Educational Data Mining Application for Estimating Students Performance in Weka Environment

IOP Conf. Series: Materials Science and Engineering **263** (2017) 032002 doi:10.1088/1757-899X/263/3/032002

EDUCATIONAL DATA MINING APPLICATION FOR ESTIMATING STUDENTS PERFORMANCE IN WEKA ENVIRONMENT

G.Shiyamala Gowri [1], Ramasamy Thulasiram [2] and Mahindra Amit Baburao [3]

[1,2]Research Scholar, Centre for Disaster Mitigation and Management
[3]Professor, School of Civil and Chemical Engineering
[1,2,3] Vellore Institute of Technology, Vellore, Tamil Nadu, India – 632014

Email: shiyamalagowri1947@gmail.com

Abstract: Educational data mining (EDM) is a multi-disciplinary research area that examines artificial intelligence, statistical modeling and data mining with the data generated from an educational institution. EDM utilizes computational ways to deal with explicate educational information keeping in mind the end goal to examine educational inquiries. To make a country stand unique among the other nations of the world, the education system has to undergo a major transition by redesigning its framework. The concealed patterns and data from various information repositories can be extracted by adopting the techniques of data mining. In order to summarize the performance of students with their credentials, we scrutinize the exploitation of data mining in the field of academics. Apriori algorithmic procedure is extensively applied to the database of students for a wider classification based on various categorizes. K-means procedure is applied to the same set of databases in order to accumulate them into a specific category. Apriori algorithm deals with mining the rules in order to extract patterns that are similar along with their associations in relation to various set of records. The records can be extracted from academic information repositories. The parameters used in this study gives more importance to psychological traits than academic features. The undesirable student conduct can be clearly witnessed if we make use of information mining frameworks. Thus, the algorithms efficiently prove to profile the students in any educational environment. The ultimate objective of the study is to suspect if a student is prone to violence or not.

1. Introduction

Every educational institution generates a vast quantity of data every year. Raw data can be meaningfully transmitted with the help of data mining. The data obtained from an educational institution undergoes scrutiny of various data mining methodologies [10]. The methods employed detects the environment where a student can be better motivated to lead a meaning life so that they can contribute to the society [1]. Weka, an efficient data mining apparatus is utilized in order to generate the results.

Apriori algorithm works by mining the rules in order to extract patterns that are similar along with their associations in relation to various set of records. K-means procedure is connected to a similar arrangement of databases with a specific end goal to gather the students into a particular category in an extremely proficient way. The remainder of this paper is organized as follows: The apriori algorithm

that finds the frequent item sets is explained in the first part of the section 2, followed by the k-means clustering analysis. The case study concerned with the educational background is dealt in section 3 along with the results of the algorithms. The conclusion is discussed in Section 4.

2. Methodology

2.1 Apriori Algorithm: Finding frequent item sets

Apriori algorithm works by mining the rules in order to extract patterns that are similar along with their associations in relation to various set of records [3]. In our case we are dealing with the student's database. If a student satisfies criteria A, then he/she is likely to satisfy criteria B also. An extraordinary student can be classified by following the below criteria. If a student scores marks in the range from 80 to 100 in midterm test, practical session marks between 21 to 30, project team score is grade A and attendance is also good, then the performance of the student is very good. But not only the academic records prove his performance, other psychological factors also affect their credentials.

Let us consider the set of items S = {s1, s2, sm}. R is the transaction set, in which each transaction U has an itemset where U ⊆ S. Whenever a transaction U has A, then A ⊆ U, the association rule takes the shape A => B, which means that when A intersects with B the itemset would be null provided A is a subset of S and B is also a subset of S. The support u is applied in the association rule C => D which originally belongs to the transaction Z along with p percent of item sets takes the form A => B. In order to prove that transactions of C also contain transactions d, the confidence c rule is applied. The apriori algorithm works as below [4][7]:

```
Hi: Candidate itemset of size i
Ti: visit itemset of size i
T1= {frequent items};
for(i= 1; Ti!=∅; i++) do start

Hi+1= canditate produced from Ti;
for every transaction t in database do:
increase the count of all candidates Hi+1 that are contained in t

Ti+1=candidates in Hi+1with min_support
end

return ∪iLi;
```

2.2 K-means Cluster analysis:

The main aim of the k-means cluster analysis is to generate a group of students based on their personal traits in addition to their academic record. The teachers can effectively frame a customized educational network, with the help of the generated clusters of students. This can further be utilized to develop an active team learning. The machine learning is a part of artificial intelligence, which can be classified into supervised learning and unsupervised learning. classification comes under supervised and cluster analysis is a part of unsupervised analysis. Clustering is the allotment of group of observations into a subset called clusters. Observations present in the same group are similar to each other, whereas observations present in other clusters are dissimilar to each other. The grouping of students can be accomplished by using various clustering methodologies such as, Hierarchical clustering, k-means, c-means clustering [2].

Here we choose k-means clustering algorithm in order to effectively cluster the students who possesses similar characteristics. The final outcome would be to classify the students as good, average, poor. The algorithm works by randomly picking up objects, where every object denotes a cluster centroid. Whichever object comes close to the cluster centroid, those objects are selected and assigned

to one particular group/cluster. After assigning the new object to a cluster, again the centroid for that particular cluster alone will be calculated. Once the standard capacity is met, the procedure stops. The squared error criterion is characterized as follows:

$$W(b)=\frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(j)=k}\left\|x_i-x_j\right\|^2=\sum_{k=1}^{K}N_k\sum_{C(i)=k}\left\|x_i-m_k\right\|^2$$

Where X_1, X_N are data points or vectors or observations. In this paper, we are considering the various academic parameters like exam scores, attendance, practical marks, midterm scores and personal characteristics like family history, past record, behavior etc. which finally summarizes the student performance into various clusters namely good, satisfactory and poor. $C(i)$ indicates cluster number for the i^{th} observation, where mk is the mean vector of the k^{th} cluster, N_k is the number of observations in k^{th} cluster.

The below mentioned procedure explains the methodology of K-means [5]:

Analysis being carried out based on quantity of clusters and count of the parameters.

The students are classified as good, average and poor based on the lease squared error.

i) Start.

ii) The initial clusters are formed by picking up k items.

iii) Reiterate

iv) Allocate out every item the cluster to which the protest is most comparable, with respect to the centroid of the group.

v) The centroid of the group in incremented.

vi) Until the values of the centroids in the group remains constant.

vii) Stop.

The algorithm endeavors to decide k segments that minimize the squared error work. All the clusters get minimized as opposed to very much isolated from each other. The strategy is moderately adaptable and productive in preparing vast datasets on the grounds that the computational intricacy of the framework is O(nkt) where n represents the aggregate quantity of items, k signifies quantity of clusters, and finally t embodies the quantity of reiterations. Regularly, k<<n and t<<n. The technique frequently ends at the stage of a neighborhood ideal [7].

3. Execution and Output

For the purpose of identifying the students' traits, the manual extraction of student's records would be a huge task [7]. Considering a higher number of parameters will help to predict more detailed information about the student. The database generated from various educational institutions are used for the analysis. One of the major problems that the education system facing is predicting the behavior of students from large database. Evaluating the weakness and strengths of individual student and an analysis on effort given by the teacher to an individual student improves the performance of the educational institution. Instead of considering only academic parameters, we could also consider

personal characteristics, behavior, family history, past records in order to predict the performance of a student [8]. The whole procedure can be pictorially depicted in figure 1.

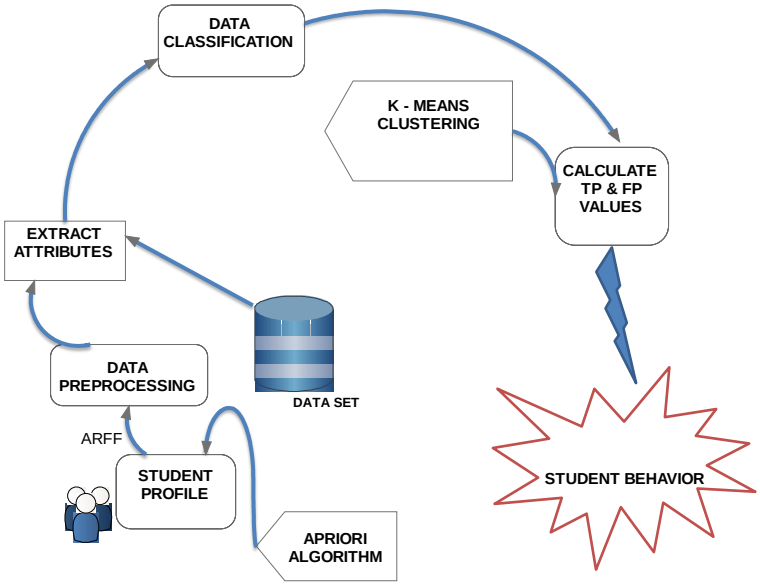

Figure 1. Pictorial Representation of the entire process

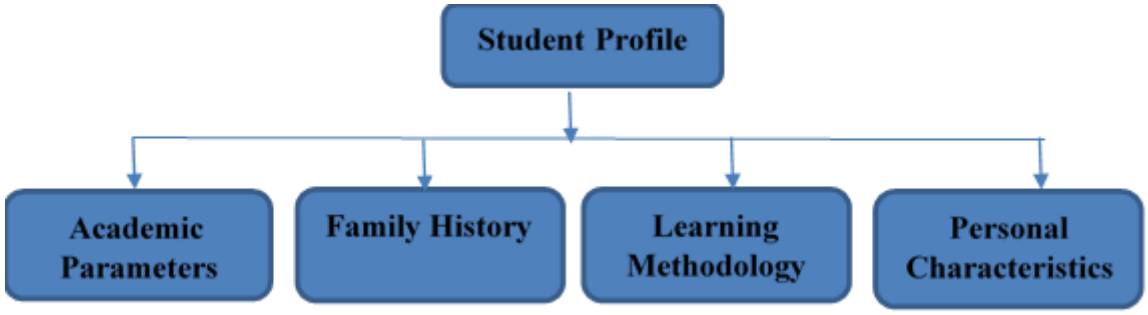

Figure 2. Parameters Considered

The different parameters considered are presented in figure 2. Weka apparatus is used to categorize the students based on their performance as good, satisfactory or poor [6]. Apriori algorithm works by mining the rules in order to extract patterns that are similar along with their associations in relation to various set of records. K-means clustering analysis groups the students based on the academic and psychological parameters. The final output is generated using weka apparatus. Any educational institution can utilize this tool in order to analyze the students and predict their behavior. Thus, the educational data mining plays a vital role in determining the performance of the students.

3.1 Case Study:

The analysis is carried out with the information that was generated by enquiring 100 students from the below listed schools. Four schools were chosen from Vellore district in the state of Tamil Nadu. Our analysis concentrates on village students. Target sector was government schools rather than private institutions. Name of the schools used for our analysis are listed in table 1.

Table 1. List of schools

S.no	Name of the school	No of students
1	Government Higher Secondary School, Kaniyambadi, Vellore	25
2	Government Higher Secondary School, Pennathur, Vellore	25
3	Government Higher Secondary School, Munjurpet, Vellore	25
4	Government Higher Secondary School, Chozhavaram, Vellore	25

3.2 ARFF document

Attribute Relationship File Format (ARFF) is the content arrangement document utilized by Weka to store information in a database. The ARFF document contains two segments: the header and the information segment. The first line of the header lets us know the relation name. At that point, there is the list of the characteristics (@attribute...). Every property is connected with an exclusive name and a category. The last depicts the sort of information contained in the variable and what values it can have. The factors sorts are: numeric, minimal, string and date. This sort of document is organized as follows ("students record" database) in view of academic parameters: [7] The parameters considered for our analysis are listed in table 2.

Table 2 . Parameters considered for analysis

S.No	Academic parameters
1	AttendanceA75
2	MarkA80
3	MarkA40
4	MarkA0
Family history	
5	Parents
6	ParentStudy
7	ParentReadWrite
8	ParentEmployed
9	FamilyIncomeL3000
10	FamilyIncomeB3Kto6K
11	ParentStrict
12	ParentCare
Learning methodology	
13	InterestinStudy {0,1}
14	Understanding {0,1}
15	MemorizeLesson {0,1}
16	SentenceOwnWords {0,1}
17	ExtraCourses {0,1}
Personal characteristics	
18	Emotional {0,1}
19	FailToleranceCapacity {0,1}
20	SocialService {0,1}
21	SportsInterest {0,1}
22	ExtraCurricular {0,1}
23	Anger {0,1}

24	Patience {0,1}
25	RespectElder {0,1}
26	PoliticalInterest {0,1}
27	FightwithFriends {0,1}
28	BadHabit {0,1}
29	BadHabitfromAdolescent {0,1}
30	BadHabitfromFriend {0,1}
31	BadHabitfromFamily {0,1}
32	BadHabitAddiction {0,1}
33	BadHabitinEveryday {0,1}
34	BadHabitinweek {0,1}
35	StartedCuriosity {0,1}
36	WantRidofBadHabbit {0,1}
37	Pocketmoney {0,1}
38	SpendMoneyUseful {0,1}
39	SpendMoenyBuyCigarAlcohol {0,1}
40	StealMoney {0,1}
41	PoliceCompliant {0,1}
42	RoamwithFriend {0,1}
43	BecomeScientist {0,1}
44	PartTimeJob {0,1}
45	DamagePublicProperty {0,1}

Based on these parameters we can classify the students into various clusters which can be used to predict the characteristics of a student. A minimum support of 20% is applied to the student record that was generated from four schools of Vellore district, where the best association rules can be extracted by using apriori algorithm [7]. We have collected 100 students' data obtained from four government schools in Vellore district. The higher the minimum support we provide, the stronger the association rules we obtain. The output of this analysis is shown in figure 3. along with the best association rules obtained that we can use to profile the students' performance. Family plays a major role in shaping a student's life. Equally a teacher also does the role in motivating a student in a right path. Most of the students who were given right motivation at right time, shine well in their life. But those who were shown wrong path are prone to violence [9].

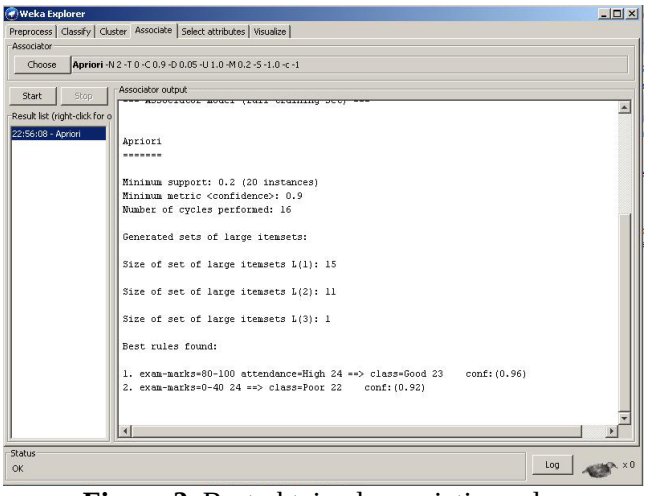

Figure 3. Best obtained association rules

Clustering intends to divide n perceptions into k clusters in which every perception has a place with the cluster to the closest mean. Figure 4a and Figure 4b. clearly explains the output of the k-means clustering analysis which was carried out using Weka apparatus. The students' performance is well understood in the below figures.

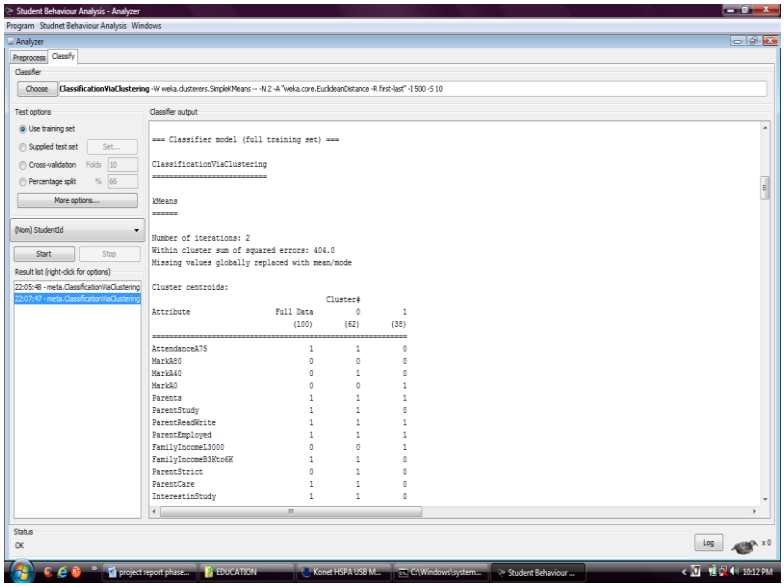

Figure 4a. Students' Performance

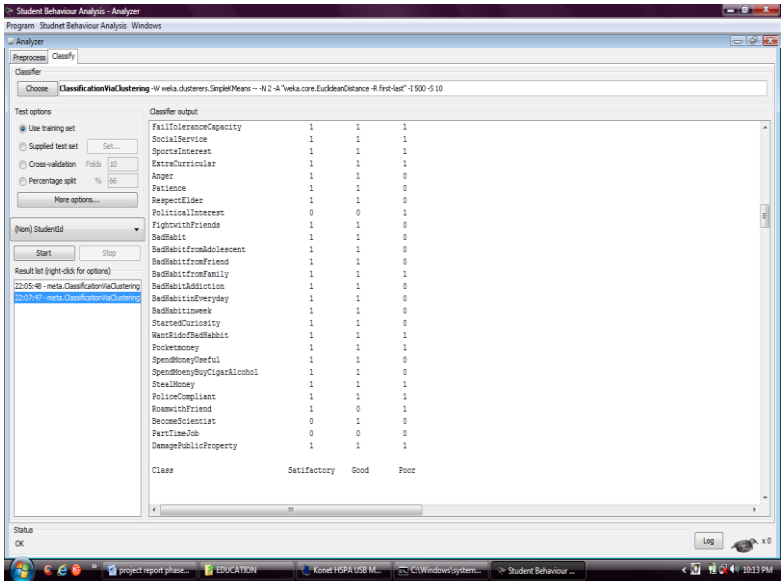

Figure 4b. Students Performance

In our analysis of 100 students , weka apparatus classifies our students into two groups with 62 and 38 in each group respectively. The final result of our student behaviour analysis proves to be satisfactory. The first cluster categorizes the students behaviour as Good where the attendance is above 75, marks between 40 to 80, student's parents are alive, educated, employed, family income is above 3000, parents are strict and they care their children very well. Students are interested in studying, they understand their lessons, memorize and also write on own sentences. They are not emotional and have failure tolerence capacity. Students are interested in social service, they excel in sports and also in

other extra curricular activities. They do not get anger unnecessarily and they have patience. Students respect elders and do not fight with their friends and also not interested in politics. They do not involve in any bad habits either through friends or through family background. Students spend their pocket money usefully. Most of the students aim to become a scientist.

The second cluster categorizes the students behaviour as Poor. Student's attendance is less than 75%. Marks are scored below 40 %. Student's parents are illiterate but they know to read and write. They are employed with poor economical background of income less than 3000 per month. Parents are not strict and they never care their children. Students are not interested in studying and they hate to come to schools. Most of the students come to school for the sake of free mid day meals and to obtain other free accessories like laptop from government. They are not able to understand their lessons and they simply memorize their lessons. They do not learn any additional courses. Students are very emotional and some even lack failure tolerence capability which may even lead to suicide . Students are interested in social service and also excel in sports. They do participate in co curricular activities out of their own interest. Their level of anger is high which reduces patience and they simply fight with their friends for silly reasons which may sometimes lead to suspension of their education. They do not respect teachers or elders and they move on their own way. They show their interest in the field of politics. If not moulded properly their career may even reach level 0.

Bad habits like smoking , drinking alcohol are common among these students. Lack of parental care and affection move them towards such habits on curiosity. Some students get involved through friends and some through family background. Few students even go to the level of stealing money , without their parents knowledge for the sake of smoking etc. Some steal public money and was also arrested by police. Those students definitely need counseling.

4. Conclusions

Educational data mining application for predicting the performance of the students using weka apparatus is demonstrated through a case study on students in four government schools of Vellore district, Tamil Nadu. The result yielded two clusters which classified the students as good and poor. Hence the overall performance of the students is predicted to be satisfactory after analyzing 100 students using 45 parameters. Overall, the results obtained in the case study points out that educational data mining using weka apparatus, where the analysis is carried out using apriori algorithm and k-means algorithm outperforms previously used utility mining in which comparison is done effectively rather than classification. As a result of the analysis, the students can be classified based on their academic data, family history, learning methodology, personal characteristics where large records are present. Also 20:1 Student Staff ratio is suggested. Parents Teachers meeting can be conducted every month. Daily updates to parents regarding students' behavior, attendance, examination schedule, marks and any other means through mobile / internet. Regular counseling with a professional psychiatrist is a must for students as well as teachers. Meditation and yoga can also be made compulsory for teachers and students. Thus, by following the above principles, definitely the performance of an Educational Institution will grow higher.

References

[1] Behrouz Minaei-Bidgoli, Deborah A. Kashy, Gerd Kortemeyer, William F. Punch 2003 Predicting student performance: an application of data mining methods with an educational web-based system 33rd ASEE/IEEE Frontiers in Education Conference

[2] Cristobal Romero, Sebastian Ventura. 2010 Educational Data Mining: A Review of the State of the Art IEEE transaction on systems man and cybernetics part c: applications and reviews vol. **40**

[3] WANG Pei-ji, SHI Lin, BAI Jin-niu, ZHAO Yu-lin. 2009 Mining association rules based on apriori algorithm and application International Forum on Computer Science-Technology and Applications

[4] Anita Wasilewska. (undated). [Online]. APRIORI Algorithm Available: http:/www.cs.sunysb.edu/~cse634/lecture_notes/07apriori.pdf
[5] Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques 1st edition Morgan Kaufmann Publishers 2000
[6] Vasile Paul Bresfelean 2007 Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment in Proceedings of the ITI 2007 29th Int. Conf. on Information Technology Interfaces Cavtat Croatia.
[7] Parack, Suhem, Zain Zahid, and Fatima Merchant 2012 Application of data mining in educational databases for predicting academic trends and patterns IEEE International Conference on Technology Enhanced Education (ICTEE)
[8] C. Antunes 2008 Acquiring background knowledge for intelligent tutoring systems in *Proc. Int. Conf. Educ. Data Mining* Montreal QC Canada pp. 18–27
[9] I. Arroyo, T. Murray, and B. P. Woolf 2004 Inferring unobservable learning variables from students' help seeking behavior in *Proc. Int. Conf. Intell.Tutoring Syst.* Brazil pp. 782–784
[10] R. Baker and K. Yacef 2009 The state of educational data mining in 2009: A review and future visions *J. Educ. Data Mining* vol. **1** no. 1 pp. 3–17