ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

UNIQUE DATA MINING APPROACH TO PREDICT PLACEMENT CHANCE

Dr. G Suganthi*, Mr. ASHOK M V

* Associate Professor, Dept. of Computer Science Women's Christian College, Nagrcoil, Tamil Nadu
Head of the Department Dept. of Computer Science Teachers Academy, Bangalore

DOI: 10.5281/zenodo.400931

ABSTRACT

Data mining is an effective methodology, which can be used to analyze large amount of data to produce hidden patterns and relationships. The main idea proposed is analyzing the law student's historical data and predicting the appropriate, qualitative specialization chances. A student enters his/her rank, gender, sector and reservation category, and this model predicts specialization that suits for entered student as example with criminal law as E(excellent) and civil law as P(poor) based on the entered information. Different algorithms were applied for same set of data and comparison will be made in terms of accuracy, precision and truth positive rate. This paper will work for law students selecting a best specialization for them which ensures best carrier based on data entered. A model is proposed where in a latest algorithm, one from each different category of models such as clustering, classification and association models are selected and applied on the same dataset. Hence the name CCA model. Rock algorithm from Clustering, Naïve baye's from classification and frequent pattern algorithm from Association models are selected. These algorithms are applied, to predict accurately one among the various courses offered which predict better placement chances. Student will enter Rank, Gender, Category and Sector and the model will give answer in terms of Excellent [E], Good [G], Average [A] and Poor [P] for the data entered. Algorithms are compared in terms of precision, accuracy and truth positive rate. From the results obtained it is found that the Clustering model predicts better in comparison with other two models. This work will help the students in selecting a best course suitable for them which ensure better placement based on the data entered.

KEYWORDS: Data mining, Naive Bayes, Rock algorithm, frequent pattern algorithm, Confusion matrix, Prediction and modeling.

INTRODUCTION

Data mining consists of group of techniques to mine the data, such as association rule mining, classification and clustering. In this model, an algorithm is selected from clustering and two from classification models. Law profession is one of the excellent professions which have bright scope. Therefore selection a right specialization at the time of post graduate will plays an important role in his/her carrier. Selection of specialization is arrived by accessing and analyzing previous year's law college data. The main aim of doing this is to find the hidden patterns and characteristics of student's relationships. Hence it helps to predict the future scope of specialization for him/her. For this there is a need of processing and comparisons of huge data. Classification techniques will classify the data into predefined class label and clustering technique will group the data which has similar relationships.

PROBLEM STATEMENT

Every student dreams to be successful in life. For him to be successful, choosing the right courses while studying is important. Hence a prediction model is proposed which helps the students to choose a course based on type of data or information that he/she furnishes. Among the fields or attributes that he/she enters, those attributes which contribute to the result are selected. Various mining algorithms from different models are applied on the processed data and tested accordingly. Algorithms are compared based on certain criteria such as accuracy, precision and truth positive rate.

RELATED WORK: Guha, S et al., 1999[1] proposed a new concept of links to measure the similarity/proximity between a pair of data points with categorical attributes. For large synthetic data set it was found that the combination of random sampling and labeling enables ROCK's performance to scale quite well for large

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

databases. KakotiMahanta et al., 2005[2] proposed a new algorithm QROCK which computes the clusters by determining the connected components of the graph. AgnieszkaPrusiewicz et al., 2010[3] proposed a method of grouping modified ROCK algorithm is used during service execution.Balazs Racz, D 2004[5] described an implementation of a pattern growth based frequent item set mining algorithm. Ke Wang, Liu Tang et al., 2002[6] propose an efficient algorithm, called **TD-FP-Growth** (the shorthand for Top-Down FP-Growth), to mine frequent patterns. Sudheep Elayidom et al., (2011)[7] describes to help the prospective students to make wise career decisions using technologies like data mining using decision trees, Naïve Bayes and artificial neural networks. Ajay Kumar Pal et al., (2013)[8] explains data mining methodology that can analyze relevant information results and produce different perspectives to understand more about the students' activities.A. K. Pal et al., (2013)[9] describes the use of data mining techniques to improve the efficiency of academic performance in the educational institutions. B.K. Bharadwaj et al., (2011)[10] proposed that the classification task is used on student database to predict the students division on the basis of previous database.

DATA DESCRIPTION

Table 1: Data description

Variables	Description	Possible Values
Name	Name of the student	{Text}
Category	Category of the student	{String}
Age	Age of the student	{1, 2, 3, 4, 5...}
Sector	Sector of the student	{ Text }
Rank	Rank of the student	{1, 2, 3, 4, 5...}
Address	Address of the student	{String}
Ph.No	Contact number of the student	{1, 2, 3, 4, 5...}
Gender	Gender of the student	{Text}
Branch	Branch that a student chosen	{Text}

Name – it is the name student. It can take only the alphabetical values that are from A to Z.
Category – it is the category of that he /she belongings. It can the string values. The possible values that it can take are 2A, 3A, 2B, 3B, SC/ST and GM.
Age – it is the age of the student and it take only numeric values from 0 to 9.
Sector – represents the sector that the student belongs and the possible values that it can take is URBAN and RURAL.
Rank – the rank that a student got in entrance exam and it can take values from 0 to 9.
Address – it is the address of the student. It can take the alphanumeric values that are from A to Z, 0 to 9.
Ph.no – it is the contact number of the student and it should be of only 11 digits from 0 to 9.
Gender – it is the gender of the student and the possible values are male, female.
Branch – it is the branch that the student chooses and the possible values are like crime, civil, family and etc.

METHODOLOGY

Rock: It is clustering algorithm. Rock follows the bottom-up procedure. It considers the entire object as the node.

Name	Category	Age	Sector	Rank	Address	Ph. No	Gender	Branch
Ravi	2A	30	Urban	550	Bgl	211234	Male	International property law
Raj	3B	38	Urban	549	Ncl	213456	Male	International property law
Rani	2A	21	Rural	1	Bgl	214356	Female	Criminal law, Corporate law
Rahul	GM	23	Rural	100	Bgl	1234124	Male	Family law,

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

								constitutional law

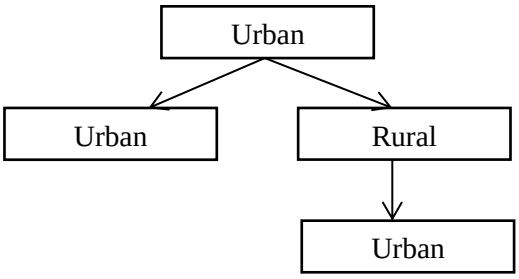
Table 2: Input for the rock algorithm

Step1: find link between the dataset whether they are dependent or independent based on which a tree is formed. In the above table rank is identified as an independent attribute based on which tree is constructed with rank as root.

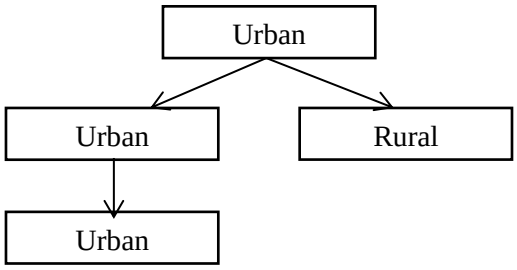
Step 2: The attribute which is at the extreme left as a leaf node in the tree is identified as the header. A bottom-up approach is followed while comparing the header with each one of the other attribute (taken as a current node) in succession in a tree, generated from step1. If a similarity is found on comparison then a link is established between them making that attribute as a child of the parent viz., header.

For sector

Before step 2



After step 2:



If (header_attribute similar to current_attribute (in any hierarchical level))
{
Reallocate the current_attribute(as a child)
Link the header_attribute (Parent) and current_attribute
}

Table 3: Representation of the knowledge database

Rank	Sector	Gender	Category	Branch
1<= =>200	Rural	Female	2A	Criminal law, Corporate law, Family law
400<= =>600	Urban	Male	2A, 3B	International property law

For the following user input (Rank, Gender, sector, category) (452, Rural, Female, 2a) the table 4 represents the possibilities of choice of specialization as the final output, after processing the knowledge data

Table 4: Output table for the user input

ID	Branch	CHANCE
1	International property law	E

NAIVE BAYES:

Naive Bayes classifier is a probabilistic classifier that works based on the Bayes theorem.

The procedure to be followed while applying this method is as follows

- Data preprocessing
- Finding positive and negative knowledge data
- Application of Bayes theorem

Table 3: Input for naïve Bayes

Name	Age	Gender	sector	Category	Rank	Branch

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Shiva	21	M	Rural	2a	200	Civil law
John	22	M	Urban	3b	100	Family Law
Rani	22	F	Rural	SC	400	Criminal Law
Ravi	22	M	Rural	SC	867	

Step 1: Data preprocessing: Filling of the missing values and the dependency check on the attributes listed in the table 8 is performed using chi-square test and Table 9 is a resultant after preprocessing.

Table 4: After preprocessing

Gender	sector	Category	Rank	Branch
M	Rural	2a	200	Civil law
M	Urban	3b	100	Family Law
F	Rural	SC	400	Criminal Law
M	Rural	SC	867	Null

Step 2:Finding positive and negative knowledge data: selection constructs are applied on a rank attribute to get a positive and negative knowledge data.
If (rank <= 800)//the maximum limit of the possible rank
{Positive knowledge data}
Else
{Negative knowledge data}
The above process is repeated for all the attributes listed in table 9 to get the positive knowledge data as given below.

Table 5: Positive knowledge data

Name	Age	Gender	sector	Category	Rank	Branch
Shiva	21	M	Rural	2a	200	Civil law
John	22	M	Urban	3b	100	FamilyLaw
Rani	22	F	Rural	SC	400	CriminalLaw

Step 3: Application of Bayes theorem on table 10 gives the resultant output table.
At the first instance data in table 10 is converted to the numeric data. Formulae listed under are used to get the below output table as the resultant.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Rank	Gender	Sector	Category	Branch	Chance
1-200	M	Rural	Any	Civil law	E
1-200	M	urban	Any	Family Law	E
1-200	F	Rural	Any	Criminal Law	E

$H_{map} = max$ (P (h/D)) where P (h) =h/n
P (D) =D/n
h=hypothesis (possibilities)
d=data set (not possible)
n=number of data set

H_{map}= max always calculate under the formula of P (D/h) P (h)/P (D)

And the maximum like hood calculate under the formula of P (D/h).

Table 6: output table

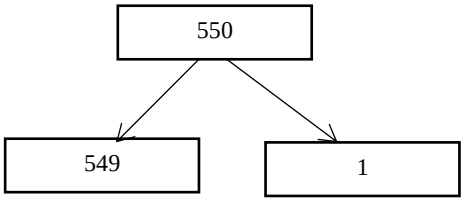
FP (Frequent Pattern) growth:
It is based on association method. It follows two pass. In the first pass, the algorithm counts occurrence of items (attribute-value pairs) in the dataset, and stores them to 'header table'. In the second pass, it builds the FP-treestructure by inserting instances. Items in each instance have to be sorted by descending order of their frequency in the dataset, so that the tree can be processed quickly.

Table 11: Input for FP growth

Name	Category	Age	Sector	Rank	Address	Ph. No	Gender	Branch
Ravi	2A	30	Urban	550	Bgl	211234	Male	International law
Raj	3B	38	Urban	549	Ncl	213456	Male	International law
Rani	2A	21	Rural	1	Bgl	214353	Female	Criminal law
Rahul	GM	33	Rural	550	Ncl	213457	Female	Constitutional law

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Step1: scan database and the repeated data in an attribute are identified which is explained in the code given below.1.1 Loop until EOF
1.2 finding the frequent dataset
e.g.: if (head_rank similar to current_rank)
In 'rank' attribute 550is compared with 549, 1 and 550 of the fourth row.One with similarity are eliminated retaining other values which forms a tree as shown below in descending order.
Below tree generated based on step1.



Apply the step1 for all the remaining attributes.
Step2: divide and conquer method is applied to the tree generated in step1, which results in a formation of a cluster as represented in table 12.

Table 12: Representation of the knowledge database form FP Growth

Rank	Sector	Gender	Category	Branch
1<= =>200	Rural	Female	2A	Criminal law
400<= =>600	Urban	Male	2A, 3B	Corporate law, Family law
400<= =>600	Rural	Female	GM	Constitutional law, International law

For the user input (Rank, Gender, sector, category) (421, Urban, Male, 2A) the table 11 represents the possibilities of choice of specialization as the final output, after processing the knowledge data.
Table 13: Output table for the user input

ID	SPECIALIZATION	CHANCE
1	Family law	E
2	Corporate law	E

All the algorithms used were implemented and the front end of the tool were developed using PHP and MYSQL as a database. Prospective Student will enter basic information like rank in the Post-graduate entrance exam, category etc., in the user interface developed and the application will predict the course suitable for the student which he/she can opt during selection of the Post-graduate course which provides better chances of placement.

RESULTS

Data mining algorithms like rock and FP growth were applied on the same dataset and the tests were conducted separately. Results obtained after the tests for each algorithm were modeled as confusion matrix. Confusion matrix explains the performance of both algorithms expressed in terms of True Positive Rate, Accuracy and Precision.

Table 14: Confusion matrix table

Algorithms	TPR	Accuracy	Precision
Rock	0.83	84.3%	0.79
FP growth	0.81	81%	0.81

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Naïve Bayes	0.80	77%	0.75

From the above table 14, it is clear that the Clustering model viz., Rock algorithm is more accurate with 84.3% compared to the FP growth (81%). Clustering algorithm leads with respect to true positive rate (TPR) with 0.83 correct instances and Precision (0.79).Thus Clustering model predicts the results better than the other models used.

CONCLUSION

Applying data mining techniques on educational data is concerned with developing methods for exploring the unique types of data; in educational domain each educational problem has specific objectives with unique characteristics that require different approaches for solving the problem.

In this study, CCA model has been used for predicting placement chances. A clustering algorithm, rock and association algorithm, FP growth were used. In the model proposed, clustering model proved to be the best predicting model for solving placement chance prediction problems. Hence, having the information generated through our study, student would be able to select the appropriate specialization with best chances of getting placed. Furthermore, the work can be extended to solve problems on predictions, using different approaches on data of different disciplines.

BIBLIOGRAPHY

[1] Guha, S.; Rastogi, R.; Kyuseok Shim "ROCK: a robust clustering algorithm for categorical attributes", Pages 512 – 521.
[2] Kakoti Mahanta, Arun K. Pujari," QROCK: A quick version of the ROCK algorithm for clustering of categorical data",Volume 26, Issue 15, November 2005, Pages 2364–2373.
[3] Agnieszka Prusiewicz, "Maciej ZiębaServices Recommendation in Systems Based on Service Oriented Architecture by Applying Modified ROCK Algorithm" Volume 88, 2010, pp 226-238.
[4] Christian Borgelt, "**An implementation of the FP-growth algorithm",** Pages 1 – 5, 2000.
[5] Balazs Racz, D: An FP-Growth Variation without Rebuilding the FP-Tree".
[6] Ke Wang, Liu Tang, Jiawei Han, "Junqiang Liu "Top down FP-Growth for Association Rule Mining", Volume 2336, 2002, pp 334-340.
[7] Sudheep Elayidom, Suman Mary Idikkula & Joseph Alexander "A Generalized Data mining Framework for Placement Chance Prediction Problems" International Journal of Computer Application (0975-8887) Volume 31- No.3, October 2011.
[8] Ajay Kumar Pal, Saurabh Pal "Classification Model of Prediction for Placement of students" I.J.Modren Education and Computer Science, 2013, 11, 49-56.
[9] K. Pal, and S. Pal, "Analysis and Mining of Educational Data for Predicting the Performance of Students", (IJECCE) International Journal of Electronics Communication and Computer Engineering, Vol. 4, Issue 5, pp. 1560-1565, ISSN: 2278-4209, 2013.
[10] B.K. Bharadwaj and S. Pal. "Mining Educational Data to Analyze Students' Performance", International Journal of Advance Computer Science and Applications (IJACSA), Vol. 2, No. 6, pp. 63-69, 2011.