Data Mining Approach For Predicting Student and Institution's Placement Percentage

Ashok M V
Professor, Dept. of Computer Science
Teacher's Academy
Bangalore, India
ashokmv@ymail.com

Apoorva A
Assistant Professor, Dept. of MCA
GIMS
Bangalore, India
a.apoorva89@gmail.com

Abstract-**Placement of students is one of the very important activities in educational institutions. Admission and reputation of institutions mainly depends on placements. Hence all institutions strive to strengthen placement department. In this study, the objective is to analyze previous year's student's historical data and predict placement chance of the current students and the percentage placement chance of the institution. A model is proposed along with an algorithm to predict the placement chance of students. Data pertaining to the study were collected form the same institution for which the placement chance prediction and percentage placement need to be found from 2006 to 2015. Data collected is divided into historic data form 2016to 1014 and test data i.e, 2014; 2016 data is considered as current data. Suitable data pre-processing methods are applied. Students having better chance of placement are characterized as good if not bad. This proposed model is compared with other classification algorithms such as Naïve bayes, Decision tree, and Neural network with respect to accuracy, precision and recall. From the results obtained it is found that the proposed algorithm predicts better in comparison with other algorithms.**

Keywords-Data mining; prediction; placement; classification; Naïve bayes; Decision tree

I. INTRODUCTION

It's a well known fact all round the world that admission of students in an educational institution depends on the placements. Placement is one of the factors considered for determining the quality of the institution. Hence every institution strives hard to provide better placements to their students. An educational institution contains a large number of student records. This data is a prosperity of information, but is too large for any one person to understand in its entirety. Finding characteristics in this data is an essential task in education research. It does not make sense to find the placement possibility of all the students in the institution as all the students will have not have good KSA(knowledge, skill and attitude) score. Hence there is a need for identifying those students among the whole set of students who have good KSA score and finding placement chance for them would help us achieve the objective and thus save lot of time. Hence input for the study is the best cluster of students having better KSA score who will have good chance of placement which is obtained by applying clustering algorithm and other necessary data preprocessing techniques.

II. PROBLEM STATEMENT

Every student dreams to be successful in life. The onus is on the institution to help them by providing good placement opportunity. Every student cannot be placed hence the intention of this study is to predict placement chance of the clustered students who have better chance of placement and thus find the percentage placement of the institution for the current academic year. This would help the institution to analyze the status of the institution in comparison with other institutions and take appropriate measures to improve it. A prediction model is proposed. Various mining algorithms are applied on the processed data, tested, and compared with the proposed model based on certain criteria like accuracy, precision and recall etc.

III. 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:

Jae H. Min et al., 2001[1] Applies support vector machines (SVMs) and used a grid-search technique using 5-fold cross-validation to find out the optimal parameter values of kernel function of SVM, they applied SVM to bankruptcy prediction problem, and showed its attractive prediction power compared to the existing methods;J.A.K. Suykens et al., 1998 [2]discussed a least squares version of support vector machine classifiers and illustrated that a least squares SVM with RBF kernel is readily found with excellent generalization performance and low computational cost.;Tung-Kuang Wu et al., 2008[3]apply two well-known artificial intelligence techniques, artificial neural network (ANN) and support vector machine (SVM), to the LD diagnosis problem;Guha, S et al., 1999[4]proposed a new concept of links to measure the similarity/proximity between a pair of data points with categorical attributes and developed a robust hierarchical clustering algorithm;KakotiMahanta et al., 2005[5]prove that under certain conditions, the final clusters obtained by the algorithm are nothing but the connected components of a certain graph with the input data-points as vertices;AgnieszkaPrusiewicz et al.,[6] 2010 proposal for services recommendation in online educational systems based

on service oriented architecture are introduced;Christian Borgelt2005[7]proposed a new data structure for frequent item set mining algorithms.BalazsRacz, D 2004[8] described an implementation of a pattern growth based frequent item set mining algorithm. The data structure presented here can accommodate the top-down recursion approach, thereby further reducing memory need and computation Time;Ke Wang, Liu Tang et al., 2002[9]propose an efficient algorithm, called TD-FP-Growth (the shorthand for Top-Down FP-Growth), to mine frequent patterns;SudheepElayidom et al., 2011[10] attempt 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 Palet al,.2013[11]suggested that Naïve Bayes classifier has the potential to significantly improve the conventional classification methods for use in placement among all the machine learning algorithm tested;K. Pal et al., 2013[12]describe the use of data mining techniques to improve the efficiency of academic performance in the educational institutions;B.K. Bharadwaj et al., 2011[13]the classification task is used on student database to predict the students division on the basis of previous database;S. K. Yadav et al., 2012[14]focusing upon methodologies for extracting useful knowledge from data and there are several useful KDD tools to extracting the knowledge.

IV. PROPOSED MODEL.

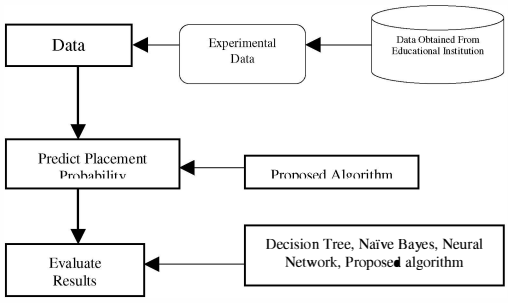

Fig 1: Proposed Model

The algorithm of the proposed model, along with its computational processes for predicting placement chance is outlined below:

Step 1: Data collection.

The goal is to find the proficient students in the college under consideration viz., XX for the year 2016. In this college there were 1,434 students. These students hailed from various courses that were operative in the college. The courses are MBA, MCA, BCA, B.Com, and BBA.

Step 2: Predict Placement chance.

This step predicts placement chance of the student and also percentage placement of institution using Proposed Classification algorithm.

Step 3: Evaluate the result

The obtained result is compared in terms of precision, accuracy, variance with other algorithms such as Decision tree, Naïve Bayes, Neural network.

V. DATA DESCRIPTION

The objective is to predict the placement chance of students identified as proficient students in the college identifies as XX. Basic requirement of any prediction problem is the existence of previous or past data based on which future is predicted. Data is collected from a college XX identified above, that offers various courses.

Data collection is divided into three types.

Historic Data: Collected for the duration of 10 years starting from 2006 to 2014

Test data: Collected for the year 2015.

Current data: Students identified as proficient students for the year 2016.

Table I: Data Description

Variables	Description	Possible Values
Year	Year for which the data is entered ,	{int}
Reg-no	Register number of the student's	{int}
Branch	Branch (MCA,MBA, BSC….etc) of the student	{1, 2, 3, 4, 5...}
Percent	Over all Percentage of the students	{65,71, 82,….,100}
Skills	Knowledge, Skills and Ability	{1, 2, 3, 4, 5...10}
Effective-score	Effective-score=percent+skills*10, it shows the overall performance of the student.	{ 0,50,99,154,..200 }
Placed	Student Placed based on her performance	{Text}

Year : Year that student completed education. Data collected were from 2006-2015.

Reg-no : Register number of the student. It takes any integer values.

Branch **:** represents the name of the Branch. It can take only text values ranging from A-Z

Percent **:** various marks scored by student in subjects. It can take only the numeric values from 0 to 100.

Skills : it shows the overall Skills of the student.. It can take only the numeric values from 0 to 10.

Effective-score: it shows the overall performance of the student Formula to calculate Effective-score is as follows

Effective score = percent + skills * 10 It can take only the numeric values from 0 to 200.

Placed : Placed based on student performance. Value is taken in the form of Yes\No,

IF Yes
Student placed,
ELSE
Student has not been placed.

VI. PROPOSED ALGORITHM.

An algorithm is proposed to achieve the objective of study. The algorithm is as follows.

Input : current Student, oldStudentList
Output : Placement chance

1. Read Student
2. Read oldStudentList
3. Select count all students in oldStudentList having score = score of current Student.
   Store it as count Selected
4. Select count all placed students in oldStudentList having score = score of current Student.
   Store it as count Placed And Selected
5. If countSelected == 0 then
6. probability =0.5
7. else
8. Calculate chance = count Placed And Selected/count Selected
   [End If at Step 5]
9. If chance >=0.4 (i.e Excellent/Good/Average)
10. set placement chance Good
    else
11. set placement chance Bad
12. [End If at Step 9]
13. Write placement chance

VII. EXPERIMENTAL EVALUATION

TABLE II :Input data(Output of clustering algorithm)

reg_no	branch	effective_score	centroids of cluster
1	MCA	53	55.33
2	MCA	72	55.33
3	MCA	110	105.0
4	MCA	41	55.33
5	MCA	129	146.66
6	MCA	146	146.66
7	MCA	106	105.0
8	MCA	100	105.0
9	MCA	104	105.0
10	MCA	165	146.66

An algorithm was proposed to estimate the number of clusters and finding the elements of the cluster using centroid based on Euclidean distance.

Table II represents output of the clustering algorithm which is used as the input to the proposed algorithm with the attributes as shown above.

For each student in the selected clusters the following operations are performed. Store the student in a variable 'S'.

$$f(x,y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

$$C = \sum_{x=1}^{n} f(x,y)$$

Where x is effective score of each historic student data and y is effective score of current student and C is count of Selected Historic Data

Count all the effective scores of historic data same as effective score 'S'.

Illustration:

If the effective score of current student is 104 then this value will be searched in the historic data and returns the count of such data.

ex: c = f(72,104)+f(121,104)+f(146,104)+f(106,104)+f(104,104)+f(100,104)+f(165,104)+f(83,104)+f(129,104)+f(110,104)
=0+0+0+0+1+0+0+0+0+0
=1

Step 4 of the algorithm.

$$f(x,y) = \begin{cases} 1 & \text{if } x = y \ \&x(\text{placed}) = \text{true} \\ 0 & \text{otherwise} \end{cases}$$

$$d = \sum_{x=1}^{n} f(x,y)$$

Where x is effective score of each historic student data and y is effective score of current studen and d is count of selected and placed historic data.

Count all the effective scores of historic data same as effective score 'S' and also having flag as placed.

Illustration:

If the effective score of current student is 105 then this value will be searched in the historic data only which has flag of placed and returns the count of such data.

ex: d = f(72,104)+f(121,104)+f(146,104)+f(106,104)+f(104,104)+f(100,104)+f(165,104)+f(83,104)+f(129,104)+f(110,104)
=0+0+0+0+1+0+0+0+0+0
=1

Step 8: p =d / c

Where p = probability of placement, d = count of selected historic students who have been placed, c= count of selected historic students,
if c=0; consider p=0.5 according to step 6 if in case c becomes 0 then p tends to infinity. Practically it is a unique case where in there is no instance of occurrence in historical data. To avoid this situation the value of p is taken as 0.5.

Step : 9, 10 and 11

TABLE III: Probability Ranges

Range of Probability	Remark	Value
p >= 0.9	Excellent	Good
0.9 > p >= 0.6	Good	
0.6 > p >=0.4	Average	
p < 0.4	Poor	Bad

According to step 8 various values of p are obtained. These values are classified as above. If p> 0.9 it is considered as excellent, if it ranges between 0.9 and 0.6 it is good, if it is between 0.6 and 0.4 it is average and if it is less than 0.4 it is poor.

TABLE IV : Placement Chance for input data

reg_no	branch	effective_score	centroids of cluster	Placement Chance
1	MCA	53	55.33	Bad (Not Selected)
2	MCA	72	55.33	Bad (Not Selected)
3	MCA	110	105.0	Bad
4	MCA	41	55.33	Bad (Not Selected)
5	MCA	129	146.66	Good
6	MCA	146	146.66	Good
7	MCA	106	105.0	Bad
8	MCA	100	105.0	Good
9	MCA	104	105.0	Good
10	MCA	165	146.66	Good

let us explain the first instance of the table IV.Since the centroid of the reg_no1 falls in the eliminated cluster in the module1, this student will not be placed. similarly the student with reg_no 2 won't be selected. consider student with reg_no 3.The centroid value is 105.0. hence it is concluded that the placement chance is bad as it falls in the row poor with p<0.4 in the above table IV. Same explanation can be given for student with reg_no5 where the p value falls in the excellent row with p>0.9.
placement percentage is calculated using
Placement percentage =number of good*100 / total number of students
As per the above calculations percentage of placement chance is **50%**

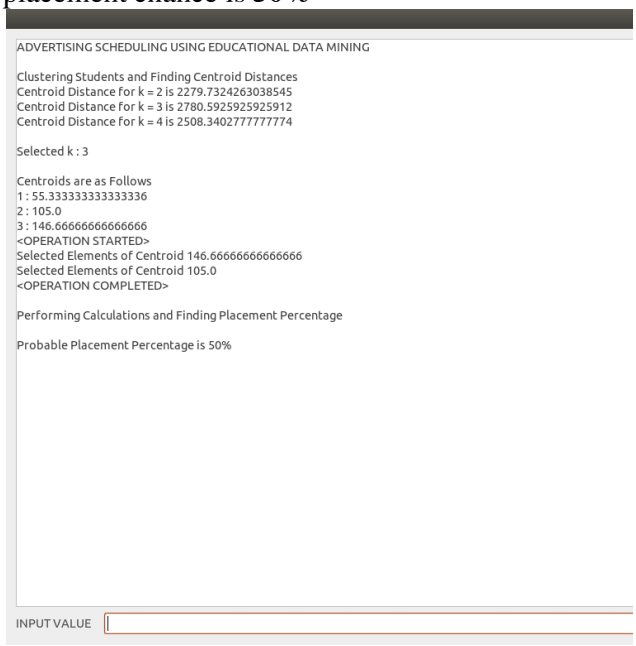

Fig 2: screen shot of placement chance prediction in percentage.

The above screenshot represents percentage placements of the institution considered.

VIII. RESULTS

CONFUSION MATRIX

Data mining algorithms like decision tree, Naïve bayes, neural network and proposed algorithm 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.

TABLE V : Comparison of Proposed algorithm with other algorithms

Algorithm	Accuracy	TPR	Precision
Decision Tree	0.84	0.95	0.74
Naïve Bayes	0.87	0.93	0.78
Neural Networks	0.83	0.88	0.69
Proposed Algorithm	0.92	0.96	0.87

The above table V gives accuracy, true positive rate, false positive rate, true negative rate, false negative rate and precision of different algorithms compared with the proposed algorithm. Precision and accuracy of the proposed algorithm is high compared with other classifying algorithms. The false negative rate of the proposed algorithm is low against all other algorithms.

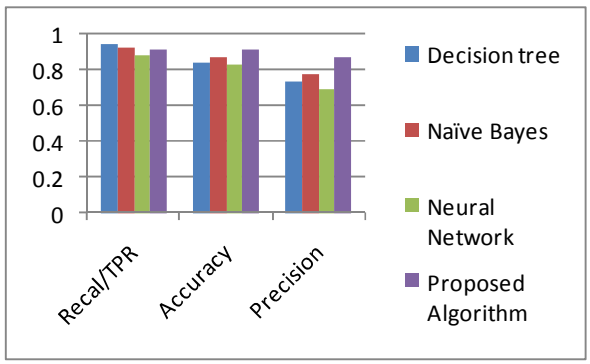

Fig 3:Comaparison of algorithms with Proposed algorithm

The above graph represents accuracy, recall, precision of various classification algorithms. Proposed algorithm has the highest precision, accuracy and recall. Decision tree, Naïve bayes, Neural network and proposed algorithm is represented by blue, red, green and purple respectively.

IX. CONCLUSION

Data mining techniques applied on educational data in 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 model was proposed along with a algorithm. This was compared with three other classification algorithms such as decision tree, naïve bayes, and neural network in terms of accuracy, precision, true positive rate(recall).The proposed model, proved to be the best predicting model for solving placement chance prediction problems compared to all other algorithms. Hence, having the information generated through our study, institution would be able to design strategies to overcome lacunae and improve placements with best chances of getting placed. Thus admission can be increased.

REFERENCES

[1] Jae H. Mina, Young-Chan Leeb, "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters", Volume 28, Issue 4, May 2005, Pages 603–614.

[2] J.A.K. Suykens and J. Vandewalle, "Least Squares Support Vector Machine Classifiers:" Volume 308, Issue 2, 27 April 2001, Pages 397–407.

[3] Tung-Kuang Wu, Shian-Chang Huang:"Evaluation of ANN and SVM classifiers as predictors to the diagnosis of students with learning disabilities", Volume 34, Issue 3, April 2008, Pages 1846–1856.

[4] Guha, S.; Rastogi, R.; Kyuseok Shim "ROCK: a robust clustering algorithm for categorical attributes", Pages 512 – 521.

[5] KakotiMahanta, 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.

[6] Agnieszka Prusiewicz, "MaciejZiębaServices Recommendation in Systems Based on Service Oriented Architecture by Applying Modified ROCK Algorithm" Volume 88, 2010, pp 226-238.

[7] Christian Borgelt, "An implementation of the FP-growth algorithm", Pages 1 – 5, 2000.

[8] BalazsRacz, D: An FP-Growth Variation without Rebuilding the FP-Tree".

[9] Ke Wang, Liu Tang, Jiawei Han, "Junqiang Liu "Top down FP-Growth for Association Rule Mining", Volume 2336, 2002, pp 334-340.

[10] SudheepElayidom, 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.

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

[12] 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.

[13] 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.

[14] 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.