A Combined Clustering and Classification Approach for Predicting Placement Chance

Syed Junaid Rahmathulla
Department of MCA
Global Institute of Management Sciences
Bangalore

***Abstract*— An Educational data mining (EDM) is an emerging practice that emphasizes on the application of data mining tools and techniques on educational data. The practice 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 Management students in selecting or choosing a right Specialization in Master of Business Administration (MBA) namely, Finance, Marketing, HR, Operations etc., based on the entrance exam ranking for admission to PG course. In this paper, A combined clustering and classification is applied. 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, Operation– E, Finance– G, Marketing –A,HR – P. 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 choosing a best course suitable for them which ensure best placement chances based on the data entered..**

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

I. 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. Management profession is one of the excellent professions which have bright scope, The Master of Business Administration (MBA) is a master's degree and the duration and of the course is two years. Students acquire both a working knowledge of management functions and the analytical skills needed to practice them. Therefore selecting a right specialization at the time of post graduation will plays an important role in his/her carrier. Selection of specialization is arrived by accessing and analyzing previous year's Management college data i.e, entrance exam rank. 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. Data collected from Management colleges is in the excel format, which was fed to MySQL in the form of queries and two database were constructed, one containing the historical data and another test data.

II. PROBLEM STATEMENT

Every student dreams to be successful in life. In order to be successful in life he/she must select a good specialization after graduation . 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.

III. RELATED WORKS

- Many researchers 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:
- SeralŞahan, Kemal Polat. (1) Explains that the fuzzy sets are introduced into the k-nn technique to develop a fuzzy version of the algorithm.Ying Huang* and YandaLi(2)in this paper, fuzzy k-NN method based on protein's dipeptide
- Composition was proposed for prediction of subcellular locations. An advantage of the new method is it's incorporating sequence-order effects into prediction. This method was performed to a new data set derived from version 41.0SWISS-PROT databank, and high predictive accuracy has been achieved in a jackknife test. Keller, J.M, Givens, J.A.(3)the fuzzy nearest classifier while not produce errors rate as low as k-nearest neighbor and it also attractive and desirable. kaimingting, zijianzheng(4)bayes algorithm to increase its instability and expect this to increase the success of boosting. yongchuan Tang, Yang Xu(5)present a method to identify a fuzzy model from data by using the fuzzy Naive Bayes and a real-valued genetic algorithm. The detection of a fuzzy model is comprised of the extraction of "if–then" rules that is followed by the estimation of their parameters.Hongjun

Lu, Rudy Setiono, and HuanLiu(6)constructing and training a network to correctly classify tuples in the given training data set to required accuracy, pruning the network while maintaining the classification accuracy, and extracting symbolic rules from the pruned network. Mark W Craven, Jude W Shavlik(7)neural network methods deserve a place in the tool box of the data miner. Markus Brameier and Wolfgang Banzhaf(8)We have presented an efficient algorithm for the detection of non-effective instructions in linear genetic programs. The elimination of these introns before fitness evaluations results in a significant decrease in runtime.SudheepElayidom, Suman Mary Idikkula& Joseph Alexander proved that the technology named data mining can be very effectively applied to the domain called employment prediction, which helps the students to decide a good branch that may fetch them placement. A generalized structure 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 conclude 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.

IV. 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.

B. *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 I. 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

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

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

c) Gender: Range (Male, female).

d) Sector: Range (Urban, Rural).

e) Specialization: Range (A to N).

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

C. *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.

TABLE II. 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

$$D = \sum^{N} (dataset\ (i) - center\ (dataset\ (i)))\ 2$$

```
If the value of D is close to 0 then
        Algorithm performance is good
Else
        Below average performance.
```

D. 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 III. INPUT FOR NAÏVE BAYES

Name	**Age**	**Gender**	**sector**	**Category**	**Rank**	**Branch**
Deepak	21	M	Rural	2a	200	Finance
Imran	22	M	Urban	3b	100	Operations
Kavya	22	F	Rural	SC	400	HR
Ravi	22	M	Rural	SC	867	Marketing

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 4 is a resultant after preprocessing.

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 4 to get the positive knowledge data as given below.

TABLE IV. AFTER PREPROCESSING

Gender	**sector**	**Category**	**Rank**	**Branch**
M	Rural	2a	200	Finance
M	Urban	3b	100	operations
F	Rural	SC	400	HR
M	Rural	SC	867	Null

TABLE V. POSITIVE KNOWLEDGE DATA

Name	**Age**	**Gender**	**sector**	**Category**	**Rank**	**Branch**
Shiva	21	M	Rural	2a	200	Marketing
John	22	M	Urban	3b	100	Operations
Rani	22	F	Rural	SC	400	HR

Step 3: Application of Bayes theorem on table 5 gives the resultant output table.

At the first instance data in table 5 is converted to the numeric data. Formulae listed under are used to get the below output table as the resultant.

Hmap =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

Hmap= 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 VI. OUTPUT TABLE

Rank	**Gender**	**Sector**	**Category**	**Branch**	**Chance**
1-200	M	Rural	Any	Marketing	E
1-200	M	urban	Any	Operations	E
1-200	F	Rural	Any	HR	E

E. 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 VII. 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 VIII. OUTPUT OF SVC APPLIED ON RANK

Rank	Branch	Chance
1 to 13	**Operations**	**E**
549 to 550	**Markating**	**A**

V. IMPLEMENTATION

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.

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 and Accuracy and Precision.

TABLE IX. 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.

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, 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.

BIBLIOGRAPHY

[1] A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis, Volume 37, Issue 3, March 2007, Pages 415–423,SeralŞahan,Kemal Polat.

[2] Prediction of protein subcellular locations using fuzzy k-NN method, Volume 20, Issue 1pages 21-28,Ying Huang* and Yanda Li

[3] A fuzzy K-nearest neighbor algorithm, Volume:SMC-15 Issue:4,pages 580 – 585,Keller, J.M, Givens, J.A.

[4] Improving the performance of boosting for Naïve Bayesian classification, volume 1574, 1999, pages 296-305, kaimingting,zijianzheng.

[5] 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, yongchuan Tang, Yang Xu.

[6] Effective Data Mining Using Neural Networks, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 8, NO. 6, DECEMBER 1996,Hongjun Lu, Rudy Setiono, and Huan Liu.

[7] Using Neural Networks for Data Mining, Future Generation Computer Systems special issue on Data Mining, Mark W Craven, and Jude W Shavlik.

[8] A Comparison of Linear Genetic Programming and Neural Networks in Medical Data Mining, IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 5, NO. 1, FEBRUARY 2001, Markus Brameier and Wolfgang Banzhaf

[9] A Generalized Data mining Framework for Placement Chance Prediction Problems International Journal of Computer Application (0975-8887) Volume 31- No.3, October 2011, SudheepElayidom, Suman Mary Idikkula& Joseph Alexander .

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

[11] 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, A. K. Pal, and S. Pal,.