Generating Association Rules To Identify Adolesence Behavior Of Students In Higher Educational Institutions

K.Arunmozhi Arasan, Dr.E.Ramaraj, S.Muthukumaran

Abstract-Education is the most powerful factor in all aspects of Human life and also plays a dominant role to create a good society. Large amount of data can be collected from the educational field and it can be used to extract new knowledge. The data mining techniques such as classification, clustering, association rule mining can be used to discover the hidden patterns from educational data. The discovered knowledge may be used in several ways, such as to improve enrollment of students towards higher education, to improve the performance of the students as well as teaching faculty, finding attitude of students and so on. This paper focuses on finding the best tool to find the relationship between external environment and the attitude of the students studying in the higher educational institutions. The data collection is done through a questionnaire which contains 27 close ended questions. The Machine Learning algorithms such as FP_Growth and Apriori are applied to find the relationship between the attributes and to create association rules based on the relationships. Two familiar machine learning tools, namely Rapidminer and Tanagra were used to generate rules and compared with each other.

Keywords: Educational Datamining, Machine learning, Support, Confidence, FP_Growth, Apriori Algorithm.

1. INTRODUCTION

The social and family environment plays a vital role in predicting the attitude and performance of the students, especially who are persuing the higher education. In this paper association rule mining algorithm was applied on two different tools to find the association between attributes such as gender, age group, nature of the school, nature of the college in which he/she studied his/her UG course, staying with parents or with relation, etc. The attributes are named as gender, age, sclnature, Under Graduate nature, parents, presentstay, etc. The data is collected from various higher education institutions located in rural/semi-rural/urban area of Villupuram, Pudukottai and Sivaganga districts of Tamil Nadu state. Also the data is collected from various colleges that are fall under different categories such as Government/Private colleges, Co-ed/Women's colleges.

2. RELATED WORK

Abu Tair, & El-Halees[1], presented a paper about various machine learning algorithms such as Naïve Bayesian classifier, k-means clustering, fp-growth algorithms to improve the students performance. Noguera, [2] described in his paper, about the role and influence of environmental and cultural factors on the academic performance of African American male students. Venkatesan, [3] discussed about the irregularity of students' behavior in e-learning process using various data mining techniques. Vijayarani, and Prasannalakshmi.[4] pointed out in their paper for generating association rules in various sized data streams. Also, in it they compared different association rule mining algorithms using a dataset with different size and they conclude frequent itemset algorithm is the best algorithm for mining association rules. Venkatesan, G.Suresh, and S.Muthukumaran, [5] presented a paper to predict the reason for students absentieesm at under graduate level. In this paper they used Apriori algorithm to predict the reason and they presents guidelines to be followed to avoid students absentieesm.

3. PROPOSED METHODOLOGY

In order to analysis the personal behavior of students at higher education institutions a questioniaire was framed by discussing with Psychologist, Educational Experts, Sports personalities and Doctors. The questionnaire was distributed to Under Graduate Students and Post Graduate students at Government Universities, Government colleges, Autonomous colleges, self-financing colleges and the Collected data was preprocessed. The dataset contains 2083 records and 27 discrete attributes having nominal values like Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree. Frequent Pattern mining algorithm called fp-growth was applied to the dataset and association rules are generated to find the most influencing attributes in the dataset. FP-Growth algorithm works by divide and conquer method, it decomposes the dataset into smaller for mining patterns hidden in the databases. Frequent patterns are identified from the dataset using the support and confidence value given to the algorithm as parameters. The association rules generated has the attributes influencing the personal behavior in its Antecendt(Left side) and gender in its Consequent(Right Side). The proposed framework is shown in the figure below.

- *K.Arunmozhi Arasan is a research scholar in Department of Computer Science in Alagappa University, Karaikudi, India. E-mail: arunlucks@yahoo.co.in*
- *Dr.E.Ramaraj is working as Professor and Head of the Department of Computer Science, Alagappa University, Karaikudi, India. E-mail: eramaraj@rediffmail.com*
- *S.Muthukumaran is a research scholar in Department of Computer Science in Alagappa University, Karaikudi, India. E-mail: muthumphil11@gmail.com*

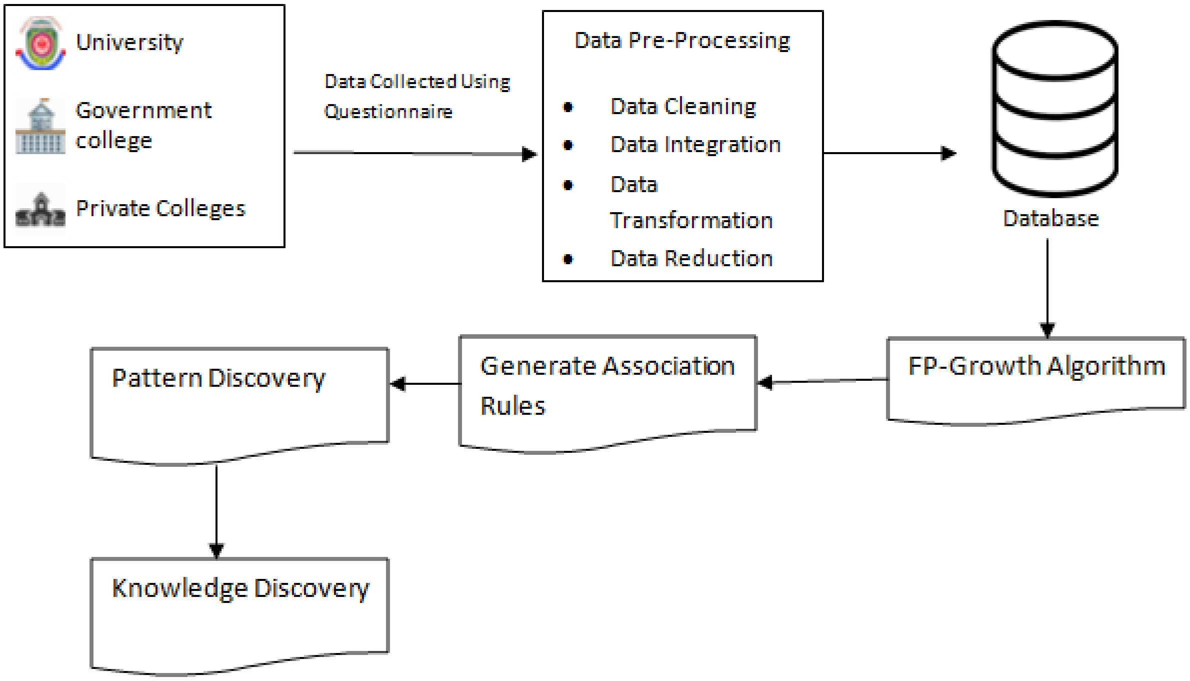

***Figure 1**. Proposed Framework*

3.1 PATTERN DISCOVERY USING FP-GROWTH ALGORITHM

The FP_Growth algorithm reduces the large database into smaller database called the FP tree. The FP-Growth algorithm works on divide and conquer method to create fp-growth tree by compressing the database which represents frequent itemset and then by dividing into conditional databases. The pseudocode for the modified FP-Growth algorithm to predict the personal behavior of students at Higher Education Institutions are as follows: Input:A student Datatbase D containing 27 categorical attributes with nominal values and 2083 records. Output: FP tree and Association Rules Generated from FP-Tree. Method: Ck: Candidate itemset of size k; Lk: Frequent itemset of size k; L1:{frequent itemsets}; for(k=1,Lk≠∅; k++) do begin count support for each item with minimum support and generate Ck+1; for each transaction t in database do begin Lk+1=Candidates support count with minimum support; Compare L1 table items wih D table items End return fp-tree enerate association rules generated from the fp-tree return highly influencing reason for student’s personal behavior end

4. RESULTS AND DISCUSSION

Apriori algorithm is applied to the dataset using Rapidminer 9.2 and Tanagra 1.4.50 data mining tools to create association rules. Support value is taken as 0.33 and confidence value is taken as 0.75. The table below shows the highly influenced rules and its measures that are generated in Rapidminer and Tanagra tools.

***Table 1**. Highly influenced rules obtained by Tanagra*.

S.No.	Antecedent	Consequent	N	Support	Confidence	Lift	Conviction
1	"currentdegree=ug" - "sclnature=Co-Ed"	"age=18-20" - "ugnature=Co-Ed"	2083	0.3644	0.7698	1.3883	1.9351
2	"age=18-20" - "sclnature= Co-Ed"	"currentdegree=ug" - "ugnature= Co-Ed"	2083	0.3644	0.7620	1.3555	1.8400
3	"age=18-20" - "ugnature= Co-Ed"	"presentstay=parents" - "currentdegree=ug"	2083	0.4714	0.8502	1.3167	2.3654
4	"age=18-20" - "ugnature= Co-Ed" - "sclnature= Co-Ed"	"currentdegree=ug"	2083	0.3644	0.9347	1.2963	4.2733
5	"age=18-20" - "ugnature= Co-Ed"	"familynature=small" - "currentdegree=ug"	2083	0.4268	0.7697	1.2961	1.7635
6	"presentstay=A" - "age=18-20" - "ugnature= Co-Ed"	"currentdegree=ug"	2083	0.4714	0.9317	1.2921	4.0831
7	"currentdegree=ug" - "familyrelation=very close bonding"	"parents=both" - "age=18-20"	2083	0.3649	0.8361	1.2796	2.1146
8	"presentstay=parents" - "age=18-20" - "sclnature= Co-Ed"	"currentdegree=ug"	2083	0.3922	0.9221	1.2788	3.5816
9	"age=18-20" - "ugnature= Co-Ed"	"currentdegree=ug"	2083	0.5113	0.9221	1.2788	3.5795
10	"age=18-20" - "sclnature= Co-Ed"	"parents=both" - "currentdegree=ug"	2083	0.3821	0.7992	1.2747	1.8576

The observations of the Tanagra output clearly show that the students who have studied in co-education schools also studying in co-education colleges for their higher studies, especially UG students from the age group between 18-20 (R1). If a student studying UG and having very close bonding with his/her family then he is accommodating with his/her both parents (R7). Since the lift value of all the above rules are greater than one, there is close dependence between the antecedent and consequent of the above rules. The following table shows the measures and their values for highly influenced rules obtained from Rapidminer for the dataset

***Table 2**. Highly influenced rules and its measures*

S.No.	Antecedent	Consequent	N	Support	Confidence	Lift	Conviction
1	[sclnature =co-ed]	[parents = both, presentstay = parents]	2083	0.474	0.751	1.002	1007
2	[parents =both, familyrelation = veryclose bonding]	[age = 18-20]	2083	0.420	0.751	1.009	1.026
3	[perbehavior = Strongly Agree]	[age = 18-20]	2083	0.349	0.754	1.012	1.037
4	[parents = both, lngstudied]	[age = 18-20]	2083	0.419	0.755	1.013	1.040
5	[lngstudied]	[age = 18-20]	2083	0.482	0.755	1.014	1.043
6	[sclnature = co-ed]	[age = 18-20]	2083	0.478	0.757	1.016	1.050
7	[parents = both, sclnature = co-ed]	[age = 18-20]	2083	0.423	0.759	1.019	1.059
8	[parents = both, presentstay = parents, age = 18-20]	[ugnature = co-ed]	2083	0.449	0.759	1.084	1.243
9	[parents = both, presentstay = parents, lngstudied]	[ugnature = co-ed]	2083	0.359	0.761	1.086	1.253
10	[familyrelation = very close bonding]	[parents=both, presentstay = parents]	2083	0.471	0.761	1.016	1.052

The rules obtained from **Rapidminer** shows that the students in their adolescent age are presently staying with their parents and also came from small family (R1). Some other rules show that students from the age group between 18 and 20 are staying with parents (R8) and also prefer the co-education institutions for their higher studies (R9). Criticizing personal behaviour is not encouraged by adolescents (R3). In the above table, rule 2 to rule 7 indicates the age factors plays the major role for students behavior. The following table shows number of rules generated for different support and confidence value

***Table 3.** No. of rules generated for different support and confidence values.*

S.No.	Supp.	Confid.	No. of rules generated in Rapidminer	No. of rules generated in Tanagra	S.No.	Supp.	Confid.	No. of rules generated in Rapidminer	No. of rules generated in Tanagra
1	0.23	0.6	1171	924	7	0.33	0.65	223	272
2	0.28	0.65	442	438	8	0.33	0.75	147	177
3	0.29	0.66	370	390	9	0.28	0.7	354	363
4	0.33	0.7	179	230	10	0.32	0.7	198	247
5	0.34	0.71	237	189	11	0.34	0.7	245	196
6	0.42	0.79	152	54	12	0.38	0.7	244	127

The above result table shows that by adjusting support and confidence threshold values, the number of rules generated will vary. The table contains the number of rules generated for different values assigned for support and confidence threshold. If the support and confidence parameter values gradually increased then rules generated by both tools are also gradually decreased. Similarly, Tanagra always generates more rules when comparing with Rapid miner when support value is 0.33 and confidence values vary between 0.65 t 0.75 for the dataset. If support values are taken as <0.33 and confidence =0.7 then Rapidminer gives the less number of rules that of in the Tanagra tool. If support values are taken as >0.33 and confidence =0.7 then Rapidminer gives more rules that of in the Tanagra tool.

4.1 EVALUATION

Support, Confidence. Lift, and Conviction are the measures to calculate the significance of a rule. Support: Support s is the percentage of transactions in D (Transaction Dataset) that contains both X and Y. S(X→Y)=P(X ∪Y). Confidence: Confidence is based on conditional probability. The transactions in D contains X and also contains Y. The Confidence C(X→Y)=P(Y/X). Lift: Lift (also called interest of a rule) is the measure to find the dependence relationship between any two variables/itemset. The occurrence of an itemset2 depends on occurrence of another itemset1 then there is a correlation between the two itemsets. If the lift value is less than 1 then there is negative dependence in between the two variables/itemsets, if it is equal to 1 then no dependence between the variables/itemsets and if it is greater than 1 then lift value indicates the positive dependence between the two variables/itemsets. Lift(X→Y)= P(X ∪Y)/ P(X)P(Y). Conviction: The conviction is also used to measure the dependence between two attributes/itemsets. If it is > 1 then there is positive dependence between antecedent and consequent (itemsets) of the rule. If the conviction value is in between 0 and 1 then there is negative dependence between the antecedent and consequent (itemsets) and the conviction value is equal to 1 indicates that the two attributes/itemsets (antecedent and consequent) are independent. conviction(X→Y)= P(X) P(¬Y) / P(X ∪ ¬Y).

5. SUGGESTION

From the obtained results, it should be noted that most of the students are studying in the co-educational institutions, especially the students who are staying with their parents and studied in co-ed schools. The managing authorities of government and private educational institutions have to create psychological awareness for the students about their personal behavior and attitude during the adolescent age.

6. CONCLUSION

The rules generated by Rapidminer using association rule generating algorithms for our data set show that the students under the age group of 18 to 20 who were studied in the co-educational institutions are enrolled in the co-educational colleges. Similarly, there is a strong relationship between the attributes present stay and sclnature and ugnature. Rapidminer produce more rules than the Tanagra tool to identify the students' behavior in higher education institutions. People in adolescent age, especially studying co-educational institutions are facing many personal issues such as confusion about their future, issues about their personal behavior, bonding with their family, etc. They need special counselling to improve their behavior and attitude and the comparison results shows that Rapidminer is the best option to get more number of association rules for this dataset.

ACKNOWLEDGMENTS

This article has been written with the financial support of RUSA-Phase 2.0 grant sanctioned vide Letter No.F.24-51/2014-U, Policy (TNMulti-Gen), Dep.ofEdn.Govt. of India, Dt.09.10.2018.

REFERENCES

[1] Abu Tair, M. M., & El-Halees, A. M. (2012). Mining educational data to improve students' performance: a case study. Mining educational data to improve students' performance: a case study, 2(2).

[2] Noguera, P. A. (2003). The trouble with Black boys: The role and influence of environmental and cultural factors on the academic performance of African American males. Urban education, Vol. 38, Issue 4, pp 431-459.

[3] Venkatesan, N. (2013). Role of Data Mining Techniques in Educational and E-learning System. Asia Pacific Journal of Research, 2.

[4] Vijayarani, D. S., & Prasannalakshmi, M. R. (2015). Comparative analysis of association rule generation algorithms in data streams. International Journal on Cybernetics & Informatics (IJCI), Vol. 4, Issue 1, pp. 15-25.

[5] Venkatesan, N., Suresh.G., Muthukumaran.S. (2015). A Survey on Student's Absentieesm at Under Graduate Level using Apriori Algorithm for Categorical Dataset. International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 1, pp. 79.