Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 5, Issue, 5, pp.1209-1213, May, 2013

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

EDUCATION DATA MINING In HIGHER EDUCATION – A PRIMARY PREDICTION MODEL AND ITS AFFECTING PARAMETERS

[1]Jaimin N. Undavia, [2]Prashant M. Dolia and Nikhil P. Shah[3]

[1]CMPICA, CHARUSAT, Changa, Gujarat
[2]Department of Computer Science and Application, Maharaja Krushnakumar Sinhji University, Bhavnagar, Gujarat
[3]MCA Department, Dharamsinh Desai University, Nadiad

ARTICLE INFO

Article History:
Received 05th February, 2013
Received in revised form
13th March, 2013
Accepted 29th April, 2013
Published online 12th May, 2013

Key words:
Data Mining,
Decision Tree,
Prediction Model, J48.

ABSTRACT

The major aspect of Data Mining is in education for various purposes. Prediction is the most widely used technique in Data Mining. In recent research, Data Mining in education is at its peak. Higher education institutes are preparing professional with expertise in discipline, high moral values and with extensive knowledge. To achieve this, they required candidates who are meant for the concern subject or institutes. So, Student Performance Prediction (SPS) become very much important for the higher education institutes. We may achieve the highest level of quality in higher education is by discovering knowledge of prediction regarding the academic history of particular student. So it is always better to have some prediction of student performance before admitted for particular course. Currently, EDM tends to focus on pattern discovery and that discovered pattern will be used in some learning analytics. The academic performance of the student may be influenced by many factors, so it becomes necessary to develop predictive data mining model for students. In this paper we have analyzed students of MCA/MBA 2012 batch of Charutar University of Science & Technology. At the end we will conclude the result against our parameters.

INTRODUCTION

Since long ago, data is mainly used for decision making and other computations for the future use. Almost all types of industries including education use complex computations on customer/student data for analytics. Here the data mining techniques are used and these data mining techniques can differentiate historical patterns and trends from data and as a result we can evolve new models that can predict future trends and patterns. Ability to store huge amount of data, complex study and finding of hidden patterns from that data of computers has pen a new area in the field of information technology. We can organize the data if variety of format regardless to the amount of data and from this organization of data we can explore new knowledge which was either impossible or time consuming approach for a person [1], [2], [3]. The technique to find patterns in data and model building is basically generating outcome probabilistically. This technique is known as Knowledge discovery form large amount of data. KDD is an interdisciplinary area focusing on methodologies for extracting useful knowledge from data. Extracting knowledge from data draws on research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing to deliver advanced business intelligence and Web discovery solutions. Knowledge can be discovered by using these techniques such association rule mining, classification and clustering. This discovered knowledge can be used for stream selection, carrier guidance and drop out ratio, effective teaching model, enrollment prediction, unfair means and many others [4].

Education Data Mining and its Learning Analytics

Learning analytics works differently than education data mining. Primarily it focuses on measurement and data collection activities. These activities are useful for the institutions for the development of methods and models to answer important questions that effect student learning and organizational learning systems. Unlike educational data mining, which emphasizes system-generated and automated responses to students, learning analytics enables human tailoring of responses, such as through adapting instructional content, intervening with at-risk students, and providing feedback.

Here is the taxonomy of Data Mining with all its impacts with techniques.

Fig. Taxanomy of Data Mining Method

Educational data mining and learning analytics are used to research and build models in several areas that can influence online learning systems. As Learning analytics refers to the interpretation of a wide range of data produced by and gathered on behalf of students in order to assess academic progress, predict future performance, and spot potential issues, we can use this method to address some key issues related to education field. One area is user modeling, which encompasses what a learner knows, what a learner's behavior and

Corresponding author: shahnikhil1983@gmail.com

motivation are, what the user experience is like, and how satisfied users are with online learning. At the simplest level, analytics can detect when a student in an online course is going astray and nudge him or her on to a course correction. Educational data mining and learning analytics research are beginning to answer increasingly complex questions about what a student knows and whether a student is engaged.

Learning analytics systems apply models to answer such questions as:

- When are students ready to move on to the next topic?
- When are students falling behind in a course?
- When is a student at risk for not completing a course?
- What grade is a student likely to get without intervention?
- What is the best next course for a given student?
- Should a student be referred to a counselor for help?

Decision Tree – A Predictive Data Mining Tool

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called a "root" that has no incoming edges [6]. Basically, Decision Tree is a part of induction class of data mining [5]. Decision tree induction is closely related to rule induction. Each path from the root of a decision tree to one of its leaves can be transformed into a rule simply by conjoining the tests along the path to form the antecedent part, and taking the leaf's class prediction as the class value. An empirical tree represents a segmentation of the data that is created by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of one input. One rule is applied after another, resulting in a hierarchy of segments within segments. The hierarchy is called a tree and each segment is called a node. The original segment contains the entire data set and is called the root node of the tree. So, we can conclude that a tree diagram contains the following.

- Root Node-Top Node: Contains all observations.
- Internal Node – Non Terminal Node: Contain the splitting nodes.
- Leaf nodes-terminal Node: Contain the final classification for a set of observations.

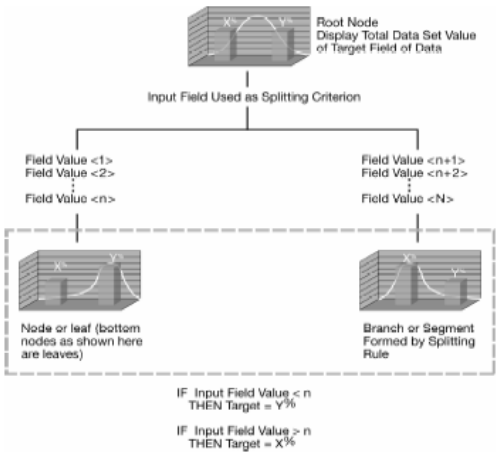


Fig. Illustration of Decision Tree

The tree techniques provide insights into the decision making process [3]. The decision tree is efficient and is thus suitable for large/small data sets. They are perhaps the most successful exploratory method for uncovering deviant data structure. Trees recursively partition the input data space in order to identify segments where the records are homogeneous. In this research we have used the same approach of decision tree to predict the master's course for the student on the basis of their academic history. Here, we kept it to very specific for two master's courses, MCA and MBA. The classification will take place by following approach. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. Specifically, we start with a root of a tree; we consider the characteristic that corresponds to a root; and we define to which branch the observed value of the given characteristic corresponds. Then we consider the node in which the given branch appears. We repeat the same operations for this node etc., until we reach a leaf. The tree techniques provide insights into the decision making process [3]. The decision tree is efficient and is thus suitable for large/small data sets. They are perhaps the most successful exploratory method for uncovering deviant data structure. Trees recursively partition the input data space in order to identify segments where the records are homogeneous.

Weka - Data Mining Software

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka.[4] Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modeling. This model, make use of the software Weka. The J4.8 algorithm (J4.8 implements a later and slightly improved version called C4.5) is used for predictive data mining.

Data Selection and Preprocessing

The model is used for the prediction of MBA and MCA courses. Here we have collected data from CHARUSAT, Changa, Gujarat for the development of the model. The CHARUSAT offers 3 years of Master of Computer Application course and 2 years of Master of Business Administration Course. Here the study is conducted for 2012 batch pass out students. The model is approached in two steps.

1. Pass out MBA/MCA students data are inserted into model. So it is said to be training of the model.
2. New variable data is inserted without MBA/MCA label and its classified by the prediction model.

Following list of parameters are considered for the primary model discovery.

Table: List of Primary Parameters

Variable	Meaning	Value
SSC	Percentage in SSC	{ A, B or C}
HSC	Percentage in HSC	{ A, B or C}
Stream	Stream in HSC	{ SC,ARTS,COMM }
Per	Percentage in HSC	
Gstream	Stream in Graduation	{ SC,CM,ARTS,COMM }
Medium	Medium of instruction in Graduation	{ ENG,GUJ }
Pg	Selected pg course	{ MCA, MBA }
Pgscore	Secured score in Masters Degree	

The selected variables are specified as follows:

SSC: Student grade in Secondary School Certificate. The grade will be assigned as per following criteria.

- >= 60 – A Grade
- 50 >= & < 60 – B Grade
- 40 >= & < 50 – C Grade

HSC: Student grade in Higher Secondary School Certificate. The grade will be assigned as per following criteria.

- >= 60 – A Grade
- 50 >= & < 60 – B Grade
- 40 >= & < 50 – C Grade

Stream: As per Gujarat Government, students are offered following stream in HSC. The stream will be assigned as per following criteria.

- SC – Science Stream
- ARTS – Arts Stream
- COMM – Commerce Stream

Per: This is percentage obtained by the students in their HSC examination.

Gstream: It indicates the graduation stream for the students. Mainly we have considered four streams. The streams are assigned as per following criteria.

- SC – Science Stream
- ARTS – Arts Stream
- COMM – Commerce Stream
- CM – Computer

Medium: It indicates the medium of instruction in opted by the students for their graduation course. The medium is assigned as per following criteria.

- ENG – English Medium
- GUJ – Gujarati Medium

PG: This is master course completed by the students. The PG course will be assigned as per following criteria.

- MCA – Master of Computer Application
- MBA – Master of Business Administration

Pgscore: It is the marks obtained by the students in their master's degree.

The data collected on the basis of the above criteria are mapped in. arff (Attributer Relationship File Format) file for the input in WEKA. Here is the same data that we have prepared for WEKA. This file is for training of the model.

```
@attribute hsc { B,A,C }
@attribute stream { SC,CM,ARTS,COMM }
@attribute per real
@attribute gstream { SC,COMP,COMM,ARTS }
@attribute medium { ENG,GUJ }
@attribute pg { MCA,MBA }
@attribute pgscore real
@data

B,B,SC,45.55,SC,ENG,MCA,6.5
B,B,CM,43.22,SC,ENG,MBA,6.8
A,A,ARTS,55.34,SC,ENG,MCA,7.9
C,A,SC,70.89,COMP,ENG,MCA,6.3
B,B,CM,60.8,COMM,ENG,MBA,5.9
A,B,CM,50.56,SC,ENG,MCA,8.5
A,A,SC,54.23,SC,ENG,MCA,8.5
B,A,SC,56.32,COMP,ENG,MBA,7.3
C,B,SC,70.89,COMP,ENG,MBA,6.8
B, A,SC, 63.8,COMP,ENG, MCA, 5.9
B, B, SC, 40.89, ARTS, ENG, MCA, 6.3
```

Analysis and Result

J48 pruned tree

```
stream = SC
|   pg = MCA: A (15.0/7.0)
|   pg = MBA: B (4.0/1.0)
stream = CM
|   medium = ENG
|   |   hsc = B: B (5.0/1.0)
|   |   hsc = A: A (6.0/1.0)
|   |   hsc = C: A (1.0)
|   medium = GUJ: A (10.0/1.0)
stream = ARTS
|   pgscore <= 6.2: C (2.0)
|   pgscore > 6.2: A (7.0/1.0)
stream = COMM: A (1.0)
```

Number of Leaves: 9

Size of the tree: 14

Time taken to build model: 0.16seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	28	54.902 %
Incorrectly Classified Instances	23	45.098 %
Kappa statistic	-0.0236	
Mean absolute error	0.3625	
Root mean squared error	0.4921	
Relative absolute error	97.8052 %	
Root relative squared error	114.8263 %	
Total Number of Instances	51	

=== Detailed Accuracy By Class ===

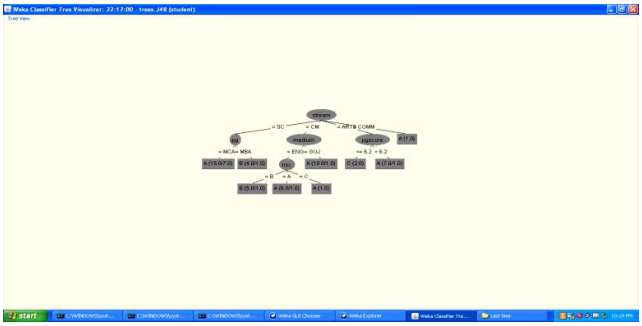
TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0	0.079	0	0	0	0.45	B
0.903	0.85	0.622	0.903	0.737	0.55	A
0	0.068	0	0	0	0.425	C
Weighted Avg.	0.549	0.546	0.378	0.549	0.448	0.508

=== Confusion Matrix ===

```
a  b  c   <-- classified as

0 11  2 |  a = B
2 28  1 |  b = A
1  6  0 |  c = C
```

This is the confusion matrix we have obtained from 51 tuples of training data set. Subsequent decision tree we have obtained is as follows.



This is the decision tree generated through Weka of 51 instances. The same training set has been used to predict a new tuple for master degree prediction.3

Here is resul along with confusion matrix for predicted data.

J48 pruned tree

```
stream = SC
|   pgscore <= 6.4
|   |   ssc = B: COMP (2.0/1.0)
|   |   ssc = A: COMM (2.0)
|   |   ssc = C: COMP (1.0)
|   pgscore > 6.4
|   |   medium = ENG
|   |   |   ssc = B: SC (5.0/1.0)
|   |   |   ssc = A
|   |   |   |   pgscore <= 8.8: SC (3.0)
|   |   |   |   pgscore > 8.8: COMP (2.0)
|   |   |   ssc = C: COMP (2.0)
|   |   medium = GUJ: SC (2.0)
stream = CM
|   per <= 66.56: COMM (11.0/2.0)
|   per > 66.56
|   |   per <= 70.89: COMP (5.0)
|   |   per > 70.89: COMM (6.0/1.0)
stream = ARTS: COMP (9.0/2.0)
stream = COMM: COMP (1.0)

Number of Leaves: 13

Size of the tree: 21

Time taken to build model: 0.02seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          0               0       %
Incorrectly Classified Instances        1             100       %
Kappa statistic                         0
Mean absolute error                     0.5
Root mean squared error                 0.6481
Relative absolute error               103.7736 %
Root relative squared error           115.9809 %
Total Number of Instances               1

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0          1          0          0          0          ?        SC
                 0          0          0          0          0          ?        COMP
                 0          0          0          0          0          ?        COMM
                 0          0          0          0          0          ?        ARTS
Weighted Avg.    0          0          0          0          0          0

=== Confusion Matrix ===
 a b c d   <-- classified as
 0 0 0 0 | a = SC
 0 0 0 0 | b = COMP
 0 0 0 0 | c = COMM
 1 0 0 0 | d = ARTS
```

Whenever, predicted data set applied on trained data set than result of the parameter is ARTS student is classified as Science student, it means particular student who is learning in ARTS faculty he may be a student of Science Faculty. So according to consideration the student will be suggested to pursue MCA not MBA.

Conclusion

As discussed in the paper, the primary model is developed against only prime parameters are considered and students are classified for only for two masters course. Upon successful invocation of this primary model, the model can be enhanced with subsidiary parameters and full spectrum of master's courses.

REFERENCES

[1] Heikki, Mannila, Data mining: machine learning, statistics, and databases, IEEE, 1996.

[2] U. Fayadd, Piatesky, G. Shapiro, and P. Smyth, From data mining to knowledge discovery in databases, AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–262 56097–6, 1996.

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

[4] Mr. M. N. Quadri, Dr. N.V. Kalyankar, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques" Global Journal of Science & Technology, 2 Vol. 10 Issue 2 (Ver 1.0), April 2010.

[5] Quinlan, J.R., 1983. Induction of Machine learning Machine learning, 1:81-106.

[6] An Introduction to the WEKA Data Mining System by Zdravko Markov & Ingrid Russell.

[7] Surjeet Kumar Yadav & Saurabh pal – "Data Mining Application in Enrollment Management: A Case Study" – An International Journal of Computer Applications, Volume 41– No.5, March 2012

[8] An Empirical Study of the Applications of Data Mining Techniques in Higher Education, Dr. Varun Kumar, Anupama Chadha- International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, March 2011.

[9] Venkatadri.M & Dr. Lokanatha C. Reddy– "A Review on Data mining from Past to the Future"- International Journal of Computer Applications (0975 – 8887)Volume 15– No.7, February 2011.

[10] SERHAT ÖZEKES and A.YILMAZ ÇAMURCU – "CLASSIFICATION AND PREDICTION IN A DATA MINING APPLICATION" , Journal of Marmara for Pure and Applied Sciences, 18 (2002) 159-174 Marmara University, Printed in Turkey.

[11] Ian H. Written & Eibe Frank, " Data Mining – Practical Machine Learning Tools & Techniques".

[12] Elena Gaudioso, Miguel Montero & Felix Hernandez-del-Olmo – "Supporting teachers in adaptive educational systems through predictive models: A proof of concept" , Expert Systems with Applications 39 (2012) 621–625.

[13] S. Anupama Kumar & Dr. Vijayalakshmi M.N – "EFFICIENCY OF DECISION TREES IN PREDICTING STUDENT'S ACADEMIC PERFORMANCE" - D.C. Wyld, et al. (Eds): CCSEA 2011, CS & IT 02, pp. 335–343, 2011.

[14] N.Ayyanathan, A.Kannammal, & A.Bavani Rekha – "Students' Communicative Competence Prediction and Performance Analysis of Probabilistic Neural Network Model" - IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 2, July 2012 ISSN (Online): 1694-0814.

[15] Monika Goyal and Rajan Vohra – "Applications of Data Mining in Higher Education" - IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012 ISSN (Online): 1694-0814.

[16] Dr. Vuda Shreenivasarao & Capt. Genetu Yohannes – " Improving Academic Performance of Students of Defense University Based on Data Warehousing & Data Mining" – Global Journal of Computer Science & Technology, Vol. 12, Issue – 2, January 2012.

[17] Prof. Dr. P. K. Srimani, Mrs. Malini M. Patil & Prof.Dr.P.K.Srivathsa – "PERFORMANCE EVALUATION OF CLASSIFIERS FOR EDU-DATA: AN INTEGRATED APPROACH" - International Journal of Current Research Vol. 4, Issue, 02, pp.183-190, February, 2012.

[18] Elena Gaudioso , Miguel Montero & Felix Hernandez-del-Olmo – "Supporting teachers in adaptive educational systems through predictive models: A proof of concept" - Expert Systems with Applications, Science Direct, ELSEVIER.

[19] C. Romero, S. Ventura – "Educational data mining: A survey from 1995 to 2005 " - Expert Systems with Applications 33 (2007) 135–146, Science Direct, ELSEVIER.

****