ISSN 1870-4069

Analyzing the Effect of Family Factors over the Scholar drop out at Elementary School using Classification and Association Rules Techniques

Silvia Beatriz González, Josué Figueroa

Universidad Autónoma Metropolitana,
Ciudad de Mexico, Mexico

{sgb,jfgo}@correo.azc.uam.mx

Abstract. Data Mining offers great opportunities for analyzing data related with several themes, one of the most interesting is the educative environment, which has a lot information about many areas which can be improved. Scholar drop out, it's one of the biggest problems that education faces, but the great amount of factors that cause it, make it difficult for analyzing. This work presents an analysis of the elementary school drop out problem using techniques like decision tree and generation of association rules for obtaining models that allow to identify the most important family aspects that causes scholar desertion.

Keywords. Association rules, decision trees, educational data mining, mining educational data, studies drop out.

1 Introduction

Data Mining (DM) focuses in analyzing big volumes of information in order to obtain patterns or knowledge about different topics for explaining, classifying or predicting some kind of phenomena and helping in making decisions. In the last decades, there has been a rise in the quantity of available information related with sectors like commerce, finances or consumer preferences. Also, the use of technology in the educative environment has also risen, so the amount of data and information related with several aspects of education. Applying technology over traditional educative systems, and other like e-learning systems have allowed gathering a lot of information about students, courses, schools and other elements; processing this information using DM techniques has created a relatively new branch of DM called Educational Data Mining (EDM). EDM includes the use of DM techniques for discovering patterns in large amount of information generated in an educative environment. This concept first appeared in year 2000 with a set of conferences, but it has been a rise in the amount of researching and works related with this topic [12]. Educational environment offers several opportunities for applying techniques for discovering knowledge, like: predicting students performance, planning courses, clustering students according certain

features, etc.; being one of the most worked areas, the prediction of students' performance over their studies, a single class or even an exam or exercise.

In any educative system, probably the biggest problem is related with the scholar drop out, which is present since the elementary, until superior level. The main issue, is that there are a lot of factors which can cause that a student leaves its studies. Family, personal, social, academic or labor factors affects all students, and identifying the most important could be really difficult due to the great amount of information which should be processed. Considering this, the main goal of this work is to apply two of the most common DM techniques, classification and generation of association rules, for obtaining predictive models and rules among events or situations which allow to understand the effect of family factors over the scholar drop out at elementary school.

2 Knowledge Discovery using Data Mining

Discovering knowledge requires a set of stages that makes easier the process of finding interesting patterns over data. This process, called Knowledge Discovery on Databases (KDD) is composed by different steps [7]: obtaining information, cleaning process or data cooking, use of DM techniques and interpreting for obtaining knowledge. In the DM stage, the applied techniques depend on the goal to be reached, in general these techniques are: classification, clustering, regression and association rules, having each on of these, different algorithms and techniques for generating results [8].

2.1 Classification

Classification or supervised learning, is on the most used DM techniques; it uses a learning scheme where, using a set of classified data (training data), a model which allows predicting new not classified data, is generated. This two steps are called training or learning stage and the classification stage. Most common algorithms used in classification technique are: decision trees, neural networks and Naive Bayes; once a model has been generated, its efficiency it's measured using another set of data (test data).

2.2 Decision Trees

A Decision Tree is one of the most used predictive models in DM. In a Decision Tree, leafs represents a decision or classification and branches a set of characteristics that lead to a particular decision. As a tree, it has a root node, called the best predictor, this means that it's the most important factor or variable for taking a decision about a classification. Trees are used in DM for obtaining models which predicts the value of a variable, called decision. Once the model is constructed, the characteristics of a non classified data are taken and depending their values, a path is followed from the root to a leaf that indicates the corresponding class.

At the moment of constructing a Decision Tree, the main problem is to decide which of the variables will be the root node, and then, decide the order of the rest of the variables. The nearer a node is to the root, represents that it is more important for assigning a classification. This problem is solved by finding a variable that better divides the target destiny considering the purity of its children set. Purity refers to how mixed are the goal values in a node. Purity measure is known as information, and the concept of impurity is known as Entropy, and is defined as follows:

$$\sum_{i=1}^{k} P(C_i|D)\, log_k\, P(C_i|D)), \tag{1}$$

where:

$$P(C_i|D) = \frac{amount\ of\ oservations\ in\ D\ with\ value\ C_i}{amount\ of\ oservations\ in\ D}. \tag{2}$$

Another concept used during the generation of a Decision Tree besides Entropy is the Information Gain and it's based in the decrease of Entropy after a set of data has been divided. Information Gain is defined as:

$$G(S, A) = Entropy(S) - \sum_{v=values(A)} \frac{S_v}{S}\, Entropy\,(S_v), \tag{3}$$

where *values(A)* represents the set of all the possible values of the attribute or property A and Sv is the sub set of elements in S from which the attribute A has a value of v.

2.3 Measuring the Classification Efficiency

Data for generating a Decision Tree are be divided in two: training or learning data and test data. The model is generated using the training data and then, it's tested using the test data for determining how efficient is. The recommended percentage for each one is 70% to 80% for training and 20% to 30% for testing. The model determines a value for the decision variable (predicted value) of a test data and this is compared with the real value. Comparing predicted value and real value for all the elements of the test data set, can be calculated the efficiency of the tree.

2.4 Association Rules

Association rules are used for showing the relationships that exists in a set of items. In a formal way, an association rule is defined as: let I = $\{I_1, I_2, \ldots, I_m\}$ a set of attributes known as items, and T a set of transactions $\{t_1, t_2, \ldots, t_n\}$ represented as t[k] = 1 if t is related with I k and t[k] = 0 otherwise. Let X a set of some of the elements in I, a transaction satisfies X if for all the elements I_k in X, t[k] = 1. An association rule is an implication represented as X ⇒ I ij where

X is a set of some of the elements in I and I_{ij} is an element of I which is not present in X. In this way, the rule X $\Rightarrow I_{ij}$ is satisfied in the set of transactions T, if certain percentage of transactions in T that satisfy X, also satisfy I_j[1].

2.5 Measuring the Importance of a Rule

It's very common that a great amount of association rules are generated, most of them can be redundant or not significant. For this reason, measures for knowing the importance of a rule have been developed [13]. From this measures, the most used are: Support, Confidence and Lift.

Support of a set of elements A represents the percentage of transactions which contains A in a set of transactions T. Support is defined as:

$$support(A) = \frac{|A|}{|T|}. \tag{4}$$

Confidence is the amount of transactions which contains A as an antecedent and B as a consequence. A can represent a single element or a set of elements. Formally, confidence is defined as:

$$confidence(A \Rightarrow B) = \frac{support(A \cup B)}{support(A)}. \tag{5}$$

Lift represents the occurrence frequency of A and B respect an expected value, considering that A and B are independent. Lift is defined as:

$$lift(A \Rightarrow B) = \frac{support(A \Rightarrow B)}{support(A) * support(B)}. \tag{6}$$

The value of lift shows how strong is the relationship among A and B according to:

- If Lift < 1, it's considered that A and B have a weak relationship and are called substitutes.
- If Lift $= 1$, it's considered that a relationship among A and B is because a random behavior.
- If Lift > 1, it's considered that A and B have a strong relationship and are called complements.

3 Related Works

One of the most recurrent researches in EDM is related with using classification techniques for generating models which classify or predict the behavior of students during their studies, a course, exam or other activities.

In [3] it's analyzed the data from students, which is obtained from a data warehouse, using a decision tree and other techniques for generating a model that predicts their performance considering their verbal and mathematical abilities.

The efficiency of all the techniques is compared. In [11] some classification techniques are used for generating models which predict the final score of students in a certain course based on their personal, social and academic characteristics. In [4] some of the algorithms for creating decision trees are used for generating models which predict the performance of students considering academic, familiar and demographic factors. In [10] generation of association rules are used for analyzing the effect of teaching in different languages over the amount of regis tered students in several courses. In [5] are identified the elements which have in common the best students in a course using association rules and other DM techniques. In [9], the phenomena of dropping out an on-line course is analyzed using decision trees. A model for identifying the students which have greater risk of dropping the course is generated.

4 Knowledge Discovery

For obtaining the prediction model and the association rules, were followed the steps considered in the KDD process. Data was obtained from the National Poll about Scholar Drop out at Medium High Level [14]. Despite of the poll was focused in medium high level students, there were a lot of surveyed which didn't finish the elementary school, so they were considered for this work. Were considered 699 people, 407 didn't finish elementary school and 292 did it.

The decision tree was generated using Classification and Regression Trees algorithm (CART) [6], and for generating the association rules, was used the Apriori algorithm [2], this was performed with the software R.

4.1 Pre-processing the Information

The goal of this stage is preparing the information for applying DM techniques. The poll had questions grouped in three aspects: personal and family, academic and labor. For this work, only were considered the personal and family factors; initially were considered 19 variables and one decision variable. For helping the interpretation of the models, it was assigned a letter for each variable, variables, including their letter and possible values are shown at Table 5 at the end of document.

4.2 Classification using Decision Trees

It was chosen the variable ZZ (Studies finished) to be the predicted variable with two possible values, Yes, if the student finished their studies or No, otherwise. 75% of the data (524) were used for the training process and 25% (175) for testing. The set of rules obtained from the model is presented in Table 1, and the Decision Tree obtained is shown in Figure 1.

The represented variable by each letter can be reviewed in Table 5. The values of S correspond to:

– NO: Didn't finish Elementary School.

Table 1. Rules obtained from Decision Tree.

Rule
IF L = "NO", THEN Finish = "NO"
IF L = "YES" AND S = "ME" OR "MES" OR "SUP", THEN ZZ = "YES"
IF L = "YES" AND S = "NO" OR "EL" AND F = "YES", THEN ZZ = "NO"
IF L = "YES" AND S = "NO" OR "EL" AND F = "NO" AND O = "YES", THEN Z = "NO"
IF L = "YES" AND S = "NO" OR "EL" AND F = "NO" AND O = "NO", THEN Z = "YES"

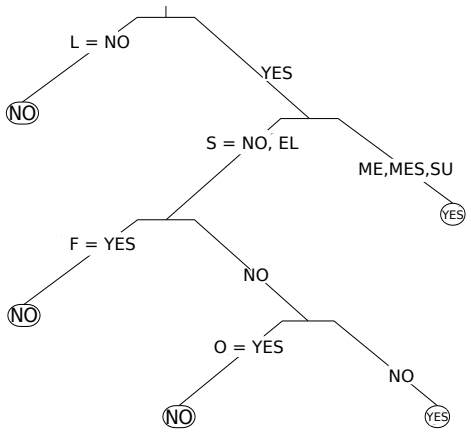

Fig. 1. Decision tree obtained after processing the information.

- EL: Finished Elementary School.
- ME: Finished Medium School.
- MES: Finished Medium High School.
- SU: Started or Finished Superior School or higher.

The efficiency of the model was measured comparing the predicted value using the tree, against the real value for the test data. The results are shown on Table 2.

Table 2. Measuring the efficiency of the model.

	No	Yes
No	80	20
Yes	27	48

This represents that, from 100 cases with "NO" value, 80 were classified in a right way, and 20 in a wrong way. 27 cases with "YES" value were assigned by the model to "NO" and 48 were predicted as "YES". Considering these results, the efficiency of the model was 73.142%.

4.3 Obtaining Association Rules

Were generated rules for the possible values of ZZ (Studies finished). The antecedents for each possible values are shown in Table 3 for Finishing and, in Table 4 for Not Finishing. The importance of a rule was measured using the lift value.

Table 3. Antecedents related with Finishing the Elementary School.

Rule	Confidence	Lift
{F="NO",L="YES",N="NO",T="ME"}	0.8301887	1.987335
{C="NO",L="YES",N="NO",T="ME"}	0.8301887	1.987335
{L="YES",N="NO",O="NO",T="ME"}	0.8269231	1.979518
{L="YES",N="NO",O="NO",T="ME"},	0.8200000	1.962945

Table 4. Antecedents related with Not Finishing the Elementary School.

Rule	Confidence	Lift
{E="NO",L="NO",O="YES",S="NO"}	0.8409091	1.444215
{B="NO",F="YES",R="NO",T="EL"}	0.8409091	1.444215
{B="NO",D="NO",F="YES",S="NO"}	0.8367347	1.437046
{L="DIRECT",B="YES",O="NO",T="NO"},	0.8367347	1.425803

The represented variable for each letter can also be reviewed in Table 5. The values for T are the same than the ones for S in the Classification with Decision Trees section.

4.4 Interpreting Models

Once the models and rules were obtained from the decision tree and the association rules, the results must be analyzed. It's interesting that the variable L (considering studies few important) with a value of "NO" has a lot of importance for dropping out the school, it is expected that the opposite occurs, it's supposed that considering the studies important leaves to continue them, but at elementary school, maybe the interest for studying of a kid it's not a decisive factor. In this case, it's more relevant aspects related with the parents, both techniques

show the importance of the studies of the parents. In the decision tree, values for S (studies of the father) that have a higher level than elementary school are related with finishing school, the same occurs with association rules, where the value of T (studies of the mother) is higher than elementary school it's related with the consequence of finishing. Having not studies or only the elementary school for both or any of the parents is related with dropping the school. The F variable (lack of money) is also related with not finishing the school in both techniques; indicating that it could be a relevant factor in the scholar drop out. The same occurs with variable O (brothers that dropped out the school),in the decision tree and the rules associated with not finishing the studies, this value is present indicating that if a brother or sister abandoned it's studies, its probable that the other can do it. Other variables which appear on the rules are not present in the decision tree, but several of the ones with the highest value of lift do it. Considering this, it can be concluded that both models are right and can help in studying the problem of scholar drop out at elementary school.

Although the results of both models are congruent, the percentage of efficiency of the Decision Tree can be considered low, a good percentage should be above 90%. This may be because of the small quantity of data, with more data, there are more combinations for generating better models.

5 Conclusions

Applying Data Mining Techniques to an educative environment, offers a big amount of opportunities for studying a lots of aspects that occurs in education. From the performed work, it can be concluded that depending the desired goal, the appropriate technique of Data Mining must be chosen. Following the the KDD process can be helpful for easing the whole process. It's important to fulfill a right cleaning and pre-processing of the data for obtaining better results independently the used technique. Also, testing different sets of data it's necessary for obtaining better models, this specially related with Decision Trees, but also testing different parameter values for applying them to the association rules algorithm helps in obtaining better results. As mentioned, the amount of information is an important topic, having more quantity would help in obtaining better models. Related with the quality of the models, having an adequate way for measuring the efficiency of each model or set of rules it's essential for acquiring the right knowledge. In the obtained Decision Tree, the efficiency can be considered low, however some of their nodes also appear as relevant rules in the association rules model, which validate the obtained results.

Having the models or rules it's not the end of the process, a stage of interpreting those results it's necessary for finally obtaining knowledge that helps in making decisions about a certain problem.According the obtained results in each model, can be concluded that promoting a higher level of studies in the parents (before, or even, having children) could reduce the elementary school desertion problem. Also politics for improving the income for homes can be a good solution.

Table 5. Personal factors for rule generation.

Key	Variable	Value
A	While studying, lived with	Alone
		Direct
		Other relatives
		Friends
		Own family
B	Desire for continuing studying	Yes, No
C	Influence of parents for dropping out	Yes, No
D	Influence of other relatives for dropping out	Yes, No
E	Influence of friends for dropping out	Yes, No
F	Lack of money in home	Yes, No
G	Low desire for studying	Yes, No
H	Bullying	Yes, No
I	Problems in home	Yes, No
J	Preference for studies of other relatives	Yes, No
K	Serious illness or decease of a relative	Yes, No
L	Considering studies few important	Yes, No
M	Low self steem	Yes, No
N	Closest friends dropped out school	Yes, No
O	Brothers or sisters dropped out school	Yes, No
P	Cigar consume	High
		Medium
		Low
		No
Q	Alcohol consume	High
		Medium
		Low
		No
R	Drugs consume	High
		Medium
		Low
		No
S	Father's level studies	No
		Elementary
		Medium
		Medium Superior
		Superior
T	Mother's level studies	No
		Elementary
		Medium
		Medium Superior
		Superior
ZZ	Studies finished	Yes
		No

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. Acm sigmod record ACM. 22(2), 207–216 (1993)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: 20th int. conf. very large data bases, VLDB, 487–499 (1994)
3. Agarwal, S., Pandey, G. N., Tiwari, M. D.: Data mining in education: data classification and decision tree approach. International Journal of e-Education, e-Business, e-Management and e-Learning, 2(2), 140–144 (2012)
4. Bhardwaj, B. K., Pal, S.: Data Mining: A prediction for performance improvement using classification. arXiv preprint arXiv:1201.3418 (2012)
5. Baradwaj, B. K., Pal, S.: Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417 (2012)
6. Fayyad, U. M., Irani, K. B.: The attribute selection problem in decision tree generation. In: AAAI. 104–110 (1992)
7. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI magazine, 17(3), 37–54 (1996)
8. Goebel, M., Gruenwald, L.: A survey of data mining and knowledge discovery software tools. ACM SIGKDD explorations newsletter, 1(1), 20–33 (1999)
9. Kotsiantis, S. B., Pierrakeas, C. J., Pintelas, P. E.: Preventing student dropout in distance learning using machine learning techniques. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Springer Berlin Heidelbergpp. 267–274 (2003)
10. Pandey, U. K., Pal, S.: A Data mining view on class room teaching language. arXiv preprint arXiv:1104.4164 (2011)
11. Ramesh, V., Parkavi, P., Ramar, K.: Predicting student performance: a statistical and data mining approach. International journal of computer applications. 63(8), 35–39 (2013)
12. Romero, C., Ventura, S.: Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 3(1), 12–27 (2013)
13. Sheikh, L. M., Tanveer, B., Hamdani, M. A.: Interesting measures for mining association rules. In: Multitopic Conference, Proceedings of INMIC 2004. 8th International IEEE. 641–644 (2004)
14. Subsecretaría de Educación Media Superior. Encuesta Nacional de Deserción en la Educación, `http://www.sems.gob.mx/sems/encuesta_nacional_desercion_ems`