ISSN:0975-9646

Application of Data Mining in Educational Database for Predicting Behavioural Patterns of the Students

[1] Elakia, [2]Gayathri, [3]Aarthi, [4]Naren J

[1,2,3] *B.Tech Student CSE, School of Computing, SASTRA UNIVERSITY, India,*
[4]*Assistant Professor CSE, School of Computing, SASTRA UNIVERSITY, India,*

Abstract **- Data mining refers to the technique of obtaining hidden, previously unknown and possibly significant knowledge from humongous amount of data. Data mining uses a combination of a vast knowledge base, advanced analytical skills, and domain knowledge to unveil hidden trends and patterns which can be applied in almost any sector ranging from business to medicine, then to Engineering. Nonetheless educational institutes can employ data mining to discover valuable information from their databases known as Educational Data Mining (EDM). Educational data mining requires transformation of existing or innovation of new approaches derived from statistics, machine learning, psychometrics, scientific computing etc. Current system is designed to justify that various data mining techniques which includes classification, can be used in educational databases to suggest career options for the high school students and also to predict the potentially violent behaviour among the students by including extra parameters other than academic details. RapidMiner has been used as Data mining tool.**

Keywords **- Data Mining, Rapid Miner, C4.5 Classification.**

I. INTRODUCTION

Educational data mining is a surfacing field which explores statistical information, machine learning and other data mining algorithms to discover interesting patterns in educational database. So far all the educational institutes have restricted themselves to predict only the academic performance of the students by considering internal, external marks and grades etc [1]. But it can be proved that various prediction models can be used for predicting the future details like career options and possibility for a teen to get violent in future. There are various data mining techniques like Association rule mining, clustering and classification that serve this purpose. Through the proposed system, it is suggested that various classifiers like ID3 (Iterative Dichotomizer) and C4.5 algorithm [2] can be used to give career suggestion and to find the violent behaviour in a student. Finally a comparative study is made with the different decision trees to find a suitable algorithm.

II. RELATED WORKS

Educational data mining is an upcoming trend and attractive discipline which accomplishes various predictions in educational institutes. There has been various works in this area involving different techniques of data mining.

A. Evaluation of Students academic performance

Shreenath Acharya et al. [1] presented an overview on Knowledge Discovery on Databases to predict the student's academic trends. Students placements were envisaged using their historical information. Apriori Algorithm has been used for the purpose of association rule mining and to generate frequent item sets.

Kalpesh Adhatrao et al.[2] have applied different decision tree algorithms like ID3(Iterative Dichotomiser) and C4.5 to predict students academic performances and checked their accuracy to choose the appropriate algorithm that could be effectively applied.

Mohammed M. Abu Tair and Alaa M. El-Halees [3] have given a case study on how to use different data mining techniques like Association Rule Mining, Clustering, Classification and Outlier Detection at various phases to improve the academic performance of graduate students.

B. Predicting School dropouts

N. Quadri1 et al [4] presented a technique to determine the list of students who are likely to drop out of the college. Various attributes like their gender, family background, attendance etc. were considered while mining. Decision tree algorithm has been suggested as to predict the dropout.

Gerben W. Dekker [5] gave a case study which included attributes that could act as a strong predictor for success. The famous Decision tree algorithms like C4.5 and CART were found suitable to predict dropouts.

C. Students behavioural prediction

M.Sindhuja et al. [6] have proposed a methodology which explored Behaviour Attitude Relationship Clustering (BARC) Algorithm to show the improvement in students performance in terms of predicting their behaviour, attitude and relationship with faculty members and Tutors. Hierarchical clustering was used to group students based on their similarity measures and the study was experimented using Weka Tool.

Ryan S.J.D. Baker [7] presented a machine-learning technique that could instantaneously find when a student using an intelligent tutoring system is off-task, i.e., engaged in things which does not include the system or a learning task. Only the system log files were considered while mining and Latent Response Models were used as the statistical basis for all of the detectors of off-task behaviour.

III. PRE-PROCESSING

Before working with the original raw dataset, proper pre-processing techniques must be applied on dataset to put them into efficient use.

1) Replacing missing values: The missing data were replaced using mean imputation method. It works by replacing the missing value for a given attribute by the mean of all known values of that attribute in the class.

2) Sampling: A subsample from the original dataset was taken for the purpose of bringing balance into training and testing datasets. Shuffled sampling without replacement is a sampling technique applied on the subsample where a selected item cannot be selected again (removed from the full dataset once selected).

3) Feature Selection: Not all the properties in a database will be used while mining so only the attributes which could be relevant for prediction is selected. A combination of forward selection and backward elimination was used to select the best attributes and to remove worst attribute.

4) Feature Creation: Creation of new attributes that could capture the important information from dataset much more efficiently than the original dataset is done using attribute selection operator found on rapid miner tool. Attributes like family economic status, attendance etc. are reduced into two instances namely yes or no, based on certain if then else conditions.

5) Set Role: The set role operator in rapid miner was used to indicate the role played by each attribute while classification. The attribute to be predicted was given a role of label while other attributes were treated as regular attributes.

IV. CAREER SUGGESTION FOR HIGH SCHOOL STUDENTS

Data Mining can be used to suggest suitable careers for high school students by mining relevant attributes. Hence Data sets were collected from the response sheet posted on the website. Various Data pre-processing techniques like replace missing values ,feature selection, feature creation, role setting(Label) were done on the data sets to make them suitable for mining. The aim of the above module is to suggest suitable career options for high school students based on every student's interests, skills, likes, hobbies etc. Firstly a part of database was used for training and testing. Training dataset was used, so that ID3 algorithm could be applied on them to learn the rules found on the training set and their corresponding labels. The learnt rules were represented in the form of a decision tree generated using impurity measures and information gain values.

$$\text{Entropy(s)} = \sum_{i=1}^{c} p_i \log_2 p_i$$

Then the learnt rules were applied on the test dataset with labeled attribute to check the performance measures like accuracy, f- measure etc. based on the TP(True Positive),TN(True Negative),FP(FalsePositive),FN(False Negative).K-fold Cross validation was used where out of K subsamples, K-1 subsamples would be used for training and remaining for testing. Finally, after getting a decent accuracy measure and low misclassification rate from the confusion matrix shown in Fig 1. The learnt model could be applied on the original unlabelled data for which the appropriate label was to be predicted. Student's career options were divided into six classes.

Class A: Enterprise
Class B: Enterprise, Conventional & Investigative
Class C: Artistic.
Class D: Realistic & Social.
Class E: Enterprising & Artistic.
Class F: Enterprise, Conventional, Investigative.

Students who were suggested with class A category would be listed a set of career options and similarly for other classes too.

Multiclass Classification Performance ○ Annotations

Table View ○ Plot View

accuracy: 92.53% +/- 9.56% (mikro: 92.54%)

	true B	true E	true C	true D	true A	true F	class precision
pred. B	12	0	0	1	1	0	85.71%
pred. E	0	15	0	0	1	0	93.75%
pred. C	0	0	8	0	0	0	100.00%
pred. D	1	0	0	16	0	0	94.12%
pred. A	0	0	0	0	7	0	100.00%
pred. F	1	0	0	0	0	4	80.00%
class recall	85.71%	100.00%	100.00%	94.12%	77.78%	100.00%	

Fig 1: The confusion matrix generated for career suggestion

The following is the general decision tree algorithm:

- 1) Establish Classification Attribute (in Table R)
- 2) Compute Classification Entropy.
- 3) For each attribute in R, calculate Information Gain using classification attribute.
- 4) Select Attribute with the highest gain to be the next Node in the tree (starting from the Root node).
- 5) Remove Node Attribute, creating reduced table R_S.
- 6) Repeat steps 3-5 until all attributes have been used, or the same classification value remains for all rows in the reduced table.

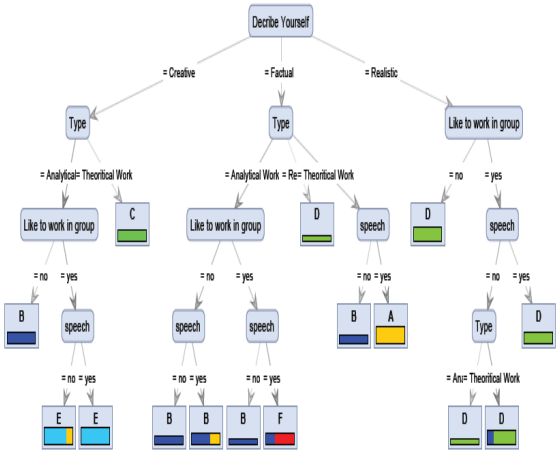

Fig 2: The generated ID3 tree for career classification

Finally the statistics of the predicted careers with respect to the number of students can be represented using a pie chart as shown in Fig 3.

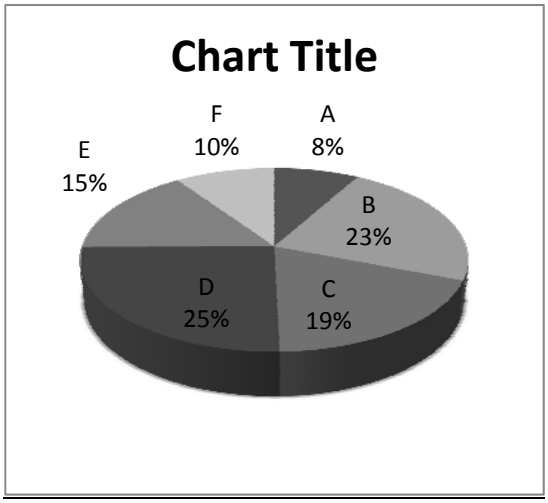

Fig 3: Statistical representation of career prediction

V. PREDICTING POTENTIALLY VIOLENT BEHAVIOUR OF A STUDENT

Even if career is suggested, discipline is an important factor to continue higher studies and pursue ones career. Hence the chance of a student to get violent in future is predicted. So that appropriate counselling methods can be applied to curb the violence in the initial stages itself. There are various factors that could be considered as causes for youth violence. The parameters like Background information, reaction of a student when irritated and interest in involvement of vigorous sports and games etc could be gathered using a response sheet and could mined to predict the potentially violent behaviour among the students. The same approach used for career prediction is applied to find the behavioural pattern. A subsample from the original dataset was used for training and testing. The learnt classification rules can be visualized using the decision tree as shown in Fig 4. And finally learnt model was applied on the original unlabelled dataset for the actual prediction .Binary classification was done using ID3 algorithm to determine whether a student needs counselling or not.

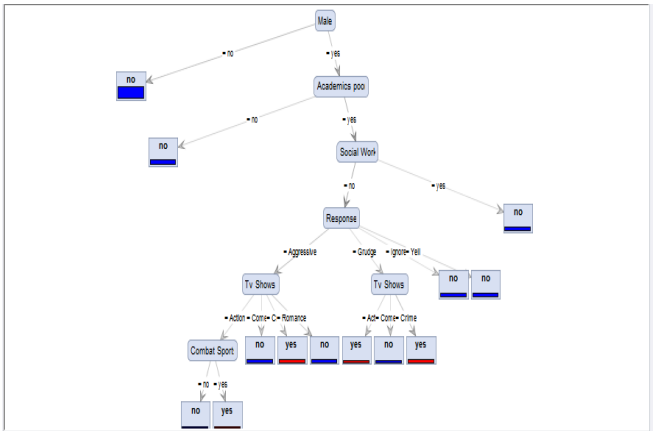

Fig 4: ID3 decision tree generated for behavioural prediction

V. VARIOUS ALGORITHM EFFICIENCY CHECK

There are numerous decision tree algorithms like C4.5, CART, CHAID and MARS that could be used for classification.ID3 could handle only categorical attributes while C4.5 can handle both numerical and categorical attributes. Hence a suitable decision tree algorithm must be selected to serve the purpose rightly. Three decision trees ID3, C4.5 and CHAID were compared for their various performance measures. In each case performance would be presented in a form of confusion matrix(Fig 6) with TN, TP, FP, FN from which evaluation measures like accuracy, precision, recall, sensitivity, F-measure, and specificity etc.(Fig 5) can be determined.

Accuracy: Proportion of predictions that are correctly classified.

Accuracy= (True Positives + True Negatives) / Total no. of samples.

Precision: Proportion of positive predictions that are right.

Precision= True Positives / (True Positives + False Positives)

Recall/ Sensitivity: The proportion of positive labeled samples that were predicted as positive.

Recall= True Positives / (True Positives + False Negatives)

Specificity: The Proportion of negative labeled samples that were predicted as negative.

Specificity = True Negatives / (True Negatives + False Positives)

F-measure: F-measure calculates the harmonic mean of precision and recall.

F=2*(Precision*Recall)/ (Precision + Recall)

Fig 5: Formulae for various efficiency measures

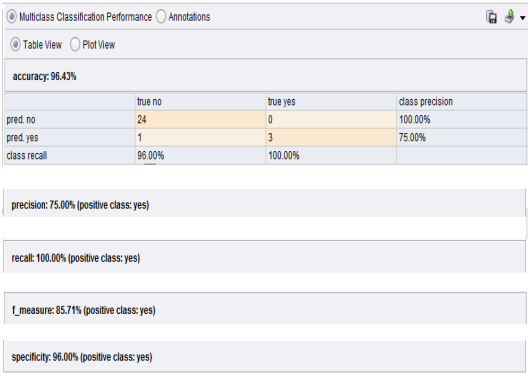

Fig 6: Accuracy measures calculated from confusion matrix

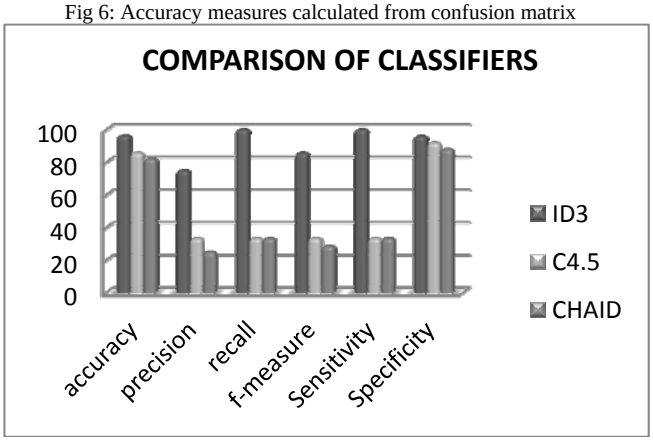

Fig 4 Comparison of classifiers

From the above chart it was evident that ID3 outperforms every other decision trees when all the performance measures were compared. Hence ID3 algorithm was considered to be an applicable algorithm.

VI. CONCLUSION

Through the paper classification as a methodology to predict a suitable career for a student is suggested. A student must be disciplined and should not be prone to violence which might affect their future (career).Hence a prediction on the chance of the student getting violent in future is done which will help the organization to offer counselling to the appropriate student in order to curb the violence in the initial stages itself. Various classifiers have been evaluated for their accuracy and performance and suitable classifier was used.

A. Future Enhancements

Future application of the techniques that has been used in the proposed system can consider higher number of attributes to make the predictions ever more accurate. Additionally, different techniques like neural networks, SVM, K-NN, Naive bayes and genetic algorithms can be experimented.

B. Limitations

In the career prediction, the predictions has been restricted to suggest career options only for those students who have taken first three major groups(Bio, comp science, commerce) in higher secondary school.

REFERENCES

[1] Shreenath Acharya, Madhu N, "Discovery of students' academic patterns using data mining techniques" IJCSE,Vol. 4 No. 06 June 2012.

[2] Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao "PREDICTING STUDENTS' PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS" IJDKP, Vol.3, No.5, September 2013.

[3] Mohammed M. Abu Tair and Alaa M. El-Halees "Mining Educational Data to improve Students' performance "JICT, Volume 2 No. 2, February 2012.

[4] Mr. M. N. Quadri1,Dr. N.V. Kalyankar "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques"GJCST, Vol. 10 Issue 2 (Ver 1.0), April 2010.

[5] Gerben W. Dekker "Predicting students drop out: a case study" 2nd Int. Conf. on Educational Data Mining, April 10, 2009.

[6] M.Sindhuja et al. "Prediction and Analysis of students Behaviour using BARC Algorithm",IJCSE, Vol. 5 No. 06 Jun 2013.

[7] Baker.R., "Modeling and understanding students' off-task behaviour in intelligent tutoring systems. In Conference on Human Factors in Computing Systems", San Jose, 2007 California, 1059-1068

[8] Suhem Parack, Zain Zahid, Fatima Merchant," Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns" IEEE Conference,2012.

[9] Smitha.T,Sundaram.V,"Comparative Study Of Data Mining Algorithms For High Dimensional Data Analysis",IJAET,Sept 2012

[10] Sandra K. Rumpel , Kathleen Lecertua , " Strong Interest Inventory with High School Profile " , Report , March 31,2012.

[11] Ron Kohavi,"A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", IJCAI, 1995.

[12] Rapid Miner, http://rapidi.com/content/view/181/190.

[13] Youth Violence Factors, http://www.cdc.gov/violenceprevention/ youth violence/riskprotectivefactors.html.

[14] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", 1st edition, Morgan Kaufmann Publishers, 2000.

[15] Dr. Matthew A. North, "Data Mining for the Masses ", 2012..