Quality IMPROVISATION OF STUDENT PERFORMANCE USING DATA MINING TECHNIQUES

SAYALI RAJESH SUYAL, MOHINI MUKUND MOHOD

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

Abstract— *Mining of gold from sand and rocks is referred to as gold mining rather than sand or rock mining. Thus data mining is nothing but the 'knowledge mining' from the data. Data mining techniques extract knowledge from large amount of data about a system. This knowledge can be used for taking various strategic decisions as well as finding the solutions towards the betterment of the system. 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. In this paper we have discussed various data mining techniques which can be used to improve the academic performance of the students. The main objective of higher education institutions is to provide quality education to their students. With the help of some data mining techniques like Association Rules, Classification, and KDD (Knowledge Discovery in Database), the institutes can evaluate the student performance and overcome the problems of low grades. The data required for this evaluation is hidden in huge educational databases. Data mining techniques not only abstract this data but also make predictions about student's performance and suggest remedies depending upon the performance.*

Key Terms *- Association Rules, Classification, Data Preprocessing, Educational Database*

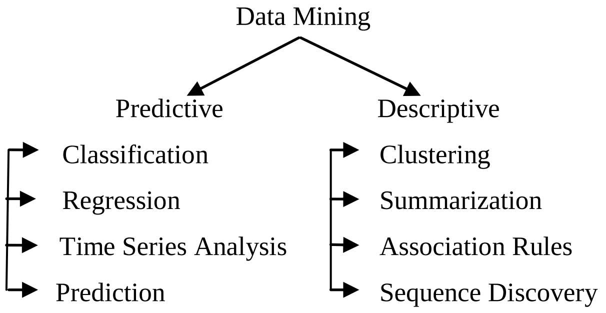
INTRODUCTION

In various institutes, the amount of educational data kept in computer files and database growing rapidly. At the same time educationalists are expecting more sophisticated information from them. An educationalist is no longer satisfied with a simple listing of student's data but want detail information about the student's past academic performance as well as the prediction of their academic future. Simple structured or query languages are not adequate to support these increased demands for information. Data mining steps in to fulfill these needs. The data mining has been called exploratory data analysis, data driven discovery and deductive learning [1].

The main objective of our paper is to improvise student's performance by studying the overall academic growth of the student and predicting her learning status for some future examinations. If the predicted performance of the student is not up to the mark, the institute can provide some remedial coaching or regular consultations to such students to boost them to cope up with their studies. The predictions are made by data mining techniques applied on the data extracted from the database of the institute. We have applied association rules mining on the data to find the relationships among various attributes. The association rules reveal the hidden patterns of the students with poor performance as well as the hidden patterns of the students with good performance. Also, we have used classification as a data mining technique and applied it on the main attributes that may affect the student's performance. The extracted classification rules predict the values measuring the future performance of the students. For example, one may predicate student's final grades using classification rules. The association rules are descriptive i.e. they characterize current scenario and classification rules are predictive i.e. they characterize future scenario.

The data required for the case study in this paper is explored from the database of 'C. K. Thakur Arts, Commerce and Science College, New Panvel affiliated with University of Mumbai'. The research is divided in two parts. The first part of the paper shows how the data is collected, pre-processed and how to apply data mining mechanisms on the data. The rest of the paper shows how can we benefitted from the discovered knowledge.

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.

Classification: Classification involves the predictive learning that classifies a data item into one of several predefined classes. It involves examining the features of an item and assigning to it a predefined class. Classification is a two-step process. First a model is built describing a predefined set of data classes and secondly, the model is used for classification.

Regression: Regression is a statistical technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line (y = mx + b) and determines the appropriate values for ‘m’ and ‘b’ to predict the value of ‘y’ based upon a given value of ‘x’. Advanced techniques, such as multiple regressions, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation.

Time Series Analysis: A Time series is a sequence of data points, measured typically at successive points in space-time at uniform time intervals. Time Series Analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the process of using a model to generate predictions for future events based on known past events.

Prediction: Prediction technique applies various functions on data set to predict the unknown or missing values.

Clustering: Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.

Summarization: It is the abstraction or generalization of data. A set of relevant data is summarized and abstracted, resulting smaller set which gives a general overview of the data and usually with aggregation information.

Association Rules: Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

Sequence Discovery: Sequence Discovery is concerned with finding statistically relevant patterns between data where the values are delivered in a sequence.

DATA COLLECTION AND PRE-PROCESSING

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. In Our Research paper, data is acquired 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 results are HSC, SSC, Attendance, Tutorial, Class Test as mentioned in Table I.

TABLE I. STUDENT DATA ATTRIBUTES

Attributes	***Description***	***Values***
SSC %	Percentage of marks obtained in Secondary School Certificate Examination.	Distinction, First, Second, Third, Fail
HSC%	Percentage of marks obtained in Higher Secondary Certificate Examination.	Distinction, First, Second, Third, Fail
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

Association Rules: Association mining is about discovering a set of rules that is shared among a large percentage of the data. Association rules mining tend to produce a large number of rules. The goal is to find the rules that are useful to users. There are two criteria of measuring usefulness viz. support and confidence. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.
The association rules problem is as follows:

Let I= $\{i_1, i_2\ldots i_n\}$ be a set of literals call items. Let D be a set of all transactions where each transaction T is a set of items such that $T \subseteq I$. Let X, Y be a set of items such that $X, Y \subset I$. An association rule is an implication in the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, $X \cap Y = \emptyset$.

Support: The rule $X \Rightarrow Y$ holds with support s if s% of transactions in D contains $X \cup Y$. Rules that have a s greater than a user-specified support are said to have minimum support.

Confidence: The rule $X \Rightarrow Y$ holds with confidence c if c% of the transactions in D that contain X also contain Y. Rules that have a c greater than a user-specified confidence are said to have minimum confidence.

The association between various data items can be found out by mining multilevel association rules, multidimensional association rules and/or quantitative association rules. Multilevel association rules involve concepts at different levels of abstraction. Multidimensional association rules involve more than one dimension or predicate. Quantitative association rules involve numeric attributes that have implicit ordering among values e.g. grades of students.[2] We have mined quantitative association rules from the pre-processed data extracted from the educational database of ‘C. K. Thakur Arts, Commerce and Science College, New Panvel’. Each quantitative association rule has quantitative attributes on left – hand side of the rule and one categorical attribute on right-hand side of the rule.

A_{quan1} ^ A_{quan2} → A_{cat}

Hence, the association rules we have mined from the aforementioned database are mentioned in Table II.

TABLE II. ASSOCIATION RULES

Association Rule	*Support (%)*	*Confidence (%)*
ssc (S1,70..100) ^ hsc (S1,70..100) ^ fy(S1,70..100) → Student(S1,"Advanced")	13.33	16.66
ssc (S2,55..69) ^ hsc (S2,40..55) ^ fy(S2, 55..69) → Student(S2,"Medium")	10	13.33
ssc (S3,40..55) ^ hsc (S3,40..55) ^ fy(S1, 55..69) → Student(S3,"Slow")	10	23.33
ssc (S4, 55..69) ^ hsc (S1, 55..69) ^ fy(S1, 55..69) → Student(S4,"Medium")	10	10
ssc (S5, 40..55) ^ hsc (S5,40..55) ^ fy(S5, 40..55) → Student(S5,"Slow")	13.33	23.33
ssc (S6,70..100) ^ hsc (S6, 70..100) ^ fy(S6, 55..69) → Student(S6,"Medium")	3.33	16.66
ssc (S7,70..100) ^ hsc (S7, 55..69) ^ fy(S7, 70..100) → Student(S7,"Advanced")	20	33.33
ssc (S8,70..100) ^ hsc (S8, 70..100) ^ fy(S8, 70..100) → Student(S8,"Advanced")	13.33	13.33
ssc (S14, 55..69) ^ hsc (S14, 40..55) ^ fy(S14, 40..55) → Student(S14,"Slow")	3.33	13.33
ssc (S20,70..100) ^ hsc (S20, 40..55) ^ fy(S20, 70..100) → Student(S20,"Medium")	3.33	3.33

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. The first step, the 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 = (x_1,x_2,x_3,x_4) where x_1,x_2,x_3,x_4 are the values of attributes Attendance, Tutorial and Class Test respectively. The corresponding class label (Advanced, Medium or Slow) is provided to each training tuple. We have obtained this label is by the association rules mining technique mentioned in the previous section of the paper. The classification rules obtained at the end of the training phase are as mentioned in Table III.

TABLE III. CLASSIFICATION RULES

IF student = advanced AND attendance = good AND tutorial = good AND class test = average THEN final_exam_performance = Medium
IF student = medium AND attendance = very good AND tutorial = good AND class test = poor THEN final_exam_performance = Medium
IF student = slow AND attendance = good AND tutorial = poor AND class test = average THEN final_exam_performance = Slow
IF student = medium AND attendance = very good AND tutorial = good AND class test = average THEN final_exam_performance = Medium
IF student = slow AND attendance = poor AND tutorial = average AND class test = average THEN final_exam_performance = Medium
IF student =medium AND attendance = very good AND tutorial = good AND class test = average THEN final_exam_performance = Medium
IF student = advanced AND attendance = very good AND tutorial = good AND class test = good THEN final_exam_performance = Advanced
IF student = medium AND attendance = very good AND tutorial = average AND class test = average THEN final_exam_performance = Medium
IF student = medium AND attendance = good AND tutorial = average AND class test = poor THEN final_exam_performance = Slow
IF student = slow AND attendance = poor AND tutorial = average AND class test = average THEN final_exam_performance = Slow
IF student = medium AND Attendance = poor AND tutorial = average AND class test = poor THEN final_exam_performance = Medium
IF student = medium AND attendance = good AND tutorial = poor AND class test = average THEN final_exam_performance = Medium
IF student = advanced AND attendance = very good AND tutorial = average AND class test = average THEN final_exam_performance = Advanced
IF student = advanced AND attendance = good AND tutorial = good AND class test = poor THEN final_exam_performance = Medium
IF student = medium AND attendance = good AND tutorial = good AND class test = poor THEN final_exam_performance = Medium
IF student = slow AND attendance = good AND tutorial = good AND class test = poor THEN final_exam_performance = Slow
IF student = medium AND attendance = very good AND tutorial = good AND class test = good THEN final_exam_performance = Advanced
IF student = slow AND attendance = poor AND tutorial = poor AND class test = poor THEN final_exam_performance = Slow
IF student = slow AND attendance = good AND tutorial = average AND class test = poor THEN final_exam_performance = Slow

The classification rules mentioned in Table III predict student's performance in their final examination. The second step of classification process estimates the predictive accuracy of the classifier. If we were to use the training set to measure the accuracy of the classifier, this estimate would likely be optimistic, because the classifier tends to overfit the data. Therefore, a test set is used made up of test tuples and their associated class labels. These tuples are independent of training tuples. The associated class label of each test tuple is compared with the trained classifier's class prediction for that tuple. This comparison has proved that the accuracy of the classification process we followed is acceptable and 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. The outcome of the classification process is the set of classification rules which predict the future performance of any student in the institute. The instructors in the institute can take the remedial action based on this prediction. They can provide the special coaching in advance, using various ICT (Information and Communication Technologies) methods to enhance the student's performance in their final examination.

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 descriptive data mining technique called association rules mining is used to describe the student's current performance and a predictive technique called classification is used to predict student's future performance. 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.

We are also working on investigation of similar patterns in student's withdrawal (which also includes student's socio-economic status parallel with the academics) from any course and help students as well as the institute in student retention along with the upgraded performance.

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] Agathe Merceron , Kalina Yacef , "Educational data mining A Case Study".

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

[6] Kalyani Moroney, Dr. Sanjay Makh, "Examination System For Performance Improvisation of Students using Regression" ISBN 978-81-923393-3-7.

[7] Rajan chattamvelli, "Data Mining Methods" Narosa 2009.

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

[9] Bing Liu, Wynne Hsu, Yiming Ma, "Integrating Classification and Association Rule Mining", KDD-98 Proceedings.

[10] 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.

Authors

First Author – Sayali Rajesh Suyal, M.Sc.(Computer Science) Department of Information Technology, Changu Kana Thakur A.C.S. College, New Panvel, India, sayali.suyal@gmail.com

Second Author – Mohini Mukund Mohod, M.C.A., Department of Information Technology, Changu Kana Thakur A.C.S. College, New Panvel, India, mohini_mohod@rediffmail.com