Success is hidden in the students' data

Dimitrios Kravvaris[1], Katia Kermanidis[1], and Eleni Thanou[2]

[1] Department of Informatics, Ionian University, Corfu, Greece
jkravv@gmail.com, kerman@ionio.gr
[2] Pedagogical Department of Preschool Education, University of Ioannina, Ioannina, Greece
elenithanou@gmail.com

Abstract. The contribution of data mining to education as well as research in this area is done on a variety of levels and can affect the instructors' approach to learning. This particular study focuses on problems associated with classification and attribute selection. An effort to forecast the results takes place before the educational process ends in order to prevent a potential learning failure.

The methodology used during the experiments excluded the case of overfitting and ensured the completion of the study. Particular emphasis was placed on analyzing the results, which demonstrated the superiority of the Pearson VII function kernel using the Support Vector Machines algorithm to the Bagging meta-learning method. We also determined the appropriate point in the course timeline in order to get reliable results regarding students' outcome and finally, attribute selection gave us interesting results, in terms of students' data.

Keywords: Data mining, educational, classification, support vector machines, complexity, bagging, boosting, attribute selection.

1 Introduction

Our research focused on a) evaluating the success or failure of students in attending a course and b) analyzing the attributes from the dataset that we acquired in order to obtain useful information about the course. The learning procedure involved a co-educational (blended learning) training method in order for the students to obtain a computer skills certificate. Data was collected automatically using the Moodle e-learning platform [11], [12], which was used to implement the lesson plan from a human resources training organization.

The first objective of our study was to find the best classification model for our case. Previous researches by others [6], [7], [8], [9] have shown that the prediction of student's outcome attending a specific course has been primarily done within academic institutions. A variety of algorithms such as Decision Trees, Bayesian Networks, Neural Networks, K-Nearest Neighbor and Support Vector Machines (SVM) have been used and the best value for classification accuracy did not exceed 90%. The second objective was to determine at which time point of the lesson we could satisfy the classification accuracy in order to predict the students outcome. Finally the third

objective of our study was to analyze the attributes from the dataset in order to obtain useful information about the course structure and the students' profiles.

In previous work the datasets were drawn from K-12 or college students where their profiles, behaviors and goals have different characteristics from those of our study. The students' data obtained for this particular survey were provided from a Vocational Training Institution and concerned a computer skills training program in Microsoft Word. These institutions focus on training workers or those who are unemployed in an effort to enhance their existing qualifications. The characteristics and the behavior of the students that were attending a computer skills training course were not similar to K-12 or college students that were examined in previous work. With this survey we expand the research from schools and universities to companies and training institutions.

The classification done in our study used the SVM algorithm. We chose to use SVM since it is widely used within the data mining community and satisfies the objectives of our research protocol. Specifically SVM [12] a) can create a general purpose model (as opposed to a local), b) can handle non-linear class boundaries, c) is accurate on small datasets, d) deals well with spaces where the majority of data is numerical and e) can easily be adapted directly to the current state of the user.

With regards to the attributes of the dataset we found that emphasis was placed on the monitoring time per unit instead of the total time monitoring of the course as indicated by other investigators. This is of particular importance to the learning process since the student must be connected to the platform in order to monitor the course online. The reason for this is that the systems policy makes the students unable to download the course material locally to their computers. It is clear that the time devoted to studying or reading the material is useful in an educational setting and is essential in maintaining the flow of learning. The ultimate goal of the present study is to provide the necessary information tools to teachers and course administrators in order to help students successfully complete the training course.

2 Dataset

Our study used one particular dataset in which the participants were certified in Microsoft Word. The dataset has 511 examples and their attributes are listed in the table 1.

The demographic attributes of our trainees were based on Age (years old) and Education (secondary or higher education). The course attributes consisted of Word01 to Word15 which dealt with monitoring time per unit, WordTheory was the time of all theory activities, WordTest was the time that the student used to complete the Microsoft Word tests and WordSum shows the summarized time that the students were connected to the Moodle platform.

The classification system used is based on two values (binary), which are Success or Failure (WordResults) of the course in question. The reason for this classification system is based on the fact that we are not concerned with the overall grade the student receives but whether or not they passed or failed the subject of interest.

Table 1. Word Dataset Attributes

Attribute	Type	Description	Word Section
Age	numeric	Years	-
Education	nominal	Educational level	-
Word01	numeric	Time in minutes	Document handling
Word02	numeric	Time in minutes	Environment management
Word03	numeric	Time in minutes	Write text
Word04	numeric	Time in minutes	Manage text
Word05	numeric	Time in minutes	Utilities
Word06	numeric	Time in minutes	Fonts formatting
Word07	numeric	Time in minutes	Paragraphs formatting
Word08	numeric	Time in minutes	Pages settings
Word09	numeric	Time in minutes	Headers & footers
Word10	numeric	Time in minutes	Use changes
Word11	numeric	Time in minutes	Object management
Word12	numeric	Time in minutes	Create & manage tables
Word13	numeric	Time in minutes	Tables formatting
Word14	numeric	Time in minutes	Manage mass mail
Word15	numeric	Time in minutes	Printings
WordTheory	numeric	Time in minutes	-
WordTest	numeric	Time in minutes	-
WordSum	numeric	Time in minutes	-
WordResults	nominal	Success / Failure	-

3 Methodology

3.1 Classification method

Support Vector Machines. An important role in our research is the detailed configuration of the classification algorithm Support Vector Machines (SVM). The configuration was done using the software Weka, where it was possible to change the complexity as well as the kernel function parameters of SVM [2]. With regards to Weka the SVM uses the algorithm sequential minimal optimization developed by John C. Platt and will be referred to as SMO.

Kernel Functions. For our research purposes the kernel functions [3] that parameterize Weka's SMO in our experiment are the following.

Radial Basis Function (RBFKernel). Concerning the Radial Basis Function kernel [4], [5] the γ (gamma) parameter of Weka is represented by $1/2\sigma^2$ in the formula below, which has a Gaussian distribution.

$$k(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$$

Pearson VII Function (Puk). The Pearson VII function has the flexibility to change from a Gaussian to that of a Lorentzian distribution or an intermediate of the two [4]. This flexibility gives more power especially in the representation of Puk, compared to simple linear, polynomial and RBF kernel functions. The Puk function has two parameters, namely ω (omega) and σ (sigma), as depicted in the next formula.

$$k(x,y) = \frac{1}{\left[1 + \left(\frac{2\sqrt{\|x-y\|^2}\sqrt{2^{1/\omega}-1}}{\sigma}\right)^2\right]^{\omega}}$$

The ω and σ variables control the width and shape (that is the behavior of the tail) of the Pearson VII function distribution.

Polynomial (Polykernel). Using the polynomial kernel function [4] one parameter can be changed and that is the exponent p as shown in the formula below, however, there is no specific rule in choosing the exponent allowing for more than one test to be performed.

$$k(x,y) = (x,y)^p$$

Complexity. The parameter C of the SMO algorithm in Weka refers to the complexity, which determines the soft/hard margin for the SVM algorithm. The role of complexity in our survey is rather critical because the number of the examples is small and we must avoid the classification overfitting models [1]. If C is too high, then a solution with minimum error classification might be reachable, but at a high risk of overfitting, as shown below in figure 1.

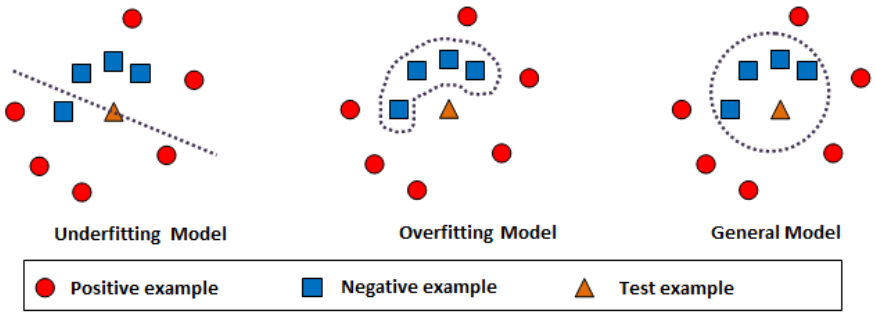

Fig. 1. The role of complexity in classification

Meta-learning algorithms. Finally, experiments involving the two more advanced classification methods Bagging and Boosting were run, which are implemented in Weka with the functions Bagging and AdaboostM1 respectively. These two methods

can use the SVM algorithm and provide us with the avoidance of adaptability to the training data. In each case the ultimate goal is to improve classification accuracy.

3.2 Early prediction

For the part of our survey concerning the early prediction of trainee outcome, we created new datasets from the original one (Word dataset). Those new datasets consists of demographic attributes and the course attributes which dealt with monitoring time per unit, adding one attribute at a time giving us 18 datasets (dw1 to dw18). The accelerating number in datasets represents the time flow of the course per section. We proceeded with classification for each dataset and we observed the accuracy values according the course flow.

3.3 Attribute selection

For the part of our survey concerning the attribute selection of the Word dataset we used a voting principal in order to obtain results using a variety of methods. We used a combination of search and evaluation methods as shown in table 2 [10]. The results were statistical analyzed choosing the top six attributes (which represents 1/3 of the total number of attributes evaluated) of each combination.

Table 2. Search and evaluation methods for attribute selection

Search Methods	Evaluation Methods
Best First	cfsSubsetEval (Consider the predictive value of each attribute individually, along with the degree of redundancy among them)
Ranker	ChiSquaredAtrributeEval (Compute the chi-squared statistic of each attribute with respect to the class)
GreedyStepwise	ConsistencySubsetEval (Project training set onto attribute set and measure consistency in class values)
Ranker	FIlteredAttributeEval (An arbitrary attribute evaluator on data that has been passed through an arbitrary filter)
GreedyStepwise	FIlteredSubsetEval (An arbitrary subset evaluator on data that has been passed through an arbitrary filter)
Ranker	GainRatioAttributeEval (Evaluate attribute based on gain ratio)
Ranker	InfoGainAttributeEval (Evaluate attribute based on information gain)
Ranker	OneRAtributeEval (Use OneR's methodology to evaluate attributes)
Ranker	ReliefFAttributeEval (Instance-based attribute evaluator)
Ranker	SVMAttributeEval (Use a linear support vector machine to determine the value of attributes)
Ranker	SymmetricalUncertAttributeEval (Evaluate attribute based on symmetric uncertainty)

4 Experiment

4.1 Classification method

With regards to the above experimental setup as described in the methodology section, we set forth to explore the configuration of the SVM algorithm.

Kernel Configuration. Tests were done in order to determine the best kernel function for our experimental data. In all experiments the complexity of SMO was set to 1 and a 10-fold cross validation technique was used.

- Parameterization of the Polykernel function concerned the exponent p, the values of which ranged from 1 to 10, and the results exhibited a classification accuracy greater than 97.0646% for p=4.
- Parameterization of the RBFKernel function concerned the γ parameter, ranging from 0.01 to 4.5, which resulted in classification accuracy greater than 96.4775% for γ=2.
- Parameterization of the Puk kernel function regarded parameters ω and σ. When ω values ranged from 1 to 10000 and σ remained constant (σ =1) the results gave us the same accuracy values. However, when ω is kept constant (ω=1) and σ varies from 1 to 5 then the results vary. Finally it was determined that the highest classification accuracy was 96.4775% using the Puk kernel function when both ω and σ were equal to 1.

Configuration Complexity. Based on the above findings with respect to the parameterization of the kernel functions, we kept the best results and experimented with complexity (C).

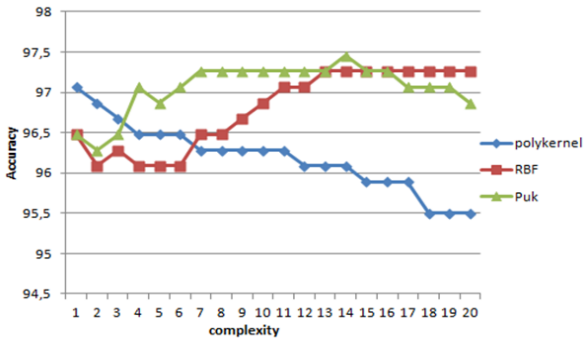

Fig. 2. Complexity parameterization

Experiments were done with values of C ranging from 1 to 20 with respect to the kernel functions that achieved the best results in the kernel configuration experiments and are shown in figure 2. The best classification accuracy rate for the Polykernel function was achieved with p=4 and C=1 and was determined to be 97.0646%. The

classification accuracy rate for the RBFKernel function was 97.2603% with γ=2 and C=13. The highest percentage of accuracy of all the functions tested was the Puk function and was determined to be 97.456% with ω=1, σ=1 and C=14. Precision, Recall and Confusion Matrix for the best results of each kernel are shown in figure 3.

Expirement	Precision	Recall	Confusion Matrix		
Polykernel, p=4, C=1	0.971	0.971	a	b	<-- classified as
			174	8	a= Failure
			7	322	b= Success
RBFKernel, γ=2, C=13	0.973	0.973	a	b	<-- classified as
			173	9	a= Failure
			5	324	b= Success
Puk, ω=1, σ=1, C=14	0.975	0.975	a	b	<-- classified as
			175	7	a= Failure
			6	323	b= Success

Fig. 3. Precision, Recall and Confusion Matrix

4.2 Bagging and Boosting

We next focused on experiments using two classification ensemble models. First we tried to increase the accuracy of the Bagging [13] and Boosting (AdaboostM1) [14] classifiers by using 10 iterations of the best algorithm we have found thus far, namely SVM using Puk were ω=1, σ=1 and C=14. The method gave a Bagging classification accuracy rate of 96.8689% and Boosting equal to 96.0861%.

Interestingly there was no increase in the accuracy when either Bagging or Boosting was used as shown above. This may be attributed to a high complexity C=14 and for this reason we chose to control the value of complexity (1-20). Our results showed that the accuracy rate of Bagging rose to 97.2603% when C=8 and Boosting rose to 96.2818% when C=1 (figure 4).

Using complexity C=14 lead to an extreme variation in the data as seen in figure 2, however if we choose a complexity C=8 this yielded the second best performance for the SVM (using Puk kernel function). With this adjustment in the complexity this leads to the same classification accuracy using the Bagging method equals to 97.2603%. This is a valuable tool for avoiding overfitting the classification method to the training data.

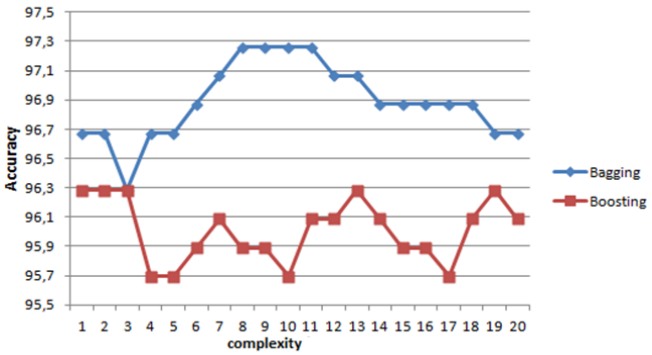

Fig. 4. Bagging and Boosting

4.3 Early Prediction

According to our findings we choose for classification the following parameterization Bagging using 10 iterations of SVM (Puk , ω=1, σ=1), this particular parameterization is best suited for the type of data that we are using according to our previous experiments. In order to avoid the classifier's data overfitting, because the datasets have a small number of attributes and examples we choose a low value of complexity C=1. The datasets we examined were arranged sequentially and represent a timeline, which depicts the monitoring of the online course (e.g. the first monitored unit is represented by dw1 dataset, the first and the second unit is represented by dw2 etc.). As shown in figure 5 the classification accuracy rises significantly earlier in the timeline rather than later and we can have a high value of classification accuracy (94.9119%) when the course timeline reaches the middle of the course (dw9 dataset).

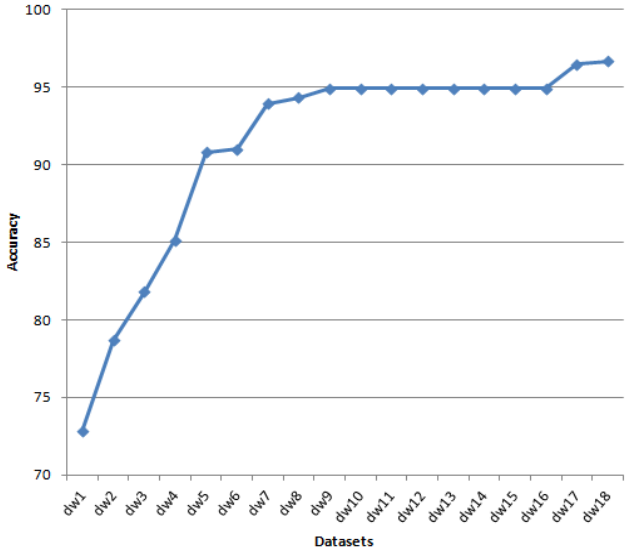

Fig. 5. Classification accuracy per dataset

4.4 Attribute selection

For the attribute selection of the Word dataset we used Weka to collect the attributes using the voting technique that we described in methodology. The collection of attributes was statistically analyzed using SPSS software and the results are shown in figure 6a, in which we observe that the demographic attribute of AGE has high rank in the Word certification course.

In addition, to the course attributes studied we next evaluated AGE and its association with the time that students spent to successfully complete the course. Groups were separated based on their age and a statistically significant difference was observed as shown in figure 6b. From this diagram we can see that students with an age greater than 40 needs on average 38% more time to successfully get certified in a computer training course compared to students that are less than 25 years old.

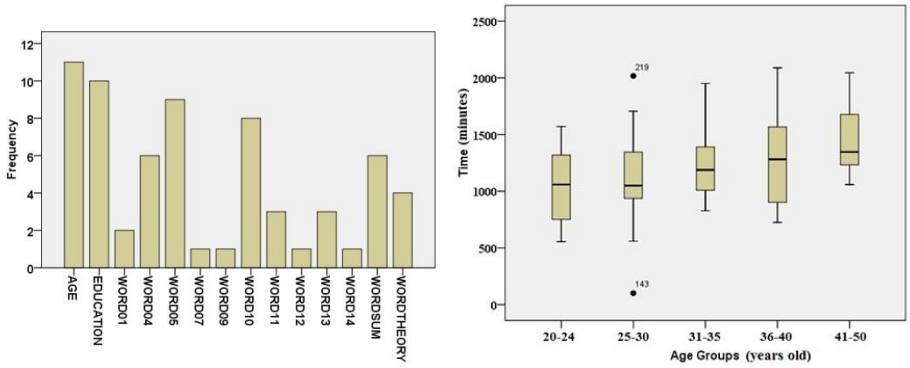

Fig. 6. SPSS results a) Attributes ranking (left) b) Mean times (right)

5 Conclusions

In studying the classification of SVM using three different function kernels (Polykernel, RBFKernel and Puk) it was determined that these functions play an important role in analyzing data obtained under our experimental conditions. Based on the configuration of the complexity within the SVM algorithm and using the different kernel functions, it was observed that Puk gave the highest classification accuracy of 97.456%.

The results from this study suggest that using the Bagging ensemble method was better suited for the data and gave a classification accuracy rate of 97.2603%, which was achieved with less complexity rather than using the classifier SVM alone. With this we attempt to reduce the margin of error with respect to the classification of success or failure, thus having a general model of classification for predicting trainee outcome in a computer skills training course.

Secondly studying at which time point of the lesson we could have satisfied classification accuracy for the prediction of the students outcome, we show that we can have a high value of 94.9119% for classification accuracy when the course timeline reaches the middle. This early prediction gives teachers and course administrators the time to find which students are inclined to fail in order to interfere and prevent them from failing, giving those students extra help and attention.

The last part of our study concerned the attribute selection for the Word dataset we followed a voting principal in order to obtain results from a large variety of methods. We observed that the attribute of AGE was always on the top six attributes for every attribute evaluation that was performed. We proceed with an analysis of the AGE attribute in compilation with the summarized time that students have spent in order to successfully complete the training course. We show that summarized time and AGE are increased together and we conclude that since the material of the lesson is common for students of all age, younger students successfully complete the training material in less time compared to older students.

Our results are of particular importance within the educational community, showing that both the students and the institutions can save time and money during the training process. The ability to predict with high level of classification accuracy the success or failure of the student before the end of the course, as well as selecting the most useful attributes for the attending course, strengthens the role of e-learning. These results do not only benefit the students but also benefits the teachers and administrators allowing them to improve existing online courses in an attempt to increase the successful completion of the course.

References

1. I.H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers, 2000.
2. M.A. Hearst, S.T. Dumais, E. Osman, J. Platt and B. Scholkopf, Support Vector Machines, Intelligent Systems and their Applications, IEEE Volume: 13 Issue:4, pp. 18-28, 1998.
3. C.J.C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, pp. 121-167, DOI: 10.1023/A:1009715923555.
4. B. Scholkopf and A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, The MIT Press, 2002.
5. B. Scholkopf, S. Kah-Kay, C.J.C. Burges, F. Girosi, P. Niyogi, T. Poggio and V. Vapnik. Comparing support vector machines with Gaussian kernels to radial basis function classifiers, In Proceedings of Signal Processing '97, pp. 2758-2765, 1997.
6. V.P. Bresfelean, M. Bresfelean, N. Ghisoiu, and C.A. Comes. Determining students academic failure profile founded on data mining methods. In Proceedings of the 30th International Conference on Information Technology Interfaces (ITI 2008), pp. 317-322, 2008.
7. W. Hamiiliiinen and M. Vinni. Comparison of machine learning methods for intelligent tutoring systems. In Proceedings of the 8th International Conference on Intelligent Tutoring Systems, Volume 4053, Lecture Notes in Computer Science, pp. 525-534, 2006.
8. W.Zang and F. Lin. Investigation of web–based teaching and learning by boosting algorithms. In Proceedings of IEEE International Conference on Information Technology: Research and Education (ITRE 2003) pp. 445-449, 2003.
9. C. Romero, S. Ventura, P.G. Espejo and C. Hervas. Data mining algorithms to classify students. In Educational Data Mining 2008: Proceedings of the 1st International Conference an Educational Data Mining, pp. 8-17. Montreal, Canada, June 20-21, 2008.
10. R. Kirkby, E. Frank and P. Reutemann, WEKA Explore User Guide, 2006.
11. C. Romero, S. Ventura and E. García. Data mining in course management systems: Moodle case study and tutorial, http://dx.doi.org/10.1016/j.compedu.2007.05.016, 2008.
12. C. Romero, S. Ventura, M. Pechenizkiy and R. Baker, Handbook of Educational Datamining, Published by CRC Press, 2011.
13. L. Breiman, Bagging Predictors, Machine learning, Volume 24, Number 2, 123-140, DOI: 10.1007/BF00058655, 1996.
14. J. Han and M. Kamber. Data Mining: Concepts and Techniques, 2nd Edition, Published by Morgan Kaufmann, 2006.