International Journal of Recent Trends in Engineering, Vol. 1, No. 1, May 2009

Applying Data mining using Statistical Techniques for Career Selection

Sudheep Elayidom, Dr. Sumam Mary Idikkula, and Joseph Alexander
Cochin University Of Science And Technology,
Computer Science and Engineering Division , School of Engineering , Kochi,India
sudheepelayidom@hotmail.com
Cochin University of Science and Technology Department of Computer Science,Kochi,India
sumam@cusat.ac.in
Project officer NODEL Center Cochin University Of Science And Technology,Kochi,India
josephalexander@cusat.ac.in

***Abstract*—For years, choosing the right career by monitoring the trends and scope for different career paths have been a requirement for all youngsters all over the world. In this paper we provide a scientific, data mining based method for job absorption rate prediction and predicting the waiting time needed for 100% placement, for different engineering courses in India. This will help the students in India in a great deal in deciding the right discipline for them for a bright future. Information about passed out students are obtained from the NTMIS (National technical manpower information system) NODAL center in Kochi, India residing in Cochin University of science and technology .**

***Index Terms*—Data mining, Time series analysis, Regression, Curve fitting, Trend lines, Prediction, Statistical analysis, Career selection.**

I. Introduction

Knowledge of statistics regarding placements in colleges has become one of the basic requirements of students all over the world. Data mining that works on statistical techniques can be used for this purpose. Need is to formulate an interface that uses a database of statistics and predicts accurately the time in which 100% placement can be achieved for a particular branch in a particular year. This would not only reduce the troubles of those responsible for displaying the statistics, but would also come up as aid to the students seeking for colleges which can guarantee them a secure future. Also the waiting time prediction is useful in the sense that more the waiting time for a branch, more will it indicate that intake for the coming years should be reduced.

Studies have been conducted in a similar area such as understanding student data as reported in [3]. 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. But in our work we use time series based statistical data mining techniques to predict job absorption rate after graduation not for an individual rather for a discipline as a whole. Time series data is a sequence of observations collected over intervals of time[2]. Each time series describes a phenomenon as a function of time. For example, daily stock prices could be used to describe the fluctuations in the stock market. In general, for a time series X with n observations, X is represented as

$$X = (v1 ,t1), (v2 ,t2), ..., (vn ,tn) \quad (1)$$

Where *vi* and *ti*, are the observation value and its time stamp, respectively. The data we are using in this work to compute the absorption rate for future is such a time series data, in which absorption rates for different years are used. Many researchers have been working on applications using time series analysis researchers. Data mining on time series data have been so important nowadays to improve the business analysis and forecasting in other social science related domains. Several Works have been done in analyzing time series data like astronomical data, business data etc. In [2] it is found that astronomical data is more periodic in nature and the author suggests that time series analysis is more suited for data in the domain which are more random in nature like the work we are doing like prediction of absorption rates etc. In [1] a time series analysis using a technique called trend calculations based on moving averages for the data on Oil Prices (in Dollars) in the Period 1900-1996 in the U.S is shown. A linear regression-based model for the calculation of short-term system load forecasts is described in [4]. The model's most significant aspects fall into the following areas: model building, including accurate holiday modeling by using binary variables, temperature modeling by using heating and cooling degree functions; and robust parameter estimation by using weighted least-squares linear regression techniques. In [5] simple markets can be used to aggregate disperse information into efficient forecasts of uncertain future events. Drawing together data from a range of prediction context show that market-generated forecasts are typically fairly accurate, and that they outperform most moderately sophisticated benchmarks.

A. Data

The data used in this project is the data supplied by National Technical Manpower Information System

(NTMIS) through its Kochi (India) Nodal center. Data is compiled by them from feedback given by graduates, post graduates, diploma holders in engineering during the year 2002 and also from various engineering colleges and polytechnics located within the state during the year 2003-2004. This survey of technical manpower information was originally done by the Board of Apprenticeship Training (BOAT) for various individual establishments.

B. Problem Statement

1. Predict the job absorption rate for a particular branch i.e. to figure out the percentage of students that will be placed in a particular branch in a particular year in the future based upon the existing trends. The developers may use the concept of linear regression and formulate equations to determine the percentage of students.

2. Calculate placement rate status for a particular batch for a period of every 3 months for each year. From this data one should be able to predict the time needed to attain 100% placement for the given batch. The developers may use the concept of curve fitting and regression modeling since it provides accurate approach to predict such rates which are based on statistical data. The outcome of the project will help the NTMIS to determine which branches need more intakes, if they have good placement rates and also will help in drawing attention to the branches which are lagging in terms of placement.

II. Concepts Used

A. Data Mining

Data mining is the principle of analyzing large database and picking out interesting Patterns. It is usually used by business intelligence organizations, and financial analysts. It is also increasingly used in sciences to extract information from the enormous data sets generated by modern experimental and observational methods.

B. Curve Fitting

Curve fitting is finding a curve which has the best fit to a series of data points.

C. Regression Analysis

1) Linear Regression

In general, the goal of linear regression is to find the line that best predicts Y from X. Linear regression does this by finding the line that minimizes the sum of the squares of the vertical distances of the points from the line.

2) Non-Linear Regression

Nonlinear regression is a general technique to fit a curve through the data. It fits data to any equation that defines Y as a function of X and one or more parameters

D. R-squared value

R-Squared is a statistical term denoting how good one term is at predicting another. If R-Squared is 1.0 then given the value of one term, the value of another term can be perfectly predicted. If R-Squared is 0.0, then prediction may not be accurate at all.

E. Trend Lines

Trend line can be defined as a straight or curved line in a trend chart that indicates the general pattern or direction of a time series data. It may be drawn visually by connecting the actual data points or (more frequently) by using statistical techniques such as 'exponential smoothing' or 'moving averages.'

F. Processing Data For Input

The Initial database provided by Nodal Center, was in FoxBASE format. FoxBASE data was converted to CSV files that could be fed into MySQL.

The Data base obtained contained records for the years 2001, 2002 2003 for Waiting-time prediction and for the years 1983-2001 for Absorption-rate prediction (degree and diploma separate). For the sake of simplicity, the required fields were extracted to separate dbf files to make the database user-friendly and the coding part easier for the programmer. An extensive sorting was performed through Microsoft Excel first using the joining year followed by the joining month.

The three attributes extracted are Rollno of the candidate, Month and year at which he/she joined the Company. The database created using these three attributes was then fed to an xls-sql converter and finally a MySQL database was created. These attributes were fed into MySQL through SQL queries and each of these entities was put under the database of the respective branch.

III. Logic

A. Absorption Rate Prediction

The attributes like year and the percentage of students absorbed into a job in that year were extracted from the database. A graph was plotted using these two fields. A pictorial representation of the trends being followed for the intake in this particular branch over the years was got. An equation for this graph was to be formulated for predicting the absorption rate in the future. A tool named “Table Curve “was used for this purpose. A line of best fit (a line that includes the maximum no. of points) was selected. Next an equation to this line was found out.. This method is very similar to linear regression method. Next we look for all the possible graphs for linear, non-linear equations for the range of our obtained graph. Each of these graphs is matched (overlapped) with the original graph and the graph which is most resembling and includes the maximum points is chosen. The equation whose graph is chosen is the required equation.

In the table R-Squared value for each of the computation were shown. Higher the R-Squared value better is the prediction. In the table R-Squared values like 0.99, 0.91 etc shows that the predicted values are having reasonably good accuracy.

TABLE I. ACCURACY CALCULATED FOR THE ABSORPTION RATE ALGORITHM.

Degree	R-square Value	Equation	Accuracy
AR	0.99	y=a+bT1(x)+cT 2(x)+dT3(x)+eT 4(x)+fT5(x)+gT 6(x)+hT7(x)+iT 8(x)+jT9(x)+kT 10(x)+lT11(x)+ mT12(x)	100%
CE	0.91	y=a+bcos(x)+csi n(x)+dcos(2x)+e sin(2x)+fcos(3x) +gsin(3x)+hcos(4x)+isin(4x)+jco s(5x) +ksin(5x)+lcos(6 x)+msin(6x)+nc os(7x)+osin(7x)	93.7%

Validation of the methods were conducted in such a way that all the available year (1983-2001) values were substituted in the equation and predicted absorption rates were compared with observed absorption rates allowing an error of 5%. Table 1a shows the equations for absorption rate obtained for EC (Electronics and communication), AR (Agriculture) and CE(Civil Engineering). The fourth column of the table shows how close the predicted values were with the actual values.

B. Waiting –Time prediction

The basic idea behind the prediction of 100% placement time was to develop an equation from the give database by plotting the graph between placement rate(Y) and the time (X). As it is clearly shown in the figure the variable Y depends on X given by the equation

$$Y = mX + c \quad (2)$$

The major issue encountered while plotting is that a straight line is never obtained whenever a graph is plotted for a given branch's database. To get a 'trend line' for the plotted graphs, a mathematical library was used. The equation thus obtained was accurate as we had a large database for each branch. Thus it provides us an appropriate linear equation which can be used to compute the statistical data by integrating it with database. Finally it draws the line that joins the data points. n the data points, and represents the result as a best fit straight line. In linear regression analysis, the data points are assumed to be related by:

$$y=m*x+c+err \quad (3)$$

Based on estimated values of m, c and err.
The uncertainties of the data point are contributed by the uncertainties in m, c, and err. The uncertainties of the data points are represented visually as a prediction band around the regression line. For example, a 95% prediction band means there is 95% probability that a data point will be in that band. The algorithm used to find 100% placement could be verified by reversing the procedure with a known result. As of from data of years 2001, 2002 and 2003 for various branches, we have 3 trend line equations for every branch corresponding to each of the above years. Now absorption rates of actual data (y value) was substituted in these trend line equations computed by the software and predicted time taken(x value) for that particular absorption rate was computed for available data of each year. It was observed that the time computed for a particular absorption rate percentage in the above method was same as the observed value with a small deviation of 3%, since the database of absorption rates follows a trend. Hence mathematically it can be said that it will be correct for 100% absorption time also.

Fig. 1 shows the 95% confidence band and prediction band of the work. This means we are 95 % sure that the curve lies within the confidence band and 95% of any new points will be in the prediction band.

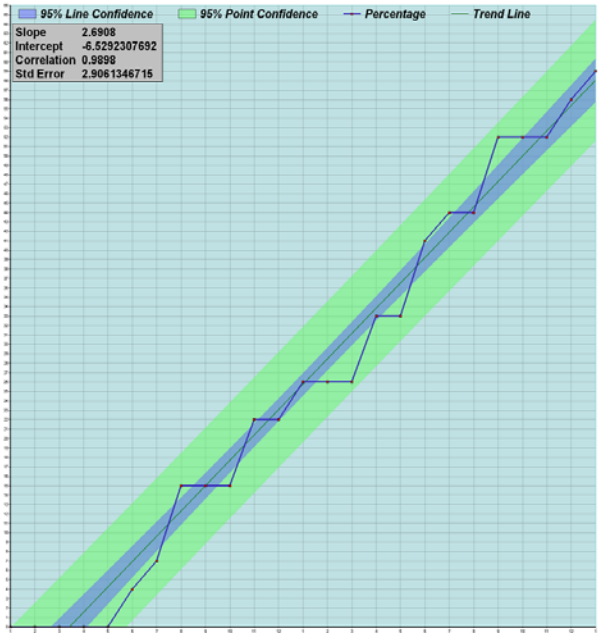
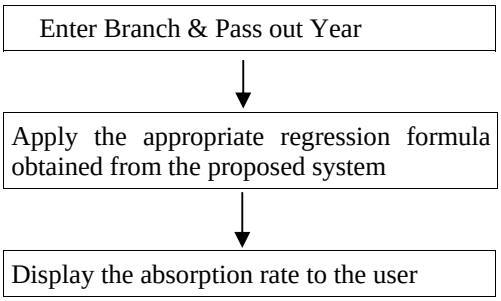


Figure 1. Confidence band and prediction band of the work

D. Proposed system

The proposed system can be made available to public as a website in future by the NODAL Centre by the concurrence of the Government, in which the admission seekers can enter the branch and the pass out year for which they have to find the absorption rate.

This website can be made as a part of the main site of the NODAL center, from which a link can be given to both the absorption rate prediction module and waiting time prediction module.



Also the government may use the waiting time prediction module to decide whether the intake for a particular branch needs to be increased next year considering the current waiting time for that branch.

E. Algorithms

The columns are sorted in the terms of joining month and joining year as shown below. To write to the database the percentage is computed for each succeeding month from passing month and year entered by the user.(for example : a database of batch 2001 will compute percentage for each month staring from July 2001). Thus a separate table is made for percentage computation of placement for each branch with columns comprising month, year and percentage (cumulative) of students got placed till date as shown in table 3.

TABLE II. REPRESENTING PLACEMENT IN JUNE 2001

YEAR	J_MONTH	%
2001	7	25
2001	8	37
2001	9	62
2001	10	75
2001	11	87
2001	12	87
2002	1	87
2002	2	100

Algorithm to develop a linear graph between the percentage of students placed and the corresponding time taken plus derive a relation between them from the line of best fit and computation of 100% placement time:

Step 1: The page redirected from the interface is supplied with arguments containing branch and the pass out year, and the program links with the required database.

Step 2: Generate array of the percentage and months for percentage database and pass these arrays.

Step 3: Using getSlope () function gets the slope of the line of best fit.

Step 4: Similarly obtain values of y intercept and error rate possible.

Step 5: Thus linear equation of two variables is formed. Substitute value of placement rate as 100 in equation and compute corresponding time in months.

Step 6: The time obtained is changed into years and months format and is the required waiting time for 100 % placement.

F. Conclusion

A statistical data mining approach to the prediction of values is always superior to one which requires manual work. The software is simple to use besides being reasonably accurate. Moreover the user friendly interface used in this project turns out to be easy to handle and avoid complications. It gives the students platform to judge what is better for them. Basically, it studies past data, follows the trend, figures out how things have been working and then gives a refined judgment on what should be the case in the future. A statistical data mining approach to the prediction of scope of various branches in future is always very useful to the youngsters to choose a prospective career.

ACKNOWLEDGMENT

We would like to acknowledge the programming support given by undergraduate students in computer engineering of Cochin University namely
Aayush Raina (aayush.raina21@gmail.com)
Althaf K Backer (althafkbacker@gmail.com)
Prateek Uniyal (prateekuniyal1@gmail.com)

REFERENCES

[1] Alaaeldin Hafez, Association mining of dependency between time series SPIE proceedings series, INIST-CNRS
[2] Jefery D Scargles, Studies in Astronomical time series analysis The astrophysical journal 263;835-853 1982 December 15
[3] Elizabeth Murray, Using Decision Trees to Understand Student Data,Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005
[4] Papalexopoulos,A.D. Hesterberg,T.C.A regression-based approach to short-term system load forecasting Pacific Gas & Electr. Co. November 1990
[5] Wolfers, E Zitzewitz,Prediction Markets,NBER Working Paper, 2004