ISSN 1870-4069

Discovering Relationships among Personal and Academic Factors with Academic Performance using Association Rules

Silvia Beatriz González Brambila, Josué Figueroa González

Universidad Autónoma Metropolitana, Ciudad de México, Mexico

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

Abstract. Educational environment produces a lot of information related with many aspects, besides the traditional system, the growth of online educative systems and the called e-learning, have increased the amount of data available for its analysis. Data Mining offers different techniques for processing information in order to discover patterns that can be used for obtaining knowledge which helps improving or taking decisions about certain topics. This work presents the concepts related with, and the use of the called Educational Data Mining, specifically the technique of Association Rules, for obtaining relationships among the academic and personal factors with academic performance.

Keywords: Academic performance factors, association rules, educational data mining, knowledge discover, mining scholarship data.

1 Introduction

Educational environment produces a lot of information that can help in the improvement of many aspects related with education if it is analyzed and processed in a right way. Data Mining allows analyzing big volumes of information searching for interesting patterns that allow making decisions and solving problems. Data Mining techniques have been applied frequently to commerce and customer preferences situations, however in the last decade, these techniques have been applied to an educative environment creating the called Educational Data Mining (EDM) [10].

There are many definitions about what EDM is, but in a general way, it can be understood as the development of techniques and processes that allow discovering knowledge from data generated in an educational environment [8]. EDM uses the same techniques that Data Mining, in general, these techniques are: classification, clustering, regression and association rules. All of these techniques use different algorithms like decision trees, neural networks, k-nearest neighbor, Bayesian classifiers and others, for processing information and which are used depending the desired goal.

Applying these techniques to the information produced in an educational environment, allows to discover knowledge, which can improve many aspects related with education, like: courses planning, students performance during their studies or even in a single course or exam or identifying risk factors which can cause a student drop up its studies or fail in approving a certain course.

One of the problems that worry the most in an educative system is the academic performance; this one can be affected by a lot of aspects, like personal, academic or even labor ones, so it could be difficult to identify the most important events or situations that can affect the academic performance of students. Considering this, the goal of this paper is to present the use of EDM, reviewing and applying their main stages, in order to identify the relationship among different personal and academic events or situations, and discovering how they affect the performance of medium high level students.

The paper is structured as follows: Section 2 presents the necessary concepts for understanding and performing the process for obtaining knowledge using Association Rules. Section 3 presents related works. Section 4 presents the development of the knowledge obtaining process following the Knowledge Discovery in Databases (KDD) steps [4]. Finally, conclusions are presented in Section 5.

2 Obtaining Knowledge using Association Rules

The formal process of KDD is composed by several steps , although the amount of stages may vary, four steps can be identified: obtaining information, cleaning information, data mining and interpretation.

2.1 Association Rules

Association Rules are used to discover the relationships among a set of items, rules are related by to measures: support and confidence. 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 let 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 if not [1]. Let X be a set of some elements of I, a transaction satisfies X if for all the elements I_k in X, t[k] = 1. An association rule is an implication represented by X $\Rightarrow$ I_{ij} where X is a set of some elements in I and I_j is an element of I which is not present in X [1]. The most common algorithms for obtaining association rules are: Apriori, Equivalence Class Transformation (ECLAT) and Frequent Pattern (FP-algorithm). Apriori algorithm is based on identifying the frequent items in an individual way from the data set, and extend them to larger sets of items but only if they appear frequently in the set [2].

2.2 Measuring the Importance of a Rule

When the association rules are generated, it is common that a huge quantity of them appear, however, not all the rules are enough significant for taking a

decision. There are a lot of rules which are not important and can be discarded. For this reason, some concepts can be used for measuring the importance of a rule [11]. These techniques include: support, confidence, conviction, lift, leverage, coverage, correlation and odds ratio, being the most commonly used support, confidence and lift. Support of a set of elements X represents de percentage of transactions which contains X in a group of transactions D and is defined as

$$support(X) = \frac{|X|}{|D|}. \tag{1}$$

Confidence for an association X $\Rightarrow$ Y is the grade of the amount which contains X as antecedent and Y as a consequence. X can represent an item or a set of items, and it can be defined as

$$confidence(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)}. \tag{2}$$

Lift indicates the occurrence frequency of X and Y respects an expected value and is defined as

$$lift(X \Rightarrow Y) = \frac{support(X \Rightarrow Y)}{support(X) * support(Y)}. \tag{3}$$

When value of lift is = 1, it means that the relationship between observed and expected values could be a random consequence. If lift is > than 1, it represents that exists a strong relationship between an antecedent and a consequence, in this case, X and Y are complements. If lift is < than 1, it means that there is not a strong relationship and X and Y can be classified as substitutes. Note that the value for lift for X $\Rightarrow$ Y it is not necessary the same for Y $\Rightarrow$ X.

3 Related Works

The amount of works about EDM has grown in the last decade and are applied to different topics [9]. The application of EDM includes traditional system, e-learning, learning systems, courses development, students performance and others. The most common use of EDM is related with student performance using classification techniques, it is applied for classifying or predicting the behavior or results for students in courses or exams, however, many works that use generation rules as main topic also can be found.

The generation of association rules can be found in topics like identifying patterns for students which can have learning problems or identifying a relationship among some common mistakes solving certain kind of exercises [7].

In [5] are presented the generation of rules for finding the relationship among three kinds of common mistakes at the moment of solving a problem. In [3] some Data Mining techniques are applied to a set of data gathered from a university, association rules are used for identifying the relationship among variables shared by students which obtain the best notes in a course.

In [6], the main goal was to analyze different ways of measuring the efficiency of some association rules about the results obtained for some students which uses or not the material given by a professor in an on-line educative system.

4 Generating Rules and Obtaining Knowledge

In order to obtain the association rules related with the academic performance, were considered the steps of the KDD process.

4.1 Obtaining Information

The data to be processed were obtained from the Medium High Level sub Secretary of Education [12]. The information was gathered from polls answered by youths. There were 13,014 students, but 4,779 were discarded because they did not start their middle high studies, so only 8,235 were considered to be processed. The poll had a lot of questions about personal, academic and labor factors, for this work, only were considered the personal and academic ones. An example of the data present in the poll can be reviewed in the Table 1.

Table 1. Sample of raw data obtained from the poll.

$p45$	$p46$	$p47$	$p48$	$p50h$	$p50m$	$p51_1$	$p51_2$	$p51_3$	$p52_1$
999	99.8	1	4	-1	15	7	1	5	2

This sample represents the data stored and were difficult for processing and understanding, so, before generating the rules, a process for cleaning the data was necessary.

4.2 Data Cleaning Process

From all the variables (questions) included in the poll, only were considered the personal and academic ones. The data cleaning, or data cooking, process, included a transformation from continue values (numeric) to discrete values (categories) for some of the data. This, because it could be obtained better results using discrete values. For example, considering the academic factor "Travel time to school", it was converted from values that represented "less than 30 minutes", "from 30 minutes to 1 hour" or "more than 1 hour" to Low, Medium or Hight. "Last year average" was converted from continue values (8.5, 9.6) to discrete values and categories like High, Medium or Low. Some variables were discarded because most of the students did not have them, for example, the entrance exam score, which was not remembered by many students, or the final medium level score. At the end, were considered 11 and 10 academic variables.

Personal factors were: Relatives which student was living, family support with scholar problems, cigar consume, alcohol consumption, bullying, problems at home, lack of money, low self esteem, serious illness or decease of a relative, parents which finished medium high level and academic performance.

Academic factors were: Type of Medium level school, travel time to school, reason for school selection, class attendance average, last year average, scholarship, scholar personal support, understanding problems, problems related with school and academic performance.

Both groups had a variable related with Academic Performance, which was the interesting variable of the work.

4.3 Using Association Rules Algorithm for Data Mining

For generating the association rules, it was used the *Apriori* algorithm and for selecting the most representative rules, the lift value was considered. The processing was performed using R. Information was managed in its list presentation, where each row is a transaction and each column represents an item (variable). When one variable had more than one possible value, this was considered as another item. Associations had two parts, the left one or antecedent, and the right one or consequence, although were generated rules with many variables in both sides, only were considered for the interpretation step the ones with the variable related to academic performance in the right side as a consequence.

For clearing the results and helping the interpretation, for each variable, a letter (key) was assigned, the letters which appear in the rules for the academic factors and their possible values correspond to

- (A) Type of Medium level school: Private, Public, Both,
- (B) Travel time to school: Low, Medium, High,
- (C) Reason for choosing school: Academic, Economic, Personal, Assigned,
- (D) Class attendance average: Low, Medium, High,
- (E) Last year average: Low, Medium, High,
- (F) Scholarship: Yes, No,
- (G) Scholar personal support: Yes, No,
- (H) Understanding problems: Yes, No,
- (I) School related problems: Yes, No,
- (Z) Academic performance: Low, Medium, High.

Rules related with Low, Medium and High academic performance considering Academic factors are presented in Table 2, Table 3 and Table 4 respectively. The interpretation of the rules can be found in the Interpreting Rules sub section.

For Personal Factors, a set of letter and variable was also created, the key letter, variable and possible values were:

- (A) Relatives which student was living with: Alone, Direct, Couple, Other,
- (B) Family support for scholar problems: Parents, Brothers, Friends, Any,
- (C) Cigar consume: Low, Medium, High,

Table 2. Association rules considering academic factors for low academic performance.

Rule	Confidence	Lift
{A=PUBLIC,B=LOW,C=PERSONAL,D=MEDIUM,E=LOW, F=NO,G=YES,I=NO}	0.6769	4.0281
{C=PERSONAL,E=LOW,F=NO,G=YES,H=YES}	0.6721	3.9996
{B=LOW,C=PERSONAL,E=LOW,H=YES}	0.6666	3.9671

Table 3. Association rules considering academic factors for medium academic performance.

Rule	Confidence	Lift
{B=HIGH,C=ACADEMIC,E=MEDIUM,F=NO,G=SI}	0.6666	1.7168
{B=HIGH,C=ACADEMIC,E=MEDIUM,F=NO,I=NO}	0.6610	1.7022
{B=MEDIUM,C=ACADEMIC,D=HIGH,E=MEDIUM, F=NO,G=NO,I=NO}	0.65346	1.6828

Table 4. Association rules considering academic factors for high academic performance.

Rule	Confidence	Lift
{A=PRIVATE,C=ACADEMIC,D=HIGH,E=HIGH, F=NO,G=YES,H=NO,I=NO}	0.9393	2.1174
{A=PRIVATE,C=ACADEMIC,D=HIGH,E=HIGH, F=NO,G=YES,H=NO}	0.9342	2.1057
{A=PRIVATE,B=MEDIUM,C=ACADEMIC, D=HIGH,E=HIGH}	0.93023	2.0968

- (D) Alcohol consume: Low, Medium, High,
- (E) Bullying: Yes, No,
- (F) Problems in home: Yes, No,
- (G) Lack of money in home: Yes, No,
- (H) Low self esteem: Yes, No,
- (I) Serious illness or decease of a relative: Yes, No,
- (J) Parents which finished medium high level: Father, Mother, Both, Any,
- (Z) Academic performance: Low, Medium, High.

Rules related with Low, Medium and High academic performance considering Personal factors are presented in Table 5, Table 6, and Table 7 respectively.

The interpretation of these rules can be found in the "Interpreting Rules" sub section.

4.4 Obtaining Knowledge from Academic and Personal Rules

Interpreting the rules is quite simple, the left side can be interpreted as a set of actions which tends to a certain result. For example, consider the following

Table 5. Association rules considering personal factors for low academic performance.

Rule	Confidence	Lift
{D=HIGH,F=YES,G=NO,J=ANY}	1	5.9506
{C=NO,D=LOW,F=YES,H=YES, J=ANY}	0.8888	5.2894

Table 6. Association rules considering personal factors for medium academic performance.

Rule	Confidence	Lift
{B=NO,C=HIGH,D=MEDIUM,E=NO,G=NO}	1	2.5752
{B=NO,C=HIGH,D=MEDIUM,G=NO,I=NO}	1	2.5752
{A=DIRECT,B=YES,C=NO,D=NO,F=YES, G=YES,H=YES,I=NO,J=ANY}	0.90909	2.3411

Table 7. Association rules considering personal factors for high academic performance.

Rule	Confidence	Lift
{D=NO,E=NO,F=YES,H=NO,I=YES,J=BOTH}	0.8888	2.00363
{D=HIGH,F=NO,J=FATHER}	0.88888	2.00363

rule which is present in Table 7, indicating Personal factor related with High academic performance: {D=HIGH,F=NO,J=FATHER} $\rightarrow$ Z=HIGH. This can be interpreted as: High alcohol consume, not having problems in home and father finished medium high school, are related to having a High Academic Performance.

But this direct interpretation is not enough, for obtaining knowledge, the most representative rules were analyzed (with the highest value of lift) in order to find interesting patterns. For example, it was found that many (70%) of the rules with a lift value from 1.85 to 2.1 for Academic factors, have as private, the school where medium level was finished. Looking for this criteria in the rules for medium and high performance, was found that 78% of the rules related with low performance have as Public this value. Medium academic performance did not consider this factor.

Something similar occurs about the reason for choosing school, all the rules related with high academic performance which have this criteria, have as value Academic, meanwhile, several of the rules associated with low performance have as value Personal.

A relevant criteria is also the average attendance, which is related directly with the performance, high, medium, and low attendance average correspond to high, medium, and low academic performance.

From this, talking about Academic factors, the kind of school where the student studied medium level, it is a decisive factor, indicating that public schools

of this level may be having problems or not an enough academic level. Also, the fact that a student chooses a school due to personal factors should not be the best decision, promoting that the criteria for choosing a school be for its academic characteristics, should be a better option in the performance of students.

A similar analysis was performed for the Personal factors. Here, a rule indicates that the fact that any of the parents have finished medium high school, it is a factor related with low academic performance. From the rules obtained about this performance, all of them contains this value for the criteria. Something similar occurs with a medium performance, but with a high performance, all of the values for this criteria, when it appeared, have the value of Both, Father or Mother.

A factor, also present in low academic performance is having low self esteem, reviewing the rules, all the ones that have this criteria with a value of Yes correspond to a Low performance, meanwhile the value of No, it's present in the rules associated with high academic performance.

Some values are a little difficult to identify, for example, consider Alcohol consumption, which has values of High in all the rules, however, analyzing a bigger set of those ones, can be found that this value was present in more rules for low performance, and in less rules for medium and high performance, which have more rules with the value of Low or No consume.

From this analysis, it can be concluded that an important factor for having a good academic performance is not related with the student, but with the parents, promoting that adult people to finish at least this level, could improve the academic performance of their children. Also aspects like self esteem and the consumption of alcohol should be considered.

5 Conclusions

Educational Data Mining offers great opportunities for processing and obtaining knowledge from data generated in an educational environment. The amount and variety of information that can be analyzed is enormous and if it's right processed and interpreted, it can be a powerful tool for improving many of the aspects that occur in education. The goal of this work was to identify relationships among academic and personal aspects with academic performance, so the association rule technique was chosen, this is the first step for obtaining knowledge, choosing the correct technique of Data Mining, according the kind of problem to treat and the desired goal. About the whole process, KDD offers a set of formal steps that helps in the application of the EDM. Before analyzing and processing the information, it is important that a correct cleaning and pre-processing of the data is performed, considering the format of the original data, it was necessary to perform this process, where it's important to know the problem and environment, so the best criteria, transformations and categories can be specified, this contributes to obtaining better results, removing variables that are not significant and creating categories also helps in getting clearer associations. At the moment of interpreting the rules, it's necessary a certain

level of knowledge about the problem and concepts that are being analyzed. Using the lift property for measuring the importance of the rules, and taking the ones with the biggest value contribute to a clearer interpretation of those ones. But it is necessary to analyze several rules (with the highest value of lift) to identify the relationship between some criteria with the academic performance.

From the obtained rules, it is clear that factors like the studies level of the parents, alcohol consumption, self esteem and attendance average are related with the academic performance. This can be used for different institutions, not necessarily educative, for taking decisions or implementing politics that can improve the academic performance of students.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. Acm sigmod record vol. 22, No. 2, pp. 207–216, ACM (1993)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: 20th int. conf. Very large data bases, VLDB, pp. 487–499 (1994)
3. Baradwaj, B.K., Pal, S.: Mining educational data to analyze students' performance. ArXiv preprint arXiv:1201.3417 (2012)
4. Fayyad, U., Piatesky-Shapiro, G., Smyth, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of the ACM. 11, pp. 27–34 (1996)
5. Merceron, A., Yacef, K.: Revisiting interestingness of strong symmetric association rules in educational data. In: Proc. of Int. Workshop on Applying Data Mining in e-Learning, pp. 3–12 (2007)
6. Merceron, A., Yacef, K.: Interestingness measures for association rules in educational data. In: Educational Data Mining (2008)
7. Merceron A., Yacef K.: Measuring correlation of strong symmetric association rules in educational data. In: Romero C, Ventura S, Pechenizkiy M, Baker RSJd, eds. Handbook of Educational Data Mining, pp. 245–256. CRC Press (2010)
8. Rajadhyax, N., Shirwaikar, R.: Data Mining on Educational Domain. International Journal of Advanced Studies in Computers, Science and Engineering, 1(1), pp. 1 (2012)
9. Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C Applications and Reviews, vol. 40 No. 6, pp. 601–618 (2010)
10. Romero, C., Ventura, S.: Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), pp. 12–27 (2013)
11. Sheikh, L.M., Tanveer, B., Hamdani, M.A.: Interesting measures for mining association rules. In: Multitopic Conference, Proceedings of INMIC 2004. 8th International, pp. 641–644. IEEE (2004)
12. Subsecretaria de Educación Media Superior. Encuesta Nacional de Deserción en la Educación, `http://www.sems.gob.mx/sems/encuesta_nacional_desercion_ems`