International Journal of Engineering & Technology, 7 (2.7) (2018) 1100-1103

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Comparative analysis on job prediction of students based on resume using data mining techniques

T. Ravi Kumar [1] *, P. Yasaswini1, G. Rafi [2], Dhulipalla Vijay Krishna [3]

[1] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur Dist., Andhra Pradesh, India
[2] Department of Computer Science and Engineering, Narasaraopeta Engineering College, Guntur Dist., Andhra Pradesh, India
[3] Department of Management Studies, VFSTR, Vadlamudi, Guntur Dist., Andhra Pradesh, India
**Corresponding author E-mail: rktata@kluniversity.in*

Abstract

The manuscript should contain an abstract. The abstract should be self-contained and citation-free and should not exceed 200 words. The abstract should state the purpose, approach, results and conclusions of the work. The author should assume that the reader has some knowledge of the subject but has not read the paper. Thus, the abstract should be intelligible and complete in it-self (no numerical references); it should not cite figures, tables, or sections of the paper. The abstract should be written using third person instead of first person.

Keywords: *Naïve Bayes Classifier; Decision Trees; K-Nearest Neighbor Classifier; Weka.*

1. Introduction

The traditional method to select students for the placements takes lot of time to monitor student's skill set from a large data base according to the requirements of that job. As it consumes lot of time, and difficult to identify, the most promising solution is implementing classification techniques in data mining which classifies the given data set and used to predict that student is eligible for the job or not. Therefore the proposed system will enable more effective way to short list the students based on resume using data mining techniques. As there are many classification algorithms in data mining it is necessary to know about the algorithm which predicts and classifies the best. So in this paper comparative analysis on different algorithms is done using Weka tool to know the best classification algorithm.
To find the best algorithm we have to analyze the results of various data mining classification techniques using data mining tools such as weka-3. 4. 9, R, python. In this paper Weka tool is implemented to compare the results for the given data set. Weka a data mining tool which is used for preprocessing, classification, regression, clustering, Visualization, and association rules. In this we considered a data set and analyzed using this weka-3.4.9 tool.

2. Literature survey

In Data Mining, the approach for student prediction and placement percentage of institution proposed a new algorithm and it is tested and compared with Decision tree, Naïve Bayes and Neural Network [1]. A Model for predicting student placement eligibility Using Data Mining Technique by proposes a model which checks current status of the student and predicts which company the student can be placed into, giving the student scope to better prepare for the company [2]. Student placements prediction using Bayesian Classification is similar to former. In this the author used Naive Bayes data mining technique is used to analyse student academic data and predicted results based on attributes which helps management authorities to improve placements of students from extracted data [3].Prediction of campus placement using Data Mining classification algorithms such as Fuzzy logic and K-nearest neighbor and compared KNN and Fuzzy logic to see which yields better results comparatively [4].A Placement prediction using K-nearest neighbor again used KNN but compares the results with logistic regression and SVM [5]. In Application of Data mining techniques in placement prediction of Students approaches using algorithms like cluster analysis, classification decision trees. In the future scope research may focus on more attributes with various datasets [6]. Predicting student placement using Data mining techniques although it is focused on Indonesian students it uses a wide range of algorithms like Simple Kart, Kstar, OneR [7].By implementing Naïve Bayes algorithm on old data base of students and predicted the student's performance [8]. In this paper quality of educational system is enhanced by evaluating student data of academics using data mining functions [9In this paper based on the final results of students by using classification techniques predicted the predication of placements [10].A Generalized Data mining framework for Placement prediction in this paper confusion matrix and decision trees are used to predict and compared the results of decision trees and confusion matrix [11].Here in this paper they used network and decision tree and also compared the performance based on the results of two techniques [12].

3. Methodology

In the data collection and pre-processing phase the resume are collected from the students and then the data is pre-processed manually by verifying the attributes.
In next phase of the system model supervised learning algorithms are implemented to data set to predict the probability of a student to get place.

Copyright © 2018 T. Ravi Kumar et. al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

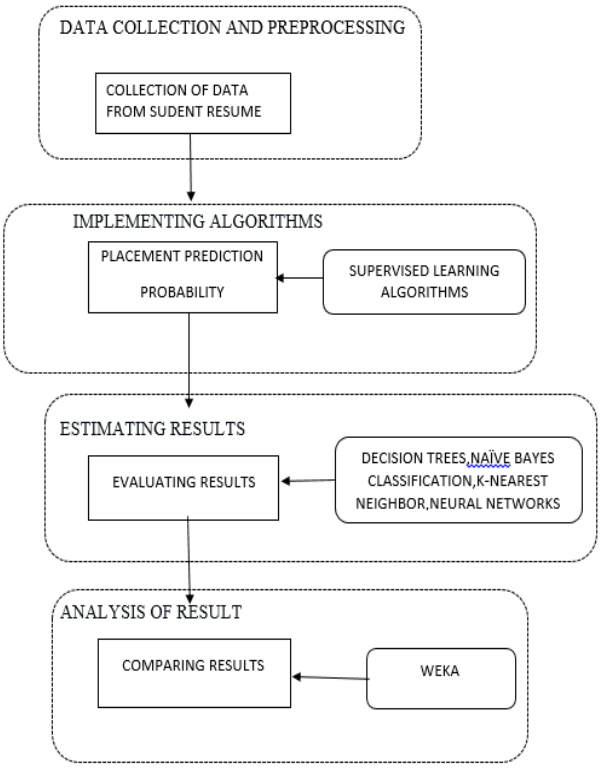


Fig. 1: Methodology

In estimating results, phase the outcome of various algorithms like

- Decision trees
- Naïve Bayes Classification
- K-Nearest Neighbor
- Neural Networks are evaluated.

In analysis phase the result from various algorithms which are implemented are compared by using Weka tool to determine the best technique for student job prediction. We make use of some data mining algorithms like K-nearest neighbors' classification, Decision trees, Neural Networks, Naïve Bayes Classification for classifying the likelihood of a student getting placed.

4. Classification algorithms

4.1. Decision trees

Decision tree is data mining classification technique used to solve problems related to classification and prediction. It is supervised learning algorithm and similar to flow chart structure, simply recursive. Decision trees are easily understood as they can be expressed in natural language by converting into IF-THEN rules. The rules are obtained by traversing every path i.e., starting from the root node to each leaf node in the tree.

Decision tree is a directed tree comprises of nodes and directed edges. Nodes are of three types root node, leaf node (terminal) which is denoted by oval and internal node (non leaf node) which is denoted by rectangles. The top most node with no incoming edge is called the root node of the tree. Each internal node has only one incoming edge and each internal node is computed that means it performs test on an attribute and specifies a path resulting to the leaf node which is a decision node. A directed edge or branch representing the outcome of the test and determines the direction of flow from the root node to leaf node and holds a class label. The following metrics used to compute time complexity are depth of the tree and the number of nodes, attributes used and the tree depth.

There are several algorithms to construct Decision trees

- ID3
- C4.5
- CART
- CHART
- CHAID

4.2. ID3 algorithm

Iterative Dichotomiser is a classification algorithm which is used to build a decision tree from the given data set. Once the tree is constructed it is applied to each and every instance and results in classification of that instance. It is a greedy approach as the decision trees are constructed in a top down recursive divide and conquer manner. The given information is partitioned into smaller subsets based on the output after performing calculation using entropy and information gain. This algorithm chooses the information gain as attribute selection and the node which has high value of information gain is considered as splitting node attribute of the present node.

4.3. Entropy

Entropy depicts the amount of information in an attribute .For a given collection S and probabilities or proportions p1, p2, p3…...ps of Entropy is calculated by using the below formula

Entropy_(S) = ∑ -pi log2 pi

4.4. Information gain

ID3 considers an attribute which has highest gain in information as splitting attribute. Value of information gain is obtained by calculating difference of the entropy values of the original data and the weighted sum of each of the subdivided data set's entropy values.

G_ (D, S) =H (D)-∑P (Di) H (Di)

The attribute, which is having largest information gain, is considered as splitting node.

4.5. K-nearest neighbour's classification

K-nearest neighbor, a data mining classification algorithm is one of the supervised learning algorithm used to classify the given data into two classes either a yes or no. The classification is done by using Euclidean distance by calculating the distance between the training data and testing data. The nearest k (positive integer) neighbors are selected depending on the similarities between the testing and training data. The neighbors which associated with labels are taken as reference and the testing data is associated to class which has high majority of the votes amongst the k nearest neighbors.

4.6. Pictorial representation

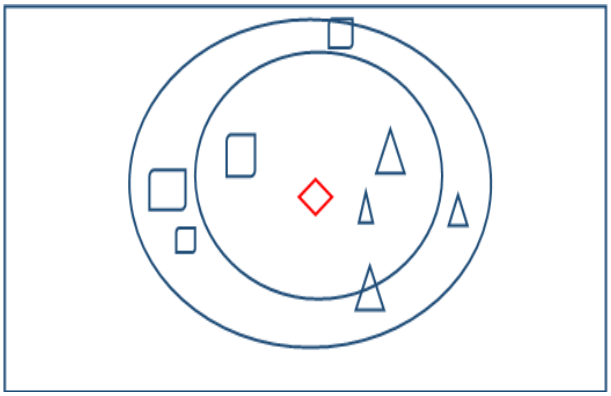


Fig.2: Pictorial Representation.

From the above representation, we have to determine the class of unknown instance by checking the majority of neighbors belong to and assign the class to the training data.
Algorithm:
Determining the k value.
Calculate the distance between training set data and testing set data using Euclidean distance.
Based on the distances sort the neighbors in ascending order.
From the sorted list select first k neighbors.
Assigning the class to the training data based on majority of neighbors belongs to.

4.7. Naive bayes classification

Bayes classification is one of the classification techniques which are based on Bayes Conditional Probability.It is applicable for data sets of large type which is discrete.This algorithm assumes each attribute to be independent to simplify the computations involved, as we consider each feature as independent so this algorithm is considered as naïve.

4.8. Bayes rule

Bayes theorem used to compute posterior probability p(c/x) from the given data, which is considered as prior probability p(c) and p(x/c).The formula used to calculate

P (c/x) = (p(x/c) * p(c))/p(x)

The algorithm mentioned below works as follows:
Consider D as training data and X=(x1, x2…, xn-1, xn) be an dimensional attribute vector which represents each instance.
Let C1, C2… Cm be the classes for prediction. Naive Bayes classifier predicts that tuple belongs to the class Ci if and only if P (Ci|X) >P (Cj|X) .Thus we maximize P(Ci|X) and the class Ci for which P(Ci|X) is maximized is defined as the maximum posteriori hypothesis.
From Bayes theorem

$$p\left(\frac{c}{x}\right) = \frac{p(x/c)*p(c)}{p(x)}$$

4.9. Neural networks

Neural networks are computational model similar to biological neural networks which resembles human neural networks. It comprises interconnected artificial neurons and by using a connectionist approach processes information. By using neural network it will assign weight to each attribute. As a result members will be shortlisted based on the final output weight.

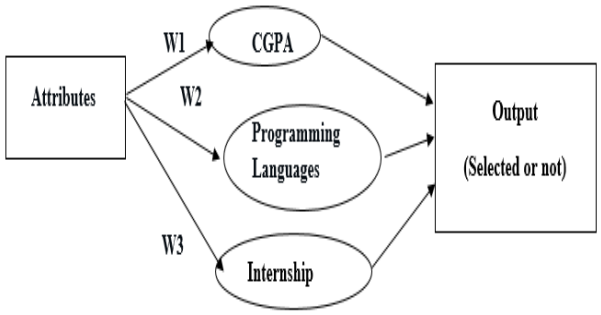


Fig. 3: Neural Network.

5. Data description

The attributes considered the classification to predict placement.

Table 1: Attributes Table

Attribute	Attribute_Range	Attribute_type
Gender	0,1	Numeric and discrete
Tenth percentage	0-100	Numeric and discrete
Inter Percentage	0-100	Numeric and continuous
Backlogs	0,1	Numeric and discrete
CGPA	0-10	Numeric and continuous
Programming Languages	0-10	Numeric and continuous
Clubs	0,1	Numeric and discrete
Internship	0,1	Numeric and discrete
Communication Skills	0-10	Numeric and continuous
Technical_Skills	0-10	Numeric and continuous
Team Work	0-10	Numeric and continuous
Achievements	0,1	Numeric and discrete

6. Results and discussions

In this paper we have conducted classification on student resume data set and compared the analysis of different data set using Weka tool and determined the best classification algorithm based on the below table mentioned. The data set we considered consists of 12 attributes. The characteristics of attributes are of discrete valued type, continuous valued type and numeric. The following table represents the algorithm, how many instances it classified correct and classified wrong, kappa statistics, mean absolute error and root mean squared error. We have implemented this data set using four algorithms those are Naïve Bayes, Decision Trees, Neural Network and K-nearest Neighbor using Weka -3.4.9 tool. The results are discussed in the table mentioned below.

Table 2: Comparison of Classification Algorithms

Data Mining Algorithms	Percentage of Correctly Classified Instances	Percentage of Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean square error
Naïve Bayes	93.0693 %	6.907 %	0.861	0.1026	0.2487
Decision Trees	94.0594%	5.9406 %	0.8812	0.1476	0.3159
Neural Network	98.1098 %	1.9802 %	0.9604	0.0315	0.01296
K-Nearest Neighbor	100 %	0 %	1	0.0097	0.0097

These algorithms are also compared with time taken to classify the student resume data Naïve Bayes classifier took 0.12 seconds to classify the data, Decision trees took 0.23 seconds to classify the data, Neural Network took 0.19 seconds to classify the data, K-Nearest Neighbor took 0.14 seconds to classify the data, which is mentioned in the below figure 2.The time taken by these algorithms are plotted which can be referred from figure 2. It can infer that OneR took less time compared with other algorithms and Decision Table took more time to classify the taken data.

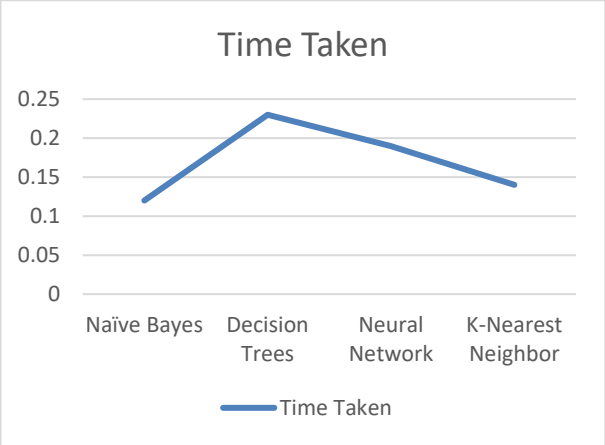


Fig. 3: Graphical Representation of Time Taken by Algorithms to Classify the Data.

7. Conclusion

As a conclusion we have analysed the student resume data and predicted the probability of getting placed into a company using classification techniques. We have also compared the results of the classification algorithms using weka-3.4.9 tool. As a result we got to know which students are eligible for placements. According to our analysis K-Nearest Neighbor classified instances with less error rate compared to other algorithms. Naive Bayes algorithm took less time to analyse the data where Decision trees took more time to analyse the data.

References

[1] Ashok M V Apoorva A. "Data Mining Approach For Predicting Student and Institution placement Percentage".International Conference on Computational Systems and Information Systems for Sustainable Solutions 2016.

[2] Mansi Gera, Shivani Goel."A Model for Predicting the Eligibility for Placement of Students Using Data Mining Technique". International Conference on Computing, Communication and Automation (ICCCA2015).

[3] Pratiyush Guleria, Manu Sood."Predicting Student Placements Using Bayesian Classification" Third International Conference on Image Information Processing 2015.

[4] Mangasuli Sheetal B, Prof. Savita Bakare."Prediction of Campus Placement Using Data Mining Algorithm-Fuzzy logic and K nearest neighbor". International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 6, June 2016.

[5] Animesh Giri, M Vignesh V Bhagavath, Bysani Pruthvi, Naini Dubey."A Placement Prediction System Using K-Nearest Neighbors Classifier". Second International Conference on Cognitive Computing and Information Processing (CCIP) 2016.

[6] Karan Pruthi, Dr. Parteek Bhatia. "Application of Data Mining in Predicting Placement of Students". International Conference on Green Computing and Internet of Things (leGCloT) 2015.

[7] Oktariani Nurul Pratiwi. "Predicting Student Placement Class using ratiwi Data Mining". IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE) 2013.

[8] B.K. Bharadwaj and S. Pal. Data Mining"A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (UCSIS), Vol. 9, No. 4, pp. 136-140,2011.

[9] AI-Radaideh,Q. A., AI-Shawakfa, E.M., and AI-Najjar, M. I., "Mining Student Data using Decision Trees", International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.

[10] 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.

[11] S. Elayidom, S. Mary Idikkula &Alexander, "A Generalized Data mining Framework for Placement Chance Prediction Problems, "International Journal of Computer Applications, (0975– 8887) vol.31, no. 3, October 2011.

[12] Sudheep Elayidom, Sumam Mary Idikkula and Joseph Alexander. "Article:A Generalized Data mining Framework for Placement Chance Prediction Problems". International Journal of Computer Applications 31(3):40-47, October 2011.

[13] Kabakchieva, Dorina. "Predicting student performance by using data mining methods for classification." Cybernetics and Information Technologies 13, no. 1 (2013): 61-72.

[14] Kevin P. Murphy, "Naive Bayes classifiers", University of British Columbia, 2006.

[15] Elizabeth Murray, "Using Decision Trees to Understand Student Data, Proceedings of the 22nd International Conference on Machine Learning", Bonn, Germany, 2005.

[16] Rakesh Kumar Arora, Dr.Dhannendra Badal. "Placement prediction through data mining, International Journal of Advanced Research in Computer Science and Software engineering". Volume 4, Issue 7, July2014.

[17] Weka Tutorials, http://www.technologyforge.netlWekaTutorials.

[18] A. K. Pal and S. Pal "Classification Model of Prediction for placement of Students", Modem Education and Computer Science Volume II, pp. 49-56, November 2013.

[19] Siddhi Parekh, Ankit Parekh, Ameya Nadkarni, Riya Mehta. "Results and Placement Analysis and Prediction using Data Mining and Dashboard".