EDUCATIONAL DATA MINING USING NURSARY DATASET WITH DIFFERENT TECHNIQUES

Manpreet kaur **Er.chamkaur singh**
Research scholor Assistant professor
Guru kashi university,Talwandi sabo

Abstract: Educational Data Mining (EDM) is an emerging research area help the educational institutions to improve the performance of their students. Feature Selection (FS) algorithms remove irrelevant data from the educational dataset and hence increases the performance of classifiers used in EDM techniques. This paper present an analysis of the performance of feature selection algorithms on student data set. .In this papers the different problems that are defined in problem formulation. All these problems are resolved in future. Furthermore the paper is an attempt of playing a positive role in the improvement of education quality, as well as guides new researchers in making academic intervention. There are many techniques being anticipated to assess the student academic performance in way of making fruit full future of a student. Predicting performance of student has been continued to a hot topic in the Educational data mining domain. Data mining is considered to be one of the best choices for the researchers to analyse student's performance. Feature Selection is very dynamic and productive field and research area of machine learning and data mining. The main goal of feature selection is to choose a subset by eliminating non-predictive data. Furthermore, it increases the predictive accuracy and reduces the complexity of learned results .The effectiveness of student performance prediction models can be increased in connection with feature selection techniques. In this paper 12960 instances are identified. In the future result is improved.

Keyword: EDM, Data, mining, feature, accuracy etc.

I. INTRODUCTION

The improvement in the quality of education is one of the most significant aspects of forming a successful member of society. The data stored in educational institutions repository plays an important role in order to extract hidden and interesting patterns to assist every stakeholder of an educational process [1]. There are many techniques being anticipated to assess the student academic performance in way of making fruit full future of a student. Predicting performance of student has been continued to a hot topic in the Educational data mining domain. Data mining is considered to be one of the best choices for the researchers to analyse student's performance. The techniques of data mining are extensively used on educational data now a day's [2, 3]. It is called educational data mining. Educational Data Mining (EDM) explores the educational data to better understand the issues of student's performance using the fundamental nature of data mining techniques [4]. EDM manipulates educational data to help educational institutions to plan educational strategies, in order to improve the educational quality. Prediction is one of the main areas in EDM. Prediction and analysis of student academic performance are essential for student academic growth. Identifying the factors affecting the student academic performance is complicated research task [5].

II. EDUCATIONAL DATA MINING

Poised to meet the growing need for pervasive assessment is the nascent field of Educational Data Mining (EDM). EDM focuses on the collection, archiving, and analysis of data related to student learning and assessment. EDM is a very new and very small academic field. The first publications to mention educational data mining were published in the last two years, and there are likely fewer than thirty people in the world that identify themselves as being a part of it. As with all new fields, EDM has grown out of existing disciplines and is spreading to overlap with new ones. Many of the researchers who are shaping EDM hail from the Intelligent Tutoring System (ITS) community, where ready access to large quantities of educational data make EDM a logical direction to advance in. EDM research shares some commonalities with the Artificial Intelligence in Education (AIED) community. The analysis performed in EDM research is often related to techniques in psychometrics and educational statistics. EDM is poised to revolutionize, or at the very least enhance and expand, the statistical methods used in education by bringing to bear the results of decades of research in data mining and machine learning. Finally, given the computational backgrounds of most EDM researchers, it is not uncommon to find data pertaining to students learning computer science. As such, it is not surprising to find some overlap between the EDM and Computer Science Education (CSE) fields. This overlap may become stronger in the next few years as CSE naturally progresses toward more quantitative research and EDM broadens away from its original ITS focus.

III. FEATURE SELECTION

Feature selection has been an active and fruitful field of research area in pattern recognition, machine learning, statistics and data mining communities [2, 3]. The main objective of feature selection is to choose a subset of input variables by eliminating features, which are irrelevant or of no predictive information. Feature selection has proven in both theory and practice to be effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learned results [4, 5]. Feature selection in supervised learning has a main goal of finding a feature subset that produces higher classification accuracy. As the dimensionality of a domain expands, the number of features N increases. Finding an optimal feature subset is intractable and problems related feature selections have been proved to be NP-hard [6]. At this juncture, it is essential to describe traditional feature selection process, which consists of four basic steps, namely, subset generation, subset evaluation, stopping criterion, and validation [7]. Subset generation is a search process that produces candidate feature subsets for evaluation based on a certain search strategy. Each candidate subset is evaluated and compared with the previous best one according to a certain evaluation.

If the new subset turns to be better, it replaces best one. This process is repeated until a given stopping condition is satisfied. Ranking of features determines the importance of any individual feature, neglecting their possible interactions. Ranking methods are based on statistics, information theory, or on some functions of classifier's outputs [8]. Algorithms for feature selection fall into two broad categories namely wrappers that use the learning algorithm itself to evaluate the usefulness of features and filters that evaluate features according to heuristics based on general characteristics of the data [7].

IV. METHODOLOGY

The main aim of the research is to evaluate the performance of different FS algorithms on different classification algorithms using student dataset. The comparison between different FS algorithms give a deep insight to new educational data miners about the performance of different feature selection algorithms on student data .To achieve the objective of the research , a student dataset is taken from a valid sources, and then different FS algorithms are applied on it , which was not used earlier on this dataset. Different classification algorithms are applied by using selected FS algorithms, and furthermore evaluated to check the best performance among all the combinations applied on student data set.

Data set Description:

The dataset used in this study is taken from the source www.kaggle.com, and is comprised of 500 students 16 features. This dataset has been used in the study [11], to check the learner's interactivity with e- learning management system, bagging and boosting methods are applied on the given dataset, however, only information gain based feature selection algorithm is used previously. In this paper, the main aim of using the dataset is to identify the best combinations of FS algorithms and classifiers, in order to identify the key performance factors on the academic achievements of students.

WEKA (Waikato Environment for Knowledge Analysis) is used as a data mining tool. It has a rich source of Machine learning algorithms. WEKA is developed by the University of Waikato in New Zealand. It is an open source software developed in JAVA language, that provides facility for developing machine learning techniques for data mining tasks.

Feature Selection Algorithm and Classifiers

In this research work six FS algorithm CfsSubsetEval, ChiSquaredAttributeEval, FilteredAttribute Eval, GainRatioAttributeEval, Principal Components, and ReliefAttributeEval are evaluated. The classification algorithm BayesNet(BN), Naïve Bayes(NB), NaiveBayesUpdateable(NBU), MLP, Simple Logistic(SL), SMO, Decision Table(DT), Jrip, OneR, OneR, DecsionStump(DS), J48, Random Forest(RF), RandomTree(RT), REPtree(RepT) are evaluated through the educational data set.

V. RESULT & DISCUSSION

In this research work different snp shorts are calculated with the help of weka tool. All these snap shorts are given below:

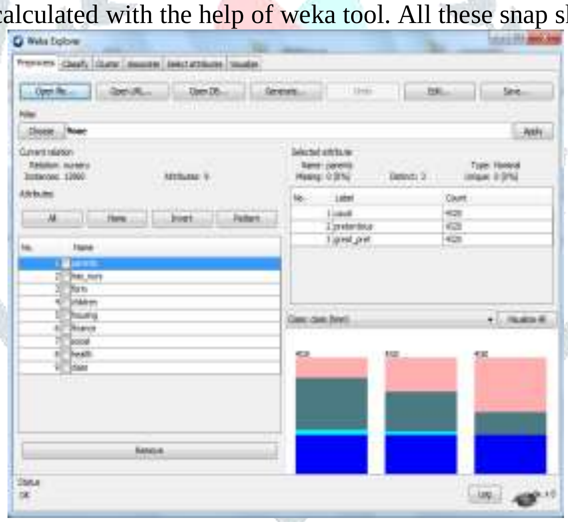

Figure 1: Database open file in weka

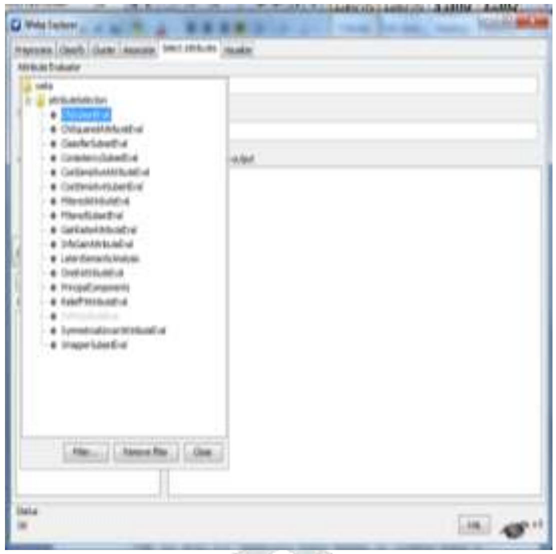

Figure 2: Select Fetures with selection attributes

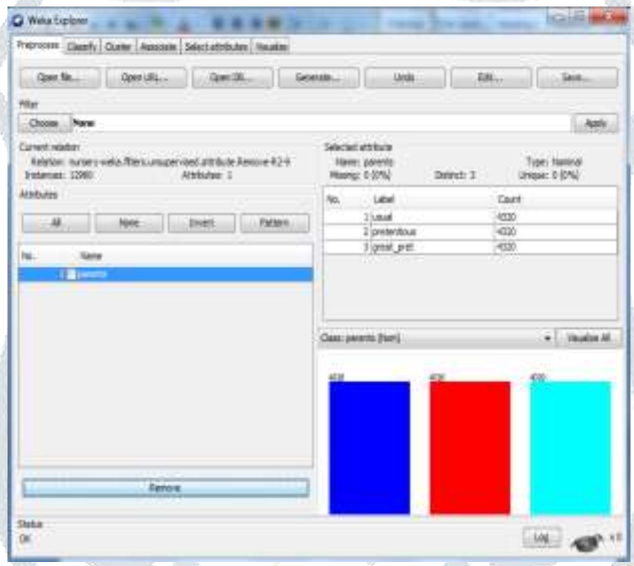

Figure 3: one feature is selected and removed all others

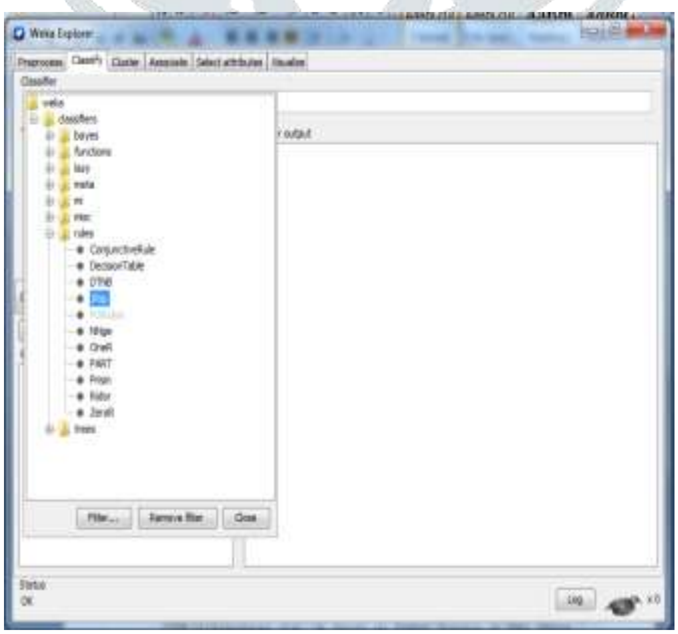

Figure 4: Classify the selected feature with Jrip

JRIP rules:
Time taken to build model: 1.1 seconds

Table 1. JRIP rules Stratified cross-validation

Correctly Classified Instances	9198	70.9722 %
Incorrectly Classified Instances	3762	29.0278 %

Table 2: Performance Parameters of JRIP rules

Kappa statistic	0.5701
Mean absolute error	0.1386
Root mean squared error	0.2632
Relative absolute error	50.7454 %
Root relative squared error	71.2424 %
Coverage of cases (0.95 level)	99.9846 %
Mean rel. region size (0.95 level)	40 %
Total Number of Instances	12960

Table 3: Detailed Accuracy By Class Using JRIP rules

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	not_recom
0.000	0.000	0.000	0.000	0.000	0.000	0.700	0.000	recommend
0.000	0.000	0.000	0.000	0.000	0.000	0.839	0.074	very_recom
0.565	0.219	0.558	0.565	0.562	0.345	0.777	0.532	priority
0.610	0.208	0.571	0.610	0.590	0.395	0.791	0.536	spec_prior

Table 4: Confusion Matrix by JRIP rules

A	**b**	**c**	**d**	**e**	classified as
4320	0	0	0	0	a = not_recom
0	0	0	2	**0**	b = recommend
0	0	0	328	**0**	c = very_recom
0	0	0	2412	1854	d = priority
0	0	0	1578	2466	e = spec_prior

Table 5. Decision Table Stratified cross-validation

Correctly Classified Instances	12273	94.6991 %
Incorrectly Classified Instances	687	5.3009 %

Table 6: Performance Parameters of Decision Table

Kappa statistic	0.922
Mean absolute error	0.1148
Root mean squared error	0.1693
Relative absolute error	42.0421 %
Root relative squared error	45.8214 %
Coverage of cases (0.95 level)	100 %
Mean rel. region size (0.95 level)	100 %
Total Number of Instances	12960

Table 7: Detailed Accuracy By Decision Table

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	not_recom
0.000	0.000	0.000	0.000	0.000	0.000	0.522	0.000	recommend
0.482	0.000	0.988	0.482	0.648	0.685	0.957	0.692	very_recom
0.883	0.021	0.953	0.883	0.917	0.879	0.976	0.962	priority
0.996	0.056	0.890	0.996	0.940	0.914	0.985	0.949	spec_prior

Table 8: Confusion Matrix Decision Table

A	b	c		e	classified as
4320	0	0	**d** 0	0	a = not_recom
0	0	2	0	0	b = recommend
0	0	158	170	0	c = very_recom
0	0	0	3766	500	d = priority
0	0	0	15	4029	e = spec_prior

Decision Stump
Time taken to build model: 0.06 seconds

Table 8: Stratified cross-validation

Correctly Classified Instances	8586	66.25%
Incorrectly Classified Instances	4374	33.75 %

Table 9: Performance Parameters of Decision Stump

Kappa statistic	0.4959
Mean absolute error	0.1429
Root mean squared error	0.2673
Relative absolute error	52.3204 %
Root relative squared error	72.3355 %
Coverage of cases (0.95 level)	97.4537 %
Mean rel. region size (0.95 level)	33.3333 %
Total Number of Instances	12960

Table 10: Detailed Accuracy By Decision Stump

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	not_recom
0.000	0.000	0.000	0.000	0.000	0.000	0.400	0.000	recommend
0.000	0.000	0.000	0.000	0.000	0.000	0.669	0.038	very_recom
1.000	0.503	0.494	1.000	0.661	0.495	0.748	0.493	priority
0.000	0.000	0.000	0.000	0.000	0.000	0.742	0.468	spec_prior

Table 11: Confusion Matrix By Decision Stump

A	b	c	d	e	classified as
4320	0	0	0	0	a = not_recom
0	0	0	2	0	b = recommend
0	0	0	328	0	c = very_recom
0	0	0	4266	0	d = priority
0	0	0	4044	0	e = spec_prior

Table 12: Stratified cross-validation Using ZeroR predicts

Correctly Classified Instances	4320	33.3333 %
Incorrectly Classified Instances	8640	66.6667 %

Table 13: Performance Parameters of ZeroR predicts

Kappa statistic	0
Mean absolute error	0.4444
Root mean squared error	0.4714
Relative absolute error	100%
Root relative squared error	100%
Coverage of cases (0.95 level)	100%
Mean rel. region size (0.95 level)	100%
Total Number of Instances	12960

Table 14: Detailed Accuracy By ZeroR predicts

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	1.000	0.333	1.000	0.500	0.000	0.500	0.333	recommended
0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.333	priority
0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.333	not_recom

Table 15: Confusion Matrix By ZeroR predicts

a	b	c	classified as
4320	0	0	a = recommended
4320	0	0	b = priority
4320	0	0	c = not_recom

VI. CONCLUSION

EDM manipulates educational data to help educational institutions to plan educational strategies, in order to improve the educational quality. Prediction is one of the main areas in EDM. Prediction and analysis of student academic performance are essential for student academic growth. If the new subset turns to be better, it replaces best one. This process is repeated until a given stopping condition is satisfied. Ranking of features determines the importance of any individual feature, neglecting their possible interactions. Ranking methods are based on statistics, information theory, or on some functions of classifier's outputs. In this paper different problems like dimensionality of a domain expands, the number of features N increases. Finding an optimal feature subset is intractable and problems related feature selections have been proved to be NP-hard. Future all the problems are resolved with Feature Selection Algorithm on student data with Naïve Bayes (NB), Decision tree and decision table. Feature selection has proven in both theory and practice to be effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learned results. Each candidate subset is evaluated and compared with the previous best one according to a certain evaluation.

REFERENCES

[1] Maryam Zaffar et.al. “Performance Analysis of Feature Selection Algorithm for Educational Data Mining” IEEE Conference on Big Data and Analytics (ICBDA)-2017.

[2] R. Sasi Regha et.al. “Optimization Feature Selection for classifying student in Educational Data Mining” International Journal of Innovations in Engineering and Technology (IJIET) , Volume 7 Issue 4 December 2016.

[3] Dr.M.Chidambaram et.al. “A Survey on Feature Selection in Data Mining ” International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN: 2347-5552, Volume-4, Issue-1, January 2016.

[4] Mital Doshi et.al. “Survey of Feature Selection Algorithms in Higher Education” International Journal of Computer Applications in Engineering Sciences [VOL IV, ISSUE I, MARCH 2014].

[5] Anal Acharya et.al. “Application of Feature Selection Methods in Educational Data Mining” International Journal of Computer Applications © 2014 by IJCA Journal Volume 103 - Number 2 Year of Publication: 2014.

[6] M. Ramaswami et.al. “A Study on Feature Selection Techniques in Educational Data Mining” Journal Of Computing, Volume 1, Issue 1, December 2009.

[7] E. Osmanbegović, M. Suljić, and H. Agić, "DETERMINING DOMINANT FACTOR FOR STUDENTS PERFORMANCE PREDICTION BY USING DATA MINING CLASSIFICATION ALGORITHMS," Tranzicija, vol. 16, pp. 147-158, 2015.

[8] A. M. Shahiri and W. Husain, "A review on predicting student's performance using data mining techniques," Procedia Computer Science, vol. 72, pp. 414-422, 2015.

[9] C. Romero and S. Ventura, "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, pp. 601-618, 2010.

[10] M. Ramaswami and R. Bhaskaran, "A study on feature selection techniques in educational data mining," arXiv preprint arXiv:0912.3924, 2009.

[11] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques," International Journal of Modern Education and Computer Science, vol. 8, p. 36, 2016.

[12] W. Punlumjeak and N. Rachburee, "A comparative study of feature selection techniques for classify student performance," in Information Technology and Electrical Engineering (ICITEE), 2015 7th International Conference on, 2015, pp. 425- 429.