Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 170 (2020) 953–958

Procedia Computer Science

www.elsevier.com/locate/procedia

The 10th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS)
April 6-9, 2020, Warsaw, Poland

# Assembly of classifiers to determine the academic profile of students

Jesus Silva[a]*, Karina Rojas[b], Alexa Senior Naveda[c], Rosio Barrios[d], Carlos Vargas Mercado[e], Claudia Medina[f]

[a, b] *Facultad de Negocios, Universidad Peruana de Ciencias Aplicadas, Lima, Peru.*
[c, f] *Universidad de la Costa, Barranquilla, Colombia*
[e] *Corporacion Universitaria Minuto de Dios (UNIMINUTO), Barranquilla, Colombia.*
[f] *Corporacion Universitaria Latinoamericana, CUL, Barranquilla, Colombia..*

**Abstract**

The assembly methods, or combination of models, arise with the purpose of improving the accuracy of predictions. An assembly contains a number of apprentices (base models) which, when of the same type are called homogeneous and if of different, heterogeneous. The characteristic is that these apprentices do not perform well. The assembly is generated using another algorithm that combines the apprentices, examples of which are the majority vote, the decision table and the neural networks [1]. This article proposes the use of an assembly of classifiers to determine the academic profile of the student, based on the student's overall average and data related to educational factors.



*Keywords:* Assembly of classifiers, decision trees, artificial neural network..

## 1. Introduction

One of the concerns in the universities is the dropout and low academic performance of students. It is known that

* Corresponding author. Tel.:+51-975857503.
*E-mail address:* jesussilvaUPC@gmail.com

1877-0509 
10.1016/j.procs.2020.03.102

this is influenced by many factors, including social, economic, educational, academic and institutional, to name a few. This study seeks to generate an assembly of classifiers to determine the academic profile of a student, based on their overall average and using educational factors such as study activities, forms of learning and study habits as discriminants. The database consists of the students' records from the University of Mumbai (UM) in India who enrolled in the years 2010, 2011 and 2012 in the different careers that are offered.

A popular strategy is to test different algorithms to generate the models, evaluate them and choose the one that provides the best results, i.e. the one with the least error in predicting the class of an unknown instance [2] [3] [4] [5]. In contrast, an assembly method builds or uses several classifiers and combines them [6]. Classifiers built from different algorithms on the same data are usually used. However, since doing it in this way did not lead to good results, the data in this study were divided by groups of factors and the same C4.5 algorithm was applied, building 3 classifiers and generating the assembly with them.

## 2. Method

### *2.1 Data selection*

The data were obtained from the socio-economic study applied to the students of the 2010-2015, 2011-2016 and 2012-2017 generations of all careers at the University of Mumbai (UM) in India. The selected data correspond to the section on educational factors, which applies to students, specifically to study activities, see Table 1, learning styles, see Table 2 and to study habits, see Table 3. These data are related to school monitoring data, specifically to the student's overall average.

The answer scale to the questions is divided into: Never, Normal, Always, except in those where a number of hours or books is requested, which are the first 3 questions on study habits.

Table 1. Study activities.

| **Activities** | |
|---|---|
| I meet with my classmates to study for an exam | Study mainly with monographs |
| I meet with my classmates to develop a task or group work | I study mainly with my class notes |
| When I start, I identify what I need to study and draw up a work plan. | I study mainly with the textbook of the subject |
| I check what I remember from what I studied. | I study mainly with the notes of my classmates |
| I identify concepts that I have not yet understood | I use encyclopedias, dictionaries or atlases |
| When I don't understand something, I look for more information. | I use a computer or the Internet to study, do homework, or solve an exam |

Table 2. Learning styles.

| **Styles** | |
|---|---|
| I learn more when I work with other colleagues | I do well on exams in most subjects |
| It is helpful for everyone to contribute ideas when working in a group | I like to work with other colleagues |
| I study to secure my future financially | I only read when I have an obligation to do so |
| I study to get a good job | Reading is one of my favorite hobbies |
| I study to learn more | I like to discuss books with other people |
| I study for a better life | It is hard for me to finish reading a book |
| I trust I can understand what I study even the most difficult texts | I like to be given books |
| I am confident that I can do an excellent job on my homework and exams | Reading seems like a waste of time to me |
| I am sure I master the skills I was taught | I enjoy visiting bookstores or libraries |
| I learn quickly in most subjects | I only read to get the information I need |
| I am competent in most subjects | I have a hard time sitting down to read for a long time |

Table 3. Study habits.

| **Habits** | |
|---|---|
| Hours a week that you study or do homework outside of school hours? | Magazines |
| Do you spend hours a week reading about what you like or are interested in? | Newspapers |
| Mention how many complete books you have read in the past 12 months without taking textbooks into account | Comics |
| Literature books (novel, theater, poetry) | Websites |
| Book son other subjects (science, technology, economics, etc.) | |

### *2.2 Preprocessing*

In the original file, the number of instances was 4600, however, because some students did not answer a large number of questions, instances with incomplete information were removed [9], leaving only 1230 instances.

### *2.3 Transformation*

The academic profile of the student was divided into 3 classes, according to the general average obtained from their entry and until the last semester. Table 4 presents this information, as well as the number of students who present this profile.

Table 4. Distribution of students by academic profile.

| Academic profile | General average | Students |
|---|---|---|
| Regular | [0.0,7.4] | 170 |
| Good | [7.5,8.6] | 852 |
| Excellent | [8.7,10] | 208 |
| | Total | 1230 |

The selected data set, 1230 instances or specimens, was divided into two sets: the training set with 984 instances, and the test set with 246 instances.

## 3. Results

The experiments carried out and the results obtained are presented below, starting with the generation of the single classifier that uses all the data, continuing with the individual classifiers and ending with the assembly.

### *3.1 Generation and testing of a classifier considering all educational factors*

The first classifier was generated using algorithm C4.5 [10]. The model was generated from the training set (984 instances). The results are presented in Table 5, in terms of percentages of success and error in the classification of the instances, as well as the confusion matrix. By using the generated model to classify the instances in the test set, it can be observed that the efficiency of the classifier drops from 91.2% to 60.23%.

### *3.3 Assembly of classifiers using an artificial neural network*

#### *3.3.1 Neural network architecture (NN).*

The output of the classifiers is represented by a 3x1 vector. P= [Class 1, Class 2, Class 3] T, so the NN has nine

inputs. The nine inputs pass to the first layer of 10 neurons, with an independent compensation input (bias).

Table 5. Results of the classifier generated with the algorithm C4.5 (J48 WEKA).

| Classes | Confusion Matrix | | | Success | Error |
|---|---|---|---|---|---|
| | A | B | C | 91.2% | **8.8%** |
| a = Good | 540 | 10 | 7 | 984 instances | |
| b = Regular | 30 | 110 | 1 | 897 correctly classified | |
| c = Excellent | 34 | 9 | 90 | 87 incorrectly classified | |

The activation function is logsig, (1). The final layer has three neurons with bias and logsig activation function. The objective is that the output represents the class by means of the vector S= [Class1, Class2, Class3]T [11] [12] [13]:

$$Logsig(x) = \frac{1}{1-e^{-x}}. \tag{1}$$

*3.3.2 Training algorithm.*

The training algorithm " Conjugate Scaled Gradient " [10] is used. The performance index is [14]:

$$H_{y'}(y) = \sum_i y_i' \log(y_i), \tag{2}$$

Where:
$y'$ is the probability of the expected class i,
$y'$ is the estimated probability.

*3.3.3 Experimental results.*

The following results were obtained when training for 52 seasons. Figure 1 shows the performance index for each training season. The calculation of this index is through (2), which in the last period has a value of 0.0012458.

In Figure 2, three sets of performance index values are shown for the set of data taken for training, validation and testing. It should be noted that at the 42nd epoch the minimum performance index equal to 0.18854 was obtained for the validation data set, this information is taken as a means of stopping the training algorithm.

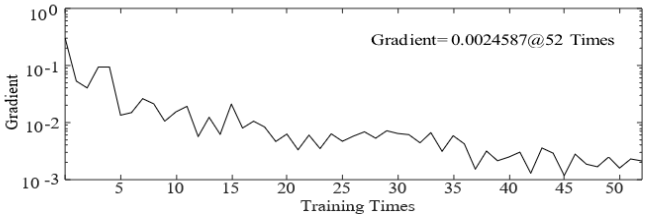


Fig. 1. Performance Index Behavior.

To assess the quality of the classifier in Figures 3, 4 and 5, there are the confusion matrices and the ROC "Receiver operating characteristic" graphs that allow the quality of the class estimate to be assessed. During the training in Figure 3, a correct classification percentage is observed for class 1 of 80.7%, for class 2 of 70.9% and for class 3 of 63.4%, resulting in an average percentage of 77.3%. When observing these values for the validation and test sets, a correct estimation percentage of 88.3% is observed, see Figures 4 and 5.

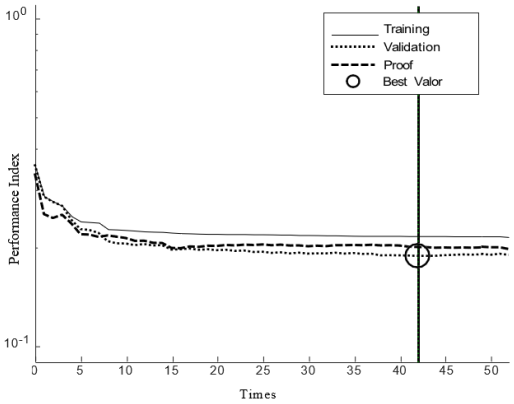


Fig. 2. Behavior of the performance index with respect to the phases of Training, Validation and Test of the Neural Network.

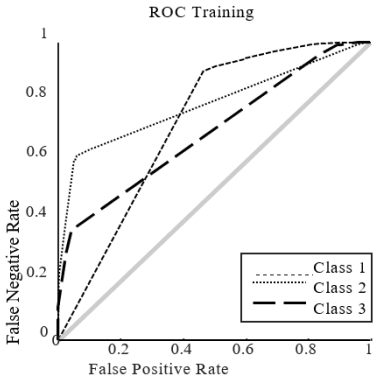


Fig. 3. Receiver operating characteristic (ROC Training) graph: Note that in all the results the sensitivity curves are at the top of the non-discrimination line (diagonal).

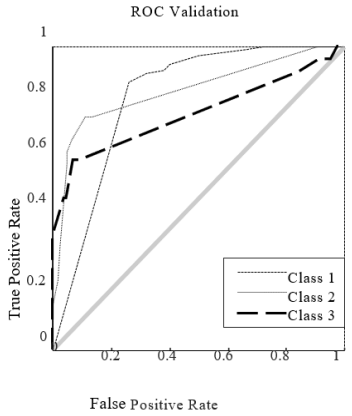


Fig. 4. Receiver operating characteristic" graph. (ROC Validation)

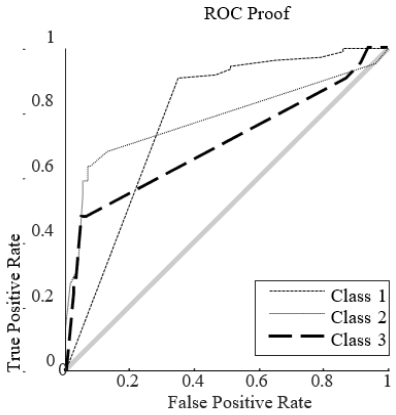


Fig. 5. Receiver operating characteristic" graph (ROC proof).

## 4. Conclusions

The classifier assembly performed better when evaluated on the test set, 88.3%, than the individual classifiers, which range between 63% and 67.5% or the classifier that considered all factors, 59.82%. From these results, it can be concluded that the classifier assembly built with the neural network performed better during the test phase.

Results are taken in the test phase because that is when the classifier is subjected to data considered during the training phase, being a better reflection of its ability to classify.

Future studies should evaluate with other assembly methods, such as: the majority vote and the decision table, among others, to assess whether performance is improved [15] [16][17][18][19].

It is also suggested to analyze the weighting given by the neural network of the assembly to identify the classifier that came closest to the correct class, and thus examine the structure of the corresponding tree, in order to detect educational factors for the excellent class and generate strategies to strengthen them.

## References

[1] 1Zhi-Hua, Z.: Ensemble methods: Foundations and Algorithms. CRC Press, Taylor & Francis Group (2012)

[2] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), pp. 37–54 (1996)

[3] Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. Morgan Kaufmann Publishers (2005)

[4] WEKA 3: Data Mining Software in Java Homepage. https://www.cs.waikato.ac.nz/ ml/weka/ (2016)

[5] Singh, Y., Chanuhan, A.: Neural Networks in Data Mining. Journal of Theorical & Applied Information Technology, 5(1), pp.37–42 (2009)

[6] Orallo, J., Ramírez, M., Ferri, C.: Introducción a la Minería de Datos. Pearson Education, (2008)

[7] Khasawneh, K., Ozsoy, M., Ghazaleh, N., Ponomarev, D.: EnsembleHMD: Accurate Hardware Malware Detectors with Specialized Ensemble Classifiers. IEEE Transactions on Dependable and Secure Computing, pp. 10 (2018)

[8] Yan, Y., Yang, H., Wang, H.: Two simple and effective ensemble classifiers for twitter sen- timent analysis. Computing Conference 2017, pp. 1386–1393 (2017)

[9] Vogado, L., Veras, R., Andrade, A., Araujo, F., Silva, R., Aires, K.: Diagnosing Leukemia in Blood Smear Images Using an Ensemble of Classifiers and Pre-Trained Convolutional Neural Networks. 30th (SIBGRAPI) Conference on Graphics, Patterns and Images, pp. 367– 373, Niteroi (2017)

[10]Hestenes, M., Stiefel, E.: Methods of Conjugate Gradients for Solving Linear Systems. Journal of Research of the National Bureau of Standards, 49(6), pp 409–436 (1952)

[11]C. & Sotelo-Figueroa, M. A. Castillo, O.; Melin, P.; Pedrycz, W. & Kacprzyk, J.: Generic Memetic Algorithm for Course Timetabling. In: ITC2007 Recent Advances on Hybrid Approaches for Designing Intelligent Systems, Springer, vol. 547, pp. 481–492 (2014)

[12]Aladag, C., & Hocaoglu, G.: A tabu search algorithm to solve a course timetabling problem. Hacettepe journal of mathematics and statistics, pp. 53–64 (2007)

[13].Moscato, P.: On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts: Towards Memetic Algorithms. Caltech Concurrent Computation Program (report 826) (1989)

[14]Frausto-Solís, J., Alonso-Pecina, F., & Mora-Vargas, J.: An efficient simulated annealing algorithm for feasible solutions of course timetabling. Springer, pp. 675–685 (2008)

[15]Joudaki, M., Imani, M., & Mazhari, N.: Using improved Memetic algorithm and local search to solve University Course Timetabling Problem (UCTTP). Doroud, Iran: Islamic Azad University (2010)

[16] Viloria, Amelec; Lezama, Omar Bonerge Pineda. Improvements for Determining the Number of Clusters in k-Means for Innovation Databases in SMEs. Procedia Computer Science, 2019, vol. 151, p. 1201-1206.

[17] Kamatkar, S. J., Kamble, A., Viloria, A., Hernández-Fernández, L., & Cali, E. G. (2018, June). Database performance tuning and query optimization. In International Conference on Data Mining and Big Data (pp. 3-11). Springer, Cham.

[18] Viloria, Amelec, et al. Integration of Data Mining Techniques to PostgreSQL Database Manager System. Procedia Computer Science, 2019, vol. 155, p. 575-580.

[19].Viloria A., Lis-Gutiérrez JP., Gaitán-Angulo M., Godoy A.R.M., Moreno G.C., Kamatkar S.J. (2018) 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. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham.