Applying Statistical Dependency Analysis Techniques In a Data Mining Domain

Sudheep Elayidom *sudheepelayidom@hotmail.com*
Computer Science and Engineering Division
School of Engineering
Cochin University of Science and Technology
Kochi, 682022, India

Sumam Mary Idikkula *umam@cusat.ac.in*
Department of Computer Science
Cochin University of Science and Technology
Kochi, 682022, India

Joseph Alexander *josephalexander@cusat.ac.in*
Project Officer, Nodel Centre
Cochin University of Science and Technology
Kochi, 682022, India

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 center in the IAMR, New Delhi, and 21 nodal centers located at different parts of the country. The Kerala State Nodal Center 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).

Key Words: Confusion matrix, Data Mining, Decision tree, Neural Network, Chi- square

1. 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, and is part of a larger task of developing programs that increase student learning. 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 analyzed the trends of placements in the colleges, keeping in account of details like rank, sex, category and location using decision tree models like naive bayes classifier, neural networks etc. chi square test is often shorthand for pearson chi square test.it is a statistical hypothesis test. spss (statistical package for social science) is most widely used programed for statistical analysis. 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. [10] Suggests a data mining approach in students results data while [11] and [12] represents association rules techniques in the data mining domain.

2. 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.

3. PROBLEM STATEMENT

To prepare data mining models and predict placement chances for students keeping account of input details like his/her Rank, Gender, Branch, Category, Reservation and Sector. Statistical dependency analysis techniques are to be used for input attributes to determine the attributes on which placement chances are dependent on. Also performances of the models are to be compared.

4. CONCEPTS USED

4.1. 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. It has been described as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" and "the science of extracting useful information from large data sets or databases".

A typical example for a data mining scenario may be “In a 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.”

4.2 Cross Tabulation

Count		*Chance*				*Total*
		1	2	3	4	
Sector	1	324	194	68	168	754
	2	416	114	69	192	791
Total		740	308	137	360	1545

TABLE 1: A Sample Cross tabulation

The row and column variables are independent, or unrelated. If that assumption was true one would expect that the values in the cells of the table are balanced. To determine what is meant by *balanced*, consider a simple example with two variables, sector and chance for example. It is to be decided on whether there is a relation between sectors (male/female) and chance (yes/no), or whether the two variables are independent of each other. An experiment is to be conducted by constructing a crosstab table as in table 2.

	chance-1	*chance-2*	*Totals*
sector-1	22	18	*40*
sector-2	26	34	*60*
Totals	*48*	*52*	*100*

TABLE 2: Cross tabulation table 2

Now the sector-1 and sector-2 are divided pretty much evenly among chance-1and chance-2 suggesting perhaps that the two variables *are* independent of each other. Suppose it is decided again to conduct the experiment and select some random sample, but, if only totals for each variable separately are considered, for example:

Number of sector-1 is 30; number of sector-2 is 70

Number of chance-1 is 40; number of chance-2 is 60

Total number of data values (subjects) is 100

With this information we could construct a crosstabs tables as in table 3.

	chance-1	*chance-2*	*Totals*
sector-1			*30*
sector-2			*70*
Totals	*40*	*60*	*100*

TABLE 3: Cross tabulation table 3

Now it is to be understood what kind of distribution in the various cells one would *expect* if the two variables were independent. It is known that 30 of 100 (30%) are sector-1 there are 40 of chance-1and 60 of chance-2- if chance had nothing to do with sector (the variables were independent) that we would expect that 30% of the 40 of chance-1are of sector-1, while 30% of the 60 chance-2 are of sector-1. Same concept may be applied to sector 2 also.

Under the assumption of independence one can *expect* the table to look as in table 4:

	chance-1	*chance-2*	*Totals*
sector-1	30/100 * 40 = 12	30/100 * 60 = 18	*30*
sector-2	70/100 * 40 = 28	70/100 * 60 = 42	*70*
Totals	*40*	*60*	*100*

TABLE 4: Cross tabulation table 4

In other words, if a crosstabs table with 2 rows and 2 columns has a row totals r1 and r2, respectively, and column totals c1 and c2, and then if the two variables were indeed independent one would expect the complete table to look as follows:

	X	*Y*	*Totals*
A	$r_1 * c_1$ / total	$r_1 * c_2$ / total	r_1
B	$r_2 * c_1$ / total	$r_2 * c_2$ / total	r_2
Totals	c_1	c_2	*total*

TABLE 5: Cross tabulation table 5

The procedure to test whether two variables are independent is as follows:

Create a crosstabs table as usual, called the actual or observed values (not percentages)

Create a second crosstabs table where you leave the row and column totals, but erase the number in the individual cells.

If the two variables were independent, the entry in i-th row and j-th column is expected to be,

(TotalOfRow i) * (totalOfColumn j) / (overallTotal)

Fill in all cells in this way and call the resulting crosstabs table the expected values table

The important point to be noted is that, if the actual values are *very different* from the expected values, the conclusion is that the variables can not be independent after all (because if they were independent the actual values should look similar to the expected values).

The only question left to answer is "how different is very different", in other words when it can be decided that actual and expected values are sufficiently different to conclude that the variables are not independent? The answer to this question is the Chi-Square Test.

4.3 The Chi-Square Test

The Chi-Square test computes the sum of the differences between actual and expected values (or to be precise the sum of the squares of the differences) and assign a probability value to that number depending on the size of the difference and the number of rows and columns of the crosstabs table.

If the probability value *p* computed by the Chi-Square test is very small, differences between actual and expected values are judged to be significant (large) and therefore you conclude that the assumption of independence is invalid and *there must be a relation* between the variables. The error you commit by rejecting the independence assumption is given by this value of p.

If the probability value p computed by the Chi-Square test is large, differences between actual and expected values are not significant (small) and you do not reject the assumption of independence, i.e. it is likely that the variables are indeed independent.

	Value	Df	Asymp. Sig (2-sided)
Pearson chi-square	32.957[a]	3	.000
Likelyhood ratio	33.209	3	.000
Linear by linear association	.909	1	.340
N of valid cases	1545		

TABLE 6: Sample Chi Square output from SPSS

4.4 SPSS

SPSS is among the most widely used software for statistical analysis in social science problems. It is used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations and others. The original SPSS manual (Nie, Bent & Hull, 1970) has been described as one of "sociology's most influential books".In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary is stored in the datafile) are features of the base software. SPSS can read and write data from ASCII text files (including hierarchical files), other statistics packages, spreadsheets and databases. SPSS can read and write to external relational database tables via ODBC and SQL. In this project data was fed as MS Excel spread Sheets of data.

4.5 Results of the Chi-Square Test

An "observation" consists of the values of two outcomes and the null hypothesis is that the occurrence of these outcomes is statistically independent. Each observation is allocated to one cell of a two-dimensional array of cells (called a table) according to the values of the two outcomes. If there are r rows and c columns in the table, the "theoretical frequency" for a cell, given the hypothesis of independence, is

$$E_{i,j} = \frac{\sum_{k=1}^{c} O_{i,k} \sum_{k=1}^{r} O_{k,j}}{N},$$

And fitting the model of "independence" reduces the number of degrees of freedom by p = r + c − 1. The value of the test-statistic is

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}.$$

The number of degrees of freedom is equal to the number of cells rc, minus the reduction in degrees of freedom, p, which reduces to (r − 1) (c − 1). For the test of independence, a chi-square probability of less than or equal to 0.05 (or the chi-square statistic being at or larger than the 0.05 critical point) is commonly interpreted by applied workers as justification for rejecting the null hypothesis that the row variable is unrelated (that is, only randomly related) to the column variable.

In our analysis, pairs of attributes namely reservation, sex, sector and rank verus the placement chance were having pearson's chi square value less than 0.05. So It could be concluded that this pair of attributes is dependent. In other words placement chances are showing dependency on reservation, sector, sex and rank.

5. DATA PRE PROCESSING

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.

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.

The data mining models were built using data from years 2000-2002 and tested using data of year 2003.

6. IMPLEMENTATION LOGIC

6.1. 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 (A) SECTOR (U) SEX (M)

CATEGORY (GEN) BRANCH (A)

With this knowledge we calculate the chance for case by calculating probability of placement for a test case:

Probability (P) = Number Placed/ Total Number

The chance is obtained by the following rules:

If P>=95 Chance='E'

If P>=85 && 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.

The important point is that, for each of the different combination of input attribute values, we compute the placement chances as shown in the above conditions and prepare another table, which is the input for building the data mining models.

7 DATA MINING PROCESSES APPLIED TO THE PROBLEM

7.1 Using 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.

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.

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

Confusion Matrix					
		P R E D I C T E D			
		E	P	A	G
AC-TUAL	E	496	10	13	0
	P	60	97	12	1
	A	30	18	248	0
	G	34	19	22	3

TABLE 7: Confusion Matrix (student data)

For training, we have used records 2000-2002 and for testing we used the records of year 2003. 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 modelled as a confusion matrix as shown in the above diagram, as its this matrix that is usually used to descrbe test results in data mining type of research works.

The confusion matrix obtained for the test data was as follows:

ACCURACY = 844/1063 = 0.7939

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 follows:

		Predicted	
		Negative	Positive
Actual	Negative	365	101
	Positive	57	540

TABLE 8: Modified Confusion Matrix (student data)

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

We have used WEKA, the data mining package for testing using Naive Bayes classifier.

7.2 Decision Tree

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. Decision tree is a model that is both predictive and descriptive. In data mining and machine learning, a decision tree is a predictive model. More descriptive names for such tree models are classification tree (discrete outcome) or regression tree (continuous outcome). The machine learning technique for inducing a decision tree from data is called decision tree learning. For simulation/evaluation we use the data of year 2003 obtained from NTMIS. The knowledge discovered is expressed in the form of confusion matrix.

Confusion Matrix					
		P R E D I C T E D			
		P	A	G	E
AC-TUAL	P	30	1	3	86
	A	7	404	4	11
	G	2	1	4	7
	E	74	6	7	416

TABLE 9: Confusion Matrix (student data)

Since the negative cases here are when the prediction was Poor /average and the corresponding observed values were Excellent/good and vice versa. Therefore the accuracy was given by *AC* = 854/1063 = 0.803. 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 **82.4%**.

TP = 0.84, *FP* = 0.19, *TN* = 0.81, *FN* = 0.16. We implemented and tested the decision tree concept in a web site using php, mysql platform by using data structure called adjascency list to implement a decision tree.

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

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

TABLE 10: Confusion Matrix (student data)

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

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** %.

TP = 0.83, FP = 0.17, TN = 0.81, FN = 0.17.

We used MATLAB to implement and test the neural network concept.

8. CONCLUSION

Choosing the right career is so important for any one's success. For that, we may have to do lot of history data analysis, experience based assessments etc. Nowadays technologies like data mining is there which uses concepts like naïve Bayes prediction to make logical decisions. This paper demonstrates how Chi Square based test can be used to evaluate attributes dependencies. 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. This methodology can very efficeiently be implemented by the governments to help the students make career decisions. It was observed that the performances of all the three models were comparable in the domain of placement chance prediction as a part of the original research work.

9. ACKNOWLEDGEMENT

I would like to acknowledge the technical contributions of Sunny(sunny@hotmail.co.in), Vikash kumar(vikashhotice2006@gmail.com),Vinesh.B(vineshbalan@gmail.com),Amit.N(amitnanda@hotmail.com), Vikash Agarwal (vikash.vicky007@gmail.com), Division of Computer Engineering, Cochin University Of Science and Technology, India.

10. REFERENCES

[1] SudheepElayidom.M, Sumam Mary Idikkula, Joseph Alexander. “Applying datamining using statistical techniques for career selection”. IJRTE, 1(1):446-449, 2009

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

[3] Fayyad, R. Uthurusamy. “From Data mining to knowledge discovery”, Advances *in data mining and knowledge discovery* Cambridge, MA: MIT Press., pp. 1-34, (1996)

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

[5] L. Breiman, J. Friedman, R. Olshen, and C. Stone. “Classification and Regression Trees”, Wadsworth Inc., Chapter 3, (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”. In Proceedings of the 14th International Conference on Machine Learning, ICML, Nashville, Tennessee, USA, 1997

[8] Lewis D. D. & Gale W. A. “A sequential algorithm for training text classifiers”. In proceedings of SIGIR, Dublin, Ireland, 1994

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

[10] Oyelade, Oladipupo, Olufunke. “Knowledge Discovery from students result repository: Association Rules mining approach”. CSC- IJCSS, 4(2):199-207, 2010

[11] Anandavalli, Ghose, Gauthaman. “Mining spatial gene expression data using association rules”. CSC - IJCSS, 3(5): 351-357, 2009

[12] Anyanwn, Shiva. “Comparitive analysis of serial decision tree classification algorithms”. CSC- IJCSS, 3(3):230-240, 2009