IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

Prediction of Campus Placement Using Data Mining Algorithm-Fuzzy logic and K nearest neighbor

Mangasuli Sheetal B[1], Prof. Savita Bakare[2]

Student, Department of Computer Science &Engg, KLE DR M S Sheshgiri College of Engg & Tech, Belgaum, India[1]

Professor, Dept of Computer Science &Engg, KLE DR M S Sheshgiri College of Engg & Tech, Belgaum, India[2]

Abstract: Data Mining is "the process of extracting useful information from a large scale data set". It is a powerful tool to be considered best in the field of education. Educational data mining involves the new methods and its approaches for discovering the knowledge by analysing the database sets to support the decision making process in educational institution. It interprets an effective method for mining the student's performance based on the database sets to predict and analyse whether a student (he/she) will be recruited or not in the campus placement. The placement of a student not only depends on his academic capabilities but also involves the attributes such as co-curricular activities, communication skills etc. Using these datasets and attributes, predictions are made using the Data Mining Algorithm "Fuzzy logic" and "K nearest neighbor (KNN)". The results obtained from each approaches are then compared with respect to their "performance" and "accuracy" levels by graphical analysis and thus the decisions are made towards the best prediction in the campus placement.

Keywords: Data Mining, Educational Data Mining, Fuzzy Logic, KNN.

I. INTRODUCTION

Placements are considered to be very important for each and every college. The basic success of the college is measured by the campus placement of the students. Every student takes admission to the colleges by seeing the percentage of placements in the college. Hence, in this regard the approach is about the prediction and analyses for the placement necessity in the colleges that helps to build the colleges as well as students to improve their placements. The model is built by using the data mining techniques. The algorithms used for building the model are "Fuzzy logic" and "K nearest neighbor". "Fuzzy logic is a logic system for reasoning that is approximate rather than exact". The "K Nearest Neighbor (KNN) is a standard classification algorithm that collects all available cases and classifies the new cases based on the distance measures". The efficiency/accuracy of each model is visualized and tested and based on the performance analysis, each model results are discussed.

A. Introduction to Data Mining:

Data Mining is the process of extracting useful information from large scale dataset. In other words, Data Mining is the process of mining knowledge from structured and unstructured data. It is also known as knowledge discovery process from large unstructured data. Data Mining is the imperative step in the process of knowledge discovery (KDD). The following are the various steps involved in the knowledge discovery process:

- Cleaning the Data set: Here, the process is to remove the noise and inconsistent data.
- Integration of Data: Here, multiple data sources are integrated.
- Selection of Data: From the database, the data relevant to the task are retrieved in this step
- Transformation of Data: It is a process to perform aggregation or summary tasks, i.e., data can be transformed into the forms which are appropriate for KDD.
- Mining the Data: In this stage, to extract useful data patterns various intelligent methods are applied.

B. Introduction to Educational Data Mining:

The uses of Data mining approaches in the education atmosphere are called as Educational Data Mining (EDM). EDM is defined by the International Educational Data Mining Society as "an unindustrialized discipline, concerned with growing methods for exploring the sole types of data that come from educational settings and using those methods to better understand scholars/students and the settings which they acquire in".

II. LITERATURE SURVEY

T. Jeevalatha, et.al used the decision tree algorithm to predict the selection of student for the placements [2]. They used Decision Tree (DT) algorithm such as C4.5, ID3, and CHAID which were developed by using Data Mining Rapid Miner software/tool. The validation for the above said three algorithms are checked and there significant accuracy was founded. The authors concluded that the ID3 is the best algorithm than the other. ID3

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

provides accuracy of 95.33% which is higher than the CHAID and C4.5.

NeelamNaik and SeemaPurohit built the model to classify the performance of the placement of students [3]. The error produced to classify validation data, result prediction classification tree was 38.46% and while for validating placement prediction classification tree was found 45.38% respectively.

Ajay Kumar Pal and Saurabh Pal collected the data for the study and analysis of the student's educational performance basically for training and placement. The authors used different classification algorithm and used WEKA data mining tool [4]. They concluded that naive Bayes classification model is the better algorithm based on the placement data with found accuracy of 86.15% and overall time taken to build the model is at 0 sec. As compared with others Naïve Bayes classifier had lowest average error i.e. 0.28.

Ajay Shiv Sharma, and et.al, used the logistic regression model and developed the placement prediction system (PSS) [5]. The accuracy of training and testing of the algorithm was 98.93% and 88.333%.

OktarianiNurulPratiwi used to classify/predict the student placement class using the prediction/classification algorithm [6]. He used the six classification algorithms which are OneR, SMO, KStar, J48 and SimpleCart. From the experiment they concluded that J48 and SimpleCart is the best classification algorithm with accuracy 79.61%.

BahenSen, EmineUcar and DursunDelen collected the large and feature rich dataset and build the model to predict the placement test results [7]. They used support vector machine, C5 Decision Tree algorithm, and artificial neural network. They resolved that C5 Decision Tree algorithm is the better prediction model with efficiency of 95%, the accuracy of support vector and artificial neural network is 91% and 89%.

Vikas Chirumamilla, BhagyaShruti T, SasidharVelpula and Indira Sunkara presented the study of two fold objective [8]. The first in which the supervision of objective is performed and predictive model is used and explore the hidden knowledge. The authors have used two data mining techniques naïve Bayes and C4.5. The [8] resolved that C4.5 is better than Naïve Bayes for the classification/prediction of placement of the student. The accuracy of C4.5 obtained is 88.89%.

Ravi Tiwari and Awadhesh Kumar Sharma built the prediction model to improve the placement of the students [9]. They used WEKA as the data mining tool to build the model using random tree algorithm. They also used ID3, Bayes Net, RBF network, J48, algorithms on the student data set. They resolved that the RT (Random Tree) algorithm is more accurate with 73% for the classification/prediction of the model. The accuracy using ID3 and J48 is 71%.Bayes Net is 70% accurate and 65% accuracy using RBF network algorithms.

Thus, from the above literature survey, it has been observed that most of the approaches are rooted on the decision tree algorithm. However, in this work the emphasis is on the algorithms used in order to increase the accuracy are Fuzzy logic and K nearest neighbor.

III. METHODOLOGY

The architectural design consists of the various strategies used in the data mining process. The detail of the system model is presented in the Figure 1.

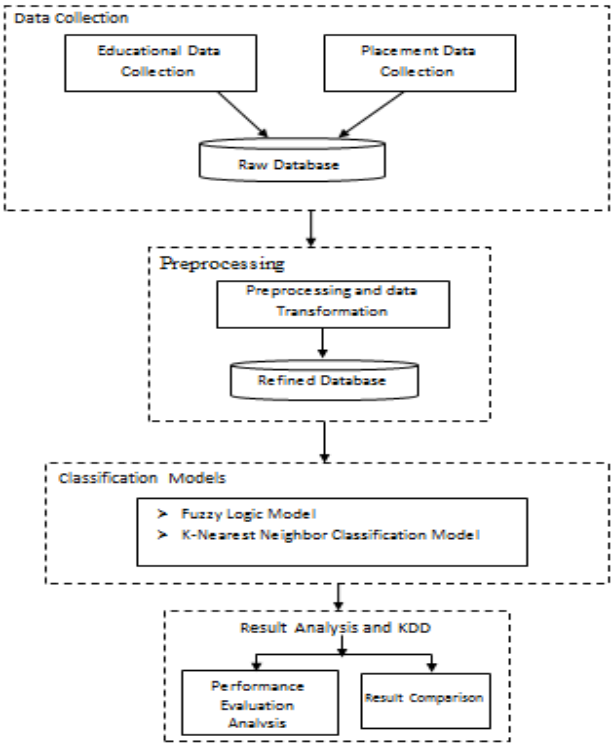

Figure 1 System Architecture

In the first phase of this model, the student's data set will be collected from the educational institutions. The data set is then cleaned and pre-processed manually by verifying all the attributes entries and making changes using Microsoft office excel. The various data mining techniques are applied to discover the knowledge.
To predict student's placement information, the following algorithms will be used to build the prediction model

- Fuzzy logic
- K nearest neighbor Classification

In the last phase of the model, the classification result is analysed and compared as a step in the process of KDD (knowledge discovery). Thus, the accuracy of each model is presented as a final outcome step.

IV. IMPLEMENTATION

A. Data Collection:
Data set involves the data of the students collected from the academic institutions. The data consists of academics and personality development skills of the students. Here

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

the attributes considered are puc marks, UG aggregate, seminar skills, practical marks, communication skill, and participation in co-curricular activities, interaction and leadership quality of the student. So here data sets of 900 students are considered. In which 600 data sets are used for training set which are used for building the model and remaining data sets are used as testing data for validating the model.

B. Pre-Processing Module:
After collection of data set, it is necessary to pre-process the data set. Pre-processing is an important phase in data mining and the dataset should be preprocessed before applying the data mining algorithm. The pre-processing tasks include cleaning, transformation and integration. The data in the dataset is cleaned and pre-processed manually by checking the attributes entries. The changes are made using the Microsoft excel format. This excel format sheet is saved in Comma Separated Value (CSV) format. The attributes considered are sslc marks, puc/diploma marks, aggregates of semester marks, practical marks aggregate, seminar marks, and participation in curriculum activities, communication skills, interaction and leadership.

C. Classification Module:
Classification of data is a two phase process. In phase one which is called training phase a classifier is built using training set of tuples. The second phase is the classification phase, where the testing set of tuples is used for validating the model and the performance of the model is analyzed. The algorithms to perform such analysis and validation are Fuzzy Logic and K Nearest Neighbor.

1. Fuzzy Logic:
Fuzzy logic is used to develop the basis as rules for inferences which include fuzzy sets. The fuzzy logic system is defined as the mapping of the non-liner input data set to the output of scalar data and is used as whole interval of real number between zero and one [10].
2. K nearest neighbour:
KNN [11] is another algorithm used for classification and prediction process in Data mining. The KNN is an illustration of instance based learning. KNN is similar to the nearest neighbor algorithm except that it searches for the k closet to the uncategorized/unclassified instance. In this, the training data set is stored, so that for a new unclassified record a classification may be detected by comparing it to the most related/similar records in the training set.

The two models output the prediction results of the student's placements and the accuracy of each model is calculated and efficiency of each model is compared in terms of accuracy where accuracy is the percentage of testing set examples correctly classified by the classifier.

V. RESULTS AND ANALYSIS:

The two approaches data set used for is further splitted into two sets consisting of two third as training set and one third as testing set. The approaches use the 300 data sets as the testing sets and the attributes include puc marks, UG aggregate, practical marks, seminar marks and personality development and the fields are actual result and prediction result. The efficiency of the two approaches is compared in terms of the accuracy.

The accuracy of the prediction model/classifier is defined as the total number of correctly predicted/classified instances. Accuracy is given by using following formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} * 100$$

Where TP, TN, FN, FP represents the number of true positives, true negative, false negative and false positive cases.

A. Accuracy of the Fuzzy Analysis:
When test data is submitted to the fuzzy logic model, the accuracy obtained is 92.67% and the execution time taken by it is 450 (msec).

TABLE I Fuzzy Analysis

Analysis of Fuzzy Accuracy	
True Positive	158
False Negative	5
False positive	17
True Negative	120
Accuracy	92.67%
Execution Time	450 (mSec)

B. Accuracy of the KNN Analysis:
When test data is submitted to the KNN classifier, the accuracy obtained is 97.33% and the execution time taken by it is 13458 (msec).

TABLE II KNN Analysis

Analysis of KNN Accuracy	
True Positive	163
False Negative	0
False positive	8
True Negative	129
Accuracy	97.33%
Execution Time	13458 (mSec)

C. Performance Analysis:

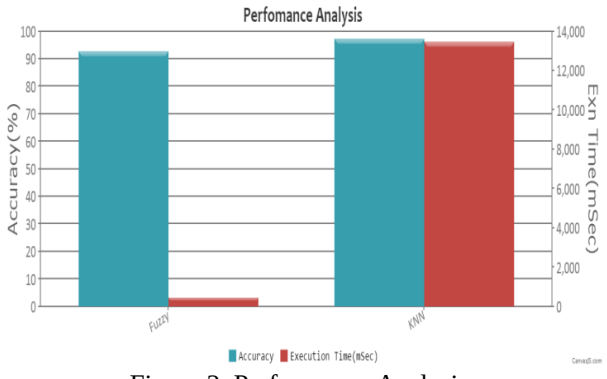

Figure 2. Performance Analysis

DOI 10.17148/IJARCCE.2016.5666

The above Figure 2, display the accuracy and execution time taken by each of the algorithm. In this, the blue colour bar represents the accuracy while the brown colour bar represents the execution time. The accuracy obtained by using the Fuzzy logic is 92.67% and the execution time is 450 (msec) while the accuracy obtained by using the K nearest neighbor is 97.33% and the execution time taken is 13458 (msec). Thus, it can be interpreted as, even though KNN algorithm takes more time compared to the Fuzzy logic but it yields a better performance.

CONCLUSION

The campus placement activity is very much important as institution point of view as well as student point of view. In this regard to improve the student's performance, a work has been analysed and predicted using the algorithms Fuzzy logic and the KNN algorithm to validate the approaches. The algorithms are applied on the data set and attributes used to build the model. The accuracy obtained after analysis for KNN is 97.33% and for the Fuzzy logic is 92.67%. Hence, from the above said analysis and prediction it would be better if the KNN is used to predict the placement results.

REFERENCES

[1] Romero, C.Ventura, S. and Garcia, "Data mining in course management systems: Model case study and Tutorial". Computers & Education,Vol. 51, No. 1. pp.368- 384. 2008.
[2] T. Jeevalatha,N. Ananthi,D. Saravana Kumar, "Performance Analysis of Undergraduate Students Placement Selection using Decision Tree Algorithms", International Journal of Computer Applications (0975 – 8887),Volume 108 – No 15, December 2014
[3] NeelamNaik and SeemaPurohit, "Prediction of Final Result and Placement of Students using Classification Algorithm", International Journal of Computer Applications (0975 – 8887), Volume 56– No.12, October 2012
[4] Ajay Kumar Pal and Saurabh Pal, "Classification Model of Prediction for Placement of Students", I. J. Modern Education and Computer Science, 2013, 11, 49-56.
[5] Ajay Shiv Sharma, Swaraj Prince, ShubhamKapoor and Keshav Kumar, "PPS - Placement Prediction System using Logistic Regression", IEEE International Conference on MOOC, Innovation and Technology in Education (MITE), 2014.
[6] OktarianiNurulPratiwi , "Predicting Student Placement Class using Data Mining", IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE), August 2013
[7] BahaSen , EmineUcar and DursunDelen , "Predicting and analyzing secondary education placement-test scores: A data mining approach", International journal of Expert system with applications, Volume 3 , 2012,Issue 10, pgno: 9468-9476.
[8] Vikas Chirumamilla, BhagyaSruthi T, SasidharVelpula and Indira Sunkara, "A Novel approach to predict Student Placement Chance with Decision Tree Induction", International journal of Systems and Technologies Double Blind Peer Reviewed Journal, Volume 7, Issue 1, 2014, pp 78-88
[9] Ravi Tiwari and Awadhesh Kumar Sharma, "A Data Mining Model to Improve Placement", International Journal of Computer Applications (0975 – 8887) Volume 120 – No.12, June 2015
[10] Jerry M. Mendel, "Fuzzy logic systems for engineering: A Tutorial", Proceedings of the IEEE, 83(3):345-377, Mar 1995.
[11] Daniel T. Larose, "Discovering Knowledge in Data: An Introduction to Data Mining".

BIOGRAPHIES

Mangasuli Sheetal B is a M. Tech student in the Department of Computer Science & Engineering, KLE Dr. M S Sheshgiri College of Engineering, Belgaum. She completed her Bachelor of Engineering in Computer science& Engineering from Basaveshwar Engineering College, Bagalkot in the year 2014.

Prof. Savita Bakare is a Faculty working for the Department of Computer Science & Engg, KLE DR M S Sheshgiri College of Engg.& Tech., Belgaum, since 2009 till date. The area of research interest is in the field of Data Mining.

DOI 10.17148/IJARCCE.2016.5666