Research Article Impact Factor: 0.621 ISSN: 2319-507X
Mohini Mukund Mohod, IJPRET, 2014; Volume 2 (9): 697-708 IJPRET

INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

PERFORMANCE ANALYSIS OF PREDICTIVE DATA MINING MODEL USING CONFUSION MATRIX

MOHINI MUKUND MOHOD, SAYALI RAJESH SUYAL

Department of Information Technology, C. K. Thakur A.C.S. College, New Panvel, India.

Accepted Date: 27/02/2014 ; Published Date: 01/05/2014

Abstract: The education system of a nation influences progressive nation building. It plays a vital role in the personal growth of a student and the social development among all. However, in real world, predicting the performance of the students is a challenging task. The main objective of higher education institutions is to provide quality education to their students. With the help of some data mining techniques the institutes can evaluate the student performance and overcome the problems of low grades. The educationalists can even predict the future performance of the students using some predictive data mining models such as classification, regression, time series analysis etc. In this paper, the classification technique is used to predict the future performance of the student. Before applying the explored classification rules to all the student data tuples in the educational database and derive conclusion about the performance of any student, it is essential to ensure the accuracy of the classifier. This accuracy is verified using confusion matrix. A confusion matrix contains information about actual and predicted classifications. The confusion matrix has proved as a useful tool for analyzing the classification model in our case study.

Keywords: Educational Database, Data Mining, Classification, Confusion Matrix.

PAPER-QR CODE

Corresponding Author: MS. MOHINI MUKUND MOHOD

Access Online On:

www.ijpret.com

How to Cite This Article:

Mohini Mukund Mohod, IJPRET, 2014; Volume 2 (9): 697-708

INTRODUCTION

We live in world where vast amount of data are collected daily. Analyzing such data is an important need. Data mining meets this need by providing techniques to discovered knowledge from data. Now a day, many educational organizations have started developing and improving the educational systems. The best decisions can be made by applying the data mining techniques on huge educational databases. The data mining has been called exploratory data analysis, data driven discovery and deductive learning [1]. A very significant tool for categorization of data for its most effective and efficient use in data mining, is classification. Classification is a two step process. In the first step, a classifier is built based on previous data. In Second step, if the classifier's accuracy is acceptable, it is used to classify the new data. The confusion matrix is useful to evaluate the performance of a classifier, showing the number per class of well classified and mislabeled instances.

The data required for this paper is explored from the educational database of 'C. K. Thakur Arts, Commerce and Science College, New Panvel affiliated with University of Mumbai'. The first half of the paper shows how the data is collected, pre-processed and how classification rules are extracted from the data. In the second half, we have verified the accuracy of our analysis using confusion matrix. This paper investigates the accuracy of classification technique for predicting student performance.

Data Mining Models

Data mining involves many different algorithms to accomplish different tasks. The purpose of these algorithms is to fit a model to the data. A data mining model can be either predictive or descriptive in nature. A predictive model makes prediction about values of data using known results found from different data and other historical data. The predictive models include classification, regression, time series analysis and prediction. A descriptive model identifies patterns or relationships in data. It explores the properties of the data being examined. It does not predict new values of the properties like predictive models. The descriptive models include clustering, summarization, association rules and sequence discovery.

Data Collection and pre-processing

For our study, we have collected the student's data from "Department Of Information Technology" of 'C. K. Thakur Arts, Commerce and Science College, New Panvel affiliated with University of Mumbai'. On the basis of collected data, some attributes are considered to predict student's performance in Final Examination. The Attributes used for forecasting the student's

performance in Final Examination are the performance in First Year of the course, Attendance, Tutorial and Class Test as mentioned in Table I.

We live in world where vast amount of data are collected daily. Analyzing such data is an important need. Data mining meets this need by providing techniques to discovered knowledge from data. Now a day, many educational organizations have started developing and improving the educational systems. The best decisions can be made by applying the data mining techniques on huge educational databases. The data mining has been called exploratory data analysis, data driven discovery and deductive learning [1]. A very significant tool for categorization of data for its most effective and efficient use in data mining, is classification. Classification is a two step process. In the first step, a classifier is built based on previous data. In Second step, if the classifier's accuracy is acceptable, it is used to classify the new data. The confusion matrix is useful to evaluate the performance of a classifier, showing the number per class of well classified and mislabeled instances.

The data required for this paper is explored from the educational database of 'C. K. Thakur Arts, Commerce and Science College, New Panvel affiliated with University of Mumbai'. The first half of the paper shows how the data is collected, pre-processed and how classification rules are extracted from the data. In the second half, we have verified the accuracy of our analysis using confusion matrix. This paper investigates the accuracy of classification technique for predicting student performance.

Data Mining Models

Data mining involves many different algorithms to accomplish different tasks. The purpose of these algorithms is to fit a model to the data. A data mining model can be either predictive or descriptive in nature. A predictive model makes prediction about values of data using known results found from different data and other historical data. The predictive models include classification, regression, time series analysis and prediction. A descriptive model identifies patterns or relationships in data. It explores the properties of the data being examined. It does not predict new values of the properties like predictive models. The descriptive models include clustering, summarization, association rules and sequence discovery.

Data Collection and pre-processing

For our study, we have collected the student's data from "Department Of Information Technology" of 'C. K. Thakur Arts, Commerce and Science College, New Panvel affiliated with University of Mumbai'. On the basis of collected data, some attributes are considered to predict student's performance in Final Examination. The Attributes used for forecasting the student's

performance in Final Examination are the performance in First Year of the course, Attendance, Tutorial and Class Test as mentioned in Table I.

TABLE I. Student Data Attributes

Attributes	*Description*	*Values*
First Year Examination %	Percentage of marks obtained in First Year Examination.	Distinction, First, Second, Third, Fail
Attendance %	Attendance of the student during the academic year.	Excellent, Very Good, Good, Poor
Tutorial	Marks obtained in Tutorial Examination	Good, Average, Poor
Class Test	Marks obtained in Class Test	Good, Average, Poor

Research Methodology

Classification: Classification is an analytical task where the classifier is constructed to predict the categorical labels as "Advanced", "Medium" and "Slow". We have divided classification of aforementioned educational data into two steps. In the first step, classifier describes the predefined set of data classes. This is the training phase where a student tuple S is represented as an attribute vector S = (x1,x2,x3,x4) where x1,x2,x3,x4 are the values of attributes FY_result, Attendance, Tutorial and Class Test respectively. The classification rules obtained at the end of the training phase are as mentioned in Table II.

Table II. Classification Rules

IF fy_result = distinction AND attendance = good AND tutorial = good AND class test = average THEN final_exam_performance = Medium
IF fy_result = first_class AND attendance = very good AND tutorial = good AND class test = poor THEN final_exam_performance = Medium
IF fy_result = third_class AND attendance=good AND tutorial = poor AND class test = average THEN final_exam_performance = Slow
IF fy_result = second_class AND attendance = very good AND tutorial = good AND class test = average THEN final_exam_performance = Medium
IF fy_result = third_class AND attendance = poor AND tutorial = average AND class test = average THEN final_exam_performance = Medium
IF fy_result = second_class AND attendance = very good AND tutorial = good AND class test = average THEN final_exam_performance = Medium
IF fy_result = first_class AND attendance = very good AND tutorial = good AND class test = good THEN final_exam_performance = Advanced

IF fy_result = second_class AND attendance = very good AND tutorial = average AND class test = average THEN final_exam_performance = Medium
IF fy_result = second_class AND attendance = good AND tutorial = average AND class test = poor THEN final_exam_performance = Slow
IF fy_result = third_class AND attendance=poor AND tutorial = average AND class test = average THEN final_exam_performance = Slow
IF fy_result = second_class AND attendance = poor AND tutorial = average AND class test = poor THEN final_exam_performance = Medium
IF fy_result = second_class AND attendance = good AND tutorial = poor AND class test = average THEN final_exam_performance = Medium
IF fy_result = distinction AND attendance=very good AND tutorial = average AND class test = average THEN final_exam_performance = Advanced
IF fy_result = first_class AND attendance = good AND tutorial = good AND class test = poor THEN final_exam_performance = Medium
IF fy_result = second_class AND attendance = good AND tutorial = good AND class test = poor THEN final_exam_performance = Medium
IF fy_result = third_class AND attendance = good AND tutorial = good AND class test = poor THEN final_exam_performance = Slow
IF fy_result = second_class AND attendance = very good AND tutorial = good AND class test = good THEN final_exam_performance = Advanced
IF fy_result = third_class AND attendance = poor AND tutorial = poor AND class test =poor THEN final_exam_performance = Slow
IF fy_result = third_class AND attendance = good AND tutorial = average AND class test = poor THEN final_exam_performance = Slow

The classification rules mentioned in Table II predict student's performance in their final examination. The second step of classification process estimates the predictive accuracy of the classifier. If the accuracy is acceptable, the classifier can be used for further data tuples for which the class label (Advanced, Medium or Slow) is not known i.e. it can be used for predicting the performance of various students other than the students listed in training set of classification.

Confusion Matrix: A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for the aforementioned classifier.

Confusion Matrix				
Actual Values	Predicted Values			
		Advanced	Medium	Slow
	Advanced	No. of stud 7 (a)	No. of stud 2 (b)	No. of stud 0 (c)
	Medium	No. of stud 1 (d)	No. of stud 12 (e)	No. of stud 2 (f)
	Slow	No. of stud 0 (g)	No. of stud 1 (h)	No. of stud 5 (i)

The entries in the confusion matrix have the following meaning in the context of our study:

- a is the number of correct predictions that a student is advanced,
- b and c is the number of incorrect predictions that a student is advanced,
- e is the number of correct predictions that a student is medium,
- d and f is the number of incorrect of predictions that a student is medium,
- i is the number of correct predictions that a student is medium and
- g and h is the number of correct predictions that a student is medium.

With the help of the entries of aforementioned confusion matrix, the accuracy of the classifier is calculated.

The accuracy (AC) is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$AC = \frac{a + e + i}{a + b + c + d + e + f + g + h + i} \quad \text{------ (1)}$$

$$= \frac{7 + 12 + 5}{7 + 2 + 0 + 1 + 11 + 2 + 0 + 1 + 5}$$

$$= 80.00\ \%$$

Assuming the confusion matrix above, its corresponding table of confusion, for the *advanced* class, would be:

Advanced -Vs-All

		Predicted Values	
		Advanced	Other
Actual Values	Advanced	11 (a)	2 (b)
	Other	1 (c)	16 (d)

The recall or true positive rate (TP) is the proportion of *advanced* cases that were correctly identified, as calculated using the equation:

$$TP_A = \frac{a}{a+b} \quad \text{---------- (2)}$$

$$= \frac{11}{11+2}$$

$$= 84.62\ \%$$

The false positive rate (FP) is the proportion of other cases that were incorrectly classified as *advanced*, as calculated using the equation:

$$FP_A = \frac{c}{c+d} \quad \text{---------- (3)}$$

$$= \frac{1}{1+16}$$

$$= 5.88\ \%$$

The true negative rate (TN) is defined as the proportion of other cases that were classified correctly, as calculated using the equation:

$$TN_A = \frac{d}{d+c} \quad \text{---------- (4)}$$

$$= \frac{16}{16+1}$$

$$= 94.11\ \%$$

The false negative rate (FN) is the proportion of *advanced* cases that were incorrectly classified as other, as calculated using the equation:

$$FN_A = \frac{b}{b + a} \quad \text{---------- (5)}$$

$$= \frac{2}{2 + 11}$$

$$= 15.38\ \%$$

Finally, precision (P) is the proportion of the predicted *advanced* cases that were correct, as calculated using the equation:

$$P_A = \frac{a}{a + c} \quad \text{---------- (6)}$$

$$= \frac{11}{11 + 1}$$

$$= 96.66\ \%$$

The corresponding table of confusion, for the *medium* class, would be:

Medium -Vs-All

		Predicted Values	
		Medium	Other
Actual Values	Medium	12 (a)	3 (b)
	Other	3 (c)	12 (d)

The recall or true positive rate (TP) is the proportion of *medium* cases that were correctly identified, as calculated using the equation:

$$TP_M = \frac{a}{a + b} \quad \text{---------- (7)}$$

$$= \frac{12}{12 + 3}$$

= 80.00 %

The false positive rate (FP) is the proportion of other cases that were incorrectly classified as *medium*, as calculated using the equation:

$$FP_M = \frac{c}{c + d} \quad \text{---------- (8)}$$

$$= \frac{3}{3 + 12}$$

= 20.00 %

The true negative rate (TN) is defined as the proportion of other cases that were classified correctly, as calculated using the equation:

$$TN_M = \frac{d}{d + c} \quad \text{---------- (9)}$$

$$= \frac{12}{12 + 3}$$

= 80.00 %

The false negative rate (FN) is the proportion of *medium* cases that were incorrectly classified as other, as calculated using the equation:

$$FN_M = \frac{b}{b + a} \quad \text{---------- (10)}$$

$$= \frac{3}{3 + 12}$$

= 20.00 %

Finally, precision (P) is the proportion of the predicted *medium* cases that were correct, as calculated using the equation:

$$P_M = \frac{a}{a + c} \quad \text{---------- (11)}$$

$$= \frac{12}{12+3}$$

$$= 80.00\ \%$$

The corresponding table of confusion, for the *slow* class, would be:

Slow -Vs-All

		Predicted Values	
		Slow	Other
Actual Values	Slow	5 (a)	1 (b)
	Other	2 (c)	18 (d)

The recall or true positive rate (TP) is the proportion of *slow* cases that were correctly identified, as calculated using the equation:

$$TP_S = \frac{a}{a+b} \qquad \text{---------- (12)}$$

$$= \frac{5}{5+1}$$

$$= 83.33\ \%$$

The false positive rate (FP) is the proportion of other cases that were incorrectly classified as *slow*, as calculated using the equation:

$$FP_S = \frac{c}{c+d} \qquad \text{---------- (13)}$$

$$= \frac{2}{2+18}$$

$$= 10.00\ \%$$

The true negative rate (TN) is defined as the proportion of other cases that were classified correctly, as calculated using the equation:

$$TN_S = \frac{d}{d + c} \quad \text{---------- (14)}$$

$$= \frac{18}{18 + 2}$$

$$= 90.00\ \%$$

The false negative rate (FN) is the proportion of *slow* cases that were incorrectly classified as other, as calculated using the equation:

$$FN_S = \frac{b}{b + a} \quad \text{---------- (15)}$$

$$= \frac{1}{1 + 5}$$

$$= 16.66\ \%$$

Finally, precision (P) is the proportion of the predicted *slow* cases that were correct, as calculated using the equation:

$$P_S = \frac{a}{a + c} \quad \text{---------- (16)}$$

$$= \frac{5}{5 + 2}$$

$$= 71.42\ \%$$

The accuracy determined using equation (1) may not be an adequate performance measure when the number of negative cases is much greater than the number of positive cases. Suppose there are 1000 cases, 995 of which are negative cases and 5 of which are positive cases. If the system classifies them all as negative, the accuracy would be 99.5%, even though the classifier missed all positive cases. Hence the precision is calculated for all the classes to cross check the accuracy.

CONCLUSION

In this paper, we explored the potential usefulness of data mining techniques in enhancing the quality of student performance. The study will help to identify those students which need special attention to reduce failure rate. A predictive data mining technique called classification

is used to predict student's future performance. The classifier extracted from this classification process is verified by confusion matrix methodology to check the efficiency of classifier. For future work, few more data mining techniques can be used to detect the outliers in the educational database for more accurate predictions about the student's performance. Also few more techniques can be applied in addition to confusion matrix such as G-Mean, F-Measure and ROC graph for profound verification of results of predictions.

REFERENCES

1. Margaret H. Dunham, "Data Mining: Introductory and advanced Topics", Pearson 2013

2. Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques", Elsevier 2009

3. E. Chandra, K. Namdini , "Knowledge Mining From Student Data", European Journal of Scientific Research, Vol. 47

4. Manpreet Singh Bhullar, "Use of Data Mining In Educational Sector", WCECS 2012 Vol- I , October 24-26, 2012

5. Rajan chattamvelli, "Data Mining Methods", Narosa 2009

6. David Cheung, Graham Williams, Qing Li, "Advances in Knowledge Discovery and Data Mining", PAKDD 2001

7. Saurabh Pal, Surjeet Kumar Yadav, " Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", WCSIT, ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012

8. Sudheep Elayidom, Sumam Mary Idikkula, Joseph Alexander, "Applying Statistical Dependency Analysis Techniques In a Data Mining Domain", (IJDE), Volume (1)

9. A. K. Santra, C. Josephine Christy, "Genetic Algorithm and Confusion Matrix for Document Clustering", IJCSI, Vol. 9, Issue 1, No 2, January 2012, ISSN (Online): 1694-0814

10. V. Ramesh, P. Parkavi, P. Yasodha, "Performance Analysis of Data Mining Techniques for Placement Chance Prediction", International Journal of Scientific & Engineering Research Volume 2, Issue 8, August-2011 1 ISSN 2229-5518