Predicting Student Performance using Classification Techniques

E.Chandra and K.Nandhini

Abstract-- **The ability to predicting the performance of a student is very essential task of all educational institutions. This will not be decided by using only the academic excellence of a student. The behaviors such as aptitude, attitude, communications, technological, interpersonally, problem solving ability etc., should be taken into care to predict the real excellence of a student. Since this is the task of prediction and mining the classification algorithms of data mining is used. The decision tree algorithms of classification are one of the fine grained methods to bring the more accuracy of prediction. The first phase of the work is collecting the questionnaire for all the testing areas according to the field the excellence is predicted. The second phase is plays vital role of the work is pruning. This is the selection of predictive attributes. The third phase applying the algorithms, it uses the Naive Bayes and tree induction of decision tree methods. The scalability of these methods has improved by perception based learning. It is not that only the student domain can be used for excellence prediction. It can be applied for any kind of domain.**

Index Terms-- **Decision Tree, Naive Bayes, Data Pruning, Data Mining**

.

I. INTRODUCTION

PREDICTING the student performance is a very needful task nowadays. Because the students can take up the further studies according to their knowledge level or the future of a student will be shaped based on the performance. The educational institutions are analyzing the performance level of individual students using the various metrics and measurements. The most commonly used measures by all the institutions are cut off marks. Setting the cut off for passing level will be differed by the institutions.

According to the literature review the mark alone not decides the performance level of students. This is the major part to think about other ways to understand the performance. This work gives the other novel approach for performance prediction.

Data mining is an emerging area consisting of techniques for prediction. It is concerned with finding new patterns in large amounts of data. Of course, Data Mining can be applied to the business of Education[1,6], for example to find out which alumni are likely to make larger donations. Data mining software allows users to analyze large databases to solve business decision problems. Data mining is, in some ways, an extension of statistics, with a few artificial intelligence and machine learning twists thrown in[8]. Like statistics, data mining is not a business solution, it is just a technology.

For example, consider a catalog retailer who needs to decide who should receive information about a new product. The information operated on by the data mining process is contained in a historical database of previous interactions with customers and the features associated with the customers, such as age, zip code, and their responses. The data mining software would use this historical information to build a model of customer behavior that could be used to predict which customers would be likely to respond to the new product. By using this information a marketing manager can select only the customers who are most likely to respond. The operational business software can then feed the results of the decision to the appropriate touch point systems (call centers, direct mail, web servers, email systems, etc.) so that the right customers receive the right offers

The goal of this work is to define how to make data possible to mine, how to discover and present patterns that are helpful for the teachers and others to judge a student.

II. DATA COLLECTION

It is obviously impossible to know everything about a student. Therefore, it becomes necessary to choose the most relevant and useful information about a student that may influence performance [4,7]. Practically, it is often difficult to model the exact values of attributes of a student with respect to some attributes. Thus, the procedure may include too much uncertainty. If those attributes are uncertain, then this uncertainty will transfer to the prediction, which may also result in poorly adapted classification of performance.

The result will be accurate only based on the way the data has been collected. The domain of student has different disciplines. The performance should me measured according to the discipline. The part of data collection can be spilt into two ways. They are *collection of existing records* of a student and *preparing questionnaire* for testing them in different practical scenarios.

The first part *collections of existing* records are collecting their performance record with respect to the marks percentage of the discipline and any other records related to their studies

Dr. E. Chandra, Asst.Professor & HOD, Department of Computer Applications, D.J.Academy for Managerial Excellence, Coimbatore – 32
crc_speech@gmail.com, 98942 55832

K. Nandhini, Lecturer, Department of Computer Applications, D.J.Academy for Managerial Excellence, Coimbatore – 32
krishnandhini@yahoo.com, 98435 19618

from the inception .The second part of preparing questionnaire is the main and most weight age work. Since this is the part which helps to identify the performance in all the aspects. Questionnaire can not be a common one. It again varies according to the level of the study. This is the work initially limits by preparing the questionnaire for only the PG students of computer science. This questionnaire mainly focuses on the following level.

1. Aptitude level

2. Analytical level

3. Logical level

4. Communication ability level

5. Technical level

The above mentioned level of each consists of around 20 to 30 questions to test the excellence. Example question of Technical level has shown below. (Total 100)

1. What is the data structure which handles the memory in LIFO order

a) Stack b) Queue c) Array

III. METHODOLOGY

A.. Bayesian Network

General Bayesian network classifiers are known as Bayesian networks, belief networks or causal probabilistic networks [7, 5]. They draw their roots from a branch of probability and statistics known as decision theory which involves the theory of how to minimize risk and loss when making decisions based on uncertain information.

Moreover, given that quite often data can not be classified with deterministic correct certainty and associated with every classification problem is a risk/loss function that indicates the severity of an incorrect classification, Bayesian learning involves the process of calculating the most probable hypothesis that would correctly classify an object or piece of data, based on Baye's rule[8].

Some attractive aspects of Bayesian learning include: each training vector can be used to update probability distributions which in turn affect the probability that a given hypothesis is true; provides more flexibility in that a hypothesis does not get completely ruled out from few examples; and prior knowledge can be easily implemented in the form of prior probability distributions.

For the purpose the experiment, technical skill, analytical and logical skills were taken as the testing area. In order to select the relevant attributes which predict performance, a list of potential attributes were identified from discussion with colleagues, expert opinion and review of literature. There are 25 questions of each of the skills.

The sample size of 20 students has asked to attend this test and their performances has stored in a separate database. The bulk of the effort was invested in assembling and integrating the data and in preparing distinct files for training dataset and test dataset. For the purpose of testing the applicability of four scale options for making the possible categories of the identified variables were calculated. The first category is of *excellent* for who scored 25 marks, the second category is of *good* for who scored the marks between 22 to 24 , the third category of *average* for who scored between 15 to 18 and finally the category of *poor* for who scored 14 and below. This category has same for all the skills.

B. Decision Tree

Decision tree learning, used in data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are **classification trees** or **regression trees [2]**. In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications.

In decision theory and decision analysis, a decision tree is a graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It can be used to create a plan to reach a goal. Decision trees are constructed in order to help with making decisions. A decision tree is a special form of tree structure. Another use of trees is as a descriptive means for calculating conditional probabilities.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making.

Decision tree learning is a common method used in data mining. Each interior node corresponds to a variable; an arc to a child represents a possible value of that variable. A leaf represents a possible value of target variable given the values of the variables represented by the path from the root.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner. The recursion is completed when splitting is either non-feasible, or a singular classification can be applied to each element of the derived subset. A random forest classifier uses a number of decision trees, in order to improve the classification rate.

In data mining, trees can be described also as the combination of mathematical and computing techniques to aid the description, categorization and generalization of a given set of data.

Data comes in records of the form:

$$\mathbf{(x, y) = (x_1, x_2, x_3..., x_k, y)} \qquad (1)$$

The dependent variable, Y, is the variable that we are trying to understand, classify or generalize. The other variables, x_1, x_2, x_3 etc., are the variables that will help with that task.

C. Gini impurity

Used by the CART algorithm, Gini impurity is based on squared probabilities of membership for each target category in the node. It reaches its minimum (zero) when all cases in the node fall into a single target category.

Suppose y takes on values in {1, 2, ..., m}, and let f(i, j) = probability of getting value j in node i. That is, f(i, j) is the proportion of records assigned to node i for which y = j.

$$I_G(i) = 1 - \sum_{j=1}^{m} f(i,j)^2 = \sum_{j \neq k} f(i,j) f(i,k) \qquad (2)$$

D. Information theory

Used by the ID3, C4.5 and C5.0 tree generation algorithms. Information gain is based on the concept of entropy used in information theory .

$$I_E(i) = - \sum_{j=1}^{m} f(i,j) \log_2 f(i,j) \qquad \text{-----}(3)$$

E. Decision tree advantages

Amongst other data mining methods, decision trees have several advantages:

- Simple to understand and interpret. People are able to understand decision tree models after a brief explanation.
- Requires little data preparation. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed.
- Able to handle both numerical and categorical data. Other techniques are usually specialized in analyzing datasets that have only one type of variable. Ex: relation rules can be only used with nominal variables while neural networks can be used only with numerical variables.
- Use a white box model. If a given situation is observable in a model the explanation for the condition is easily explained by Boolean logic. An example of a black box model is an artificial neural network since the explanation for the results is excessively complex to be comprehended.
- Possible to validate a model using statistical tests. That makes it possible to account for the reliability of the model.
- Robust, perform well with large data in a short time. Large amounts of data can be analyzed using personal computers in a time short enough to enable stakeholders to take decisions based on its analysis.

This proposed work uses both *gini* and *information theory* to construct the tree. There is the variation in prediction.

IV. RESULT DISCUSSION

From the question answering session of the program, the response of the students for each item was scored by the program. Based on the scores of each item, the system automatically calculates the values of each of the attributes. That is each record takes any one of the category of the four categories.

The program then consults the bayesian network for the probability of the student having above satisfactory, below satisfactory or satisfactory performance. The system takes the category with the higher probability and stores the information along with values of the other attributes.

The TABLE1 gives the set of the records created after giving the values for the questionnaire of 4 students.

TABLE 1
SAMPLES OF TEST PERFORMANCE

S.No	Name	Aptitude	Logical	Analytical	Technical
1	Nikila	20	20	24	25
2	Darwin	25	25	25	25
3	Menen	23	18	19	20
4	Rinu	17	12	12	13

The program for prediction variable creates the category for each record by passing the above table as input. The TABLE2 gives records with prediction variables. The calculation of prediction variable uses the simple classification rules by which tree can be constructed and make us to understand the performance of individual.

TABLE 2
PREDICTED VALUE OF TEST PERFORMANCE

S. No	Name	Aptitude	Logical	Analytical	Technical	Prediction
1	Nikila	20	20	24	25	Good
2	Darwin	25	25	25	25	Excellent
3	Menen	23	18	19	20	Good
4	Rinu	17	12	12	13	Poor

The graphical representation of numerical performance data and predicted data are presented in the Fig 1 and Fig 2 respectively.

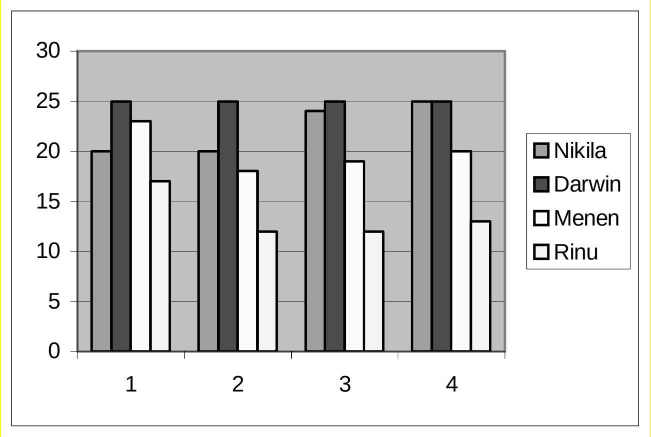

Fig 1. Test performance

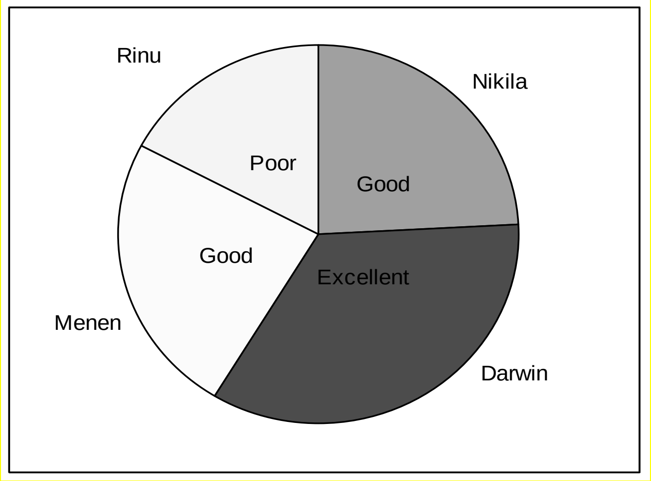

Fig 2.Predicted value of tests

This is the work should give the decision of best performer from the group. This is a more tedious job of the research. Arranging the data by acceptable data structure is itself a major work. Here tree can be built using multi dimensional arrays and also the work uses the combination of clustering, classification and association rules to make the decision among the group of students.

V. CONCLUSIONS AND FUTURE WORK

In examining the problem of prediction of performance, found that it is possible to automatically predict students' performance. Moreover by using extensible classification formalism such as Bayesian networks, it becomes possible to easily and uniformly integrate such knowledge into the learning task. Our experiments also show the need for methods aimed at predicting performance and exploring more learning algorithms. It is believed that, if put to practice, this individualized performance prediction will help the teacher a lot in giving the necessary assistance to a student. Further experiments are also being carried out to recommend student clusters based on the predicted performance. While the existing clustering algorithms are based on similarity checking, planning to explore on difference checking so that automatic composition of groups with heterogeneous nature will be possible.

VI. REFERENCES

Papers Presented in the Conference

[1] Beck, J., ed. *ITS2004 workshop on Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes*. Maceio, Brazil (2004).

[2] Iida Hakkinen, *Do University entrance exams predict academic achievement?* Working Paper Series, Department of Economics, Uppsala University, 2004.

[3] Mostow, J. "Some Useful Design Tactics for Mining ITS Data" *ITS2004 workshop Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes*, Maceio, Brazil (2004).

[4] Paul Golding & Opal Donaldson, Predicting academic performance. *Proc. 36th ASEE/IEEE Frontiers inEducation Conference* , 2006

Books

[5] Breiman L, et al *Classification and regression trees* (Wadsworth: Belmont CA, 1984) pp 56-80.

[6] Han, J. & M. Kamber, *Data mining: concepts and techniques*, San Francisco: Morgan Kaufman (2001), pp 45-78.

[7] F.V. Jensen., *An introduction to Bayesian network* (London. U.K: University College London Press, 1996), pp 90-130.

[8] Pearl J., *Probabilistic reasoning in intelligent systems: networks of plausible inference*, (Morgan Kaufmann: San Mateo CA, 1988), pp 90-120.

VII. BIOGRAPHIES

Dr.E.Chandra received her B.Sc., from Bharathiar University, Coimbatore in 1992 and received M.Sc., from Avinashilingam University ,Coimbatore in 1994. She obtained her M.Phil., in the area of Neural Networks from Bharathiar University, in 1999. She obtained her PhD degree in the area of Speech recognition system from Alagappa University Karikudi in 2007 . At present she is working as a Head and Assistant professor at Department of Computer Applications in D. J. Academy for Managerial Excellence, Coimbatore. She has published more than 20 research papers in National, International journals and conferences. She guided for more than 30 M.Phil., research scholars. Her research interest lies in the area of Data Mining, Artificial intelligence, neural networks, speech recognition systems and fuzzy logics.She is an active member of CSI, Society of Statistics and Computer Applications.

K.Nandhini received her B.Sc., from Bharathiar University, Coimbatore in 1996 and received M.C.A from Bharathidasan University,Tricy in 2001. She obtained her M.Phil., in the area of Data Mining from Bharathidasan University,Tricy in 2004. At present she is working as a Lecturer at Department of Computer Applications in D.J.Academy for Managerial Excellence, Coimbatore. She has presented more than 6 research papers in National and International conferences in the area of Data Mining. Her research interest lies in the area of Data Mining and Artificial Intelligence.