Data Mining Approach for Making Prediction of Students Success

Edin Osmanbegović[1], Haris Agić[2] and Mirza Suljic[1,3]

[1]Faculty of Economics, University of Tuzla, Bosnia and Herzegovina
[2]Pedagogical institute, Tuzla, Bosnia and Herzegovina
[3]ZD Rudnici "Kreka" d.o.o.-Tuzla, Bosnia and Herzegovina

Abstract

Although data mining represents the computational method of data processing, its use in education is still relatively new, i.e. its use is intended for discovering implicit, previously unknown, and useful knowledge out of existing data with an aim to make quality decisions in function of improvement of education system. The study was conducted by surveying the population of high school students in Tuzla Canton, Bosnia and Herzegovina (sample included about 10% of the student population, i.e. 1645 student). Using four different data mining algorithms the aim was to develop a model which can derive the conclusion of secondary level students' success.

Keywords: data mining, classification, secondary school education,

1. Introduction

The story of education can be a dramatic one in the case when the context is not respected, taken either in its narrow or broad sense [1], which does not mean that existing state in education has to be accepted. One has to struggle to reach better and high quality education. Previous experiences have shown that even small and poor countries have found the way out of the crisis by continual investment in education (Ireland, Singapur, Israel). Government that does not understand this fact and postpones investment in education is actually incapable and irresponsible. Status quo is rather being preserved, because in that case one does not require any knowledge, only bureaucratic and autocratic behavior toward subordinates, particularly toward those who do not share their point of view. A long time ago Heraklit said "nothing is eternal but changes" [2]. Ratković continues with a statement that, if nothing changes in a society, it stagnates; so standing still without progress is, according to this author, more dangerous than decline itself. No change can be planned in advance, in a way that their effects could be observed before implementation of the change [3]. There is no recipe for successful introduction of any change. It is logical, because if there is a recipe, then all changes would be successful ones. It could be referred to other experiences, when one wants to make own decisions. It is very characteristic for education.

This paper, as it is stated in the beginning, has been dealing with the prediction of students' learning results, so that the results, after being tested, could make influence on the quality of decisions in education at all levels. Results, obtained in this way, could have certain influence on the decision making process regarding creation, implementation, and continuous improvement of the policy of education. This is the case, particularly, regarding

© 2013. The authors - Published by Atlantis Press

the assessment of the achieved level of educational standards.

It is obviously clear that, based on the identification of the link between predicted successes and starting inputs from collected data, there could be developing strategies for prevention of "bad reasons" for worse success of the students. This will be used in the chapter dealing with interpretation of the collected data.

2. Introduction

This study will consider data collected during the 2011-2012 school year of twelve public schools, from the Tuzla region of Bosnia and Herzegovina. The latter with closed questions (i.e. with was designed predefined options) related to several demographic variables (e.g. mother's education, family income) and school related variables that were expected to affect student performance. All attributes are shown in Figure 1.

Attribute	Coding
sex (SP)	Nominal: M - male or Z – female
age (ST)	Numeric: from 14 to 18
type of school (TS)	Nominal: G, MS or O
address (A)	Binary: 1-urban or 0-rural
parent's cohabitation status (SR)	Numeric: coding form 1 to 4
mother's education (OM)	Numeric: coding form 1 to 5
mother's job(ZM)	Nominal: A, B or C
father's education (OO)	Numeric: coding form 1 to 5
father's job (ZO)	Nominal: A, B or C
family size (F)	Numeric: coding form 1 to 5
reason to choose this school (RI)	Numeric: coding form 1 to 5
home to school travel time (US)	Numeric: coding form 1 to 5
type of travel from home to school (DS)	Nominal: A, B or C
monthly scholarship (S)	Binary: 1-yes or 0-no
weekly study time (V)	Numeric: coding form 1 to 6
internet access at home (I)	Binary: 1-yes or 0-no
importance of grades obtained (VO)	Numeric: coding form 1 to 3
years of schooling (GS)	Numeric: coding form 1 to 4
average income of the parent's (PP)	Numeric: coding form 1 to 5
Final grade (OU)	Nominal: A, B, C or D

Fig.1: Input variables

The variable "students' grades" is divided into five levels or classes in a way that the final grade makes a class. By choosing this formulation of the output variables, they are being assigned to the problem of classification, aiming this model to recognize students' belonging to a specific classs.

Cross-validation methods are commonly used in examining the robustness of classifiers. In this study, a 10-fold cross validation was used. Initial set was divided in ten mutually different partitions of approximately equal size, according to random choice principle. Eight subsets were used for training, one subset for cross validation and one for measuring the predictive accuracy. This procedure was performed 10 times so that each subset was tested once. Test results were averaged over 10 tenfold cross-validation runs. Complexity of evaluation of the classification model directly depends on the number of iterations of the cross validation, since each iteration includes separation of construction and testing of the model.

3. Methodology

Approach to solve the problem using data mining (DM) is called techniques, or methods of modeling a data. Modeling means choice of different techniques and its application on the input set of data. The problem can be often solved by using a few different methods. Certain methods demand data in diverse forms, so, rather often it has gone go back to the phase of data preparation, where data are being prepared for application of the specific technique. In this phase, before the modeling, it has to be decided on which data the model will be trained, the model will be validated and the model will be tested.

Data classification methods represent the function which is mapping the data into one of several redefined classes. The goal of the classification techniques is to build a model that could classify future

data on the premise of the specification string. In this paper, it is being investigated the impact of four algorithms for intelligent data analysis: C4.5, Random forest (RF), Multilayer Perceptron (MLP) and Naive Bayes (NB).

The most common and probably the most used **decision tree algorithm** is **C4.5**. C4.5 has features such as handling missing values, categorization of continuous attributes, pruning of decision trees, rule derivation and others. C4.5 uses two heuristic criteria to rank possible tests: Information gain, which minimizes the total entropy of the subsets and the default gain ratio that divide information gain by the information provided by the test outcomes [4].

Random Forest algorithm (RF) is multiplied useful algorithm for data classification able to classify the enormous quantity of data with high accuracy. RF algorithm is statistics method based on the construction decision tree. The basic idea of the algorithm is to use a multitude of trees (quantifiers) instead of only one. To qualify new data out of the input vector it has to go through all the trees of the forest. During the data to qualify each classificatory (tree), has to make the decision of the class (trees vote for a class). After all the trees cast their vote the date will be classified in the class which has got the most votes. The forest makes the decision in favor of specific classification according to the number of votes.

Naive Bayes algorithm (NB) is a simple method of classification based on the theory of probability, i.e. the Bayesian theorem [5]. It is called naive because it simplifies problems of relying on two important assumptions: it assumes that the prognostic attributes are conditionally independent with familiar classification, and it supposes that there are no hidden attributes that could affect the process of prediction. This classifier represents the promising approach to the probabilistic discovery of knowledge, and it provides a very efficient algorithm for the data classification.

Multilayer Perceptron (MLP) algorithm is one of the most widely usable and popular neural networks. The network consists of a set of sensory elements that make up the input layer, one or more hidden layers of processing elements, and the output layer of the processing elements [5]. MLP is especially suitable for approximating a classification function (when we are not so much familiar with the relationship between input and output attributes) which set the example determined by the vector attribute values into one or more classes.

4. Experiment results

The aim of the research is to construct a model, out of the collected data, which will be able to predict student affiliation to the specific class in the case when the sample is unknown. The analysis has been conducted in the following way:

- to evaluate input attributes according to predicted attribute,
- to analyze each of four algorithms for data set.

The purpose of these analyses is to decide which set of the parameters gives the best results.

4.1. Evaluation of importance of the input attribute

In order to obtain a more exhaustive insight into the importance of input variables, it is common to perform an analysis of the effects of input variables, because the elimination of irrelevant attributes from the set for learning can improve the predictive quality of obtained

models. Filter methods are more practical solution for following reasons: time needed for choice and evaluation of the data is shorter, independent of the computer studying algorithms enables its application, in combination with any technique, in data modeling.

Four filter methods have been applied [5]: Hi-square test, OneR test, InfoGain test and Gain Ratio test. The average value of all the algorithms was taken as a final result of attributes ranking. Attribute weekly study time (V) has the greatest influence on output, and has shown the best performance in three of the four tests. It is followed by these attributes: GS, TS, ZO, ST, S, DS, SP, I, ZM, OM, OO, A, VO, SR, PP, F, RI and US.

The task of the third segment of the research is the choice of the most suitable data mining algorithm model. By using the previous experimental results, it has been adopted approach to remove attributes one by one from the input data base for each of four different classification algorithms of data mining. The grade and comparison of results from each of four different data mining algorithms have been conducted through ten-fold cross validation. There has been conducted nineteen experiments for each of four chosen algorithms or, in total; we have done seventy-six experiments with belonging classification analysis. Figure 2 and Table 1 show the results of the algorithm comparison in armed attribute prediction; the most accurate model of prediction is created by application of RF classificatory.

4.2. Analysis of Classification Model for Students' Success Prediction

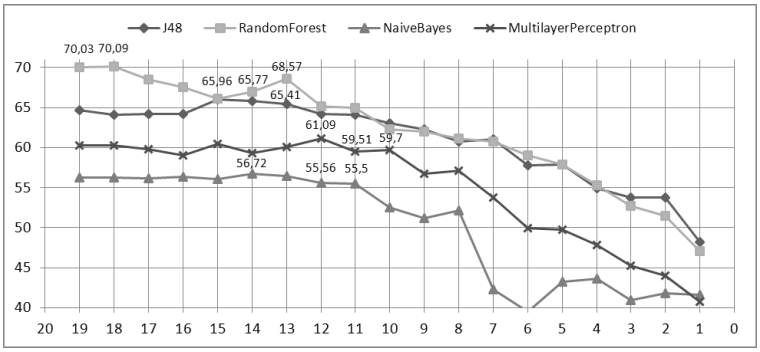

Fig.2: Algorithm comparison in aimed attribute prediction

Table 1. Classificatory comparison

EVALUATION CRITERIA	CLASSIFIERS			
	J48(13)	RF(18)	MLP(12)	NB(14)
Timing to build model (in Sec)	0,17	0,3	9,09	0,02
Correctly classified instances	1085	1153	984	933
Incorrectly classified instances	560	492	661	712
Prediction accuracy	65,96%	70,09%	59,82	56,72
TP Rate	0,66	0,70	0,60	0,57
FP Rate	0,12	0,10	0,13	0,15
Precision	0,66	0,70	0,59	0,54
Recall	0,66	0,70	0,60	0,57
ROC Area	0,81	0,89	0,80	0,81

RF classificatory has generated a model with 70.09% correctly classified examples (CCI), accuracy of 70% (0.70) and classification above the ROC curve area (0,89>0.5). It has been generated a confusion matrix for RF classificatory (Table 2). Four cases of nominal class attribute for final grades are labeled in letters A=excellent, B=very good, C=good, D=sufficient. The number of correctly classified examples can be found on the matrix diagonal while the other elements of the matrix mark the number wrongly classified instances which represent one of the classes left.

Table 2. Confusion matrix of RF classificatory

	Predicted			
Observed	a	b	c	d
A==a	292	98	27	2
B==b	139	194	67	1
C==c	51	85	273	8
D==d	4	4	6	394
Overall %	69,7%	48,4%	65,5%	96,6%

From Table 2 it can be noticed that number of the students predicted to get grade very good (class B) is 2 times smaller than the number of the students who are predicted to get sufficient (class D), so, accuracy of these classes are 48.4 and 96.6 respectively.

5. Conclusion

Education is crucial and the biggest capital of any society. Intelligent data analyses enable a high level of knowledge extraction out of data, so it offers big possibilities in the domain of education. Conducted research has been aimed to improve efficient knowledge discovery and as such could be helpful in making quality decisions in management of education.

Intelligent data analysis of the input algorithms has shown that learning time; years of schooling; type of school and paternal employment are among the most important factors in prediction of students' success (OU). From the point of view of an expert, the aim of this analysis is to present a method for reducing dimensional complexity of knowledge discovery in data sets that are often found in the analyses, and in this way, point out importance of some attributes to school managements and authorities in education. The obtained results could represent the basis for the future research. With a bigger number of input attributes and samples, it could be created more successful model that would be base for building of a support decision system at the secondary education level.

6. References

[1] A. Trnavčevič, "O kakovosti še malo drugače", Raznolikost kakovosti, pp. 9-25, 2002.

[2] M. Ratković, "Uspešan direktor škole, strategije obrazovnih reformi", Naučna knjiga, Beograd, 2000.

[3] M. Fullan, "The New Meaning of educational Change", 3th edition, Teachers College Press, NewYork, 2001.

[4] X. Wu & V. Kumar, "The Top Ten Algoritms in Data Mining", Chapman and Hall, Boca Raton, 2009.

[5] I. Witten & E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques", 2nd edn, Morgan Kaufmann, San Francisco, 2005.