Elysium Journal
Engineering Research & Management

P-ISSN: 2347-4408
E-ISSN: 2347-4734

EMPLOYMENT CHANCE PREDICTION BASED ON DECISION TREES AND NAIVE BAYES CLASSIFIER USING DATA MINING TECHNIQUES

Dr.B.Jagadhesan[1], P.Sarvanan[2], Dr.C.Pooranachandran[3]

[1]*P.G. Research and Dept of Computer Science, D.B.Jain college, Chennai, India.*
E-mail: bjagadhesan@gmail.com
[2]*P.G. Research and Dept of Computer Science, D.B.Jain college, Chennai, India.*
E-mail: npsindian@yahoo.co.in
[3]*HOD&Dept of Computer Science, Govt Arts College, Nandanam, Chennai, India.*
E-mail: cpooranachandran@hotmail.com

Abstract: Data Mining is the non-trivial extract information from a data set and transforms it into an understandable structure for further use. Data mining is the search for the relationship and global pattern that exist in large databases but are hidden among vast amount of data. Major strength of Decision trees are construction of decision tree classifiers doesn't require any account knowledge for exploratory knowledge discovery database, handle high dimensional(Lookup) data simple and fast, used for many applications such as medicine, manufacturing, financial analysis, astronomy etc and basis of several commercial rule induction systems. An another important model used in data mining for Naïve Bayesian classifiers assume that the effect of an column values on given class is self-determining of the values of the other columns. This is class conditional independence called. It's made to simplify the computations involved and, in this sense, is considered called "Naive". This paper is cover to help prospective student community to make wise career decisions using these data mining tools. A student enters his/her Entrance Rank, Gender (Male/Female), Area (rural/urban) and Reservation (GEN /MBC/SC/SCA/ST) category. Based on the entered information the model will return which Course of study is Outstanding, Excellent, Distinction, very good, good, average, satisfactory and Re-appear for him/her based on history data analysis using data mining techniques. Also in this paper we compare the performance of decision trees and Naive Bayes classifier on the same training and test data for this problem.

Keywords: Data mining, Decision Tree, Information Gain Theory Data Mining, Naive Bayes Classifier, Adjacency List and Prediction.

1. INTRODUCTION

An interested most of students join a course in Arts and Science College for securing a good job. Therefore taking a wise career decision regarding the selection of a particular course or department is crucial in a student's life. Higher educational institution contains a large amount (or) number of student records. Therefore finding patterns and characteristics in this large amount of data is a difficult task for social values. We apply data mining techniques using Naive Bayes classifier and decision tree to interpret potential and useful knowledge. With the help of this knowledge a student enters his/her rank, department, Sector etc. and on the basis of which the placement opportunity for various level of study are calculated. Now a student on the basis of this inference may decide to select for course excellent opportunity of placement.

The data preprocessing for this problem have been discuss in detail in [1], The problem of placement chance prediction can be implement using decision trees. [4] Survey 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 are conducted in similar area such as understand student data as in [2] & [10]. In [2] they apply and evaluate a decision tree algorithm to college records, producing graphs that are useful both for predict degree, and finding factors that lead to degree. In [10] they propose a way to understand student performances. But here the goal is to help the students to choose a good course to help them in a good career choice. It's always been an active debate over which college course is in demand .So this work gives a scientific solution to answer these. [3] give an overview of emerging field clarify how to data mining and knowledge discovery in databases are related both to each other and to related fields. [5] Suggests method to classify objects or predict outcomes by selecting from a large amount number of variables, the important one in determining the outcome variable. To test the naïve Bayes classifier software package WEKA has been used where as for decision tree concept it is implemented through PHP programming itself in a web site with a view that it can be later implemented for use by the public.

2. DATA

The data used in process is the data given by Placement Cell, D.B.Jain College, Chennai. Data is compiled by them from feedback by graduates, post graduates, from various Arts and Science colleges and located within the state during the year 2008-2010.

3. PROBLEM STATEMENT

The problem is to model and predict the placement chances for various courses in colleges keeping account of details like rank, sex, Reservation and Sector for a student who seeks admission to the various courses. For this data mining techniques like Naive

P-ISSN: 2347-4408
E-ISSN: 2347-4734

Bayes classifier and decision trees have to be used whose performances have to be compared.

4. CONCEPTS USED

4.1. DATA MINING

Data mining having six common classes of tasks. Anomaly detection (Outlier/Deviation detection), Association rule learning (Dependency modeling), Clustering, Classification, regression and summarization.

4.2. BAYESIAN CLASSIFICATION

Bayesian Classifier is statistical classifiers. It can predict class membership probability, such as the probability that a given row belongs to a particular class. Bayesian classification is based on Bayes theorem. It also exhibited high correct and speed when applied to large amount of databases.

Navie Bayesian Classifiers assume that the effect of column values on a given class is independent of the values of the other Column.

Bayesian Belief Networks are graphical models, which unlike naïve Bayesian classifiers, allow the representation of dependencies among subsets of.

4.2.1. Bayes Theorem:

Bayes theorem plays a critical role in probabilistic learning and classification.

Let X be a data tuple which class label is unknown. Let H can be some hypothesis, the data tuple X belongs to a specified class C. Then classification problem is determined by P(H | X), the probability that hypothesis H holds given the observed data tuple X.

Probability is classified as two types

Posteriori probability and Prior probability.

Posteriori probability - P (H | X) is the posterior probability, or a posteriori probability, of H conditioned on X.

Prior probability- P (H) is the prior probability, or a priori probability, of H.

Probability Estimation: P (H), P (X | H) and P (X) may be estimated from the given data, It is useful and provide a way of calculate the posterior probability, P (H | X), from P (H), P (X | H) and P (X).

Bayes theorem - P (H | X) = P (X | H) * P (H)/P (X)

Each term in Bayes' theorem has a conventional name:

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

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

-P (X|H) is the conditional probability of X given H.

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

4.3. WEKA

WEKA is nothing but collection of machine learning algorithms for data mining tasks. It contains tool used for, classification, data pre-processing, clustering, association rules, regression, and visualization. It's also well-suited for developing new machine learning schemes. Main strengths of Very portable because it is fully implemented in the Java programming language and runs on almost any modern computing platform. It contains a inclusive collection of data pre-processing and modeling techniques, and easy to use GUI.

4.4 CLASSIFICATION BY DECISION TREE INDUCTION

Decision tree is the learning of decision trees from class-labeled training tuples. It is a flowchart-like tree structure, Each internal node (nonleaf node) denotes a test on an attribute, each branch contains an outcomes of the test, and each leaf node(or terminal node) holds a class label. The topmost node in a tree is the root node. Internal nodes are denoted by rectangles, and leaf nodes are denoted by ovals. Some decision tree algorithms produce only binary trees, whereas others can produce non binary trees.

In data mining and machine learning, a decision tree is a analytical model; a mapping from comments about an item to conclusions about its target value. In these tree structures, leaves represent classification and branch represent conjunction of features that lead to those classifications. 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. Consider the following example:

Table.1 Sample Partial Dataset used to construct Decision Tree

COURSE	SECTOR	SEX	MARKS	GRADE	DESCRIPTION
Computer Science	Rural	M	90-100	O	Outstanding
Computer Science	Urban	F	80-89	D+	Excellent
Commerc	Rural	M	75-79	D	Distinction

P-ISSN: 2347-4408
E-ISSN: 2347-4734

e					
Commerce	Urban	F	70-74	A+	Very Good
Computer Science	Rural	F	60-69	A	Good
Computer Science	Urban	M	50-59	B	Average
Commerce	Rural	F	40-49	C	Satisfactory
Commerce	Urban	M	00-39	F	Re-appear

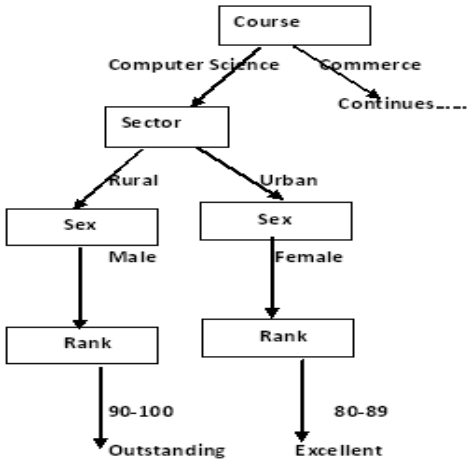

Fig.1 A partial view of a decision tree for this account

4.5. DATA PREPROCESS OF DECISION TREE

The Initial database provided by Placement Cell, D.B.Jain College, Chennai, was in FOXBASE format. It was to be converted to 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 XLSTOMYSQL converter it was converted to MYSQL format.

The individual database files (DBF format) for the years 2008-2009 were obtained and converted to MYSQL format as explained above. 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. List of attributes extracted:
CATEGORY: Social background. Range: {General(GEN), Most Backward Class(MBC), Scheduled Cast(SC), Scheduled Cast/Arunthathiar(SCA), Scheduled Tribe(ST)}
SEX: Range {Male, Female} *SECTOR:* Range {Urban, Rural}
COURSE: Range {A-Z}
{A-Computer Science, B-Commerce,….. etc.,)
ACTIVITY: Indicator of whether the candidate is placed.

4.6. MODELING

The implementation began by extracting the attributes RANK, SEX, CATEGORY, SECTOR, and BRANCH from the master database for the year 2008-2009 at the Placement Cell , D.B.Jain College, Chennai. The database was not intended to be used for any purpose other than 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 (90-100), SECTOR (U), and SEX (M). With this knowledge, the chance for a typical case may be calculated by computing the probability of placement for a test case:

Probability (P) = Number Placed/ Total Number

The chance is obtained by the following rules:

If P>=90 Grade='O' , If P>=80 && P<89 Grade='D^{+}', If P>=75 && P<79 Grade='D', If P>=70 && P<74 Grade='A^{+}', If P>=60 && P<69 Grade='A', If P>=50 && P<59 Grade='B', If P>=40 && P<49 Grade='C'
Else Grade='Re-appear';

Where O, D^{+} , D, $A^{+,}$ A, B, C and F stand for Outstanding, Excellent, Distinction, Very good, Good, Average, Satisfactory and Reappear respectively. So 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.

P-ISSN: 2347-4408
E-ISSN: 2347-4734

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 an 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. Each time, 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. Table 2 describes the typical structure of such an adjacency list that is created.

Hierarchical data(information) a super-sub relationship that is not naturally represented in a relational database table. For this purpose 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 Fig.1, the table for an adjacency list would look like in table 2.

Table.2 Adjacency list (partial view)

ID	NODE	TYPE	PARENT
1	Computer Science	Course	0
2	R	Sector	2
3	M	Sex	3
4	1	Rank	4
5	O	Grade	5
6	U	Sector	2
(Cont inues.)			

Retrieval from Adjacency list: The User enters his search criteria in the user interface screen with details such as sex, category, rank etc 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 COURSE (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 has ID 2 as parent. The algorithm will find ID =3 and ID=7 as child nodes. It then find 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 which means Excellent.

4.7. Confusion Matrix

A confusion matrix is a visualization tool typically used in supervise learning. It is used to represent the test result of a prediction model. Each column of the matrix contains the instances in a predicted class, while each row contains 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.

The entries in the confusion matrix have the following meaning in the context of our study:

- U - Number of **correct** prediction that a case is **negative (-ve)**.
- *V* - Number of **incorrect** prediction that a case is **positive (+ve)**.
- *W* - Number of **incorrect** of prediction that a case **negative (-ve).**
- *X*- Number of **correct** prediction that a case is **positive (+ve)**.
-

Confusion Matrix		Predicted	
		Negative	Positive
Actual	Negative	**U**	**V**
	Positive	**W**	**X**

Several standard terms have been defined for the 2 class matrix:

· ***Accuracy*** (*ACC*) - total number of predictions that be accurately. It is

Determined using the equation [1]: ACC = U+V/ U+ V+ W +X → (1)

· The ***recall*** or ***true positive rate*** (*TPR*)- positive cases that be accurately identified, as calculated using the equation[2]:

TPR = X / W + X → (2)

The ***false positive rate*** (*FPR*)- negatives cases that be inaccurately classified as positive, as calculated using the equation[3]: FPR = V / U + V → (3)

· The ***true negative rate*** (*TNR*) - negatives cases that be classified accurately, as calculated using the equation[4]:

TNR = U / U+ V → (4)

· The ***false negative rate*** (*FNR*) - positives cases that be inaccurately classified as -ve, as calculated using the equation[5]: FNR = W / W+ X → (5)

The truth determined using equation 1 may not be acceptable performance measure when the number of -ve cases is much greater than the no. of positive cases (Kubat et al., 1998). Suppose we have 800 cases, 795 of negative cases and 5 of positive cases. If the system classifies them all as negative, the accuracy will be 99.5%, even though the classifier missed all +ve

P-ISSN: 2347-4408
E-ISSN: 2347-4734

cases. So we have other perforation measures as indicated from [2] to [5]

5. TESTING

Testing was conducted separately for the Naive Bayes based prediction project as well as the decision tree based prediction project and accuracy and confusion matrix were computed. We used the same test data set for both Naive Bayes based project as well as the decision tree based project. We used year 2008-2009 records for building the models and year 2010 records for testing both the model.

5.1 TESTING BASED ON THE DECISION TREE BASED PREDICTION

Table.3 Confusion Matrix (student data)

Confusion Matrix

		PREDICTED			
		C & F	**A & B**	**D & A+**	**O & D+**
ACTUAL	**C&F**	30	1	3	86
	A & B	7	404	4	11
	D & A+	2	1	4	7
	O & D^+	74	6	7	416

The accuracy was given by

ACC = 854/1063 =**0.80339**

To obtain more accuracy measures, we club the field Excellent and Good as positive and Average and Poor as negative. The modified Confusion matrix obtained is as follows in table 4:

Table.4 Modified Confusion Matrix (student data)

		Predicted	
		Negative	Positive
Actual	Negative	442	104
	Positive	83	434

TPR = 0.84 *FPR* = 0.19

TNR = 0.81 *FNR* = 0.16

5.2 TESTING FOR THE NAIVE BAYES BASED PREDICTION

Table.5 Confusion Matrix (student data)

Confusion Matrix

	PREDICTED			
		D		
		O & D^+	C & F	A & B	D & A^+
ACTUAL	O & D^+	496	10	13	0
	C & F	60	97	12	1
	A & B	30	18	248	0
	D & A^+	34	19	22	3

For training, we have used records 2008-2009 and for testing we used the records of year 2010. We compared the predictions of the model for typical inputs from the training set and that with records in test set, whose actual data are already available for test comparisons.

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 was as follows:

ACCURACY(ACC) =844 = **79.398%**

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%.**

6. COMPARISON

To compare two models for predictions, a statistic (based on classical hypothesis testing) is used, which uses the formula as shown below: $P = \frac{|E1-E2|}{\sqrt{q(1-q)(2/n)}}$

Where E1= error rate for model M1

E2= Error rate for model M2

q = (E1+E2)/2

n=number of instances in test set

In our case E1=0.2, E2=0.21,n=1063,q=0.205

So applying these values in the formula **P** becomes **0.56**

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.

7. CONCLUSION AND FUTURE SCOPE

A successful career depends upon the branch the study chooses. This wise decision can be made by searching from the record history. Data mining solves these problems and has wide application in predictive problems in social sciences. Decision

P-ISSN: 2347-4408
E-ISSN: 2347-4734

tree is a well accepted tool in predictive modeling in data mining. This paper illustrates how well Naive Bayes classifier and decision tress are used as predictive tools in the data mining domain and after comparing their performances, it is observed that their efficiencies are comparable in the domain of our problem of career selection. As a summary, this paper demonstrates the interdisciplinary application of data mining tools like decision trees and Naive Bayes classifier in a social science problem. The problem may be implemented using other data mining models like neural network as in [9] and more studies on data preprocessing and dimensionality reductions for this type of domain may be done as in [11].

REFERENCES

[1] SudheepElayidom.M, Sumam Mary Idikkula, Joseph Alexander-Applying Data mining using statistical techniques for career selection, IJRTE ,ISSN 1797-9617, Vol. 1, No. 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, SIGIR'94.

[9] Shuxiang Xu, "Data Mining Using Higher Order Neural Network Models With Adaptive Neuron Activation Functions", IJACT, Vol. 2, No. 4, 168-177, 2010.

[10] [Jinlong Wang, Shunyao Wu, Yang Jiao, Huy Quan Vu, "Study on Student Score Based on Data Mining", JCIT, Vol. 5, No. 6, 171-179, 2010.

[11] Yong Shi, "A Dimension Reduction Approach Using Shrinking for Multi-Dimensional Data Analysis", IJIIP, Vol. 1, No. 2, 86-98, 2010.

[12] G. Madhu, Dr. Keshava Reddy E, "Data Mining for Genetics: A Genetic Algorithm Approach", JCIT, Vol. 3, No. 3, 39-45, 2008.

[13] Yang Hai, Wei He, Xin Liu, Lei Fan, "An Extracting Algorithm for Classification Rule based on Frequent Concept Set", AISS: Advances in Information Science and Service Sciences, Vol. 3, No. 1, 41-46, 2011.