IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 8, August 2016

Educational Data Mining for Evaluating Students Performance

Sampreethi P.K[1], VR. Nagarajan[2]

Research Scholar, M.Phil, Computer Science, Sree Narayana Guru College, K.K. Chavadi, India[1]

Assistant Professor, Department of Computer Science, Sree Narayana Guru College, K.K. Chavadi, India[2]

Abstract: Now a day's many organizations/institutions provides different courses to students. It provides quality education to its students. So such institutions keep large database with them to evaluate student's performance in future. Database may contains information such as students personnel information, their secured marks for every subjects and their internal scores based on their academic performance etc. "Educational Data Mining For Evaluating Students Performance" research paper makes use of different kinds of data such as attendance details, previous semester marks, seminar performance etc to evaluate student's academic performance. Classification methods are used here to evaluate the student's performance. Here uses data mining algorithm to calculate performance metrics. In this paper using ID3 decision tree algorithm for predicting the student performance.

Keywords: Datamining, ID3 Algorithm, Gain Ratio.

I. INTRODUCTION

A "data warehouse" is an organization-wide snapshot of data, typically used for decision-making. Data warehousing technology aims at providing support for decision making by integrating data from various heterogeneous systems in the data warehouse. The data collected from different applications require proper method of extracting knowledge from large repositories for better decision making. The term Knowledge Discovery in Databases, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods.
The four goals of EDM are:
1. Predicting student's future learning behavior
2. Discovering or improving domain models
3. Studying the effects of educational support
4. Advancing scientific knowledge about learning and learners

Using the techniques in EDM many kinds of knowledge can be discovered. The discovered knowledge can be used for prediction regarding detection of unfair means used in online examination, detection of abnormal values in the result sheets of the students, prediction about student's performance etc.

II. DATA WAREHOUSING AND DATA MINING

A data warehouse is a subject-oriented, integrated, time-variant and Non-volatile collection of data in support of management's decision making process.

Subject-Oriented: A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.

Integrated: A data warehouse integrates data from multiple data sources.

Time-Variant: Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse.

Non-volatile: Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), is the computational process of discovering patterns in large data sets ("Big Data") involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems.

III. ALGORITHMS AND TECHNIQUES

Various algorithms and techniques are used for knowledge discovery from databases.

Neural Network

Generally neural networks consist of layers of interconnected nodes where each node producing a non-linear function of its input and input to a node may come from other nodes or directly from the input data.

Classification Algorithms

Classification is one of the Data Mining techniques that is mainly used to analyze a given data set and takes each instance of it and assigns this instance to a particular class such that classification error will be least

DOI 10.17148/IJARCCE.2016.5849

Attribute	Description	Possible attribute values
PSM	Previous Semester Marks	{ Ft- Above 80% Sd –Between 60% and 80% Td –Between 35% and 60% F < 35%}
PWM	Project Work Mark	{E, B}
SEMP	Seminar Performance	{Pr , Avg, Gd}
ATTE	Attendance	{Pr , Avg, Gd}
ASSI	Assignment	{E, B}
CTM	Class Test Marks	{Pr , Avg, Gd}
SPM	Slide Presentation Mark	{E, B}
LSM	Last Semester Marks	{Ft- Above 80% Sd –Between 60% and 80% Td –Between 35% and 60% F < 35%}

ID3 Algorithm

Id3 calculation starts with the original set as the root hub. On every cycle of the algorithm it emphasizes through every unused attribute of the set and figures the entropy (or data pick up IG(A)) of that attribute. At that point chooses the attribute which has the smallest entropy value.

C4.5 Algorithm

C4.5 is an algorithm used to produce a decision tree which is an expansion of prior ID3 calculation. The decision trees created by C4.5 can be used for grouping and often referred to as a statistical classifier. C4.5 creates decision trees from a set of training data same way as Id3 algorithm.

K Nearest Neighbours Algorithm

M. Cover and P. E. Hart purpose k nearest neighbour (KNN) in which nearest neighbour is computed on the basis of estimation of k that indicates how many nearest neighbours are to be considered to characterize class of a sample data point.

Clustering

Clustering is a useful technique for the discovery of data distribution and patterns in the underlying data. There are two main approaches to clustering-hierarchical clustering and partitioning clustering.

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.

Nearest Neighbour 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. Sometimes called the k- nearest neighbour technique.

IV. DATA MINING PROCESS

Data preparation

The data preparation typically consumes about 90% of the time of the project. The outcome of the data preparation phase is the final data set.

Data selection and transformation

All the predictor and response variables which were derived from the database are given below.
The domain values for the variables are given below:

STUDENT ACADEMIC DETAILS

The ID3Algoritm

ID3 (Examples, Target_Attribute, Attributes)

```
➢ Create a root Node for the tree
➢ If all examples are positive, Return the single-Node tree Root, with label = +.
➢ If all examples are negative, Return the single-Node tree Root, with label = −.
➢ If number of predicting attributes is empty, then Return the single Node tree Root, with label = most common value of the target attribute in the examples.
➢ Otherwise Begin
    • A = The Attribute that nest classifies examples.
    • Decision Tree attribute for Root = A.
    • For each possible value, $v_i$ , of A,
        o Add a new tree branch below Root, corresponding to the test A = $v_i$.
        o Let Examples($v_i$) ne the subset of examples that have the value $v_i$ for A
        o If Examples($v_i$) is empty
            ▪ Then below this new branch add a leaf Node with label = most common target value in the examples
        o Else below this new branch add the subtree ID3(Examples($v_i$),Target_Attribute, Attributes – {A})
➢ End
➢ Return Root
```

ID3 uses information gain to help it decide which attribute goes into a decision node.

Measuring Impurity

Most well known indices to measure degree of impurity are entropy, gini index, and classification error.

Entropy

We measure the entropy of a dataset, S, with respect to one attribute, in this case the target attribute, with the following calculation:

$$\mathbf{Entropy(S) = \sum_{j=1}^{c} -P_j \log_2 P_j}$$

where Pj is the proportion of instances in the dataset that take the ith value of the target attribute, which has c different values

Gini index

The Gini index measures the inequality among values of a frequency distribution . A Gini index of zero expresses perfect equality where all values are the same A Gini index of one expresses maximal inequality among values

$$\textbf{Gini Index} = 1 - \sum_{j} P_j^{\,2}$$

The value of classification error index is always between 0 and 1.

$$\textbf{Classification Error} = 1 - \{\max P_j\}$$

Splitting Criteria

Some attributes split the data up more purely than others.

Information Gain

Constructing a decision tree is all about finding attribute that returns the highest information gain.

$$Gain\ (S, A) = Entropy\ \ (S) - \sum_{v \in A} \frac{|S_v|}{|S|} Entropy\ \ (S_v)$$

where v is a value of A, |Sv| is the subset of instances of S where A takes the value v, and |S| is the number of instances.

$$\textbf{Split Information (S, A)} = -\sum_{i=1}^{n} (|S_i| / |S|) \log_2 (|S_i| / |S|)$$

Gain Ratio(S, A) = (Gain(S,A) / Split Information (S, A))

The data set of 50 students doing the course M.Sc. Computer Science are used in this study .

DATA SET WITH SOME ACADEMICAL DETAILS

(TABLE I)

S. N.	PSM	PWM	ATTE	CTM	SPM	SEMP	ASSI	LSM
1.	Ft	B	Gd	Avg	E	Gd	B	Ft
2.	Ft	B	Gd	Gd	E	Avg	E	Ft
3.	Ft	B	Avg	Gd	B	Avg	B	Ft
4.	Ft	E	Gd	Avg	E	Avg	B	Ft
5.	Ft	B	Avg	Pr	E	Avg	B	Ft
6.	Ft	E	Gd	Gd	E	Gd	E	Ft
7.	Ft	B	Pr	Pr	E	Avg	B	Sd
8.	Ft	E	Avg	Avg	B	Pr	E	Ft
9.	Ft	B	Pr	Pr	B	Pr	B	Td
10.	Ft	E	Gd	Avg	B	Avg	E	Ft
11.	Sd	E	Gd	Gd	E	Gd	E	Ft
12.	Sd	E	Gd	Gd	E	Avg	E	Ft
13.	Sd	B	Gd	Gd	B	Avg	E	Ft
14.	Sd	E	Gd	Avg	B	Gd	E	Ft
15.	Sd	E	Avg	Gd	E	Avg	E	Ft
16.	Sd	E	Pr	Gd	E	Avg	E	Sd
17.	Sd	E	Gd	Avg	E	Avg	E	Sd
18.	Sd	E	Pr	Avg	E	Avg	E	Sd
19.	Sd	E	Gd	Pr	E	Avg	B	Sd
20.	Sd	B	Avg	Avg	E	Pr	E	Sd
21.	Sd	E	Pr	Pr	B	Avg	B	Td
22.	Sd	E	Avg	Pr	E	Pr	E	Td
23.	Sd	B	Avg	Pr	E	Pr	B	Td
24.	Sd	E	Gd	Pr	E	Pr	E	Sd
25.	Sd	E	Pr	Pr	E	Pr	E	Td
26.	Sd	B	Pr	Pr	E	Pr	B	F
27.	Sd	E	Gd	Gd	E	Gd	E	Ft
28.	Td	E	Gd	Avg	E	Gd	E	Sd
29.	Td	E	Gd	Gd	E	Avg	E	Sd
30.	Td	E	Avg	Gd	E	Gd	E	Sd
31.	Td	B	Gd	Gd	E	Gd	B	Sd
32.	Td	E	Gd	Avg	E	Avg	E	Sd

33.	Td	E	Avg	Avg	E	Avg	B	Td
34.	Td	B	Gd	Avg	E	Gd	B	Td
35.	Td	E	Avg	Gd	E	Avg	B	Td
36.	Td	B	Avg	Avg	E	Pr	B	Td
37.	Td	B	Avg	Pr	E	Avg	E	Td
38.	Td	E	Pr	Pr	E	Avg	B	F
39.	Td	E	Pr	Avg	E	Avg	B	Td
40.	Td	B	Gd	Pr	B	Pr	B	Td
41.	Td	E	Pr	Pr	E	Pr	B	F
42.	Td	B	Pr	Pr	B	Pr	B	F
43.	F	E	Gd	Gd	E	Gd	E	Sd
44.	F	E	Avg	Gd	E	Gd	E	Sd
45.	F	E	Avg	Avg	E	Gd	E	Td
46.	F	E	Avg	Pr	B	Pr	E	F
47.	F	E	Pr	Gd	E	Pr	B	F
48.	F	B	Pr	Pr	E	Pr	B	F
49.	F	E	Gd	Avg	E	Avg	E	Sd
50.	F	B	Pr	Pr	B	Gd	B	F

Entropy (S) = $-p_{Ft}\log_2(p_{Ft}) - p_{Sd}\log_2(p_{Sd}) - p_{Td}\log_2(p_{Td}) - p_F\log_2(p_F)$

$= -(14/50)\log_2(14/50) - (15/50)\log_2(15/50) - (13/50)\log_2(13/50)$
$-(8/50)\log_2(8/50)$
$= 1.965$

Gain(S,PSM)=Entropy(S) – $((|S_{Ft}|/|S|)\,\text{Entropy}(S_{Ft}) - ((|S_{Sd}|/|S|)\,\text{Entropy}(S_{Sd}) - ((|S_{Td}|/|S|)\,\text{Entropy}(S_{Td}) - ((|S_F|/|S|)\,\text{Entropy}(S_F)$

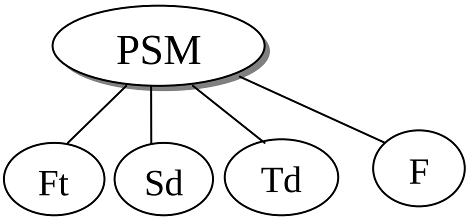
Gain(S, PSM)=0.577 Gain(S, PWM)= 0.044

Gain(S, ATTE)= 0.452 Gain(S, CTM)= 0.515

Gain(S, SPM)= 0.456 Gain(S, SEMP)= 0.366

Gain(S, ASSI)= 0.219

PSM has the highest gain, therefore it is used as the root node as shown in figure 4.



(Figure 4) Split Information is shown below

$$\textbf{Split Information (S, A)} = -\sum_{i=1}^{n}(|S_i|/|S|)\log_2(|S_i|/|S|)$$

SplitInfo(S, PSM)= 1.387 SplitInfo(S, PWM)= 1.919
SplitInfo(S, ATTE)= 1.512 SplitInfo(S, CTM)= 1.449
SplitInfo(S, SPM)= 1.510 SplitInfo(S, SEMP)= 1.598
SplitInfo(S, ASSI)= 1.745

Gain Ratio(S, A) = (Gain(S, A) / Split Information (S, A))

Gain Ratio(S, PSM) = 0.577/1.387=0.416
Gain Ratio(S, PWM) = 0.044/1.919=0.023
Gain Ratio(S, ATTE) = 0.452 /1.512=0.299
Gain Ratio(S, CTM) = 0.515 /1.449 =0.356
Gain Ratio(S, SPM) = 0.456 /1.510 =0.300
Gain Ratio(S, SEMP) = 0.366 /1.598 =0.229
Gain Ratio(S, ASSI) = 0.219 /1.387 =0.125

It is very easy to transform from decision tree to IF THEN rules. Some rules are given below.

Set of IF THEN rules generated by Decision Tree:
Rule1 ► IF PSM = “Td” AND ATTE = “Avg” AND ASSI = “B” THEN LSM = “Td”.
Rule2 ► IF PSM = “Ft” AND ATTE = “Gd” AND CTM = “Gd” OR “Avg”.
Rule3 ► IF PSM = “Ft” AND ATTE = “Gd” OR “Avg” AND CTM = “Gd” THEN LSM = “Ft”.
Rule4 ► IF PSM = “Td” AND CTM = “Gd” OR “Avg” AND ATTE = “Gd” OR “Avg” THEN LSM = “Sd”.
Rule5► IF PSM = “Sd” AND ASSI = “E” AND ATTE = “Gd” THEN LSM = “Ft”.
Rule6 ► IF PSM = “Sd” AND CTM = “Avg” AND SPM = “E” THEN LSM = “Sd”.

V. CONCLUSION

In this paper, the classification task is used on student database to predict the students division on the basis of previous database. The decision tree method is used here for classification. In student’s previous database saved some attribute values. Information’s like Project Work Mark, Attendance, Class test mark, Slide Presentation mark, Seminar and Assignment marks were collected from the student’s previous database, to predict the performance at the end of the semester. This study will help to the

DOI 10.17148/IJARCCE.2016.5849

students and the teachers to evaluate performance of each student. Teachers can use this value to identify those students which needed special attention to reduce failure rate and taking appropriate action for the next semester examination.

ACKNOWLEDGMENT

I take this opportunity to express my deep sense of gratitude and profound thanks to our principal **Dr. K.V. SURENDRAN M,.A., PGDTE., M.Litt., Ph.D,** of Sree Narayana Guru College, K.G. Chavadi, Coimbatore, for providing me all the necessary facilities in the college for successful completion of this dissertation.

I am thankful to our **H.O.D. Mrs. R. PRIYA MCA, M.Phil, (Ph.D)** for giving me chance to undergo the course in this Department.

A great debt owed to my guide **Mr. VR.NAGARAJAN PGDCS, M.C.A., M.Phil.** for accepting me and guiding me with valuable advice and kindness.

I also wish to thank all those who have been helpful to me in some way or the other to carry out this innovative work successfully.

REFERENCES

[1] Data Mining: The Textbook, Springer, May 2015, Charu C. Aggarwal.

[2] Tan P.-N., Steinbach M. and Kumar V., Introduction to Data Mining, Addison Wesley, 2006.

[3] Hand D., Mannila H. and Smyth P., Principles of Data Mining, MIT Press, 2001.

[4] Larose D.T., Discovering knowledge in data: an introduction to data mining, Wiley-Interscience, 2005.

[5] "Introduction to data mining" by Tan, Steinbach & Kumar (2006)

[6] Data Mining: Concepts and Techniques, Third Edition by Han, Kamber & Pei (2013)

[7] Data Mining and Analysis Fundamental Concepts and Algorithms by Zaki & Meira (2014)

[8] "The Elements of Statistical Learning" by Freidman et al (2009).

[9] Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems) by Jiawei Han, Micheline Kamber.

[10] Website: https://en.wikipedia.org/wiki/Information_gain_ratio

[11] Website: https://webdocs.cs.ualberta.ca/~aixplore/learning/Decision Trees /InterArticle/5-DecisionTree.html

[12] Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel Algorithms Hardcover – Import, 1 Jan 2001 by Jean-Marc Adamo

[13] Website: https://en.wikipedia.org/wiki/ID3_algorithm.

DOI 10.17148/IJARCCE.2016.5849