ISSN: 2321-7782 (Online)
Volume 3, Issue 9, September 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

A Review on Using Clustering and Classification Techniques to Predict Student Failure with High Dimensional and Imbalanced Data

Sanket M. Bhandari[1]
Department of Computer and Science and Engineering
Parul Institute of Technology
Vadodara, India

Krunal Patel[2]
Department of Computer and Science and Engineering
Parul Institute of Technology
Vadodara, India

Abstract: ***Currently there is an increasing interest in data mining and educational systems, making educational data mining as a new growing research community. Predicting student failure at school has become a difficult challenge due to both the high number of factors that can affect the low performance of students and the imbalanced nature of these types of datasets, Educational data mining concerns the prediction of student failures in different levels. So this paper is about an early prediction of students' failure using different data mining techniques that may help the management provide timely counselling as well coaching to increase success rate and student retention.***

Keywords: Educational Data Mining, Predicting Student Performance, Student Failure, Classification, Clustering.

I. INTRODUCTION

Data mining is the iterative and interactive process of discovering valid, novel, useful, and understandable knowledge (patterns, models, rules etc.) in Massive databases. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology and data dredging. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (outliers detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps [8].

Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in.

II. LITERATURE SURVEY

There are some literature surveys on predicting student failure.

S.Taruna, Mrinal Pandey [1] compares five classification algorithms namely Decision Tree, Naïve Bayes, Naïve Bayes Tree, K-Nearest Neighbor and Bayesian Network algorithms for predicting students' grade particularly for engineering students. This is a four class prediction problem. Student's marks are classified into four classes A, B, C and F respectively. Initially complete data set is used to build the classifiers then Bootstrap method is used to improve the accuracy of the each classifier. Bootstrap method is a resample function available in WEKA tool kit. The excellent results of this function can be seen through

IBK, Decision Tree and Bayes Net algorithm. However the overall results of all four algorithms are good but the results of individual classes for Naïve Bayes and NB Tree is not sufficient enough for the individual class prediction particularly for this study. This paper also presents a comparative study of the previous work related to student's performance predictions.

This research can be enhanced by improving the results of Naïve Bayes and NB Tree algorithm by using other evaluating methods such as Boosting and Bagging etc. The study can be inclined toward the binary class predictions and various methods of binary class classifications can be studied for predicting the future results of the students.

Tripti Mishra, Dr. Dharminder Kumar, Dr. Sangeeta Gupta [2] use different classification techniques to build performance prediction model based on students' social integration, academic integration, and various emotional skills which have not been considered so far. Two algorithms J48 (Implementation of C4.5) and Random Tree have been applied to the records of MCA students of colleges affiliated to Guru Gobind Singh Indraprastha University to predict third semester performance. Random Tree is found to be more accurate in predicting performance than J48 algorithm.

The future research direction will include professional courses of B.Tech as well as the development of a decision support system to help authorities identify the weak students and take timely measures.

J.K. Jothi Kalpana, K. Venkatalakshmi [3] categorize the college student's academic performance for Villupuram district. Based on the clustering methods such as centroid based, distribution based and density based clustering. Cluster includes groups with small distance among the cluster members. The performance of student's multi-level of optimization formulated by using clustering. In centroid based clustering, clusters are represented by a central vector. The number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem. The clustering model most closely related to statistics is based on distribution model. Experiments attempts to improve the accuracy by using the method of Gaussian mixture model. The data set is modeled with a fixed number of Gaussian distribution that is initialized randomly and the parameters are iteratively optimized to fit better to the data set. The density based clustering method is a linkage based clustering. The range parameter ε produces a hierarchical result related to that of linkage clustering. Clustering can be represents in a large range of classifications and applications. K-means algorithm categorizes the large dataset. In this analysis use genetically improved particle swarm optimization algorithm to model the students level. The GAI-PSO algorithm searches the solution space to find the optimal result. The processing of refining use the k-means algorithm.

In future work, they aim to apply data mining techniques on an expanded data set with more distinctive attributes to get more accurate results.

Komal S. Sahedani, Prof. B Supriya Reddy [4] provides a review of the available literature on Educational Data mining, Classification method and different feature selection techniques that we should apply on Student dataset. The knowledge is hidden among the educational data set and it is extractable through data mining techniques.

Carlos Márquez-Vera, Cristóbal Romero Morales, and Sebastián Ventura Soto [5] proposes to apply data mining techniques to predict school failure and dropout. They use real data on 670 middle-school students from Zacatecas, México, and employ white-box classification methods, such as induction rules and decision trees. Experiments attempt to improve their accuracy for predicting which students might fail or dropout by first, using all the available attributes; next, selecting the best attributes; and finally, rebalancing data and using cost sensitive classification. The outcomes have been compared and the models with the best results are shown.

As future work, they aim to carry out more experiments using more data and also from different educational levels (primary, secondary, and higher) to test whether the same performance results are obtained with different DM approaches. Another aim to develop algorithm for classification/prediction based on grammar using genetic programming that can be compared versus classic algorithms. To predict the student failure as soon as possible. The earlier the better, in order to detect

students at risk in time before it is too late. To propose actions for helping students identified within the risk group. Then, to check the rate of the times it is possible to prevent the fail or dropout of that student previously detected.

Carlos Márquez-Vera, Alberto Cano, Cristóbal Romero, Sebastián Ventura [6] proposed a genetic programming algorithm and different data mining approaches for solving problems using real data about 670 high school students from Zacatecas, Mexico. Firstly, they select the best attributes in order to resolve the problem of high dimensionality. Then, rebalancing of data and cost sensitive classification have been applied in order to resolve the problem of classifying imbalanced data. They also propose to use a genetic programming model versus different white box techniques in order to obtain both more comprehensible and accuracy classification rules. The outcomes of each approach are shown and compared in order to select the best to improve classification accuracy, specifically with regard to which students might fail.

As future work, they aim to carry out more experiments using more data and also from different educational levels (primary, secondary, and higher) to test whether the same performance results are obtained with different DM approaches (feature selection, data balancing, and cost-sensitive classification) and their Interpretable Classification Rule Mining (ICRM) algorithm.

Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah [7] investigate the problem of website phishing using a developed AC method called Multi-label Classifier based Associative Classification (MCAC) to seek its applicability to the phishing problem. Experimental results using real data collected from different sources show that AC particularly MCAC detects phishing websites with higher accuracy than other intelligent algorithms. Further, MCAC generates new hidden knowledge (rules) that other algorithms (CBA,PART, C4.5, JRip, MCAR) are unable to find and this has improved its classifiers predictive performance.

III. EXPERIMENTS

In this section we carried out several experiments to compare the three versions of Interpretable Classification Rule Mining (ICRM) with the 10 classical classification algorithms.

These several experiments are carried out using real data about 670 high school students from Zacatecas, Mexico.

A. Classification results using all attributes [6]

Algorithm	TP rate	TN rate	Acc	GM	#Rules	#Conditions per rule	#Conditions
IBk	97.7	78.3	96.0	87.5	8.0	1.5	12.0
NNge	98.5	73.3	96.3	85.0	31.0	76.0	2356.0
OneR	98.9	41.7	93.7	64.2	2.0	0.5	1.0
Prism	99.5	25.0	93.1	49.9	76.0	1.4	110.0
Ridor	96.6	65.0	93.7	79.2	4.0	1.7	7.0
ADTree	99.7	76.7	97.6	87.4	21.0	1.7	36.0
J48	97.4	53.3	93.4	72.1	31.0	3.1	98.0
RandomTree	95.7	48.3	91.5	68	212.0	4.9	1041.0
REPTree	98.0	56.7	94.3	74.5	44.0	1.8	83.0
SimpleCart	97.7	65.0	94.8	79.7	5.0	12.8	64.0
ICRM v1	94.3	90.0	93.9	91.9	2.0	1.5	3.1
ICRM v2	97.5	75.0	95.5	85.0	7.6	1.9	14.7
ICRM v3	84.4	93.3	85.2	88.5	4.0	1.1	4.5

B. Classification results using the best attributes [6]

Algorithm	TP rate	TN rate	Acc	GM	#Rules	#Conditions per rule	#Conditions
IBip	97.0	81.7	95.7	89.0	5.7	1.5	8.7
NNge	98.0	76.7	96.1	86.7	22.2	14.0	310.8
OneR	98.9	41.7	93.7	64.2	2.0	0.8	1.6
Prism	99.2	44.2	94.7	66.2	55.6	1.7	93.8
Ridor	95.6	68.3	93.1	80.8	4.0	1.2	5.4
ADTree	99.2	78.3	97.3	88.1	21.0	3.0	63.0
J48	97.7	55.5	93.9	73.6	19.9	2.1	43.0
RandomTree	98.0	63.3	94.9	78.8	278.6	3.3	912.2
REPTree	97.9	60.0	94.5	76.6	30.0	1.9	68.4
SimpleCart	98.0	65.0	95.1	79.8	6.9	4.1	29.4
ICRM v1	92.0	93.3	92.1	92.5	2.0	2.4	4.9
ICRM v2	97.2	71.7	94.9	82.8	8.2	2.1	17.9
ICRM v3	75.9	85.0	76.7	79.0	4.0	0.9	3.8

C. Classification results using all attributes [6]

Algorithm	TP rate	TN rate	Acc	GM	#Rules	#Conditions per rule	#Conditions
IBip	97.7	78.3	96.0	87.5	8.0	1.5	12.0
NNge	98.5	73.3	96.3	85.0	31.0	76.0	2356.0
OneR	98.9	41.7	93.7	64.2	2.0	0.5	1.0
Prism	99.5	25.0	93.1	49.9	76.0	1.4	110.0
Ridor	96.6	65.0	93.7	79.2	4.0	1.7	7.0
ADTree	99.7	76.7	97.6	87.4	21.0	1.7	36.0
J48	97.4	53.3	93.4	72.1	31.0	3.1	98.0
RandomTree	95.7	48.3	91.5	68	212.0	4.9	1041.0
REPTree	98.0	56.7	94.3	74.5	44.0	1.8	83.0
SimpleCart	97.7	65.0	94.8	79.7	5.0	12.8	64.0
ICRM v1	94.3	90.0	93.9	91.9	2.0	1.5	3.1
ICRM v2	97.5	75.0	95.5	85.0	7.6	1.9	14.7
ICRM v3	84.4	93.3	85.2	88.5	4.0	1.1	4.5

D. Classification results using the best attributes [6]

Algorithm	TP rate	TN rate	Acc	GM	#Rules	#Conditions per rule	#Conditions
IBip	97.0	81.7	95.7	89.0	5.7	1.5	8.7
NNge	98.0	76.7	96.1	86.7	22.2	14.0	310.8
OneR	98.9	41.7	93.7	64.2	2.0	0.8	1.6
Prism	99.2	44.2	94.7	66.2	55.6	1.7	93.8
Ridor	95.6	68.3	93.1	80.8	4.0	1.2	5.4
ADTree	99.2	78.3	97.3	88.1	21.0	3.0	63.0
J48	97.7	55.5	93.9	73.6	19.9	2.1	43.0
RandomTree	98.0	63.3	94.9	78.8	278.6	3.3	912.2
REPTree	97.9	60.0	94.5	76.6	30.0	1.9	68.4
SimpleCart	98.0	65.0	95.1	79.8	6.9	4.1	29.4
ICRM v1	92.0	93.3	92.1	92.5	2.0	2.4	4.9
ICRM v2	97.2	71.7	94.9	82.8	8.2	2.1	17.9
ICRM v3	75.9	85.0	76.7	79.0	4.0	0.9	3.8

E. Average ranking of classification results [6]

Algorithm	TP rate	TN rate	Acc	GM	#Rules	#Conditions per rule	#Conditions
All attributes	1.92	3.38	2.46	2.46	1.85	2.23	2.38
Feature selection	2.62	2.69	2.46	2.69	1.62	1.92	2.00
Data balancing	2.23	1.54	2.08	1.54	3.00	2.85	3.15
Cost sensitive	3.00	2.15	2.54	2.23	2.00	2.31	2.23

IV. CONCLUSION

As we have seen, predicting student failure can be a difficult task not only because it is a multifactor problem (in which there are a lot of personal, family, social, and economic factors that can be influential) but also because the available data are normally imbalanced (most of the students pass to the next course). To resolve these problems, we have shown the use of different DM algorithms and approaches for predicting student failure. Our main focus is experiment on more data, also from different educational level with different data mining approaches and to predict the student failure as soon as possible.

References

PAPERS

1. S.Taruna, Mrinal Pandey, "An Empirical Analysis of Classification Techniques for Predicting Academic Performance", IEEE International Advance Computing Conference (IACC), 2014.
2. Tripti Mishra, Dr. Dharminder Kumar, Dr. Sangeeta Gupta, "Mining Students' Data for Performance Prediction", IEEE Fourth International Conference on Advanced Computing & Communication Technologies, 2014.
3. J.K. Jothi Kalpana, K. Venkatalakshmi, "Intellectual Performance Analysis of Students by Using Data Mining Techniques", IEEE International Conference on Innovations in Engineering and Technology, Volume 3, Special Issue 3, March 2014.
4. Komal S. Sahedani, Prof. B Supriya Reddy, "A Review: Mining Educational Data to Forecast Failure of Engineering Students", IJARCSSE, Volume 3, Issue 12, December 2013.
5. Carlos Márquez-Vera, Cristóbal Romero Morales, and Sebastián Ventura Soto, "Predicting School Failure and Dropout by Using Data Mining Techniques", IEEE journal of Latin-American learning technologies, VOL. 8, NO. 1, FEBRUARY 2013.
6. Carlos Márquez-Vera, Alberto Cano, Cristóbal Romero, Sebastián Ventura, "Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data", Springer 2013, 38, 315-330.
7. Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah, "Phishing detection based Associative Classification data mining", Springer 2014, Expert Systems with Applications 41 (2014) 5948–5959.

BOOKS

1. Jiawei Han, Micheline Kamber, Data mining: concepts and techniques, 2/e. Morgan Kaufmann; 2006.

