Analysis And Implementation Of Data Mining Techniques , Using Naive-Bayes Classifier And Neural Networks

GJCST Classification
H.2.8 I.2.6

[1]Sudheep Elayidom.M, [2]Sumam Mary Idikkula, [3]Joseph Alexander

***Abstract*--Taking wise career decision is so crucial for anybody for sure. In modern days there are excellent decision support tools like data mining tools for the people to make right decisions. This paper is an attempt to help the prospective students to make wise career decisions using technologies like data mining. In India technical manpower analysis is carried out by an organization named NTMIS (National Technical Manpower Information System), established in 1983-84 by India's Ministry of Education & Culture. The NTMIS comprises of a lead centre in the IAMR, New Delhi, and 21 nodal centres, located at different parts of the country. The Kerala State Nodal Centre is located in the Cochin University of Science and Technology. Last 4 years information is obtained from the NODAL Centre of Kerala State (located in CUSAT, Kochi, India), which stores records of all students passing out from various technical colleges in Kerala State, by sending postal questionnaire. Analysis is done based on Entrance Rank, Branch, Gender (M/F), Sector (rural/urban) and Reservation (OBC/SC/ST/GEN). Using this data, data mining models like Naïve Bayes classifier and neural networks are built, tested and used to predict placement chances, given inputs like rank, sector, category and sex.**

Keywords- Data mining, WEKA, Neural networks, Naïve Bayes classifier, Confusion matrix.

I. Introduction

The popularity of subjects in science and engineering in colleges around the world is up to a large extent dependent on the viability of securing a job in the corresponding field of study. Appropriation of funding of students from various sections of society is a major decision making hurdle particularly in the developing countries.

An educational institution contains a large number of student records. This data is a wealth of information, but is too large for any one person to understand in its entirety. Finding patterns and characteristics in this data is an essential task in education research.This type of data is presented to decision makers in the State Government in the form of tables or charts, and without any substantive analysis, most analysis of the data is done according to individual intuition, or is interpreted based on prior research. This paper is an attempt to scientifically analyze the trends of placements keeping in account of details like Entrance Rank, Sex, Category and Reservation using Naive Bayes classifier and Neural Networks.

The data preprocessing for this problem has been described in detail in articles [1] & [9], which are papers published by the same authors. The problem of placement chance prediction may be implemented using decision trees. [4] Surveys a work on decision tree construction, attempting to identify the important issues involved, directions which 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 [2]. 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 [3] 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. [5] 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. The method in [6] is used for performance evaluation of the system using confusion matrix which contains information about actual and predicted classifications done by a classification system. [7] & [8] suggest further improvements in obtaining the various measures of evaluation of the classification model.

II. Data

The data used in this project is the data supplied by National Technical Manpower Information System (NTMIS) via Nodal center. Data is compiled by them from feedback by graduates, post graduates, diploma holders in engineering from various engineering colleges and polytechnics located within the state during the year 2000-2003. This survey of technical manpower information was originally done by the Board of Apprenticeship Training (BOAT) for various individual establishments. A prediction model is prepared from data during the year 2000-2002 and tested with data from the year 2003.

Sudheep Elayidom
Computer Science and Engineering Division, School Of Engineering
Cochin University of Science and Technology, Kochi, India
+91-04842463306,sudheepelayidom@hotmail.com
Sumam Mary Idikkula
Department of Computer Science
Cochin University of Science and Technology, Kochi, India
+91-04842577605, sumam@cusat.ac.in
Joseph Alexander
Project Officer, Nodel Center
Cochin University of Science and Technology, Kochi, India
+91-04842862406, josephalexander@cusat.ac.in

III. PROBLEM STATEMENT

Modeling and predicting the chances of placements in colleges keeping account of details like Rank, Gender, Branch, Category, Reservation and Sector. It also analyses and compares the performances of Naive Bayes classifier and neural networks for this problem.

IV. CONCEPTS USED

A. *Data Mining*

Data mining is the principle of searching through large amounts of data and picking out interesting patterns. It is usually used by business intelligence organizations, and financial analysts, but it is increasingly used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods. A typical example for a data mining scenario may be ―Ina mining analysis if it is observed that people who buy butter tend to buy bread too then for better business results the seller can place butter and bread together.”

B. *Naive Bayes classifier*

In simple terms, a 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.

Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for Naive Bayes model uses the method of maximum likelihood. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Recently, careful analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of naive Bayes classifiers.

An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

The classifier is based on Bayes theorem, which is stated as

$$P\,(A|B) = \frac{P\,(B|A)P(A)}{P\,(B)}$$

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.

Bayes' theorem in this form gives a mathematical representation of how the conditional probability of event A given B is related to the converse conditional probability of B given A.

C. *Weka*

WEKA is a collection of machine learning algorithms for data mining tasks. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

The main strengths of WEKA are that it is

- Very portable because it is fully implemented in the Java programming language and thus runs on almost any modern computing platform,

-contains a comprehensive collection of data pre-processing and modelling techniques, and is easy to use by a novice due to the graphical user interfaces it contains.

WEKA's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. There is also the Experimenter, which allows the systematic comparison of the predictive performance of WEKA's machine learning algorithms on a collection of datasets.

D. *Confusion Matrix*

A confusion matrix is a visualization tool typically used in supervised learning (in unsupervised learning it is typically called a matching matrix). It is used to represent the test result of a prediction model. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. One benefit of a confusion matrix is that it is easy to see if the system is confusing two classes (i.e. commonly mislabelling one as another).When a data set is unbalanced (when the number of samples in different classes vary greatly) the error rate of a classifier is not representative of the true performance of the classifier. This can easily be understood by an example: If there are for example 990 samples from class A and only 10 samples from class B, the classifier can easily be biased towards class A. If the classifier classifies all the samples as class A, the accuracy will be 199%. This is not a good indication of the classifier's true performance. The classifier has a 100% recognition rate for class A but a 0% recognition rate for class B.

E. *Data Pre-Processing*

The Initial database provided by Nodal Center was loaded to MS Excel and converted to CSV files (Comma Separated files) .This file was loaded in WEKA Knowledge Flow interface and converted into ARFF files (Attribute-Relation File Format). The individual database files(DBF format) for the years 2000-2003 were obtained and one containing records of students from the year 2000-2002 and another for year 2003, were created.

List of attributes extracted:

RANK: Rank secured by candidate in the engineering entrance exam. Range: 1-25

CATEGORY: Social background. Range: {General, Scheduled Cast, Scheduled Tribe, Other Backward Class}

SEX : Range{Male, Female}

SECTOR : Range {Urban, Rural}

BRANCH : Range{A-J}

PLACEMENT: Indicator of whether the candidate is placed.

Data from the year 200-2002 are used to model and that from the year 2003 will be used to evaluate the performance of the model.

V. IMPLEMENTATION LOGIC

A. Data Preparation

The implementation begins by extracting the attributes RANK, SEX, CATEGORY, SECTOR, and BRANCH from the master database for the year 2000-2003 at the NODAL Centre. The database was not intended to be used for any purpose other maintaining records of students. Hence there were several inconsistencies in the database structure. By effective pruning the database was cleaned. A new table is created which reduces individual ranks to classes and makes the number of cases limited. All queries will belong to a fix set of known cases like:

RANK (1) SECTOR (U) SEX (M)
CATEGORY (GEN) BRANCH (A)

For every combination of these attributes, we calculate the corresponding placement chances from the history data by calculating probability of placement and storing in a separate data sheet which is used to build the model.

Probability (P) = Number Placed/ Total Number for this particular input combination

The chance is obtained by the following rules:

If P>=95 Chance='E'
If P>=75 && P<95 Chance='G'
If P>=50 && P<75 Chance='A';
Else Chance='P'; Where E, G, A, P stand for Excellent, Good, Average & Poor respectively.

B. Conversion For The Naive Bayes Classifier

The Explorer interface of WEKA has several panels that give access to the main components of the workbench. The Pre-process panel has facilities for importing data from a database, a CSV file, etc., and for pre-processing this data using a so-called filtering algorithm. These filters can be used to transform the data (e.g., turning numeric attributes into discrete ones) and make it possible to delete instances and attributes according to specific criteria. The Classify panel enables the user to apply classification and regression algorithms
(indiscriminately called classifiers in WEKA) to the resulting dataset, to estimate the accuracy of the resulting predictive model .The panel —select attributes” provides algorithms for identifying the most predictive attributes in a dataset.

WEKA input must be in ARFF format:

```
A relation name
– E.g., @relation student
A list of attribute definitions
– E.g., @attribute RANK numeric
     @attribute SECTOR {U, R}
     @attribute SEX {F, M}
     @attribute CATEGORY {GEN, OBC, SC, ST}
 @attribute BRANCH {A, B, C, D, E, F, G, H, I, J}
@attribute CHANCE {E, P, A, G}
– The last attribute is the class to be predicted
A list of data elements
@data
1, U, F, GEN, D, E
```

VI. PRINCIPLE OF DATA MINING BASED ON NEURAL NETWORK

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.

A. 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. For ex: Symbolic data types are converted into numerical form understood by system. Also the data is scaled onto 0 to 1 or -1 to 1 scale which is acceptable by Neural Networks. Some sample mappings are shown in table I.

TABLE I. ATTRIBUTE VALUES MAPPED TO 0 TO 1 SCALE

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 II shows a snippet of sample data used to train neural network.

TABLE II. 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

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 classification that is, a 4 value code has been assigned with each record. 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.

The process of computing the placement chances of each possible attribute combination is same as that of the naïve Bayes classifier and its final assignment to codes are shown in table III.

MATLAB is used to model the neural network and scripts in MATLAB were used to model and test the neural network. The same work may be done using WEKA, but MATLAB gives more options and programmability for a neural network.

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

Table iii. Assigning Output Codes To Records

B. 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.

One of the most popular Neural Network model is Perceptron, but this model is limited to Classification of Linearly Separable vectors[4]. The input data which is obtained from NTMIS, may contain variations resulting in Non-Linear data. For example a student with good rank may opt for a weak branch and still may have a placement. To deal with such inputs data cleaning alone is not sufficient. Therefore we go for multilayer perceptron network with supervised learning which gives back propagation Neural Network. A BP neural network reduces the error by propagating the error back to the network. Appropriate model of BP neural network is selected and repeatedly trained with the input data until the error reduces to a fairly low value. At the end of training we get a set of thresholds and weights which determines the architecture of the neural network. A back propagation neural network model is used consisting of three layers: input, hidden and output layers as shown in fig. 1. 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. A small number may not be able to solve a given problem and a large number of neurons increases the computation required. The number of neurons in output layer should equal the number of output code. Here we are performing 4 fold classifications, therefore four neurons are required at output layer. The total no. of records used for training is 4091 and total no. of records used for testing is 1063.

Transfer Functions- The transfer function used in the Hidden layer is Log- Sigmoid while that in the output layer is Pure Linear.

Learning- Training is done by using one of the Conjugate Gradient algorithm, Powell-Beale Restarts. This algorithm provides faster convergence by performing a search along the conjugate direction to determine the step size which minimizes the performance function along that line.

VII. TEST RESULTS

A. Naive Bayes Classifieer

Table Iv. Confusion Matrix (Student Data)

Confusion Matrix

		PREDICTED			
		E	P	A	G
ACTUAL	E	496	10	13	0
	P	60	97	12	1
	A	30	18	248	0
	G	34	19	22	3

For training, records of 2000-2002 are used and for testing the records of year 2003 are used. The predictions of the model for typical inputs from test set, whose actual data are already available for test comparisons, were compared with predicted values.

The results of the test are modeled as a confusion matrix as shown in the above diagram, as its this matrix that is usually used to describe test results in data mining type of Research works.

The confusion matrix obtained for the test data is as shown in table IV and the accuracy is computed as below:

$$AC = 844/1063 = 0.7938$$

To obtain more accuracy measures, we club the field Excellent and Good as positive and Average and Poor as negative. In this case we got an accuracy of **83.0%**. The modified Confusion matrix obtained is as shown in table V.

		Predicted	
		Negative	Positive
Actual	Negative	**365**	**101**
	Positive	**57**	**540**

TP = 0.90 *FP* = 0.22
TN = 0.78 *FN* = 0.09

B. *Neural Networks*

For simulation/evaluation we use the data of year 2003 obtained from NTMIS. The knowledge discovered is expressed in the form of confusion matrix in table VI.

Since the negative cases here are when the prediction was Poor /average and the corresponding observed values were Excellent/good and vice versa.

Table Vi. Confusion Matrix (Student Data)

Confusion Matrix

		PREDICTED			
		P	A	G	E
ACTUAL	P	31	4	1	91
	A	5	410	2	9
	G	1	1	6	12
	E	72	13	4	401

Therefore the accuracy is given by

AC = 848/1063 = 0.797

To obtain more accuracy measures, we club the field Excellent and Good as positive and Average and Poor as negative. Then the observed accuracy was 82.1 %. The modified Confusion matrix obtained is as shown in table VII.

Table Vii. Modified Confusion Matrix (Student Data)

		Predicted	
		Negative	Positive
Actual	Negative	**450**	**103**
	Positive	**87**	**423**

TP = 0.83 *FP* = 0.17
TN = 0.81 *FN* = 0.17

From these test results, now a methodology has to be found by which we can compare the model performances on this particular domain. The following section explains the techniques that are followed for comparing these two models.

VIII. PERFOMANE COMPARISON WITH OTHER MODELS

The Kappa statistic values for neural networks and naive Bayes classifiers were found to be 0.56 and 0.69 respectively. The ROC values were 0.86 and 0.89 respectively.

For comparing model performances there exists a statistic which uses the classical hypothesis testing concept which may give a good measure on performance comparison,

$$P = \frac{|E1-E2|}{\sqrt{q(1-q)(2/n)}}$$

Where E1= error rate for model M1(Neural Network)
E2= Error rate for model M2 (Naive Bayes)
q= (E1+E2)/2
n=number of instances in test set

In our case E1=0.203, E2=0.206, n=1063,
q=0.2045

So applying these values in the formula P becomes 0.176, for the four variable output case.

According to classical hypothesis testing as **p<2** the difference in performance between the two models is not significant and is comparable or similar in their predictive capabilities with respect to this domain and test set.

IX. CONCLUSION

Choosing the right career is so important for any one's success. For that we may have to do a lot of history data analysis, experience based assessments etc. Nowadays technologies like data mining is there which uses concepts like Naïve Bayes prediction and neural networks, to make logical decisions. Hence this work is an attempt to demonstrate how technology can be used to take wise decisions for a prospective career. The methodology has been verified for its correctness and may be extended to cover any type of careers other than engineering branches. The work may be extended to analyze how data of other disciplines also may be modeled with these concepts.

X. REFERENCES

1) SudheepElayidom.M, Sumam Mary Idikkula, Joseph Alexander-Applying data mining using statistical techniques for career selection, IJRTE ,ISSN 1797-9617, volume 1, number 1, May 2009.
2) Elizabeth Murray, Using Decision Trees to Understand Student Data, Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005.
3) U Fayyad, R Uthurusamy - From Data Mining to Knowledge Discovery in Databases , 1996.
4) Sreerama K. Murthy, Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, Data Mining and Knowledge Discovery, 345-389 1998.
5) L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees, Chapter 3,Wadsworth Inc., 1984.
6) 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.
7) M. Kubat, S. Matwin, Addressing the Curse of Imbalanced Training Sets: One-Sided Selection, Proceedings of the 14th International Conference on Machine Learning, 179-186, ICML'97.
8) Lewis D. D. & Gale W. A., A sequential algorithm for training text classifiers,3-12, in SIGIR'94.

9) SudheepElayidom.M, Sumam Mary Idikkula, Joseph Alexander –—Applying Data mining techniques for placement chance prediction”. Proceedings of international conference on advances in computing, control and telecommunication technologies, India, December 2009.