Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 10, October-2016

A Prediction Model based on Data Mining to Forecast the Expectations of Passing from a College Student

Pedro R. Acosta De La Cruz
Universidad Nacional de Ingeniería
Lima, Perú

Miguel A. Meza Pinto
Universidad ESAN
Lima, Perú

José A. Flores Salinas
Universidad ESAN
Lima, Perú

Freddy C. Tineo Córdova
Universidad ESAN
Lima, Perú

***Abstract* - The present work has as objective to apply data mining techniques to develop a predictive model to forecast the chance of passing that will have a college student at the time of enrolling in a particular subject. Given that the academic record of the student can be known, and based on that information, we propose an Artificial Neural Network (ANN) that allows, using various configurations, to predict and assess our goal. The model has been applied to a compulsory subject of higher education of a University and given the results obtained. This model can be applied to any other subject analogous with satisfactory results.**

Keywords - Artificial Neural Networks, Data Mining, Higher Education, predictive techniques.

I. INTRODUCTION

In the high-level education is increasing the need for students to meet their chances of passing in the new courses they wish to enroll. We believe that it is possible to use the information in the academic performance or the student's school record and use some indicators of the course in which they want to register, in order to predict that objective. Many works show the interest of several authors in this matter, and the authors of [1] provide students with recommendations to enroll in certain courses, based on the experience of students who have taken these courses and that had academic features similar to these. In the same way, the authors of [2] pose as the fundamental objective, improving the quality of higher education, extracting relevant information using data mining (DM) techniques and use it for mentoring of guidance to students who require it.

In this work different from [1] and [2], whose results are based on models that use decision trees, we propose the use of artificial neural networks (RNA). An Artificial Neural network is a paradigm of learning and automatic processing inspired by the way the nervous system works, and is formed by a set of processing elements of information that are called "neurons" highly interconnected with "synaptic weights", with the ability to learn (in a supervised or not supervised way) and that collaborate to produce a stimulus of output. The main feature of this technique is that it applies to a large amount of problems, which may be real and complex or theoretical models sophisticated, such as recognition of images, analysis and encryption of signals, prediction, etc. The neurons or nodes of the ANN are organized in groups called layers. There are usually three types of layers: an input layer, one or more hidden layers and a layer of output. Connections are established between nodes of adjacent layers. The input layer that presents data to the network consists of input nodes that receive the information directly from the outside. The layer of output represents the response of the network to an entry and this information is transferred to the outside. Hidden layers or intermediate are responsible of processing the information and are interposed between the input layer and outputs see Fig 1.

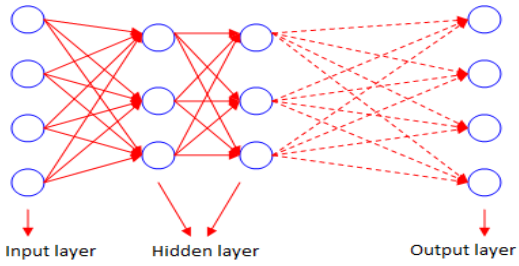

Fig. 1. An elementary ANN

The objective of this work is the development of a predictive model that allows a student of higher education to know their possibilities of passing a course. The data used in this model correspond to a single course and are mainly based on the academic performance of the students and some indicators of the course in which they want to register.

In addition, in this work we use the “open source” tool called Weka, which allows the construction of the predictive model that serves to guide a student in the process of registration. Shows the development of the predictive model ANN that uses a Multilayer Perceptron classifier.

II. MULTILAYER PERCEPTRON MODEL CLASSIFICATION

The description of the model of classification involves a description of the data and the design of the model of RNA to use, which will serve to achieve our goal. The data used in

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 10, October-2016

the model correspond to the course of Mathematics III, which is required in the third semester of the academic program in Chemical Engineering at the Faculty of Chemical and Textile Engineering (FCTE) of the National University of Engineering, Lima Peru [3]. The total number of records considered was 1276 records corresponding to the periods academics in the first half of 1993 to the second half of 2010 (17 academic years). These data were collected through the Statistical Office and academic records the FCTE.

A. *Data Structure*

The design of the proposed model considers a data structure composed by data obtained from students and course data to analyze. Each subject has a number of credits assigned in accordance with the number of hours of theory and practice of the same one. The data of the model is showed in the Table 1 with some example records, followed by the description of the attributes.

It is necessary to mention that the rating scale in Peru is from 0 to 20 and that with 11 the student passes a subject. In addition, to enroll in a course is required to have passed at least one subject previous called Pre requirement.

TABLA 1. MODEL DATA

ppa	*notPreReq1*	*sumCred*	*antAlum*	*pGD*	*Spgd*	*Cond*
12.95	14.9	14	7	9.11	34.11	A
11.03	11.0	22	2	9.48	55.62	D
12.33	11.8	24	2	9.48	63.19	A

The data of the student considers the following attributes:

1. *ppa*: Represents the weighted average accumulated by the student at the time of enrolling in a new subject. This data is calculated by adding the product of the marks (of the subjects carried out by the student) by their respective credits, and this result is divided by the sum of all credits taken by the student.
2. *notPreReq*1: Represents the course's mark considered as pre-requisite 1. In the event of more than one pre-requisite is entered all pre-requisites. In our model we employ only a pre-requisite.
3. *sumCred*: It represents the entire sum of all the credits of the subjects in which the student enrolled in the semester.
4. *antAlum*: It represents the seniority of the student at the University. This data is measured in years from the moment in which the student enrolled for the first time.
5. *Cond*: Is the categorical variable to predict, indicates whether the student passes (A) or disapproves (D) the subject in which intends to enroll.

The data of the course considers the following attributes:

6. *pGD*: Represents the average of the Grade of difficulty of the subject, the Grade of difficulty by each semester (pGDs) is defined as twenty less the average of marks of the students enrolled in that course. The main point of pGD is calculated considering the average of the guest.
7. *spgd*: It represents the sum of pGD corresponding to the subjects in which the student is enrolling.

B. *Features of the model of ANN proposed*

The proposed RNA model has the following characteristics:

- Use of the MultilayerPreceptron option for the election of the function of the classification with a multilayer Perceptron network.
- For testing, are carried out 10 executions of cross-validation (option: Test mode: 10-fold cross-validation with Weka).
- An input layer (7 neurons, one for each attribute), two hidden layers (the number of neurons is determined by trial and error) and an output layer (2 neurons, one for the ones which approved and another one for those that disapproved).
- The neuron that characterizes the attribute to predict, is not in the input layer. However it is used because otherwise it could not train the network.
- The choice of the number of hidden layers in the network model considers various configurations. Which show superior results are the configurations [6, 4], [6, 6] and [6, 8], indicating the number of neurons in the hidden layers of the network. Such configurations are shown in Fig. 2, 3 and 4.

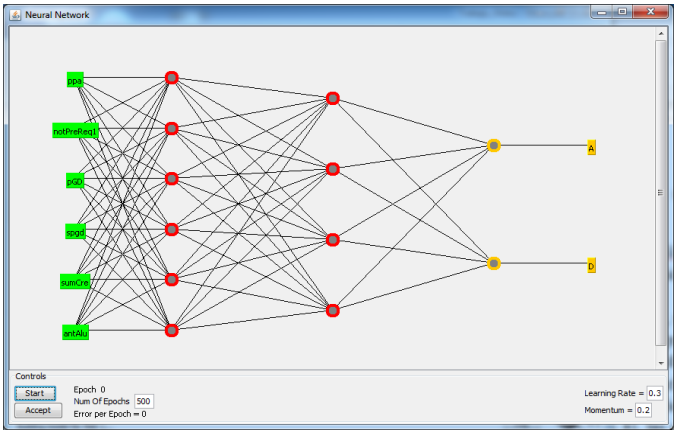

Figura 2. Neural Network Model [6, 6, 4, 2]

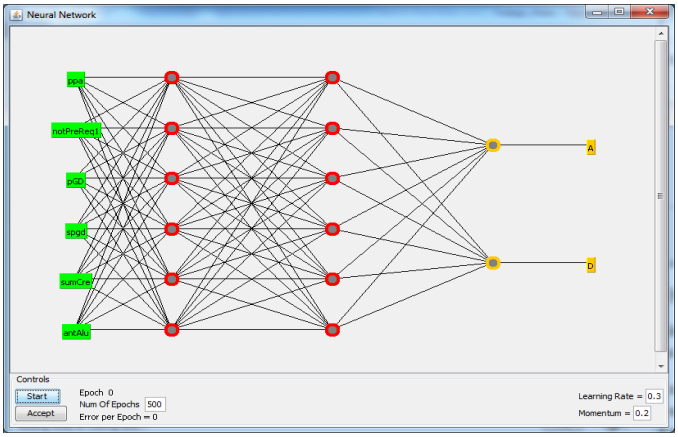

Figura 3. Neural Network Model [6, 6, 6, 2]

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 10, October-2016

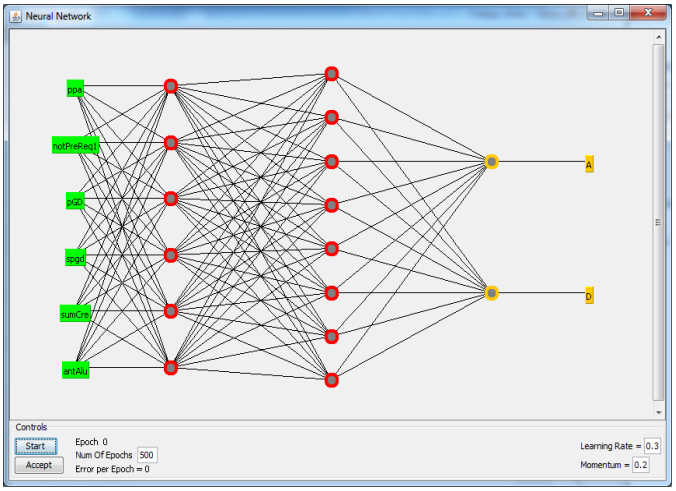

Figura 4. Neural Network Model [6, 6, 8, 2]

III. MODEL PREDICTION'S RESULTS

Table 2 shows the results obtained with the different tests with various configurations. The results were obtained using a training rate (learningRate) of 0.3 and 500 times of training (trainingTime). From this table we see that the successes are stabilized in a 70.85% and there is more variation in the last three configurations.

TABLA 2. RESULTS (LEARNINGRATE = 0,3; TRAININGTIME = 500)

Configuration	Attributes considered in section 2.1	Hits	No hits
[4, 4, 4, 2]	1, 2, 3, 4 y 5	69,59%	30,41%
[6, 6, 4, 2]	1, 2, 3, 4, 5, 6 y 7	70,45%	29,55%
[6, 6, 6, 2]	1, 2, 3, 4, 5, 6 y 7	70,85%	29,15%
[6, 6, 8, 2]	1, 2, 3, 4, 5, 6 y 7	70,85%	29,15%

Varying rates of 0.2 and 0.1 learning respectively, we obtain the results shown in Tables 3 and 4.

TABLA 3. RESULTS (LEARNINGRATE = 0,2; TRAININGTIME = 500)

Configuration	Attributes considered in section 2.1	Hits	No hits
[4, 4, 4, 2]	1, 2, 3, 4 y 5	69,91%	30,09%
[6, 6, 4, 2]	1, 2, 3, 4, 5, 6 y 7	71,08%	28,92%
[6, 6, 6, 2]	1, 2, 3, 4, 5, 6 y 7	71,00%	29,00%
[6, 6, 8, 2]	1, 2, 3, 4, 5, 6 y 7	70,77%	29,23%

TABLA 4. RESULTS (LEARNINGRATE = 0,1; TRAININGTIME = 500)

Configuration	Attributes considered in section 2.1	Hits	No hits
[4, 4, 4, 2]	1, 2, 3, 4 y 5	70,53%	29,47%
[6, 6, 4, 2]	1, 2, 3, 4, 5, 6 y 7	71,24%	28,76%
[6, 6, 6, 2]	1, 2, 3, 4, 5, 6 y 7	71,08%	28,92%
[6, 6, 8, 2]	1, 2, 3, 4, 5, 6 y 7	71,00%	29,00%

Comparing the results in Tables 3 and 4 with Table 2, we note that the percentages of correct answers do not vary significantly.

Subsequently, by changing the number of times we obtain the results shown in Tables 5 and 6.

TABLA 5. RESULTS (learningRate = 0,3; trainingTime = 750)

Configuration	Attributes considered in section 2.1	Hits	No hits
[4, 4, 4, 2]	1, 2, 3, 4 y 5	69,59%	30,41%
[6, 6, 4, 2]	1, 2, 3, 4, 5, 6 y 7	70,38%	29,62%
[6, 6, 6, 2]	1, 2, 3, 4, 5, 6 y 7	70,69%	29,31%
[6, 6, 8, 2]	1, 2, 3, 4, 5, 6 y 7	70,06%	29,94%

TABLA 6. RESULTS (learningRate = 0,3; trainingTime = 750)

Configuration	**Attributes considered in section 2.1**	**Hits**	**No hits**
[4, 4, 4, 2]	1, 2, 3, 4 y 5	69,51%	30,49%
[6, 6, 4, 2]	1, 2, 3, 4, 5, 6 y 7	69,75%	30,25%
[6, 6, 6, 2]	1, 2, 3, 4, 5, 6 y 7	69,83%	30,17%
[6, 6, 8, 2]	1, 2, 3, 4, 5, 6 y 7	69,36%	30,64%

Tables 5 and 6, note that the number of times has no significant impact in relation to the results shown in Table 2.

In all tests performed model configuration [4, 4, 4, 2], does not consider PGD attributes and SPGD, as these are not direct results of a calculation. In all other configurations these attributes are considered, but the percentage of correct answers has minimal variation, so its impact is minimal.

IV. RESULTS OF THE PREDICTION MODEL WITH DATA RESTRICTED

Given the results in Section 3 and considering that 70.85% is a low percentage for a prediction, we need to do more studies analyzing other variables.

Table 7 is considered a restriction on the mark obtained in the pre-requisite (notPreReq1 attribute) course and in Table 8 restriction is applied to the weighted average cumulative (ppp attribute). For both configuration restrictions [6, 6, 6, 2] (with learningRate = 0.3; trainingTime = 500) is used, since from it we have a stabilization of the successes (see Table 2).

TABLA 7. RESULTS OF THE RESTRICTION OF THE ATTRIBUTE notPreReq1

Restriction of notPreReq1	Number of records used	Hits	No hits
>=11	700	74,42%	25,58%
>= 12	377	75,33%	24,67%
>= 13	205	80,97%	19,03%
>= 14	88	86,36%	13,64%

Tables 7 and 8 note that the percentage of correct answers obtains significant increases as the value of the restriction increases even if the number of records used decrease.

TABLA 8. RESULTS OF THE RESTRICTION OF THE ATTRIBUTE ppa

Restriction of *ppa*	Number of records used	Hits	No hits
>=11	516	78,49%	21,51%
>= 12	209	85,65%	14,35%
>= 13	65	93,85%	06,15%

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 10, October-2016

V. CONCLUSIONS

From the execution of the model and the analysis of itself, we establish the following results or conclusions:

1. The method used is a valid way to predict the future development of a student based on their academic record and historical characteristics of the subject that he will enroll.
2. The incorporation of attributes in the data structure improves the prediction model, even though this is not very significant. According to Table 2 the addition of two attributes (PGD and SPGD) gives us improvements up to 1.24%. From there, very little new configurations improve the predictive aspect, which becomes almost steady at 70%.
3. Varying the learning rate (learningRate) or learning times (trainingTime) have very little impact, and even decreases in some cases below 70%.
4. The most influential attributes are constituted by the mark of the pre-requisite and the mark of the weighted average accumulated in an independent way. This means that the level of a student's success is closely related to the preconditions with entering the subject.
5. The attributes or indicators of the subject do not have a big impact on the model. This allows us to conclude that the success of students in a given subject mainly depends on their preparation.
6. Both attributes of the student and the number of credits in which you enroll or seniority student at the University, has very little influence on the final behavior.

From the analysis of the results we conclude that the proposed model is reliable with acceptable predictive results. Therefore the model is a predictive tool to manage the learning expectations of a student in a particular subject, and also the learning expectations of the students.

Future works involves introducing new attributes, such as age of the student, the easier access to books or specialized information, and other data on the socioeconomic status of the student, to analyze their behavior with the ANN model proposed.

REFERENCES

[1] Vialardi C., Bravo J., Shafti J., Ortigosa A.: Recomendation in Higher Education Using Data Mining Techniques. II International Conference on Educational Data Mining, pp. 190 - 199, 2009.

[2] Baradwaj B. y Pal S.: Mining Educational Data to Analyze Students' Performance. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[3] Facultad de Ingeniería Química y Textil – Universidad Nacional de Ingeniería, Oficina de Estadística.: Base de datos del curso Matemática III, periodos del 1993-I al 2010-II. Lima Perú 2011.

[4] Kinnebrew J., Biswas G.: Identifying Learning Behaviors by Contextualizing Differential Sequence Mining with Action Features and Performance Evolution. Proceedings of the 5th International Conference on Educational Data Mining, pp 57- 64, 2012.

[5] Thai-Nghe N., Horváth T., Schmidt-Thieme T.: Factorization Models for Forecasting Student Performance. Proceedings of the 4th International Conference on Educational Data Mining, pp 11- 20, 2011.

[6] Bouckaert R., Frank E, Hall M., Kirkby R., Reutemann P., Seewald A., Scuse D. : Weka Manual for Version 3-7-2. The University of Waikato, Hamilton, New Zealand, 20