Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 160 (2019) 573–577

Procedia Computer Science

www.elsevier.com/locate/procedia

The 6th International Symposium on Emerging Information, Communication and Networks (EICN 2019)
November 4-7, 2019, Coimbra, Portugal

Bayesian Classifier Applied to Higher Education Dropout

Amelec Viloria[a]*, Omar Bonerge Pineda Lezama[b] , Noel Varela[c]

[a,c] *Universidad de la Costa, Cl. 58 # 55 – 66, Barranquilla 080001, Colombia*
[6] *Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula 21101, Honduras*

Abstract

The research proposes a new simple Bayesian classifier (SBND) with Markov from the class variable to a network structure. Experimental tests are carried out by working a dropout analysis on students enrolled in the Faculty of Engineering Sciences of Mumbai University, in India in the period 2017-2018 on the basis of socioeconomic data. The Weka tool is then used to perform the classification and the proposed model is statistically compared with other Bayesian classifiers.

Keywords: Bayesian networks; Bayesian classifier; educational analysis.

1. Introduction

Probabilistic graphical models in the field of teaching are analyzed by [1] using clustering, grouping individuals or objects in clusters according to their similarities, maximizing the homogeneity of objects within clusters while maximizing the heterogeneity between aggregates. Another case study to predict the probability of a student leaving the educational institution was conducted using data mining techniques. Among these studies, [2] presented a research based on the use of knowledge, association rules, and the Top Down Induction of Decision Trees (TDIDT)

* Corresponding author. Tel.: +57-3046238313.
E-mail address: aviloria7@cuc.edu.co

10.1016/j.procs.2019.11.045

approach based on data from the academic management of a university consortium, which allows an interesting analysis to find the behavior rules.

[3] uses a dropout prediction method in e-learning courses based on three popular automatic learning techniques: feedforward neural networks, vector support machines, and simplified fuzzy ARTMAP methods. [4] compare different models to predict dropout rates during the first term of undergraduate studies at the University of Eindoven, by using classification trees, Naive Bayes, logistic regression, and tree forests, obtaining success rates between 75 and 80%.

It is important to mention the study of [5], which analyzes the relationship between the academic performance of students who enter the Faculty of Exact and Natural Sciences of a group of universities in Argentina during the first year of their careers and their socio-educational characteristics. A binary logistic regression model was adjusted, which adequately classified 75% of the data [6].

The objective of this work is to present the design of a new simple Bayesian classifier (SBND) that relates the variables of the class and its Markov frontier.

2. Method

This section develops experiments with a knowledge base of 12,247 students enrolled in the period 2017-2018 in the Faculty of Engineering Sciences of the Quevedo Technical University, using the socio-economic and academic data for classification with the Weka tool [7] [8]. The variables are illustrated in Table 1. The different values assumed by each of these variables are shown in Table 2.

Table 1. Variables and Description

Variable	Description
A	Career
B	Course
D	Disability
E	Costo of education
F	Lives separately from the family
G	Type of family housing
H	Owner of the house
I	Cable TV service
J	Credit card service
K	Internet service
L	Basic utilities
M	Private transportation
N	Phone plan service
O	Own car service
P	Comes in own car
Q	Currently working
R	Approved
S	Dropout

Table 2. Values and Description of Values

Variable	Description
E	X<200=0;200>X<800=1;X>800=2
G	HALF WATER=0; HOUSE/VILLA=1; APARTMENT=2; TENANCY ROOM=3; OTHER=4; HUMBLE HOUSE=5
H	FATHER AND MOTHER=0; FATHER=1; MOTHER=2; OTHER RELATIVE=3; OTHER=4

3. Results

Results were obtained using Naive Bayes and BayesNet classifiers with different alternatives such as K2, TAN, Hill Climber with one parent and also with a maximum of 5 parents [9] [10] [11].

In order to obtain these values in the Weka tool, the data was classified by using a cross validation of 10. As can be seen in Table 3, 12,247 cases were considered, standing out the BayesNet with K2 and maximum 5 parents as the one that best classified correctly (91.3658%), additionally indicating the rate of true negatives (TN) and the rate of true positives (TP), with sensitivity at 33.35%. This is the percentage of students who were correctly classified among those who drop out. These are the ones that should receive some attention and on which special actions should be applied to decrease the index. Although it is not a very high rate, it is important to point out that it is a difficult problem to predict and this procedure detects mainly one third of the students who drop out. In addition, the cost in terms of false positives is very low [12].

The rate of true negatives or specificity corresponds to the probability that a student who is doing well in his or her academic process will have a negative test result. In this case, only 1.49% are detected as false positives (see Table 3).

Table 3. Results obtained with different classifiers

Classifier	Correctly classified	TN Rate	TP Rate
NaiveBayes	87.0125	0.912	0.217
BayesNet with K2-1 parent	88.1463	0.963	0.217
BayesNet with K2-5 parents	91.3658	0.975	0.333
BayesNet with TAN	88.9952	0.952	0.321
BayesNet with Hill Climber-1	88.1254	0.962	0.186
BayesNet with Hill Climber-5	89.963	0.912	0.262

Figure 1 shows that all the variables are directly related to the dropout class (S). The course variable (B) also depends on the career (A) and influences the academic result (R). On the other hand, it can also be considered that the variable having cable tv service (I) directly influences (J; K; M), cellular plan service (N), and these on own car service (O) and currently working (Q) [13].

The results obtained using J48 and Random Forest as tree classifiers can be seen in Table 4.

When working with a J48 tree classifier, correctly classified cases are equivalent to 87.8862%. In addition, the percentage of sensitivity and specificity is indicated. These values do not improve on the BayesNet classifier with K2 and 5 parents. Similarly, it can be observed that working with a Random Forest tree classifier, correctly classified cases improve in relation to J48. It should be noted that it is a random forest of 100 trees where each is constructed with 5 characteristics.

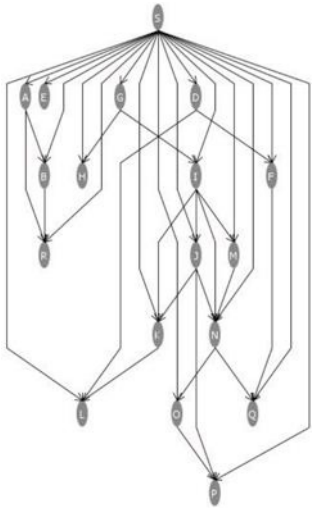

Fig 1. Network obtained with BayesNet classifier with K2 and a maximum of 5 parents

Table 4. Results obtained with the tree classifiers

Classifier	Correctly classified	TN Rate	TP Rate
J48	87.8862	0.994	0.012
RandomForest	89.7785	0.952	0.244

A comparison of the results obtained with the different state of the art algorithms is made with the database of socio-economic variables of the students from the Faculty of Engineering Sciences of the University of Mumbai as shown in Table 5.

Table 5. Results with student database

Data	SBND BDE	SBND BIC	SBND AK	SBND K2	BAN BDEu	BAN BIC
Socioeconómico	88.562	88.762	88.398	89.899	87.499	87.766
Data	**BAN K2**	**RPDag BDEu**	**RPDag BIC**	**RPDak K2**	**TAN**	**NaiveBayes**
Socioeconómico	87.685	87.579	87.998	89.686	88.766	87.598

As shown in the table, the algorithm that provides the best results is SBND with K2 with wide difference from the others that have been compared, while the worst results are provided by BDEu metric and BAN with the BIC metric. Since the value obtained with Friedman's test is greater than 0:05 [14] [15], the null hypothesis is not rejected and indicating that there are no significant differences between the distributions, so it is not necessary to continue testing. These results emerge from the fact that few databases were compared.

4. Conclusions

The student desertion problem is complex and difficult and has required the use of methods with a combination of factors (Bayesian classifiers) in order to obtain improvements over the trivial classifier that determines that no student drops out. Although the success rate is not very high, it was determined that by using a Bayesian classifier (BayesNet with K2 and a maximum of 5 parents), 33.35% of students who are going to drop out can be detected in order to apply method to avoid student's dropout. The cost measured as the percentage of students who are considered potential dropouts among non-dropouts is very low and is equivalent to 1.49 percent which does not allow the accurate identification of dropouts.

References

[1] Torres-Samuel, M., Vásquez, C., Viloria, A., Lis-Gutiérrez, J.P., Borrero, T.C., Varela, N.: Web Visibility Profiles of Top100 Latin American Universities. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, Springer, Cham, vol 10943, 1-12 (2018).

[2] Zhang, G.P.: Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50 (1), 159-75 (2003).

[3] Duan, L., Xu, L., Liu, Y., Lee, J.: Cluster-based outlier detection. Annals of Operations Research 168 (1), 151–168 (2009).

[4] Haykin, S.: Neural Networks a Comprehensive Foundation. Second Edition. Macmillan College Publishing, Inc. USA. ISBN 9780023527616 (1999).

[5] Haykin, S.: Neural Networks and Learning Machines. New Jersey, Prentice Hall International (2009).

[6] Oviedo, B. a. (2015). Análisis de datos educativos utilizando redes bayesianas, Latin American and Caribbean Conference for Engineering and Technology LACCEI 2015.

[7] Abhay, K. A., Badal, N. A.: Novel Approach for Intelligent Distribution of Data Warehouses. Published in Egyptian Informatics Journal-Elsevier, Egypt 17 (1), 147-159, (October, 2015).

[8] Vasquez, C., Torres, M., Viloria, A.: Public policies in science and technology in Latin American countries with universities in the top 100 of web ranking. J. Eng. Appl. Sci. 12(11), 2963–2965 (2017).

[9] Vásquez, C., Torres-Samuel, M., Viloria, A., Lis-Gutiérrez, J.P., Crissien Borrero, T., Varela, N., Cabrera, D.: Cluster of the Latin American Universities Top100 According to Webometrics 2017. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, Springer, Cham, vol 10943, 1-12 (2018).

[10] Sevim, C., Oztekin, A., Bali, O., Gumus, S., Guresen, E.: Developing an early warning system to predict currency crises. European Journal of Operational Research 237(1), 1095-104 (2014).

[11] Viloria, A., Lis-Gutiérrez, J.P., Gaitán-Angulo, M., Godoy, A.R.M., Moreno, G.C., Kamatkar, S.J.: Methodology for the Design of a Student Pattern Recognition Tool to Facilitate the Teaching – Learning Process Through Knowledge Data Discovery (Big Data). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data.

[12] Soca, E.B., El trabajo independiente en el proceso de enseñanza-aprendizaje, ISSN: 1684-1859, Revista Cubana de Informática Médica, 7(2), 122-131 (2015)

[13] Vanyolos, E., I. Furka, I. Miko y otros tres autores. How does practice improve the skills of medical students during consecutive training courses? doi; https://dx.doi.org/10.1590/s0102-865020170060000010. Rev. Acta Cirurgica Brasileira, 32(6), 491-502 (2017)

[14] Isasi, P., Galván, I.: Redes de Neuronas Artificiales. Un enfoque Práctico. Pearson. ISBN 8420540250 (2004).

[15] Haykin, S.: Neural Networks and Learning Machines. New Jersey, Prentice Hall International (2009).