Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Technology 1 (2012) 262 – 267

Procedia Technology

Evaluating the achievements of computer engineering department of distance education students with data mining methods

Baha Sen[a] *, Emine Ucar[b]

[a]Karabük University, BaliklarKayasi Mevkii, Karabük, 78050, TURKEY
[b]Ministry of National Education, Bakanlıklar, Ankara, 06420, TURKEY

Abstract

Recently, the internet technology has become an indispensable part of life, a very useful application that cannot be earlier have made it possible. One of these is distance learning technologies. Due to limitations of traditional learning-teaching methods in classroom activities and practitioners who intend to conduct training activities in the absence of the possibility of communication and interaction among learners with special education units are prepared and provided a wide range of media center through a certain method of teaching. According to a further recognition of Distance Education, although far away from each other with the student who teaches the same time (synchronous) or different time (asynchronous) communications with a tool as training system established. The aim of this study is to compare the achievements of Computer Engineering Department students in Karabük University according to criteria such as age, gender, type of high school graduation and whether the students studying in distance education or regular education using data mining techniques. Also discussing the differences of the techniques according to the results and to make suggestions for which technique would be more effective.

Keywords: Distance Education; Data Mining; Decision Trees; Artificial Neural Network

1. Introduction

Rapid developments in information societies has also changed interests and needs of today people while causing continuous changing and developments on lives of individuals and cultural, social and economic structure of societies. Now, people who can reach, use and produce information are needed. Therefore, development of Distance Learning has been well accelerated.

Distance Learning has many advantages. The most important of these are considered to be reproducible, distributable and accessible easily. Beside these advantages, integration of computer-aided systems, utilization of multimedia tools and techniques, reaching contents quickly and cost-efficiently over internet, increasing user interaction with help of new Technologies has provided the acceptance of distance learning sometimes as a support to formal education and sometimes as an education technique itself.

One of the first studies on data mining applied in education was published in 1995 by Sanjeev and Zytkow. Researchers gathered the knowledge discovery as terms like “P pattern for data in the range R” from university database [1].

Another study on data mining applied in education was published in 2000 by Becker and his friends who are performed for defining and understanding the impact of changes in curriculum on students at a university in Brasil [2].

* Baha Sen. Tel.: +90-370-433-2021; fax: +90-370-433-3290.
E-mail address: baha.sen@karabuk.edu.tr.

2212-0173
doi:10.1016/j.protcy.2012.02.053

A data mining application in which defining of student characteristics are used for measuring the satisfaction of students at higher education was performed by Luan in 2002 [3].
Maltepe University students identifying characteristics had been clustered using K-means algorithm in 2005 by Erdoğan and Timor. In that study 722 students' data was used and the relationship between the university entrance exam results and achievements was examined [4].
Vranić and Skoćır was examined how to improve some aspects of educational quality with data mining algorithms and techniques by taking a specific course students as target audience in academic environments [5].
In the second part of this study traditional and distance education concepts were examined. In the third section a data mining application was developed with using data from the Karabük University Computer Engineering Department students. In the conclusion sharing the experiences and findings obtained from this application is intended.

2. Formal and Distance Education

Formal education is a regular education that uses programs prepared in accordance with a purpose for the same level of certain age group and individuals at a school building. Formal education includes institutions of preschool, primary, secondary and higher education [6]. Distance education is an education that is realized with educator and students without being in the same place. This feature of distance education provides opportunity of learning for anyone at any age, place, time and speed [7].
The most obvious difference between distance education and classical education is completing their education (primary, secondary and higher education) without going to school, leaving their jobs and leaving their private lives.

3. Methodology

Data mining is relatively a new technique to the world of information sciences. Successful implementation of this technique requires a sound methodology built on best practices. In this research study, we followed a popular data mining methodology called Cross Industry Standard Process for Data Mining (CRISP-DM), which is a six-step process [8]:

- **Problem description:** Involves understanding project goals with business perspective, transforming this information into data mining problem description and making project plan to reach the related goals.
- **Understanding the data:** Involves identifying the sources of data, obtaining an initial set of data to assess the information coverage of the data for the problem on hand.
- **Preparing the data:** Involves pre-processing, cleaning, and transforming the relevant data into a form that can be used by data mining algorithms.
- **Creating the models:** Involves developing a wide range of models using comparable analytical techniques (i.e., selecting the appropriate modeling technique and setting the parameters related to the model to optimal values).
- **Evaluating the models:** Involves evaluating and assessing the validity and the utility of the models against each other and against the goals of the study.
- **Using the model:** Involves in such activities as deploying the models for use in decision making processes (i.e., making it a part of the decision support system/process).

A graphical representation of the methodology used in this study is shown in Figure 1.

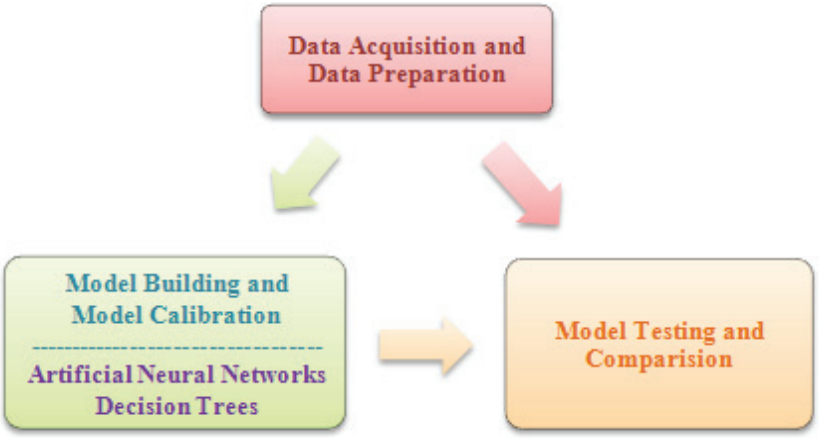

Fig.1. A graphical illustration of the methodology employed in this study

3.1. Data

In this study 3047 records were used which is taken by Karabük University Computer Engineering Department. Dataset have students' information such as age, gender, type of secondary school graduation, whether the students study in distance education or regular education and their lesson scores. And also dataset has information about the lesson taken by students in vocational lessons or cultural lessons.

Table 1. The list of independent variables used in this study

Variable Name	Data Type	Description
Gender	Text	Students' gender
Age	Number	Students' age
Type of High School Graduation	Text	Students' high school type
Distance/Regular Education	Text	Students' education type
Lesson Type	Text	Type of lessons

Scores of students which are studying in Karabük University are represented by the letter system. Score ranges of these letters are shown in Table 2.

Table 2. The output variable used in the study

Raw-Score	Nominal Representation
90-100	A1
80-89	A2
70-79	B1
65-69	B2
60-64	C
0-60	F

3.2. Data Mining Methods

In this study, two popular prediction/classification methods are used (and compared to each other): artificial neural networks, and decision trees. These prediction methods are selected because of their superior capability of modeling classification type prediction problems and their popularity in recently published data mining literature. What follows is a brief description of these modeling techniques.

- **Artificial Neural Networks:** Artificial neural networks (or NN, in short) are commonly known as biologically inspired mathematical techniques, capable of modeling extremely complex nonlinear functions [9]. In this study, we used a popular NN architecture called multilayer perceptron (MLP) with back-propagation type supervised-learning algorithm. MLP is capable of producing both classification and regression type prediction models, where the only difference is the output variable being nominal or numeric for classification or regression estimations. MLP is shown to be a strong function approximator for

prediction problems, that is, given the right size and the structure, MLP is shown to be capable of learning highly complex nonlinear relationships between input and output variables [10].

- **Decision Trees:** As the name implies, this technique recursively separates observations in branches to construct a tree for the purpose of achieving the highest possible prediction accuracy. In doing so, different mathematical algorithms (e.g., information gain, Gini index, Chisquare statistics, etc.) are used to identify a variable (from the available variable pool) and the corresponding threshold for that variable to split the pool of observations into two or more subgroups. This step is repeated at each leaf node until the complete tree is constructed. The most popular decision tree algorithms include Quinlan's ID3, C4.5, C5 and Breiman's CART (Classification and Regression Trees) algorithms. In this study, we choose to use Quinlan's C5 algorithm, which is an improved version of C4.5 (a very popular decision tree algorithm used by researchers and practitioners since early 1990s) [11, 12, 13].

4. Results and Conclusions

The prediction results of the two modeling methods are presented in Table 3. The results presented in Table 3 are the 10-fold cross validation results. Since the output variable had six nominal values, the confusion matrixes show 6x6 square matrix. In the confusion matrixes the rows represent the actual and the columns represent the predictions. The right most columns show the prediction accuracies for each of the six output variable values. The overall accuracy of each model is presented at the bottom of the right most columns.

As the results indicate, all of the classification methods performed reasonably well in predicting the six-value nominal variable. Among the two model types, decision tree algorithms produced the best prediction results with 97.8107% overall accuracy on 10 fold holdout dataset. Decision tree models followed by artificial neural networks with an overall accuracy of 94,3752%.

Table 3. Prediction results for classification methods (presented in confusion matrixes)

Artificial Neural Network							
	A1	A2	B1	B2	C	F	Accuracy
A1	**171**	21	0	0	0	0	
A2	24	**334**	15	0	0	0	
B1	0	19	**536**	22	0	0	
B2	0	0	16	**322**	13	0	
C	0	0	0	0	**519**	18	
F	0	0	0	0	19	**920**	
						Overall	**94.3752%**
Decision Trees							
	A1	A2	B1	B2	C	F	Accuracy
A1	**197**	10	0	0	0	0	
A2	12	**353**	4	0	0	0	
B1	0	8	**555**	13	0	0	
B2	0	0	5	**333**	2	0	
C	0	0	0	0	**531**	3	
F	0	0	0	0	8	**935**	
						Overall	**97.8107%**

The students' ages range from 18-38 and the success chart of students based on the age is shown in Figure 2. As show in the graph students' success rate has inverse ratio with students' age and the success score decreases with increasing age.

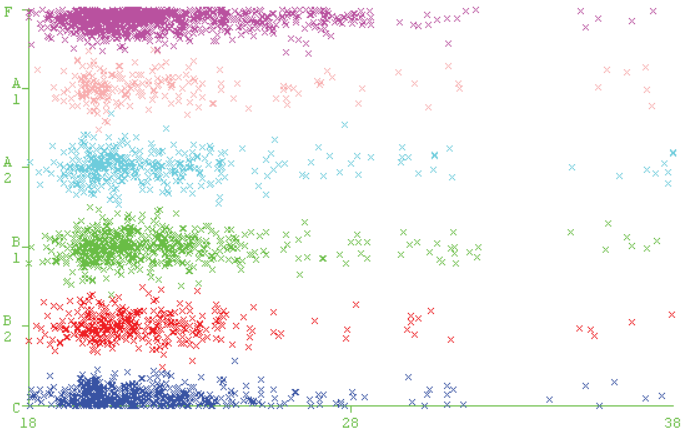

Fig. 2. Success graphic based on age

Figure 3 shows that the students' success is much better in the distance education or formal education. When we analyzed the graphic we can see that the students' scores between 65-80 are studying in the distance education and the students' scores between 80-100 are studying in the formal education. Also the students' scores less than 60 are the most in the distance education.

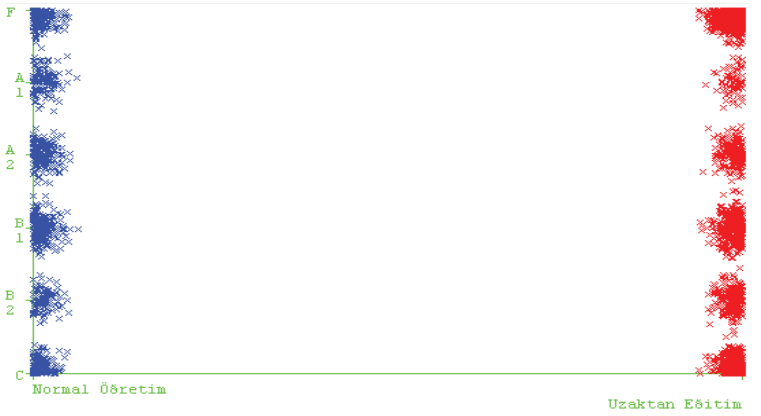

Fig. 3. Success graphic based on the type of education

Looking at the students' school type, the students which come from vocational high school are the 5% of total. Therefore, as shown in figure 4 students are more successful in the cultural lessons than the vocational lessons.

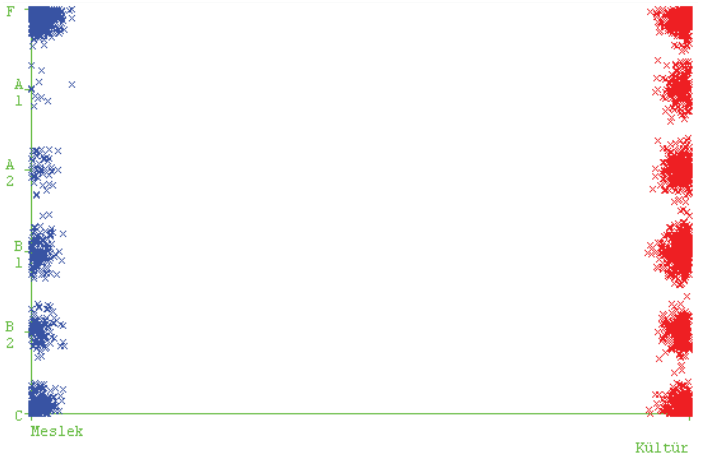

Fig. 4. Success graphic based on the type of lesson

References

1. A. P. Sanjeev ve J. M, Zytkow. "Discovering Enrollment Knowledge in University Databases," 1th Conference on KDD (Montreal. 20-21 August 1995), 246.
2. K. Becker, C. Ghedini ve E.L. Terra, "Using KDD to analyze the impact of curriculum revisions in a Brazilian university," SPIE 14th Annual International Conference (Orlando. April 2000), 412.
3. J. Luan, "Data Mining, Knowledge Management in Higher Education, Potential Applications", 42nd Associate of Institutional Research International Conference (Toronto,Canada: 2002), 1.
4. Ş.Erdoğan, M. Timor, "A Data Mining Application in a Student Database," Havacılık ve Uzay Dergisi. Cilt No 2,Sayı 2: 57-64, (July 2005), 57.
5. M.Vranić, D. Pintar, Z.Skoćır, "The Use of Data Mining in Education Environment," ConTEL 2007 (Zagrep 13-15 June 2007), 243.
6. Internet: Eğitim Sisteminin Genel Yapısı, http://www.meb.gov.tr/Stats/Apk2002/3_2.htm
7. H. E.Koçer, "Web tabanlı uzaktan eğitim", Yüksek Lisans Tezi, Selçuk Üniversitesi Fen Bilimleri Enstitüsü, Konya, 1-100 (2001)
8. C. Shearer, "The CRISP-DM model: The new blueprint for data mining" Journal of DataWarehousing, (2000). 5: 13-22.
9. S. Haykin, Neural Networks and Learning Machines (3rd Ed.). (2008). New Jersey: Prentice Hall.
10. K.Hornik,, M.Stinchcombe and H.White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward network" Neural Networks, (1990). 3: 359-366.
11. L.Quinlan, C4.5: Programs for machine learning, Morgan Kaufmann, (1993). San Mateo, CA.
12. L. Quinlan, "Induction of decision trees" Machine Learning, (1986). 1: 81–106.
13. L.Breiman, J.H.Friedman, , R.A. Olshenm and C.J.Stone, Classification and regression trees, Wadsworth & Brooks/Cole Advanced Books & Software, (1984). Monterey, CA.