ISSN: 2278-0181

Unique Data Mining Approach to Predict Placement Chance

Arin S
6th Semester
Department of MCA
Global Institute of Management Sciences
Bangalore
Arinrockcy7@gmail.com

Sandeep Kumar G.K
6th Semester
Department of MCA
Global Institute of Management Sciences
Bangalore
sandeep.gk25@gmail.com

Shivanagaraj S
6th Semester
Department of MCA
Global Institute of Management Sciences
Bangalore
shivanagaraju5@gmail.com

***ABSTRACT*- Educational data mining (EDM) is an emerging discipline that emphasizes on the application of data mining tools and techniques on educational data. The discipline focuses on extracting and analyzing educational data to develop models for improving learning experiences and institutional effectiveness. If this technology is made use for the benefit of the common man, then the purpose is served. The purpose of this paper is to help the prospective medical students in selecting or choosing a right post graduate course namely, Dermatology, neurology, etc., based on the entrance exam ranking for admission to PG course. In this paper, unique approaches were in algorithms from two different mining models such as cluster and classification is being proposed. Two clustering algorithms viz., K-means and Support vector Clustering and a classification algorithm Naïve Bayes are applied on the same data set. These algorithms are implemented, to predict accurately one among the various courses offered that 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. Each and every course offered is associated with one of the above answers viz., E, G, A, P. Such as, Dermatology with – E, Neurology with – P and so on. Algorithms applied are compared in terms of precision, accuracy and truth positive rate. From the results obtained it is found that the cluster algorithm viz., K-means predicts better in comparison with other algorithms. This work will help the students in selecting a best course suitable for them which ensure best placement chances based on the data entered.**

Keywords*: *Data mining, Naive Bayes, K-means, Support Vector Clustering, Confusion matrix, Prediction and modeling.

I. INTRODUCTION:

Medical profession is considered to be noble as it is service oriented and highly respected field. Hence there is lot of demand for the specialization. Success relies on choosing the right specialization in the Post-graduation. Decision in this regard is arrived by accessing previous year's admission records of Medical College and manually going through the database. The objective of doing this is to find the patterns and characteristics of students admitted and to predict the future choice of the course. So huge data needs to be processed and patterns need to be compared manually, which is tedious and cumbersome. Hence data mining models are used to mine the large data using algorithms like Naïve Bayes, K-means, and Support vector machines to interpret potential and beneficial data. Data was obtained from medical college in excel format from 2009 to 2013.Data in the excel format were fed to MYSQL in the form of queries and two databases were constructed, One containing historic data from 2009 to 2012 and another test data i.e., 2013.

A. BACKGROUND AND RELATED WORKS:

Many scientists have been working to explore the best mining techniques for solving placement chance prediction problems. Various works have been done in this regard. Few of the related works are listed below:

Krishna K, Murty M N [1] Propose a novel hybrid genetic algorithm (GA) viz., genetic K- means algorithm that finds a globally optimal partition of a given data into a specified number of clusters. It is also observed that GKA searches faster than some of the other evolutionary algorithms used for clustering. Zhexue Huang [2] focuses on the technical issues of extending the k-means algorithm to cluster data with categorical values. Attractive property of the k-means algorithm in data mining is its efficiency in clustering large data sets. However, it only works on numeric data limits its use in many data mining applications because of the involvement of categorical data. Leon Bottou, Yoshua Bengio [3] Studies the convergence properties of the well-known K- means clustering algorithm. It minimizes the quantization error using the very fast Newton algorithm. Kai mingting, zijian zheng [4] introduce tree structures into naive Bayesian classification to improve the performance of boosting when working with naive Bayesian classification. Yong wang, Hodges.J, Botang [5] focuses upon three aspects of this approach: different event models for the naive Bayes method, different probability

smoothing methods, and different feature selection methods. In this paper, we report the performance of each method in terms of recall, precision, and F-measures. Yongchuan Tang, Yang Xu [6] presents a method to identify a fuzzy model from data by using the fuzzy Naive Bayes and a real-valued genetic algorithm. The identification of a fuzzy model is comprised of the extraction of "if–then" rules that is followed by the estimation of their parameters. Asa Ben-Hur, David Horn, Hava T. Siegelmann, and Vladimir Vapnik [7] Proposes a novel clustering method, SVC, based on the SVM formalism. This method has no explicit bias of either the number, or the shape of clusters. A unique advantage of our algorithm is that it can generate cluster boundaries of arbitrary shape. *Usama Fayyad, Padhraic Smyth, Padhraic Smyth* [8] Provides an overview of this emerging field clarifying how data mining and knowledge discovery in databases are related both to each other and to related fields, such as machine learning, statistics, and databases. Sudheep Elayidom, Suman Mary Idikkula & Joseph Alexander [11] Proved that the technology named data mining can be very effectively applied to the domain called employment prediction, which helps the students to choose a good branch that may fetch them placement. A generalized framework for similar problems has been proposed. Ajay Kumar Pal, Saurabh Pal [12] presents a proposed model based on classification approach to find an enhanced evaluation method for predicting the placement for students. This model can determine the relations between academic achievement of students and their placement in campus selection. A K Pal, and S Pal [13] frequently used classifiers are studied and the experiments are conducted to find the best classifier for predicting the student's performance. B K Bharadwaj , S Pal [14] Provides work to identify those students which needed special attention to reduce fail ration and taking appropriate action for the next semester examination. S. K. Yadav, B K Bharadwaj and S Pal [15] Focuses on identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising/counselling.

B. 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.

II. DATA MINING ALGORITHMS APPLIED:

A. Brief Description of the K-means algorithm:

K-means: K-means is the clustering algorithm. Concept used is partitioned clustering. It traverses each column headed by a category completely in the database and clusters it as a group. While traversing the column denoted by category, N objects with attributes are identified. Among the objects identified; objects with similar attributes are grouped to form a cluster. After traversing the entire database the number of K- partitions formed will always be less then number of N objects. K–Means algorithm divides database in to partitions or Clusters which are disjoint subsets. If N is the data sets then the k will be disjoint subsets and each subset will be having its own id. Here k is always a positive integer number.

a) Data Pre-processing:

The attributes that were present in the data provided to us; Name, age, gender, Rank, Category, sector, Address, register number, phone number.

Number of attributes that were found to be contributing to the result, after applying the chi-square test is as follows

Table 1: Mapping input values to numeric values.

Category	Input Values	Numeric values
Gender	Female, Male	0 and1
Category	2A,2B,3A,3B OBC,GM, SC,ST	0 and 1
Rank	1 to N	0 and 1
Sector	Rural, Urban	0 and 1
Branch	A to N	0 and 1
Chances	E, G, A, P	0 and 1

Rank: obtained by student in PG entrance examination Range: (1 to 5000)

Category: social background Range (2A, 2B, 3A, 3B, GM, SC, ST, OBC).

Gender: Range (Male, female).

Sector: Range (Urban, Rural).

Specialization: Range (A to N).

All the input values would be mapped between 0 and 1 as given in the table below.

b) Application of K-Means algorithm on the data set:

Step 1: Initialize the value of k either manually or systematically.

Step 2: The database provided will be divided into number of groups based on the attributes as follows: Rank, Category, Gender, specialization and Sector.

Step 3:

Table 2: Input for K-Means Algorithm

Rank	Sector	Gender	Category
0.25	0	1	0.25
0.35	1	0	0.50
0.30	1	1	0.75
0.90	0	0	1

Table 2 is obtained from table 1 based on the application of numerical formulae. In the above table 0.25 in the rank attribute will be compared with all other values in the same column and the differences between the values are noted (0.25 – 0.35). Similarly the second value i.e., 0.35 will be compared with the rest of the values in the column (0.35-0.30). Same process continues for other values also and differences are obtained in each case. A column headed by difference is obtained and values which are closed to each other are grouped as cluster which forms centroid.

0.25-0.35=0.10, 0.35-0.30=0.5

0.10 And 0.5 forms a centroid provided there are no numbers less than the above values mentioned.

Step 3: while (! EOF)

{If (next value in the difference column is nearest to centroid) {

Include in the cluster

} Else {Form a new cluster}

After each step the cluster sets gets updated which results in the formation of classified knowledge dataset. The student enters the data which would be compared with the classified knowledge data set which predicts the specialization to be selected.

Distribution (D) for the k-Means algorithm is calculated using

$$D = \sum_{1}^{N} (\text{dataset (i)} - \text{center (dataset (i)))}^2$$

If the value of D is close to 0 then

Algorithm performance is good

Else

Below average performance.

B. NAIVE BAYES:

A 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
Shiva	21	M	Rural	2a	200	Dermatology
John	22	M	Urban	3b	100	Radiology
Rani	22	F	Rural	SC	400	Neurology
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	Dermatology
M	Urban	3b	100	Radiology
F	Rural	SC	400	Neurology
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	Dermatology
John	22	M	Urban	3b	100	Radiology
Rani	22	F	Rural	SC	400	Neurology

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.

$H_{map} =_{max} (P (h/D))$

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

Rank	Gender	Sector	Category	Branch	Chance
1-200	M	Rural	Any	Dermatology	E
1-200	M	urban	Any	Radiology	E
1-200	F	Rural	Any	Neurology	E

C. Support Vector Clustering (Svc):

It classifies or clusters the data set based on the complex pattern in the data. SVC follows machines; machines are the objects of a class as defined by the algorithms. SVC explores minute details regarding connections between data points in a dataset. Svc group data into resultant attribute space, obtained from the knowledge database. e.g., the knowledge database obtained from naïve Bayes in our case acts as reference for SVC.

Duality:- The function should be defined, which checks for the connection between the attributes in a dataset on the basis of which classification will be done.

Application of the algorithm on a dataset

Table 7: Input for Svc algorithm

Name	Category	Age	Sector	Rank	Gender
William	3B	30	Urban	550	Male
Michal	3B	38	Urban	549	Male
Bhagwan	2A	21	Rural	1	Male
Emmanvel	2A	32	Rural	13	Male
Reddy	3A	23	Urban	11	Male
Balu	3A	23	Urban	10	Male

Pre-processing: After the dependency check the name and age column will be removed from the input table.

Working of the algorithm: The value of an attribute in the given table 6 is compared with the value of the knowledge database. If it is found true, then it is formed as one cluster. E.g. if (rank is equal to Knowledge dataset rank and the group exists) then include in same group.

Table 8: output of SVC applied on rank

Rank	Specialization	Chance
1 to 13	Radiology, Dermatology	E
549 to 550	Neurology, Orthopedic	A

III. IMPLEMENTATION:

As this was the problem assigned to us for our 6th semester MCA project, we were expected to implement the algorithms used to solve the problem. Hence 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.

IV. TESTING:

Data mining algorithms like K-Means, Naïve Bayes, and Support Vector Clustering 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 three algorithms expressed in terms of True Positive rate, Accuracy and Precision.

Table 9: Confusion matrix table

Algorithms	TPR	Accuracy	Precision
K-means	0.83	83%	0.83
Naïve Bayes	0.80	77%	0.75
Svc	0.81	81%	0.81

From the above table 9 it is clear that the K-means algorithm is more accurate with 83% compared to the other algorithms viz., SVC (81%) and Naïve Bayes (77%).K-Means algorithm leads with respect to true positive rate (TPR) with 0.83 correct instances and Precision (0.83).Thus the clustering algorithm K-Means predicts the results better than the other algorithms used.

V. CONCLUSION AND FUTURE ENHANCEMENT:

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, a unique approach where in algorithms from classification and clustering models has been used for predicting placement chances. Two clustering algorithms viz., K-Means and SVC and a classification algorithm, naive bayes were applied. Among these algorithms, K-Means proved to be the best predicting algorithm representing cluster model, for solving placement chance prediction problems. The results obtained after application of the algorithm viz.,K-Means (83%) were compared with results obtained using Rapid miner (83.05%). 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.

VI. REFERENCES:

[1] Krishna.k, Murty M.N "Genetic k-means algorithm", volume 29, issue 3, 1999 pages 435-439.

[2] Zhexue Huang "Extensions to the k-means algorithm for clustering large data sets with categorical values", volume 2, issue 3, pages 283-304, 1998.

[3] Leon Bottou, Yoshua Bengio "Convergence properties of the k-means algorithms", 1995.

[4] Kai mingting,zijian zheng "Improving the performance of boosting for Naïve Bayesian classification", volume 1574,1999, pages 296-305.

[5] Yong wang, Hodges.J,Botang "Classification of web documents using a naïve Bayes method" ,2003,pages 560-564, Germany, 2005.

[6] Yongchuan Tang, Yang Xu "Application of fuzzy Naïve Bayes and a rel-valued gentic algorithm in identification of fuzz model", volume 169, issue 3-4, 2005, pages 205-226.

[7] Asa Ben-Hur,David Horn, Hava T. Siegelmann, Vladimir Vapnik "Support Vector Clustering", Journal of Machine Learning Research 2 (2001) 125-137.

[8] *Usama Fayyad, Padhraic Smyth, Padhraic Smyth,* "From Data Mining to Knowledge Discovery in Databases" volume 17.

[9] Quinaln, J.R., C4.5: Programs for machine learning, Morgan Kaufmann, San Francisco, 1993.

[10] *Wu, X. & Kumar, V., the Top Ten Algorithms in Data Mining, Chapman and Hall, Boca Raton. 2009.*

[11] 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.

[12] Ajay Kumar Pal, Saurabh Pal "Classification Model of Prediction for Placement of students" I.J.Modren Education and Computer Science, 2013, 11, 49-56.

[13] A. 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.

[14] 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.

[15] S. K. Yadav, B.K. Bharadwaj and S. Pal, "Data Mining Applications: A comparative study for Predicting Student's Performance", International Journal of Innovative Technology and Creative Engineering (IJITCE), Vol. 1, No. 12, pp. 13-19, 2011.