Applying data mining to compare predicted and real success of secondary school students

LULE AHMEDI
ELIOT BYTYÇI
BLERIM REXHA
VALON RAÇA
Faculty of Electrical and Computer Engineering
University of Prishtina
Kodra e Diellit p.n., 10000 Prishtina
KOSOVO
lule.ahmedi@uni-pr.edu, eliot.bytyci@uni-pr.edu, blerim.rexha@uni-pr.edu, valon.raca@uni-pr.edu
http://www.uni-pr.edu

Abstract: - Schools and academic institutions may improve their curricula by checking student's data and their interoperability. Besides that, checking different student attributes may result in acquiring new knowledge for future improvements. The challenge of many schools however remains in gathering complete data into a single structured database. Working with "in-complete" data, may decrease the reliability of acquired results. In this paper students' success when resulting from real data is compared, with predicted students' success when data mining algorithms are applied. Results show that, even if there is lack of attributes, one may still apply certain data mining algorithms over school data to gain knowledge on the mainstream flow.

Key-Words: - Educational Data Mining, Cluster Analysis, Classification, Student Success

1 Introduction

Technology is used more than ever in education around the world. Kosovo is trying to catch up by integrating new technology in education. Primary and secondary education in Kosovo is organized in public and private institutions and as result of the ongoing reform, now the system in use is 5+4+3 years [1]. Even though, there are a big number of students passing through these institutions, student data in schools is saved in hand written diaries and not in electronic format and the only statistic derived from the diaries is the overall student success in a class. There are only few schools which have started using electronic records for storing student and teacher data. But, there is a tendency to change this situation and make all data in primary and secondary schools, both public and private, available in electronic form. That would greatly help in gathering information in a centralized database, which would result in easy statistical review and also in prediction of student success, among other. When such a database would become available, one might use Educational Data Mining algorithms and tools, to predict and also support teachers in their everyday work. Besides that, that database would help us to better understand the students' learning process and their involvement in it, as well as help us find the way to improve the quality of teaching and learning in general.

Many research papers are produced aiming to predict students' performance or other interesting facts regarding relation student-school. In [2] students are classified according to the features gathered from educational system in the web. Classification algorithms used in the research resulted in highest performance and best accuracy. In [3] regression algorithms for predicting student grades are used and the M5rules algorithm is considered as most appropriate for that research. Further in [4], four distinct data mining models (decision tree, random forest, neural networks and support vector machines) are used. The research assessed that in order to have better and more accurate results, one need to have more information regarding student grades but also background information (school related and personal ones – family situation) on the student.

2 Data mining over school data

Before describing the process itself, one should mention that the most time consuming part of Data Mining is dealing with data [5]: finding data as well as data preprocessing. Data used in this research was found after days of meeting with different institutions and even when that resulted in success, still one can argue that data used are not the of the highest quality data or do not have enough attributes describing them (especially attributes related to their family background). According to [6], quality of data is substantial but not always can one find qualitative data. Reason for that can be the lack of suitable databases, loss of data while merging databases etc. Lack of suitable databases is the case that follows our research, because school had only one database with all students' data: name, date of birth, place of birth, class and success per subject integrated together.

Student database, intended to be used in the research, had cardinality of 52936, but every student success in a subject was written as a single record. Besides that, student had five grades per subject during a school year (two for the semester and one final grade). Due to this all subjects related to a certain student had to be gathered, which resulted in smaller number of records.

Preprocessing of data is done in order to have more complete data (no attributes missing), to eliminate noise data and to be more consistent [7]. To achieve that, one most first "clean" its data. Method used was ignoring rows that have no data, in order not to add data and lose their originality. This was seen as the best solution for the specific database. Some other possible methods, that were not used, would be: a) adding data according to a specified "mean value" that would be taken – not used by us due to fact that one cannot be familiar with reason why data is missing – unless had worked with data during its preparation, and if used, result could be taken as an assumption, and b) putting a constant instead of missing data – not used by us due to fact that constant can be taken as a common thing between students who possess that constant. In some cases, missing data do not impose a mistake; just simply there are no data: e.g. data for the next semester are missing due to fact that when the data were collected, summer semester hasn't started yet.

Data that contain noise can be removed by using few principles: binning – values close to each other take same value as their mean value; regression and/or clustering. Data used in the research had a specific domain (success measured by grade 1 to 5 or by percentage 0 to 100) so there were no noises present. If one would encounter a grade greater than 5, then it would come to attention immediately that it is an inconsistency and that would mean removal of the data.

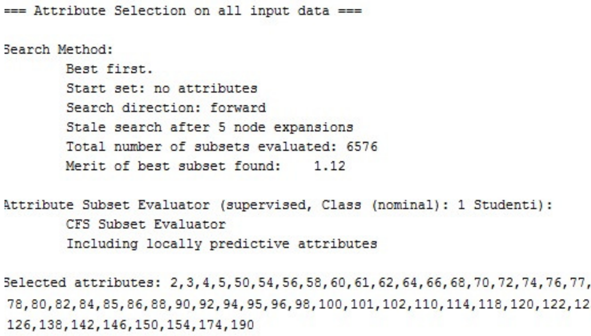
Maybe one of the most important tasks in preprocessing of data, is data integration, that because of the possibility of losing specific data. According to [8], systems developed today, use intercession architecture to achieve their goal: gathering information from different sources. In [8], LDAP and ontologies are used to integrate XML data. This research was fortunate to already have data integrated in one database.

In the end, even if have a well formed and structured database with complete data, it is highly recommended to use a portion of that data for research purposes. In the case of this research, specific class of attributes was used according to the values that they have. After the preprocessing, the database has 201 attributes and the only missing data, were data from the next semester.

Tool used in the process – WEKA [9], was chosen due to few of its attributes (graphic interface, JAVA implementation, support of a big number of Data Mining algorithms) but also due to fact that it is free under GNU GPL license.

3 Use of data mining prediction algorithms

In the research, two data mining algorithms were used: Naïve Bayes and C4.5 algorithm. Both are classification algorithms and the reasons why these algorithms were chosen to be used are: our database is a mixture of numeric and nonnumeric data; they can process more data and have minimum error rate compared to other algorithms [5]. Before applying these algorithms, one most find the most important attributes of the database. For that an algorithm that is part of WEKA tool [9] was used. The algorithm is called CFS attribute subset evaluator, which will evaluate the value of a group of the attributes while taking into consideration their prediction ability [10].



```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 6576
        Merit of best subset found:    1.12

Attribute Subset Evaluator (supervised, Class (nominal): 1 Studenti):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,3,4,5,50,54,56,58,60,61,62,64,66,68,70,72,74,76,77,
 78,80,82,84,85,86,88,90,92,94,95,96,98,100,101,102,110,114,118,120,122,123,
 126,138,142,146,150,154,174,190
```

Fig. 1. Attribute selection according to their importance

In Fig. 1 attributes found and arranged starting from first found, since it was used search method called Best First are presented. Therefore, resulted data can be considered as equally important and one of the attributes for the research can be chosen.

The attribute chosen to use in the research is attribute number 77: Eng 2 10n – where extensions have following meaning: Eng stands for English language, 2 describes that data is part of second semester, 10 is for class level (from 1 to 12) and n is for final grade (since in the database there exist columns for separated grades for the first and second semester).

The chosen attribute is evaluated with two above mentioned algorithms: Naïve Bayes and C4.5, and that in different ways: by using cross validation 2 and 10 (same data are divided for training and testing) and also training set. Results will be compared with real data, gained by separating database according grades (1 to 5), shown in Table 1, and according to groups (group 0 – holding student with grades 1 and 2; group 1 holding student with grades 3 and 4 and group 2 holding student with grades 5), shown in Table 2 and called Group012.

Tables 1 and 2, are a overview of the student but also school success.

After using Naïve Bayes and C4.5 algorithms, data gained and presented in Table 3. Table 3 shows that when using Naïve Bayes algorithm, a higher percentage of accuracy was found while using 2 fold cross validation than 10 fold cross validation, and it is the contrary when using C4.5 algorithm. The C4.5 algorithm using cross validation 10 fold, resulted more successful, with 100 % accuracy. Maybe this result is dedicated to algorithm C4.5 because it has the best results in cases when attributes are not related one to another [8].

Table 1. Dividing students by grades 1 to 5

Grade	Nr. of student	Percentage
1	2	0.73 %
2	17	6.23 %
3	36	13.18 %
4	87	31.87 %
5	131	47.99 %

Table 2. Dividing students by Group012

Group012	Nr. of student	Percentage
0	19	7%
1	123	45%
2	131	48%

Table 3. Algorithms used on database

Used algorithm	Grades 1 to 5	Group012 (1 and 2, 3 and 4, and 5)
Naive Bayes training set	87.91%	86.45%
Naive Bayes cross validation 2 fold	65.90%	76.56%
Naive Bayes cross validation 10 fold	64.48%	75.09%
C4.5 training set	94.14%	100%
C4.5 cross validation 2 fold	71.80%	99.30%
C4.5 cross validation 10 fold	82.05%	100%

The results have shown that even though there is considered not such a good correlation between attributes, still one can use data mining algorithms to classify and predict student success. Table 3 shows that dividing students in three groups will get better results and it suggests that dividing students only in two groups (pass or fail) should result in even a better approximation to real results. Better results would be possible, if the database itself was more complete, which will be future domain – securing a more complete student database to mine and also dividing students according to pass or fail prediction.

Fig 2 visualizes the accuracy of the algorithms used in the student database, both when all grades are separately considered (grades 1 to 5) as well as when grades' grouping (i.e. Group012) are used.

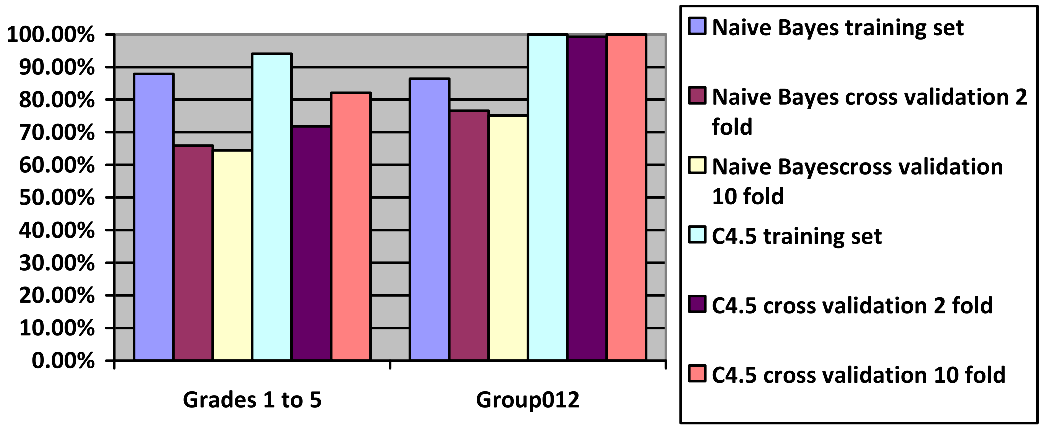


Fig. 2. Accuracy using data mining algorithms for students based on all grades and grouped grades

4 Conclusion

Data mining over school data is generally a very complicated and time consuming process but the most time consuming part of it is finding data. Even in case that the data are found and made available, there is still no guarantee that those sound and complete. However, even with not as complete data as expected, one may still get an overview of general data flow and their tendencies. In our research, the C4.5 algorithm has shown to perform best for predicting students' success, but that doesn't mean that it would remain the same if data were more complete and the database would provide more attributes. As a future work, one needs to find a more complete database (maybe filled with data gathered from eventual questionnaires) and then by using the same or different data mining algorithms as used in this paper, try to compare results of a complete and less complete database.

References:

[1] Primary and secondary school data in Kosovo, http://www.rks-gov.net/sq-AL/Qytetaret/Edukimi/Pages/ArsimiFillorDheMesem.aspx, retrieved on May 30th 2012.

[2] Behrouz Minaei-Bidgoli, William F. Punch. Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System. *Genetic and Evolutionary Computation Conference*, July 2003 Chicago. Springer-Verlag 2252-2263.

[3] S.B. Kotsiantis, P.E. Pintelas. Predicting Students Marks in Hellenic Open University. *Proceedings of 5th IEEE International Conference on Advanced Learning Technologies*, July 5-8, 2005 Kaohsiung, Taiwan, pp. 664 – 668

[4] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. *Proceedings of 5th FUture BUsiness 64 TEChnology Conference (FUBUTEC 2008) pp. 5-12*. Porto, Portugal, April, 2008.

[5] Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques, 2nd ed. *The Morgan Kaufmann Series in Data Management Systems.* Kaufmann Publishers. March 2006

[6] Eshref Januzaj, Visar Januzaj. An Application of Data Mining to Identify Data Quality Problems. *Proc. 3nd International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP 2009).* October 11-16, 2009 - Sliema, Malta, IEEE Computer Society, October 2009.

[7] Dorian Pyle. Data preparation for data mining. *Morgan Kaufmann Publishers Inc.* San Francisco, 1999.

[8] Lule Ahmedi – Global Access to interlinked XML data using LDAP and ontologies, Logos Verlag Berlin, 2004.

[9] Documentation of Weka. www.cs.waikato.ac.nz/ml/weka. Retrieved 25th May 2012.

[10] Mark A. Hall. Correlation-based feature selection for machine learning. *Ph.D. thesis, Department of Computer Science, University of Waikato.* Hamilton, New Zealand. 1999

[11] J. R. Quinlan. Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research*, 1996