Unconventional Computing for Estimating Academic Performance

Carolina Fócil Arias, Amadeo José Argüelles Cruz, and Itzamá López Yáñez

Instituto Politécnico Nacional, Centro de Investigación en Computación,
Mexico City, Mexico
Cfocil_a13@sagitario.cic.ipn.mx, jamadeo@cic.ipn.mx, ilopezy@ipn.mx

Abstract. One of the relevant tasks of the teaching-learning process is that of evaluation. In this sense, estimating the academic performance exhibited by the final evaluation of a student, at the end of the current semester, has become of particular interest for students, parents, educators, educative managers and researchers alike. More specifically, the current paper is focused in determining whether a first semester engineering student at the Universidad Tecnológica de Pereira, Colombia, will pass or fail the Mathematics I course. This paper proposes the use of the Gamma Classifier together with Wilson Editing to solve the problem. Results observed are competitive against classic models of pattern recognition.

Keywords: Gamma classifier, academic performance estimation, Wilson Editing.

1 Introduction

The target of the institutions of higher education is to provide quality education to the students [1], yet high rates of dropouts and poor performance are a current problem for universities throughout the world.According to [1] despite all efforts to prevent freshmen desertion in Colombia, there are records that student desertion in the period 1994 to 2004 is on average 52%.

Based on this number from 2005; various tests were done to new students [1-2]. The aim is to record all information obtained from the tests and grades that students obtained at the end of semester in Universidad Tecnológica de Pereira.

We propose to use this dataset, which is based in a thorough study of student performance, and implement an associative model known as Gamma Classifier together with the Wilson Editing to determine with high probability whether the student will pass or fail the Mathematics I course.

The paper is structured as follows; the second section lists selected work related with classification. The third section presents materials and methods for the analysis. The results obtained from the experiments are presented in the fourth section. The conclusions and future work are presented in the last section.

2 Related Work

There are many works about education with different approaches to evaluate the academic performance. The University in Rajasthan (located in India) used a decision tree method to develop a classification task and evaluate the student performance. The University used information such as test, seminars and attendance [1].

Another interesting work was presented in [2]. In this work the researchers developed a system that predicts the success of students in online courses. This system is able to predict the student's performance (with a 70% accuracy) by mining the data recorded on 8 days.

An important problem related with the student's performance was presented in the work [3], where different algorithms were used such as One-R, C4.5, ADTrees, Naive Bayes and Bayes Net. Results indicate the factors such as family background and family's social-economical status, high school GPA and test scores impact in the student's decision to continue or drop out of college.

Studies in Colombia determine the factor, which influence in the final score in Mathematic I course. The results indicate with a 70.4% accuracy if a student passes or fails the course, by using a multiple logistics regression model [4].

Another classification method used in education is Support Vector Machines (SVMs). Dursun Delen [5] developed analytical models to predict and to explain the reasons behind freshmen student attrition. The results were that SVM's produced the best results with an overall prediction rate of 87.23%.

3 Materials and Methods

In this study, we propose use of the Gamma classifier and Wilson Editing to classify whether the students pass or fail the Mathematics I course and to improve classification of student's final performance.

Firstly, the tools to be used are hardware: Processor: 2.9 GHz Intel Core i5 and Memory: 8GB 1600 MHz DDR3 and Software: Eclipse 4.3.2, Python pydev 3.3.3, OS X 10.9.3, WEKA 3.6.10. Secondly, dataset is requested Universidad Tecnológica de Pereira in Colombia. Next, instance selection process is applied using data preprocessing and Wilson Editing. Finally, classification models are used and compared to versus Logistic Model Regression.

3.1 Data set

The data set for this study was collected and consolidated by Universidad Tecnologica de Pereira located in Colombia with an enrollment of 919 students, 29 attributes and 2 classes (class 0 which corresponds to fail has 515 patterns and class 1 which corresponds to pass has 404 patterns).

The dataset contains variables related to student's performance, social characteristics and student's health for a semester. Table 1 details a complete list of variables obtained from the student dataset.

However based on work done [4], we decided to use the same variables. (See table 2).

Table 1. Avalaible attributes (Source [4])

Factor	***Variable***	***Description***
Personal	*Sex*	*Sex*
	Age	*Age*
Socioeconomic	*Tipocole*	*Type of school*
	Estrato	*Social stratum*
Academic	*ICFES*	*Score from Instituto Colombiano para la Evaluación de la Educación*
	Subject1	
	Subject2	
	Subject3	
	Subject4	
	Subject5	
	Average	*Average*
	Vliteral	*Quantitative Literal Reading*
	Cliteral_1	*Quantitative Literal Reading*
	Cliteral	*Qualitative Literal Reading*
	Cinferen	*Qualitative Inferential Reading*
	Ccritico	*Qualitative Critical Reading*
	Cabstract	*Abstract Logical Thinking*
	Vinferen	*Quantitative Inference Reading*
	Vcritico	*Quantitative Critical Reading*
	Cverbal	*Verbal Logical Thinking*
	Clogico	*Logical Thinking*
Institutional	*Codprog*	*Program code*
Risks	*Rsalud1*	*Health coverage*
	Rsalud2	*Physical health*
	Rsalud3	*Nutrition*
	Rsalud4	*Mental Disturbance*
	Rsalud5	*Factor for Psychoactive Substance*
	Rsalud6	*Free time*
	Rsalud7	*Other responsibilities*

3.2 Preprocesssing data

Data preprocessing is an important step for improving data quality. This step helps to increase the accuracy and efficiency of the classifer when the data tend to be incomplete and inconsistent [6]. Thanks to this process, the data with which the classification will be carried out is through qualified data processing [7]. Data preprocessing methods are divided into the following categories [8]: Data cleaning, Data integration, Data transformation, and Data reduction. In our case, we have applied Data cleaning.

Data set has incomplete and inconsistent data, therefore, we replace all missing values using the mean or the mode of each attribute that belongs to a certain class. The measures used for each attribute are given in Table 2.

Table 2. Attributes obtained from student's records and the measures used to replace missing values in each attribute

Attribute	***Type***	***Measure used***	***# Replacement***	
			Class 0	**Class 1**
ICFES	Number	Mean	37	5
Codprog	Categorical	Mode	0	0
Casbtract	Categorical	Mode	116	43

In this study we have decided using Wilson Editing to delete patterns that are misclassified by the KNN rule (K=3) and to increase the accuracy of classifier [9]. The Wilson Editing algorithm is shown below [10].

```
Initialization: S←X
For each pattern xi∈X do:
         If it is misclassified using the KNN (k=3) rule with proto-
types in X-{xi}
         then S←S-{xi}
```

Wilson Editing is based on the distance between patterns to determine their similitude [11]. For this study, we used Manhattan distance (Eq. 1).

$$d(x,y)=\sum_{i=1}^{n}\left|x_i-y_i\right| \tag{1}$$

3.3 Classification

In this work, we chose an algorithm called Gamma Classifier [12]. This algorithm is based in the Gamma similarity operator. The sequences steps of Gamma Classifier are shown in figure 1 and figure 2.

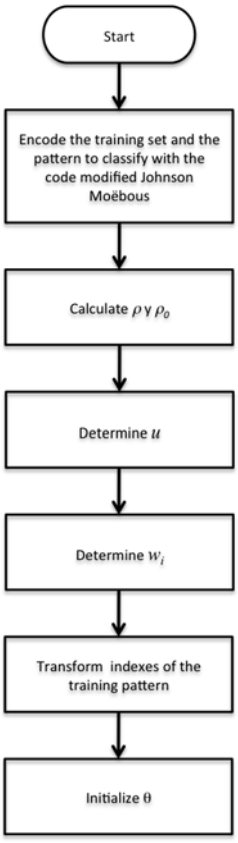

Fig. 1. Functions of Gamma Classifier part I (Source [12-13])

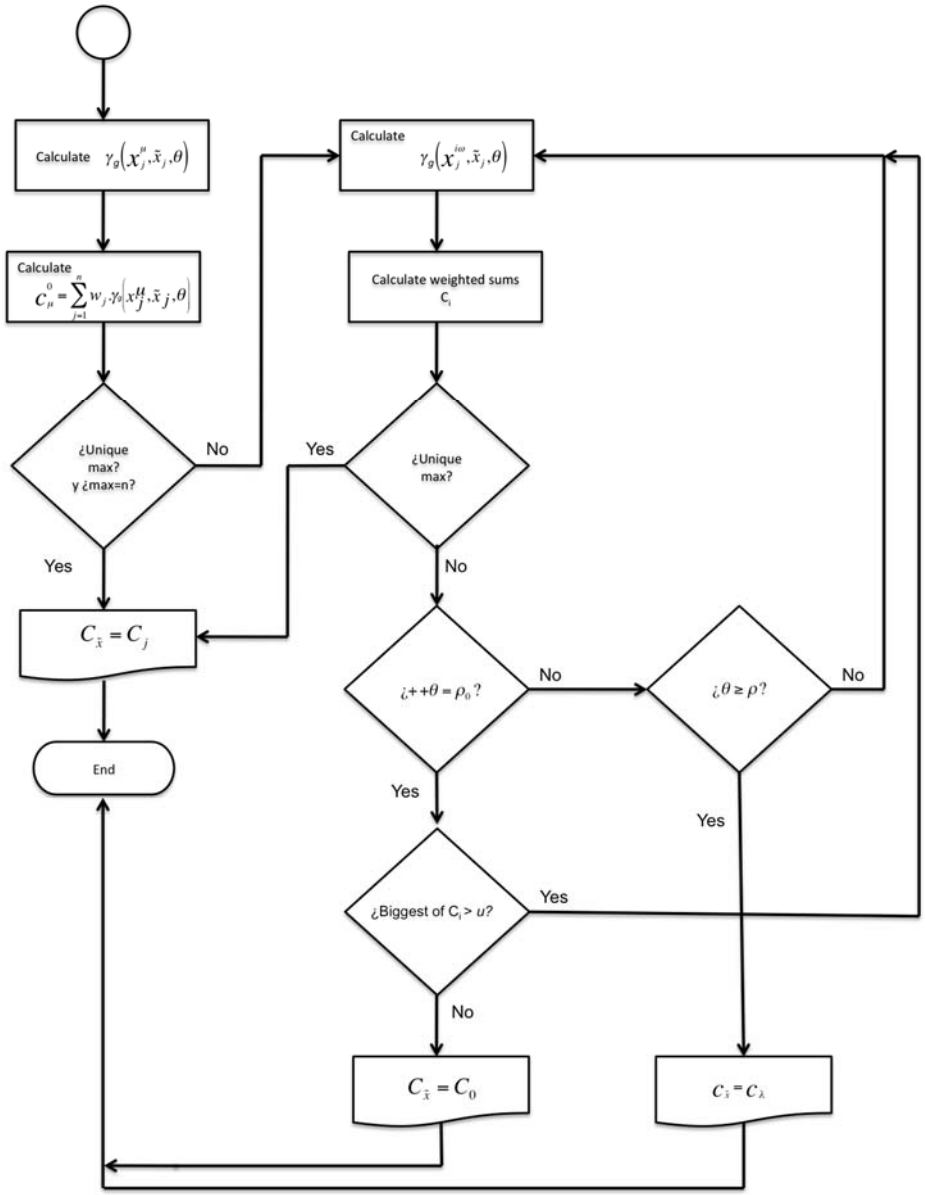

Fig. 2. Functions of Gamma Classifier part II (Source [12-13])

For more details about this classifier it is necessary to consult works [12] and [13].

4 Experiments

In this section, we present the experiments using Gamma classifier and Wilson Editing. We have also included experiments made in WEKA [14] using algorithms such as Naive Bayes, Bayes Net, Support Vector Machines, Simple logistic, and J48. These algorithms were selected based in the background (see section 2).

Firstly, we replaced each data, which has a missing value at least in a feature using mode or mean for each class because Wilson Editing does not accept tuples with Null value (See table 2).

Secondly, to improve the classifier's performance, it is necessary the weight assignment to the features. Based on work done [13] and the feature's graphs, we have decided to assign the following weights.

In Figure 3.A, we can observe that class 0 (students do not pass Mathematics I course) and class 1 (students pass the Mathematics I course) are not separated in cabstract feature because patterns almost have the same values, however, we can not ignore this feature then we assign the value 0.1. In codprog feature occurs the same situation as cabstract feature, classes are not separated, we can see that patterns, which belong to different classes, have very similar values and thus the classes are not separated and we can not ignore this feature the we assign the value 0.1 (See figure 3.B.)

The last one ICFES feature has some patterns, which it is possible to separate, thus, we decide to assign the value 1. This feature is shown in figure 3.C.

However, we performed several experiments varying each feature's weights until we obtained the best results.

Thirdly, using Wilson Editing to improve classifier performance indicates that 662 out of 919 tuples (class 0 has 302 patterns and class 1 has 362 patterns) were selected. Thus, were ignored 257 patterns of dataset.

The results using algorithms such as Gamma, Naive Bayes, Bayes Net, SMO, IBK, Logistic and Simple Logistic and Logistic multiple regression of this experiment are shown in figure 4.

5 Results

In this work, we used the data set, which was composed of three features such as ICFES, *codprog* and *cabtsract* and 662 records. Based on the stratified 10-fold cross validation, Gamma classifier and Wilson Editing produced the result with an overall classification rate of 77.3413% below of the support vector machines, Bayes Net, J48 and KNN. However, Gamma classifier obtained better results than Logistic Multiple Regression with an overall classification rate of 70.4%. The last one result was published in [4].

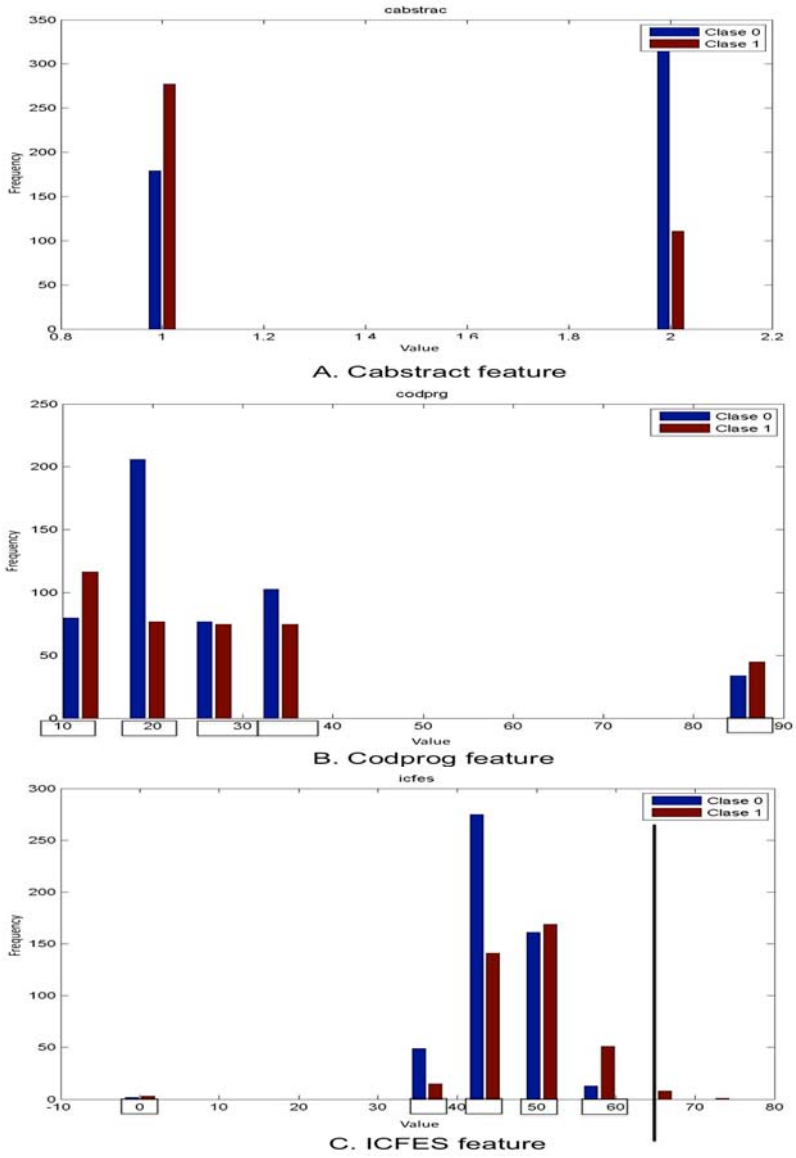

Fig. 3. Histograms of each feature used in this study

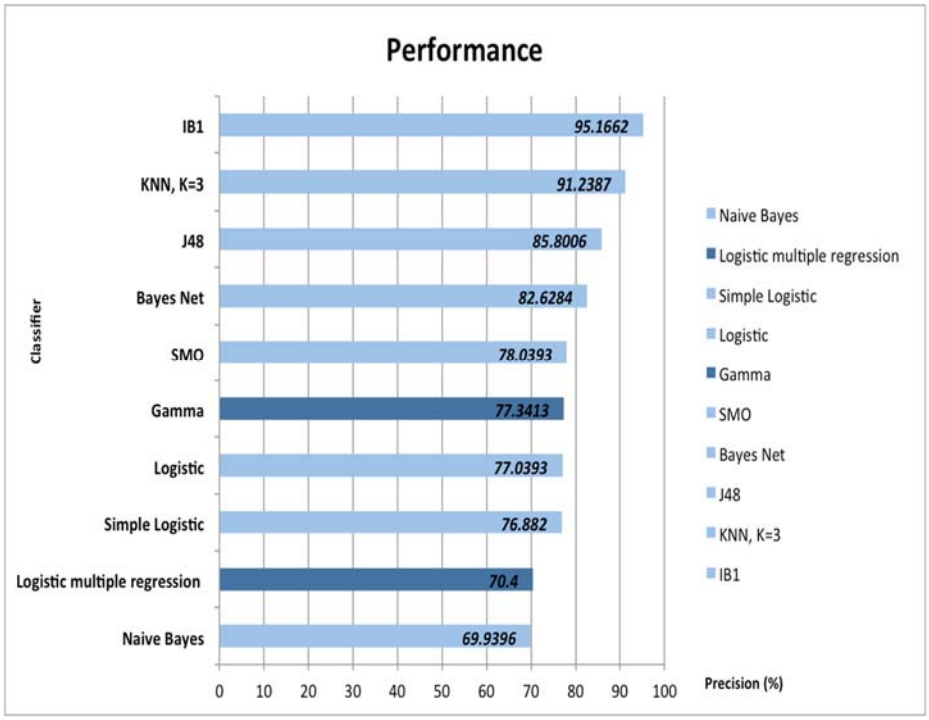

Fig. 4. Comparison graph between result of Gamma classifier and 8 algorithms in WEKA

6 Conclusions

Previous work [4] is outperformed from 70.4% to 77.3413% using Gamma classifier and Wilson Editing for identifying the students who are likely to fail Mathematics I course and to drop out the school in the first semester; regardless of unbalanced dataset. As a future work, other methods to deal with unbalanced dataset should be explored, as well as using other data preprocessing data techniques and use feature selection techniques to improve performance of Gamma classifier.

Acknowledgments

The authors would like to thank the Instituto Politécnico Nacional (Secretaría Académica, COFAA, SIP, CIC, and CIDETEC), the CONACyT, and SNI for their economic support to develop this work.

References

1. Baradwaj, B. and Pal, S.: Mining educational data to analyze student's performance. In: Int. J. od Adv. Comput. Sci. Appl., vol. 2, no. 6, pp. 63–69, Jan. 2012
2. Johnson, J. A.: RMAIR Best Paper : Ethics of Data Mining and Predictive Analytics in Higher Education . In: Association for Institutional Research Annual Forum, Long Beach, California, 2013, pp. 1–25
3. Nandeshwar, A., Menzies,T., and Nelson, A.: Learning patterns of university student retention. Expert Syst. Appl., vol. 38, no. 12, pp. 14984–14996, Nov. 2011
4. Carvajal, P., Mosquera, J. C., and Artamonova, I.: Modelos de predicción del rendimiento académico en Matemáticas I en la Universidad Tecnológica de Pereira. In: Sci. Tech., vol. 43, no. 43, pp. 258–263, Dec. 2009
5. Denle, D.: A comparative analysis of machine learning techniques for student retention management. In: Decis. Support Syst., vol. 49, no. 4, pp. 498–506, Nov. 2010
6. Hernández G., C. and Rodríguez R., J.: Preprocesamiento de datos estructurados. In: Vínculos, vol. 4, no. 2, pp. 27–48, Apr. 2008
7. Buldu, A. and Ücgiin, K.: Obtaining The Relation Between The Courses By Using Data Mining Application. In: Tech. Technol. Educ. Manag., vol. 7, no. 2, pp. 532–538, Feb. 2012
8. Olson, D. L. and Denle, D.: Advanced Data Mining Techniques. 1era. ed. Estados Unidos. Springer, 2008, pp. 141–143
9. C. F. Eick, N. Zeidat, and R. Vilalta: Using Representative-Based Clustering for Nearest Neighbor Dataset Editing. In: Fourth IEEE International Conference on Data Mining (ICDM'04), 2004, pp. 375–378
10. Vázquez, F., Sánchez, J. S., and Pla, F.: A Stochastic Approach to Wilson ' s Editing Algorithm. Springer, vol. 8, no. 4, pp. 35–42, 2005
11. I. Triguero, J. a. Sáez, J. Luengo, S. García, and F. Herrera: On the characterization of noise filters for self-training semi-supervised in nearest neighbor classification. In: Neurocomputing, vol. 132, pp. 30–41, May 2014
12. López-Yáñez, I.: Clasificador Automático de Alto Desempeño. Tesis de Maestría. Centro de Investigación en Computación, IPN, México, D.F., 2007
13. López-Yáñez, I.: Teoría y aplicaciones del clasificador Asociativo Gamma.Tesis de Doctorado. Centro de Investigación en Computación, IPN, México, D.F., 2011
14. Witten, I. H. and Frank, E.: Data Mining Practical Machine Learning Tools and Techniques. 2nd. ed. San Francisco: Morgand Kaufmann Publishers, 2005, pp. 62–76