International Journal of Computer Science and Network (IJCSN)
Volume 1, Issue 4, August 2012 www.ijcsn.org ISSN 2277-5420

Classification of Student's data Using Data Mining Techniques for Training & Placement Department in Technical Education

[1]Samrat Singh, [2]Dr. Vikesh Kumar

[1] Ph.d Research Scholar, School of Computer Science & IT, Singhania University,Rajshtan, India

[2] Professor & Director, Neelkant Institute of Technology, Meerut (India)

Abstract

Data Mining is new approach for technical education. Technical institute like engineering & other can use data mining techniques for analysis of different performances in student's qualifications. In our work, we collected enrolled student's data from engineering institute that have different information about their previous and current academics records like students roll no., name, date of birth, 10th, 12th, B.tech passing percentage & other information and then apply decision tree method for classifying students academics performance for Training & placement department can be identify the final grade of student for placement purpose. In future this study will be help to develop new approaches of data mining techniques in technical education.

Keywords – *Data Mining , discover knowledge, Technical Education, Educational data*

1. Introduction

Data Mining is a process of extracting previously unknown, valid, potentional useful and hidden patterns from large data sets (Connolly, 1999). As the amount of data stored in educational databases is increasing rapidly. In order to get required benefits from such large data and to find hidden relationships between variables using different data mining techniques developed and used (Han and Kamber, 2006). There are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data come from educational environments [1]. The data can be collected form historical and operational data reside in the databases of educational institutes. The student data can be personal or academic. Also it can be collected from e-learning systems which have a vast amount of information used by most institutes [2][3]. Educational data mining used many techniques such as decision trees, neural networks, k-nearest Neighbor, Naive Bayes, support vector machines and many others. Using these methods many kinds of knowledge can be discovered such as association rules, classifications and clustering. The discovered knowledge can be used to better understand students' behavior, to assist instructors, to improve teaching, to evaluate and improve e-learning systems , to improve curriculums and many other benefits [4] [1].

Performance monitoring involves assessments which serve a vital role in providing information that is geared to help students, teachers, administrators, and policy makers take decisions.[5] The changing factors in contemporary education has led to the quest to effectively and efficiently monitor student performance in educational institutions, which is now moving away from the traditional measurement & evaluation techniques to the use of DMT which employs various intrusive data penetration and investigation methods to isolate vital implicit or hidden information. Due to the fact that several new technologies have contributed and generated huge explicit knowledge, causing implicit knowledge to be unobserved and stacked away within huge amounts of data. The main attribute of data mining is that it subsumes Knowledge Discovery (KD) which according to [6] is a nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data processes, thereby contributing to predicting trends of outcomes by profiling performance attributes that supports effective decisions making. This paper deploys theory and practice of data mining as it relates to student's performance in their qualifications.

The main objective of this paper is to use data mining methodologies to study students‟ performance in their qualifications. Data mining provides many tasks that could be used to study the student performance. In this research, the classification task is used to evaluate student's performance and as there are many approaches that are used for data classification, the decision tree method is used here. Information's like student's course Branch, passing % of 10th, passing % of 12th and passing % of B.Tech were

collected from the student's database, to predict the performance grade. This paper also investigates the accuracy of Decision tree techniques for predicting student performance.

2. Data Mining Definition & Techniques

Data mining, also popularly known as Knowledge Discovery in Database, refers to extracting or "mining" knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. While data mining and knowledge discovery in database are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The sequences of steps identified in extracting knowledge from data are shown in Figure 1

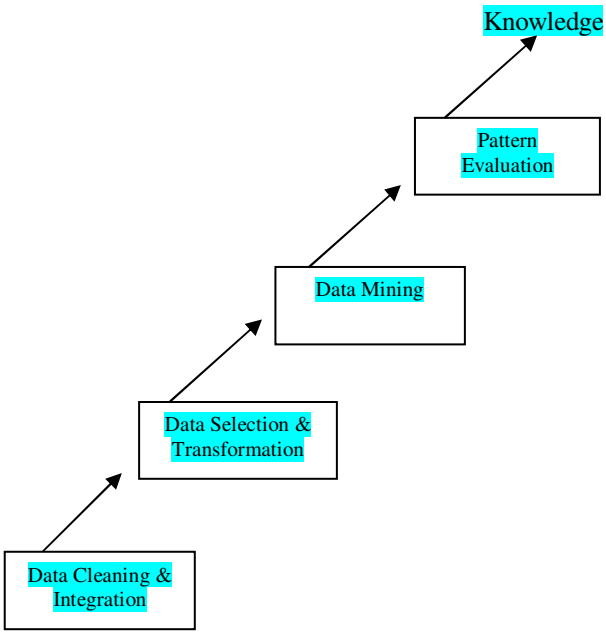

Figure -1 The steps of extracting knowledge from data

A. Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification .In learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

B. Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification.

C. Predication

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets)maybe necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

D. Association rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

E. Neural networks

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and

can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

F. Decision Trees

Decision tree is tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

G. Nearest Neighbor Method

A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k is greater than or equal to 1). Sometimes called the k-nearest neighbor technique.

3. Related Work

Data mining in higher education is a recent research field and this area of research is gaining popularity because of its potentials to educational institutes.Data Mining can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students as described by Alaa el-Halees [7]. Mining in educational environment is called Educational Data Mining.

Data mining applications in higher education given in [11], they concluded with that the Data mining is a powerful analytical tool that enables educational institutions to better allocate resources and staff to proactively manage student outcomes and improve the effectiveness of alumni development. Han and Kamber [8] describes data mining software that allow the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process. Pandey and Pal [9] conducted study on the student performance based by selecting 600 students from different colleges of Dr. R.M.L. Awadh University, Faizabad, India. By means of Bayes Classification on category, language and background qualification, it was found that whether new comer students will performer or not.

Al-Radaideh, et al [10] applied a decision tree model to predict the final grade of students who studied the C++ course in Yarmouk University, Jordan in the year 2005. Three different classification methods namely ID3, C4.5 and the Naïve Bayes were used. The outcome of their results indicated that Decision Tree model had better prediction than other models. Varsha, Anuj, Divakar, R.C Jain [13] applied four classification methods on student academic data i.e Decision tree (ID3), Multilayers perceptron, Decision table & Naïve Bayes classification method. Brijesh kumar & Saurabh Pal [14] study the data set of 50 students from VBS Purvanchal University, Jaunpur (U.P). As there are many approaches that are used for data classification, the decision tree method is used here. Information‟s like Attendance, Class test, Seminar and Assignment marks were collected from the student‟s previous database, to predict the performance at the end of the semester.

4. Proposed Work

A. Data Collection & Preparations

The data set used in this study was obtained from the different branches of the Bansal Institute of Engineering & Technology, Meerut (Uttar Pradesh, India) of B.Tech course (Bachelor of Technology). Initially size of the data is 40. The data sets have four attributes like student's Branch, passing percentage (%) in 10^{th} class, passing percentage (%) in 12^{th} class and passing percentage (%) in B.Tech course for analysis. We discretized the numerical attributes to categorical ones. For example, variable X (X = x_0, x_1, x_2 Where $x_{0=}10^{th}$ %, $x_{1=}12^{th}$ %, $x_{2=}$B.Tech %) is common variable of student's passing percentage (%) in 10^{th}, 12^{th} & B.Tech. We grouped all grades into three groups Excellent, Good, Average as described in table below.

TABLE- I
VALUES OF FINAL GRADE

Final_Percentage	Final_Grade
$X \geq 60\%$	Excellent
$X \geq 45\%$	Good
$X \geq 35\%$	Average

In the same way, we descretized other attributes such as student's course Branch, passing % of 10^{th}, passing % of 12^{th}, passing % of B.Tech. Finally the most significant attributes presented in following table:-

TABLE- II
THE SYMBOLIC ATTRIBUTE DESCRIPTION

Attribute	Description	Possible Values

Branch	Student's branch in B.Tech course.	{CS, IT, EC, EN}
10^{th} %	Percentage of marks obtained in 10^{th} class examination.	{ First > 60% Second > 45 & < 60 % Third > 35 & < 45 % }
12^{th} %	Percentage of marks obtained in 12^{th} class examination.	{ First > 60% Second > 50 & < 60 % }
B.Tech %	Percentage of marks obtained in B.Tech course.	{ First > 60% Second > 50 & < 60 % }
Final_Grade	Final Grade obtained after analysis the passing percentage of 10^{th} ,12^{th} ,B.Tech .	{ Excellent, Good, Average }

The domain values for some of the variables were defined for the present investigation as follows:

- **Branch –** Student's branch in they are enrolled in B.Tech Course. Branch split in four classes: *CS IT, EC, EN.*
- **10^{th} % --** Student's passing percentage (%) in 10^{th} class. 10^{th} % is split into three classes: *First- >60% Second - >45% and <60%, Third - >35% and < 45%.*
- **12^{th} % --**Student's passing percentage (%) in 12^{th} class. For admission in B.Tech course minimum 50% marks is compulsory in 12^{th} class. So 12^{th} % is split into two classes: *First- >60% Second - >50% and <60%.*
- **B.Tech% --**Student's passing percentage (%) in B.Tech Course. In B.Tech course Minimum 50% marks is compulsory for passing. So B.Tech % is split into two classes: *First- >60% Second - >50% and <60%.*
- **Final_Grade –**The value of final grade (X) will be finding after analysis of rule sets of Student's passing percentage (%) in 10^{th} (x_0), 12^{th} (x_1), B.Tech (x_2). The final grade is divided into three categories: *Excellent, Good, Average*

B. Decision Tree

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision.Decision tree are commonly used for gaining information for the purpose of decision -making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome. The three widely used decision tree learning algorithms are: ID3, ASSISTANT and C4.5.

C. The ID3 Decision Tree

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan [12]. The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given sets, we introduce a metric - information gain.

5. Result and Discussion

The data set of 40 students used in this study was obtained from Bansal Institute of Engineering & Technology, Meerut (India) of B.Tech course (Bachelor of Technology).

TABLE –III
RULE SET GENRATED BY DECIESION TREE

IF 10^{th} % ="First" AND 12^{th} % ="First" AND B.Tech % = "First" THEN Final_Grade = "Excellent"
IF 10^{th} % ="Second" AND 12^{th} % ="First" AND B.Tech % = "First" THEN Final_Grade = "Good"
IF 10^{th} % ="Third" AND 12^{th} % ="First" AND B.Tech % = "First" THEN Final_Grade = "Average"
IF 10^{th} % ="First" AND 12^{th} % ="Second" AND B.Tech % = "First" THEN Final_Grade = "Good"
IF 10^{th} % ="Second" AND 12^{th} % ="Second"AND B.Tech % = "First" THEN Final_Grade = "Average"
IF 10^{th} % ="Third" AND 12^{th} % ="Second" AND B.Tech % = "First" THEN Final_Grade = "Average"
IF 10^{th} % ="First" AND 12^{th} % ="First" AND B.Tech % = "Second" THEN Final_Grade = "Average"
IF 10^{th} % ="Second" AND 12^{th} % ="First" AND B.Tech % = "Second" THEN Final_Grade = "Average"
IF 10^{th} % ="Third" AND 12^{th} % ="First" AND B.Tech % = "Second" THEN Final_Grade = "Average"
IF 10^{th} % ="First" AND 12^{th} % ="Second" AND B.Tech % = "Second" THEN Final_Grade = "Average"
IF 10^{th} % ="Second" AND 12^{th} % ="Second"AND B.Tech % = "Second" THEN Final_Grade = "Average"

IF 10th % ="Third" AND 12th % ="Second" AND B.Tech % = "Second" THEN Final_Grade = "Average"

TABLE –IV
STUDENT'S DATA FOR ANALYSIS OF FINAL GRADE
TABLE- V
BRANCHWISE STUDENT'S FINAL GRADE DETAILS

S.N	Branch	10th %	12th %	B.Tech%	Final Grade
1.	CS	First	First	First	Excellent
2.	CS	First	Second	First	Good
3.	CS	First	First	First	Excellent
4.	CS	First	Second	First	Good
5.	CS	First	First	First	Excellent
6.	CS	First	First	First	Excellent
7.	CS	Third	Second	First	Average
8.	CS	First	First	First	Excellent
9.	CS	First	First	First	Excellent
10.	CS	Second	First	First	Good
11.	IT	First	First	First	Excellent
12.	IT	Second	First	First	Good
13.	IT	First	Second	First	Good
14.	IT	Second	Second	First	Average
15.	IT	First	First	First	Excellent
16.	IT	First	First	First	Excellent
17.	IT	First	First	First	Excellent
18.	IT	Third	First	First	Average
19.	IT	First	First	First	Excellent
20.	IT	First	Second	First	Good
21.	EC	Second	Second	First	Average
22.	EC	Second	First	First	Good
23.	EC	First	First	First	Excellent
24.	EC	First	First	First	Excellent
25.	EC	First	First	First	Excellent
26.	EC	First	First	First	Excellent
27.	EC	Third	Second	First	Average
28.	EC	Second	First	First	Good
29.	EC	First	Second	First	Good
30.	EC	Second	First	First	Good
31.	EN	First	First	First	Excellent
32.	EN	First	First	First	Excellent
33.	EN	First	Second	First	Good
34.	EN	Second	Second	First	Average
35.	EN	First	First	First	Excellent
36.	EN	Third	Second	Second	Average
37.	EN	Second	First	First	Good
38.	EN	First	First	First	Excellent
39.	EN	Second	First	First	Good
40.	EN	First	First	First	Excellent

SN	Branch	No. of Students	No. of students Excellent	No. of students Good	No. of students Average
1.	CS	10	6	3	1
2.	IT	10	5	3	2
3.	EC	10	4	4	2
4.	EN	10	5	3	2
	Total →	40	20	13	7

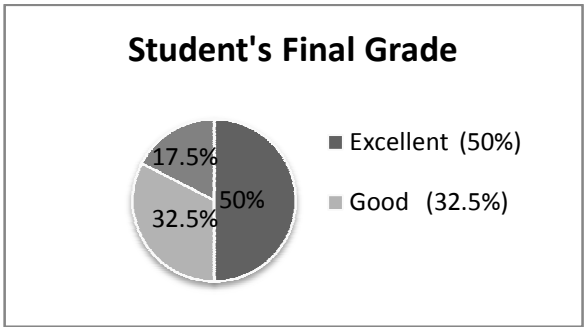

Figure -2 The Analysis chart which shows overall Percentage of Student's Final Grade in all Branches.

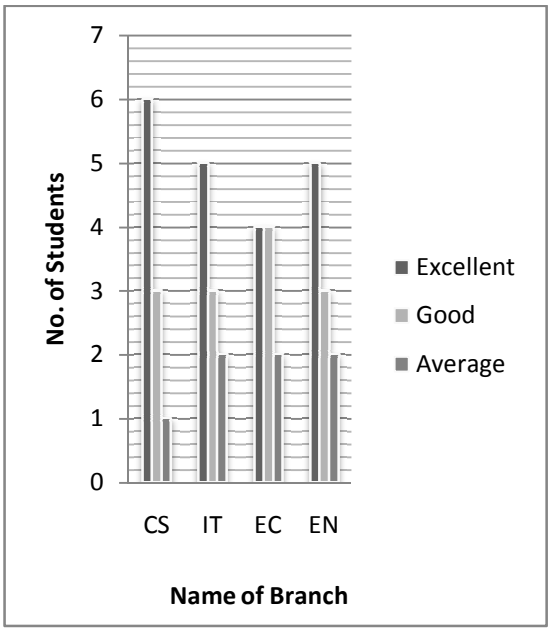

Figure- 3 The Analysis chart which shows Branch wise Student's Grades.

6. Conclusion & Future Work

In this work we make use of data mining process in a student's database using classification data mining techniques (decision tree method). The information generated after the analysis of data mining techniques on student's data base is helpful for executives for training & placement department of engineering colleges. This work classifies the categories of student's performance in their academic qualifications.

For future work, this study will be helpful for institutions and industries. We can be generating the information after implementing the others data mining techniques like clustering, Predication and Association rules etc on different eligibility criteria of industry recruitment for students.

References

[1] Romero, C. , Ventura, S. and Garcia, E., "Data mining in course management systems: Moodle case study and Tutorial". Computers & Education, Vol. 51, No. 1. pp. 368- 384. 2008

[2] Machado, L. and Becker, K. "Distance Education: A Web Usage Mining Case Study for the Evaluation of Learning Sites". Third IEEE International Conference on Advanced Learning Technologies (ICALT'03), 2003.

[3] Mostow,J and Beck , J., "Some useful tactics to modify, map and mine data from intelligent tutors".Natural Language Engineering 12(2), 195- 208. 2006.

[4] Romero,C. and Ventura, S. ,"Educational data Mining: A Survey from 1995 to 2005".Expert Systems with Applications (33) 135-146. 2007.

[5] Council N. "Knowing What Student Knows. The Science and Design of Educational Assessment ".National Academic Press. Washington, D.C. 2001

[6] Frawley, W.J., Piatetsky-Shapiro, G and Matheus, C.J.), "Knowledge Discovery databases: An overview In": Piattetsky-Shapiro and Frawley, W. J. (eds) Knowledge Discovery in Databases, AAAI/MIT.1991. pp 1-27.

[7] Alaa el-Halees "Mining students data to analyze e-Learning behavior: A Case Study", 2009.

[8] J. Han and M. Kamber, "Data Mining: Concepts and Techniques,"Morgan Kaufmann, 2000.

[9] U . K. Pandey, and S. Pal, "Data Mining: A prediction of performer or underperformer usingclassification",(IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN:0975- 9646, 2011.

[10] Q. A. AI-Radaideh, E. W. AI-Shawakfa, and M. I. AI-Najjar, "Mining student data using decision trees", International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan, 2006.

[11] Jing Luan "Data mining application in higher education" . Chief planning and Research Officer, Cabrillo College founder knowledge Discovery,2006.

[12] J. R. Quinlan, "Introduction of decision tree: Machine learn", 1: pp.86-106, 1986.

[13] Varsha, Anuj, Divakar, R.C Jain , "Result analysis using classification techniques", International Journal of Computer Applications (0975-8887) Volume 1-No. 22, 2010.

[14] Brijesh kumar & Saurabh Pal , " Mining educational data to Analyze students performance ", International Journal of Advanced Computer Science & Applications Volume 2, No- 6, 2011.

First Author

Samrat Singh is Ph.D Research Scholar in the Deptt of Computer Science in India .His area of specialization in Educational Data Mining. He did complete his master degrees M.Tech from KSOU, Mysore (India), M.phil from Alagappa University, Tamilnadu (India) in 2008 and MCA from UPTU, Lucknow (India) in 2006. Presently working as Associate Professor in Computer Sc & Engg Deptt at BIET, Meerut (India). He published many research papers in reputed conferences and journals on different issues.

Second Author

Dr. Vikesh Kumar is working as Professor & Director in NIT, Meerut (India). He did complete his doctorate degree from Gurukul kangri Vishwavidyalaya, Haridwar (India). He has more than 40 research papers in reputed international & national journals. He got completed more than 10 candidates of M.Phil and M.Tech degree under his supervision. Currently he also guides to many candidates of Ph.d program under his supervision.