SUBJECT DISTRIBUTION USING DATA MINING

Rakesh Kumar Arora[1], Dharmendra Badal[2]

[1]*Faculty, Dept. of Computer Science, Krishna Engineering College, Ghaziabad, UP, India,* ***rka1211@gmail.com***
[2]*Faculty, Dept. of Mathematical Science & Computer Applications, Bundelkhand University, Jhansi, UP, India*

Abstract

The main objective of higher education institutions is to provide quality education to its students. The faculties employed by the educational institute's plays the dominant role to achieve highest level of quality in higher education. The faculty having excellent subject knowledge and teaching skills have the major impact upon the performance of students resulting in good academic results, placements and hereby increasing the quality intake of students. This paper will assist the academic planners in distribution of subjects among the faculties in the department such that the students can make the optimum use of faculty knowledge, experience and teaching skills to reach the new heights.

Keywords: *Data Mining, Business Intelligence, WEKA, Data Visualization, Decision Tree, J48.*

----------***----------

1. INTRODUCTION

A large number of self financing private institutes have opened over the last decade with the objective of providing quality education to students in various fields of engineering and other professions. The factors affecting the quality of education include faculty profile, placements, infrastructure, working environment and vision of the institute.

Out of all the above factors the most important factor is faculty profile, Most of these self financing institutes compromise on the quality of faculty to cut down the cost of salaries and recruit inexperienced, untrained and less qualified teachers. These less qualified and inexperienced teachers are not able to make the optimum use of the institute's resources and are not able to provide quality teaching to the students. As a result the performance of the students remains below satisfactory; this also affects the placement of the institute. This adversely affects the quality of intake in the institute and causes further deterioration in the performance of the institute and slowly the institute reaches on the verge of closure.

A large number of foreign universities have also got approval from the ministry of Human Resource and Development to compete with the Indian Universities. After the entry of these foreign universities, the survival of these self financed private educational institutes has become further challenging. Since the motive of most of the self financed institutions is to maximize the profit, hence they are not able to compete with the foreign institutes resulting in the closure of these institutes.

In order to compete with the foreign universities and Government aided Indian Institutes these self financed private institutes should increase their budget on hiring experienced and qualified faculty so as to provide excellent subject knowledge and make optimum use of the institute's resources. This will have major impact upon the performance of the students, resulting in good academic results, placements and increased quality intake in the institutes. Like this these institutes will be able to sustain their existence competing with the good Indian and Foreign institutes.

An assessment about the faculty's subject knowledge and teaching skills should be made and based on which the faculty should be allocated the subject for teaching the students. Apart from the subject knowledge, faculty's qualification, feedback and the result should be taken in to consideration while allocating the subjects in future academic planning.

This paper uses Educational Data Mining Technique (EDM) to improve the distribution of subjects among faculties in the department such that students can gain benefit from knowledge and experience of faculty members to improve their performance. Data mining, the extraction of hidden predictive information from large databases is a powerful technology with great potential to help head of departments in the institutes in distribution of subjects. It discovers information within the data that queries and reports can't effectively reveal. After gathering data from the resume submitted by the faculties at the time of recruitment and feedback form filled by the students over the years, data mining technique need to be applied to determine set of patterns for allocation of subject among faculties.

With the help of data mining techniques, such as clustering, decision tree or association analysis it is possible to discover the key characteristics from the details of faculties and possibly use those characteristics for future prediction. This paper presents decision tree algorithm as a simple and efficient

tool to analyze the faculties details and allocation of subjects in subsequent semesters.[1]

2. METHODOLOGY

Decision trees are a simple, but powerful form of multiple variable analysis. A decision tree is a special form of tree structure. The tree consists of internal nodes where a logical decision has to be made, and connecting branches that are chosen according to the result of this decision. The nodes and branches that are followed constitute a sequential path through a decision tree that reaches a leaf node (final decision) in the end.[2]

In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range. Each leaf is assigned to one class representing the most appropriate target value. [3]

The decision tree algorithm is simple top down greedy algorithm. The major step of algorithm is to continue to divide leaves that are not homogeneous into leaves that are as homogeneous as possible until no further division is possible. The algorithmic steps for decision tree algorithm is as follows:[4]

Let the set of training data be S. If some of the attributes are continuous-valued, they should be discretized. Once that is done, put all of S in single tree node.
If all the instances in S are in same class, then stop.
Split the next node by selecting an attribute A from amongst the independent attributes that best divides or splits the objects in the node into subsets and create decision tree node.
Split the node according to the values of A
Stop if any of the following conditions are met, otherwise continue with step 3

Fig 1: Steps for Decision Tree Algorithm

Pruning is very important technique to be used in tree creation because of outliers. It also addresses overfitting. Datasets may contain little subsets of instances that are not well defined. To classify them correctly, pruning can be used. There are two types of pruning:

1. Post pruning (performed after creation of tree)
2. Online pruning (performed during creation of tree) [5].

The steps to extract classification rules from tree are mentioned below:

1. Represent the knowledge in the form of IF-THEN rules.
2. One rule is created for each path from the root to a leaf.
3. Each attribute-value pair along a path forms a conjunction.
4. The leaf node holds the class prediction

The analysis using decision tree is being done with the help of WEKA tool. WEKA, formally called Waikato Environment for Knowledge Learning supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. WEKA is an open source application that is freely available under the GNU general public license agreement. Originally written in C the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces. [6]

3. ANALYSIS

The study was carried out on the faculties working in computer science department of reputed engineering college of Ghaziabad. The attributes considered for analysis of faculties along with their description are reflected in Table 1.

Table 1: Parameters used for analysis

Parameters	Description
Faculty	Name of Faculty
PhD	Whether faculty holds PhD degree or not
Experience	Total Experience of faculty
No_of_times_subject _taught	Number of times subject taught by faculty
Average_Feedback	Average feedback by students in previous semesters
Results_in_Per	Result of students in previous semesters
Above_75	No. of students passed with marmarks above 75
50_75	No. of students passed with marks from 50 to 75
Below_50	No. of students obtaing less than 50 marks

The data file normally used by WEKA is in ARFF (Attribute-Relation File Format) file format, which consist of special tags to indicate different things in the data file. Figure 2 shows the sample view of dataset and Figure 3 shows the ARFF format of desired dataset. To convert an Excel format into ARFF

format an Excel to ARFF convertor is being used. The ARFF format dataset is represented in Fig 3.

	A	B	C	D	E	F	G	H	I
1	Faculty	PhD	Expierence	No_of_Times_subject_taught	Average_Feedback	Result_In_Per	Above_75	50_T0 _75	Bellow_50
2	AD	Yes	16	8	8	78	5	42	13
3	LS	No	7	4	9	80	6	41	12
4	MS	Yes	5	2	8	81	4	45	11
5	PS	No	9	5	9	84	3	48	10
6	ST	No	12	7	9	72	7	36	17
7	RT	No	10	4	8	75	5	40	15
8	BT	Yes	14	6	9	83	2	47	11
9	VP	No	12	5	8	80	5	43	12
10	NS	No	11	4	8	80	4	44	12
11	GK	NO	6	6	9	86	3	48	9
12	VG	NO	8	6	9	80	2	46	12
13	VK	NO	8	5	9	84	6	44	10

Fig2: Sample Dataset

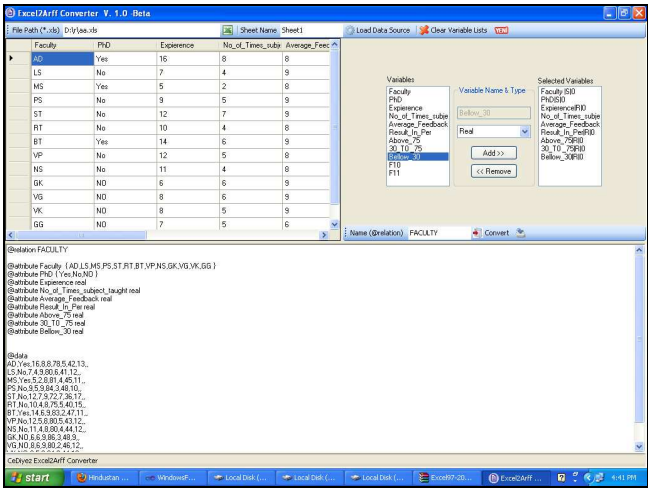

Fig 3: ARFF format of dataset

After collecting and cleaning the data, the classification of data is done using J48. J48 is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The output generated is displayed in Fig 4.

```
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances          6               46.1538 %
Incorrectly Classified Instances        7               53.8462 %
Kappa statistic                          0.4167
Mean absolute error                      0.0828
Root mean squared error                  0.2035
Relative absolute error                 58.3333 %
Root relative squared error             76.3763 %
Total Number of Instances               13
=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1          0.083      0.5         1       0.667      0.958    AD
                 1          0.083      0.5         1       0.667      0.958    LM
                 1          0.083      0.5         1       0.667      0.958    MS
                 1          0.083      0.5         1       0.667      0.958    PS
                 0          0          0           0       0          0.958    ST
                 1          0.083      0.5         1       0.667      0.958    RT
                 0          0          0           0       0          0.958    BT
                 0          0          0           0       0          0.958    VP
                 1          0.167      0.333       1       0.5        0.917    NS
                 0          0          0           0       0          0.958    GK
                 0          0          0           0       0          0.958    VG
                 0          0          0           0       0          0.917    VK
                 0          0          0           0       0          0.917    GG

Weighted Avg.    0.462    0.045    0.218    0.462    0.295    0.949
```

```
=== Confusion Matrix ===
a  b  c  d  e  f  g  h  i  j  k  l  m   <-- classified as
 1  0  0  0  0  0  0  0  0  0  0  0  0 | a = AD
 0  1  0  0  0  0  0  0  0  0  0  0  0 | b = LM
 0  0  1  0  0  0  0  0  0  0  0  0  0 | c = MS
 0  0  0  1  0  0  0  0  0  0  0  0  0 | d = PS
 1  0  0  0  0  0  0  0  0  0  0  0  0 | e = ST
 0  0  0  0  0  1  0  0  0  0  0  0  0 | f = RT
 0  0  1  0  0  0  0  0  0  0  0  0  0 | g = BT
 0  0  0  0  0  1  0  0  0  0  0  0  0 | h = VP
 0  0  0  0  0  0  0  0  1  0  0  0  0 | i = NS
 0  1  0  0  0  0  0  0  0  0  0  0  0 | j = GK
 0  0  0  1  0  0  0  0  0  0  0  0  0 | k = VG
 0  0  0  0  0  0  0  0  1  0  0  0  0 | l = VK
 0  0  0  0  0  0  0  0  1  0  0  0  0 | m = GG
```

Fig 4: Output

The accuracy is around 46%. The kappa statistic measures the agreement of prediction with the true class where value 1.0 signifies complete agreement. The confusion matrix or contingency table in this example has thirteen classes, and therefore a 13x13 confusion matrix is being displayed. The number of correctly classified instances is the sum of diagonals in the matrix; all others are incorrectly classified.

The True Positive (TP) rate is the proportion of examples which were classified as class x, among all examples which truly have class x, i.e. how much part of the class was captured. It is equivalent to Recall. The False Positive (FP) rate is the proportion of examples which were classified as class x, but belong to a different class, among all examples which are not of class x. The Precision is the proportion of the examples which truly have class x among all those which were classified as class x. The F-Measure is simply 2 * Precision * Recall / (Precision + Recall), a combined measure for precision and recall.[10]

As per J48 Algorithm, parameters that reflect noise or outliers need to be removed, hence only those targeted node are shown by tree which have some value of precision and recall. The tree generated is represented in Fig 5.

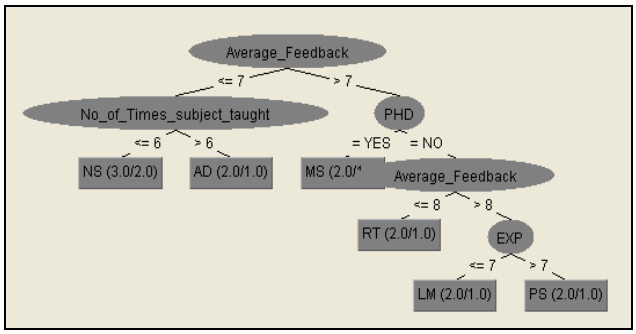

Fig 5: Decision Tree

The classification rules extracted from tree are:

1. If Average_Feedback <= 7 and No_of_Times_subject_taught is <=6 then Faculty is NS
2. If Average_Feedback >7 No_of_Times_subject_taught is > 6 then Faculty is AD

Tree has another branch which represents following rules:

1. If Average_Feedback > 7 and having PhD then Faculty is MS
2. If Average_Feedback is <=8 and having no PhD and then Faculty is RT
3. If Average_Feedback is >8 and having Experience > 7 then answer is PS
4. If Average_Feedback is >8 and having Experience <= 7 then faculty is LM

Disadvantages of J48 algorithm: The run-time complexity of the algorithm matches to the tree depth, which cannot be greater than the number of attributes. Tree depth is linked to tree size, and thereby to the number of examples. So, the size of C4.5 trees increases linearly with the number of examples. C4.5 rules slow for large and noisy datasets Space complexity is very large as we have to store the values repeatedly in arrays [8].

CONCLUSIONS

In this paper, a simple methodology based on decision tree algorithm is being used to analyze the faculty details for subject distribution within the department. This methodology will assist the Head of Departments in distribution of subjects among faculties with ease and in minimum time frame. This model will play important role in allocation of subjects in a way such that students can make efficient use of faculty knowledge and experience to enhance their career. This will have significant effect on the placements and improved quality intake in the subsequent

REFERENCES

[1]. Arora K. Rakesh, Badal Dharmendra,“ Admission Management using Data Mining using WEKA”, IJARCSSE Vol. 3, Issue 10, October 2013

[2]. [Online]http://www.estard.com/decisiontree/decision_trees_definition.asp

[3]. [Online]http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf

[4]. Gupta K G. “Introduction to Data Mining with case studies” , PHI

[5]. Moertini, Veronica S. "Towards the use of C4.5 algorithm for classifying banking dataset." Vol. 8 No. 2, October 2003 (2003): 12. Web. 24 Jan. 2013

[6]. [Online] Available: http://www.gtbit.org/downloads/dwdmsem6/dwdmsem6lman.pdf

[7]. Jing Luan, PhD Chief Planning and Research Officer, Cabrillo College Founder, Knowledge Discovery Laboratories “Data Mining Applications in Higher Education”.

[8]. Juneja, Deepti, et al. "A novel approach to construct decision tree using quick C4.5 algorithm." Oriental Journal of Computer Science & Technology Vol. 3(2), 305-310 (2010) (2010): 6. Web. 18 Feb. 2013.

[9]. [Online]http://www.creighton.edu/fileadmin/user/HSL/docs/ref/Searching__Recall_Precision.pdf

[10]. [Online] http://weka.wikispaces.com/Primer

[11]. Arora K Rakesh, Badal Dharmendra, “Location wise student admission analysis”, International Journal of Computer Science, Information Technology and Security, Dec 2012.

[12]. Arora K. Rakesh, Gupta K. Manoj ,“Data Mining: Scope Out Valuable Resources From Mountains Of Information”, IITM Buisness Review Journal, July 10

BIOGRAPHIES

Rakesh Kumar Arora is currently working in Department of Computer Science at Krishna Engineering College, Mohan Nagar, Ghaziabad, U.P, India. He has more than 11 years of teaching expierence in reputed institutes. He has no. of papers in International Journals and Conferences to his credit.

Dr. Dharmendra Badal is currently working in Department of Mathematical Sciences and Computer Applications at Bundelkhand University, Jhansi, U.P, India. He has more than 20 years of expierence at Bundelkhand University. He is also handling the additional responsibilities of Computer Head and Controller of Examination at Bundelkhand University. He was director at SRI group of institutions, Datia. He had presided no. of conferences and has no. of papers in International Journals and Conferences to his credit.