A Generalized Data mining Framework for Placement Chance Prediction Problems

Sudheep Elayidom
Associate Professor, CUSAT
Kochi, 682022, India

Sumam Mary Idikkula
Professor, CUSAT, Kochi,
682022, India

Joseph Alexander
Project officer,
NODAL Center
CUSAT, Kochi, 682022, India

ABSTRACT
Data Mining is such a promising technology whose worth becomes evident when it can be applied to a domain where a common man is benefited. This paper is an 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. A student enters his Entrance Rank, Gender (M/F), Sector (rural/urban) and Reservation category. Based on the entered information the Network or the decision tree will return which branch of study is Excellent, Good, Average or poor for him/her. Also in this paper we compare the performance of the models on the same data and propose a generalized data mining framework for problems of similar nature.

General terms
Data mining, classification, framework, performance comparison

Keywords
Confusion matrix, Data mining, Decision trees, neural networks, Placement chance prediction.

1. INTRODUCTION
Majority of students join a course in engineering for securing a good job. Therefore taking a wise career decision regarding the selection of a particular course or branch is crucial in a student's life. An educational institution contains a large number of student records. Therefore finding patterns and characteristics in this large amount of data is a difficult task. We apply data mining techniques using neural network, Decision tree and Naïve Bayes classifier to interpret potential and useful knowledge.

With the help of this knowledge a student enters his/her rank, branch, location etc. and on the basis of which the placement chances for different streams of study are calculated. Now a student on the basis of this inference may decide to opt for branch giving excellent chances of placement.

There are some algorithms for extracting comprehensible representations from neural networks. [1] Describes research to generalize and extend the capabilities of these algorithms. The application of the data mining technology based on neural network is vast. One such area of application is in the design of mechanical structure.[2] introduces one such application of the data mining based on neural network to analyze the effects of structural technological parameters on stress in the weld region of the shield engine rotor in a submarine. Prediction of Beta-Turns using global adaptive techniques from multiple alignments in Neural Networks has been studied in [3] in study of proteins. This also introduces global adaptive techniques like Conjugate gradient method, Preconditioned Conjugate gradient method etc. This paper is an attempt that uses the neural network based on back propagation training for placement prediction which uses the above said concepts with more application in the domain of data mining.

Decision trees have proved to be valuable tools for the description, classification and generalization of data. Work on constructing decision trees from data exists in multiple disciplines such as statistics, pattern recognition, decision theory, signal processing, machine learning and artificial neural networks. [5] Surveys existing work on decision tree construction, attempting to identify the important issues involved, directions the work has taken and the current state of the art. Studies have been conducted in similar area such as understanding student data as in [6]. There they apply and evaluate a decision tree algorithm to university records, producing graphs that are useful both for predicting graduation, and finding factors that lead to graduation. It's always been an active debate over which engineering branch is in demand .So this work gives a scientific solution to answer these. Article [7] 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. [8] Suggests methods to classify objects or predict outcomes by selecting from a large number of variables the most important ones in determining the outcome variable. [10] & [11] are publications by the same author describing the data preprocessing using decision trees and regression. The method in [9] is used for performance evaluation of the system using confusion matrix which contains information about actual and predicted classifications done by a classification system.

1.1 Problem Statement

To choose a suitable career for a bright future is a common need for any student. So a data mining model has to be built which can predict the most suited branch for a student who supplies his information. The problem also includes deciding the important attributes that decides the placement chance. Various data mining models are to be trained and tested for this problem. Their performances are to be compared based on statistical measures. Also a generalized design framework for a typical placement chance prediction problem has to be formulated.

2. DATA MINING BASED ON DECISION TREES

A decision tree is a popular classification method that results in a tree like structure where each node denotes a test on an attribute value and each branch represents an outcome of the test. The tree leaves represent the classes. The decision tree can be modeled using WEKA package, but since the NODEL CENTRE authorities are interested to implement a web site it have been decided to use conventional Web technologies like mySQL, php etc itself to model a decision tree. Given a set of examples (training data) described by some set of attributes (ex. Sex, rank, background) the goal of the algorithm is to learn the decision function stored in the data and then use it to classify new inputs. Let's work through an example:

TABLE 1: SAMPLE PARTIAL DATASET TO CONSTRUCT DECISION TREE

BRANCH	SECTOR	M/F	RANK	CHANCE
CS	RURAL	M	1 to 199	E
CS	URBAN	F	200 to 399	A
EC ..continues	RURAL	M	400 to 599	G

2.1 Data Pre-processing

The Initial database provided by Nodal Center, was in FoxBase format, converted to some latest DBMS like mySQL to make the approach efficient and faster. First FoxBase data was converted to CSV files (Comma Separated files) and this file was loaded to MS Excel. Then from this Excel format using xls-mySQL converter it was converted to mySQL format.

These attributes were fed into mysql through sql queries and each of these entities and two databases, one containing records of students from the year 2000-2002 and another for year 2003, were created.

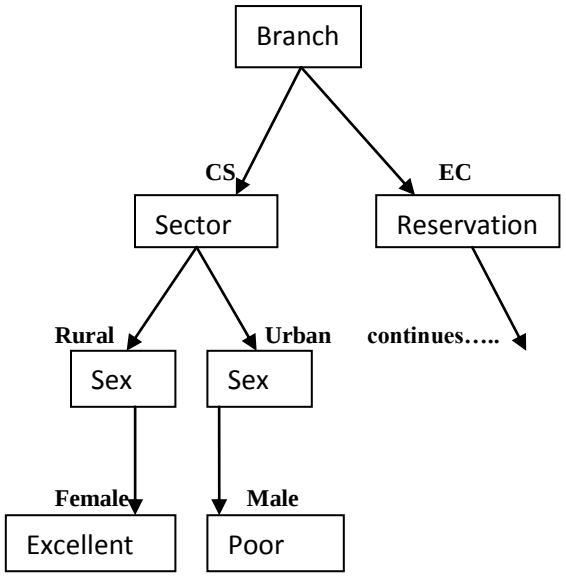

Fig. 1: A Partial view of Decision Tree for our domain

List of attributes extracted:

RANK: Rank secured by candidate in the engineering entrance Range: 1-3000
CATEGORY: Social background. Range: {General, Scheduled Cast, Scheduled Tribe, OBC}
SEX: Range{Male, Female}
SECTOR: Range{Urban, Rural}
BRANCH: Range{A-J}
ACTIVITY: Indicator of whether the candidate is placed.

All these attributes have been found to be deciding the placement chance which has been analyzed using chi-square based statistical dependency analysis.

Information Gain

A decision tree can be constructed top-down using the information gain in the following way:

1. Let the set of training data be S. Since the attribute rank is continuous valued it is discretized in intervals of nearly 200(1-199,200-399 etc). Once this is done put all of S in a single tree node.

2. If all instances in S are in same class, then stop

3. Split the next node by selecting an attribute A, for which there is maximum information gain

4. Split the node according to the values of A

5. Stop if either of the following conditions is met, otherwise continues with step 3:

(a) If this partition divides the data into subsets that be long to a single class and no other node needs splitting.

(b) If there are no remaining attributes on which the sample may be further divided.

Intuitively, the attribute which will yield the most information should become our first decision node. Then, the attribute is removed and this process is repeated recursively until all examples are classified.

2.2 Modeling

From the history data (2000-2002), a new table is created in which the placement chance for each possible input combination is stored. For e.g, if RANK (1-200) SECTOR (U) SEX (M) CATEGORY (GEN), we compute how much percentage of students having these criteria are placed in history database. If this percentage is greater than 95%, it can be called "Excellent" chance. Similarly for other grades like average, poor etc. there are other ranges which are described in neural networks modeling section. An

intermediate dataset that looks like in table 1 is prepared from which the decision tree is constructed using the decision tree construction algorithm. This decision tree is physically stored as an adjacency list in table 2. The data of year 2003 is used for testing.

Using a set of recursive queries Information gain is calculated over all attributes. The attribute with the maximum Gain is chosen. For our case, if SEX is found to be the attribute with max gain, it is added to the Adjacency list and the database is split into unique set of records with common values for sex. This process is recursively repeated for all cases. For all iteration we append the attribute field to the adjacency list. One key point here is that the attribute field names, the results etc are all treated as nodes and the list can identify the node only by its corresponding TYPE. The ID field stores the unique id number of the node, while the parent stores the id of the parent of the node.

Here Adjacency List is an ideal mechanism to store the decision tree. The Adjacency List Model is an elegant approach and needs just one, simple function to iterate through a Decision Tree.

For the decision tree in Figure 1, the table for an adjacency list would look like as in table 2.

TABLE 2: Adjacency List Partial View

ID	NODE	TYPE	PARENT
2	CS	Branch	0
3	R	Sector	2
4	M	Sex	3
5	1	Rank	4
6	E	Chance	5..continues

2.3 Retrieval from Adjacency list

The User enters his search criteria in the user interface screen with details, from which a query string is constructed, which are parsed to get individual attributes which are used to search the decision tree

If a query needs to be made, for example:

What is the chance for BRANCH (A) SECTOR(R) SEX (F) CATEGORY (GEN)?

The query proceeds from ID 2 as in the adjacency list in table 2 and the algorithm searches for all the nodes which have ID 2 as parent. The algorithm will find ID =3 and ID=7 as child nodes. It then finds that ID=3 is the right path which needs to be taken to arrive at the result. This process continues and the Chance value is found at ID=6 as E (Excellent)

3. DATA MINING BASED ON NEURAL NETWORKS

Neural Network has the ability to realize pattern recognition and derive meaning from complicated or imprecise data that are too complex to be noticed by either humans or other computer techniques. The data mining based on neural networks can generally be divided into 3 stages: data preparation, modeling and knowledge discovery as shown in fig 2.

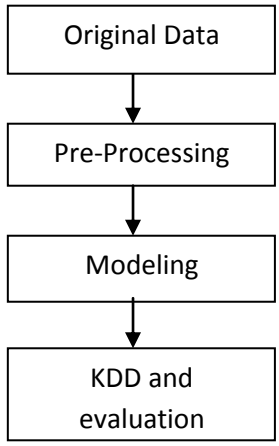

Figure 2: Stages in data mining

3.1 Data preparation

Data preparation is an important step in which the mined data is made suitable for processing. This involves cleaning data, data transformations, selecting subsets of records etc. Data selection means selecting data which are useful for the data mining purpose. Data transformation or data expression is the process of converting the data into required format which is acceptable by data mining system

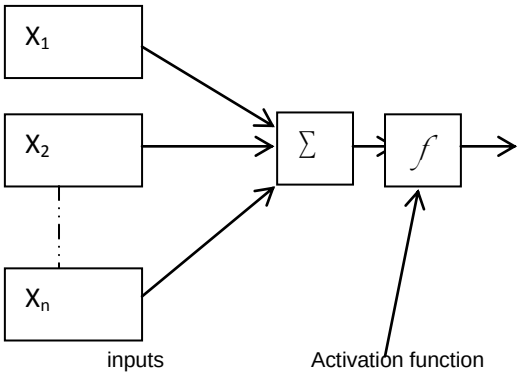

Fig. 3 Architecture of Neural Network

TABLE 3: ATTRIBUTES VALUE MAPPED TO 0/1

ATTRIBUTE	RANGE	MAPPED TO
RANK	1 to 4000	0 to 1
SEX	1 to 2	0 to 1
CATEGORY	1 to 4	0 to 1
SECTOR	1 to 2	0 to 1
BRANCH	A to J	0 to 1
ACTIVITY	1 to 2	One of the four values 'E', 'G', 'A' and 'P'.

TABLE 4: SAMPLE OF DATA USED AS INPUT TO TRAIN THE NETWORK

SEX	RESERVATION	LOCATION	RANK	BRANCH
0	0	1	0.72	0.47
1	0	1	0.72	0.47
0	1	0	0.59	0.33
0	0	1	0.4	0.66
0	1	0	0.72	0.47
1	1	1	0.27	0.38

Tables 3 and table 4 shows sample mapping to 0/1 scale and data used to train neural net. The output data used for training is derived from ACTIVITY attribute. Instead of representing the output on 0 to 1 scale basis, we have used four fold classifications that are; a 4 value code has been assigned with each record. The data processing is very similar to that used for decision trees. But since neural networks need numeric inputs slight modifications were done.

A code value of 1000 represents a 'excellent' chances of getting a student placed, a code value of 0100, 0010, 0001 represents 'good', 'average' and 'Poor' chances of placement of a student respectively these codes are calculated by following the steps:

Step 1:

Calculate the probability of each test case for getting a student placed. It is calculated as

Probability (P) = Number Placed/ Total Number

Where 'Number Placed' is the number of students placed in a particular class of inputs and 'Total Number' is the total number of students in that

class. For example let a class of records has: RANK = 0.72, SEX = 1, CATEGORY = 1, SECTOR = 0, BRANCH = 0.47.
Let the number of records be 98 and the number of records having ACTIVITY as 1 (i.e. student is placed) be 79, then probability is given by P = 79/98 = 0.80.

Step 2: Assign the output code to each record by using table 5.

TABLE 5: Assigning Output Codes to Records

Range of Probability	**Output Code**	**Chances of Placement**
P >= 0.95	1000	Excellent
0.75 <= P < 0.95	0100	Good
0.50 <= P < 0.75	0010	Average
P < 0.50	0001	Poor

3.2 Modelling

This is the most important step in the data mining. A proper selection of algorithm is made on the basis of the required objective of the work. MATLAB was used to model the neural network.

A back propagation neural network model is used consisting of three layers namely input, hidden and output layer. Structure of a typical neural network is shown in figure 3. The number of input neurons is 5, which depends upon the number of the input attributes. The number of neurons used in hidden layer is 5; this number is obtained by value based on observations. The transfer function used in the Hidden layer is Log-Sigmoid while that in the output layer is Pure Linear. Training is done by using one of the Back propagation Conjugate Gradient algorithm, Powell-Beale Restarts [4]. This algorithm provides faster convergence in comparison to conventional basic Back propagation algorithm by performing a search along the conjugate direction to determine the step size which minimizes the performance function along that line.

4. DATA MINING BASED ON NAÏVE BAYES CLASSIFIER

Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. The classifier is based on Bayes theorem, which is stated as:

P (A|B) =P (B|A)*P (A)/P (B)

Each term in Bayes' theorem has a conventional name:

*P (*A*) is the prior probability or marginal probability of *A*. It is "prior" in the sense that it does not take into account any information about *B*.

*P (A|B) is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B.

*P (B|A) is the conditional probability of B given A.

*P (B) is the prior or marginal probability of B, and acts as a normalizing constant.

5. TESTING

Testing was conducted separately for the projects based on the neural network, decision tree and Naïve Bayes classifies models. Accuracy, confusion matrix, and all performance parameters were separately computed.

We used the same test data set for the above three data mining technique. The results of the test are modeled as a confusion matrix. The negative cases here are when the prediction was Poor /average and the corresponding observed values were Excellent/good and vice versa. A performance comparison is shown in table 6, which shows that the accuracies of the data mining models are comparable with each other. The difference in accuracies is statistically comparable.

But as it is well known Naïve Bayes classifier is the fastest algorithm among all and it yielded best True positive and precision measures too.

6. A GENERALIZED DATA MINING FRAMEWORK

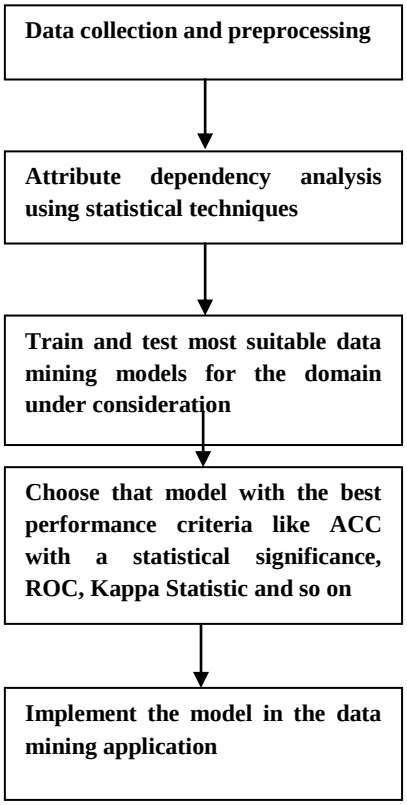

Figure 4. A general data mining framework

As shown in figure we propose a generalized design framework for a placement chance prediction data mining problem. Here history data is collected and data cleaning is done. After that attribute dependency analysis using chi square statistical analysis is done to clearly understand which all attributes have importance in deciding the placement chance. Once this is properly understood attribute reduction and porting to model dependent data formats takes place. For Weka based data mining projects we have to port it to arff based formats, for MATLAB based projects we have to port it to CSV/XLS formats and so on. Then the data mining models are built using training data and are tested using test data to decide which model performs in the best way for this domain. For those comparisons various statistical measures like ROC area, Kappa statistic, True positive, accuracy etc may be used. Once the best model is chosen that model can be implemented for actual predictions for future data.

7. COMPARISON

TABLE 6: COMPARISON OF DIFFERENT DATA MINING TECHNIQUES

No	Model	Accuracy	Precision	TP
1	Decision tree	80.30%	0.81	0.84
2	Neural Network	79.70%	0.8	0.84
3	Naïve Bayes	79.40%	0.84	0.9

From the above table it is clear that decision trees are having a slight advantage in terms of accuracy when compared to other models, but it can be statistically verified that the accuracy of other models are also comparable as the difference is insignificant. So as the data mining researchers say one cannot find the best model that can be universally accepted, but only one can suggest the best model suited for a domain. In this case the three models have given comparable performances based on accuracy. But depending on the platform used for implementation and other performance characteristics one can choose any one model from one of these.

8. CONCLUSION AND FUTURE SCOPE

Any technological research becomes useful when it helps the mankind for a better living. In this work it has been 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. Also the work can be extended to test in different domains and different type of data of different disciplines such as medicine, law and so on.

9. ACKNOWLEDGMENTS

I wish to express my gratitude to the following personnel for their support in documentation and implementation namely Rajeev Ranjan, Rajeev Kumar Singh and Pawan Kumar Thakur.

10. REFERENCES

[1] Antony Browne, Brian D. Hudsonb, David C. Whitley, Martyn G. Ford, Philip Picton, Biological data mining with neural networks : implementation and application of a flexible decision tree extraction algorithm to genomic problem domains, Elsevier, October 2003.

[2] L. Wang, T. Z. Sui, Application of Data Mining Technology Based on Neural Network in the Engineering, IEEE 1-4244-1312-5/07.

[3] Zarita Zainuddin, Chan Siow Cheng, Lye Weng Kit, Prediction of B-Turns Using Global Adaptive Techniques from Multiple Alignments in Neural Networks. Malaysian Journal of Mathematical Sciences 185-194(2008).

[4] Yoav Freund, Robert E. Schapire, Large Margin Classification using Perceptron Algorithm. Machine Learning 37(3):277-296, 1999.

[5] Sreerama K. Murthy, Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, Data Mining and Knowledge Discovery, 345-389 1998.

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

[7]] U Fayyad, R Uthurusamy, From Data Mining to Knowledge Discovery in Databases, 1996.

[8] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees, Chapter 3,Wadsworth Inc., 1984.

[9] Kohavi R. and F. Provost, Editorial for the Special Issue on application of machine learning and the knowledge of discovery process, Machine Learning 30, 271-274, 1998.

[10] SudheepElayidom.M, Sumam Mary Idikkula, Joseph Alexander, "Applying Data mining techniques for placement chance prediction". Proceedings of ACT , India. , 2010.

[11] SudheepElayidom.M, Sumam Mary Idikkula, Joseph Alexander, "Comparison of data mining techniques using decision trees and neural nets for placement chance prediction". Proceedings of ICONCEPT, India, 2010.