statistics (2016)
24. Stapel, M., Zheng, Z., & Pinkwart, N.: An ensemble method to predict student performance in an online math learning environment. In Proceedings of the 9th International Conference on Educational Data Mining, International Educational Data Mining Society, pp. 231-238 (2016)
25. Triguero, I., García, S., & Herrera, F.: Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. Knowledge and Information Systems, vol. 42(2), pp. 245-284 (2015)
26. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2016)
27. Zhang, H.: The optimality of naive Bayes. AA, vol. 1(2), pp. 3, (2004)
28. Zhou, Z. H.: Learning with unlabeled data and its application to image retrieval. In Pacific Rim International Conference on Artificial Intelligence, pp. 5-10. Springer Berlin Heidelberg (2006)
29. Zhu, X., Lafferty, J., & Ghahramani, Z.: Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining, vol. 3 (2003)