Rank Analysis Through Polyanalyst using Linear Regression

N. Sandhya †, K. Anuradha†, Sk.Althaf.H. Basha†, P. Premchand††, A.Govardhan†††

† Gokaraju Rangaraju Institute of Engineering &Technology, Hyderabad
†† University College of Engineering,Osmania University, Hyderabad
††† University of College of Engineering, JNT University, Hyderabad

Abstract
Data Mining is defined as the process of discovering significant and potentially useful patterns in large volumes of data[2]. It is the exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover Knowledge Discovery in Databases (KDD). The study reported in this paper is concerned primarily with thematic information related to admission criteria of an Institution. It aims at discerning trends in admission with reference to ranks into an institution. Here the database has been mined using the PolyAnalyst software package. The predicted and real vs. counter graph illustrates how closely the PolyAnalyst prediction follows the actual value of the attribute over the range of the dataset. The application of data mining techniques helps the decision maker to meet the goal. The specific application of PolyAnalyst gave a clear scope for evaluation and comparison of predicted and real values.

Keywords: *Data Mining, PolyAnalyst, Linear Regression*

1. Introduction

Education system everywhere is primarily based on examinations[10]. The scenario generally is that the number of aspirants would be more than the number of seats available in an institution. Entries into educational institutions are based on the competitive exams. The aspirants also will have plethora of options in front of them in the form of various courses in various institutions that are concerned with admission through a particular competitive exam[12]. The aspirant's preference is affected by various factors like the reputation of the institution, the faculty at the institution, the alumni of the institution etc., [13] in other words the type of students admitted into the institutions is influenced by the above factors. The study here focuses on this aspect where with the help of PolyAnalyst the variations in the standard of students joined into the institution are identified.

The paper is organized as follows: Section 2 presents an introduction to data mining. Section 3 deals with PolyAnalyst tool and describes its workspace. Section 4 presents the description of data used, problem formulation and the approach followed to obtain the analytical solution. Section 5 presents the results and related discussion. Section 6 presents conclusion and the future work.

Manuscript received September 5, 2009
Manuscript revised September 20, 2009

2. Data Mining

Data mining, or Knowledge Discovery in Databases (KDD) is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data[1]. KDD is the process of identifying a valid, potentially, useful and ultimately understandable structure in data. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems.

2.1 Issues and Challenges in Data Mining

Data mining system depends upon databases to supply the raw input and these rise problems such as that database tend to be dynamic, incomplete, noisy, and large. Other problems arises as a result of the inadequacy and irrelevance of the information stored .The difficulties in data mining can be categorized as

- Limited Information
- Noise or Missing Data
- User Interaction and Prior Knowledge
- Uncertainty
- Size, Updates and Irrelevant Fields.

The above mentioned issues can be solved by preprocessing[2] the data as mentioned in Section 4.

3. Polyanalyst

PolyAnalyst[9] is a powerful multi-strategy data mining system that implements a broad variety of mutually complementing methods for the automatic data analysis. It works with data extracted from flat files or relational databases, and can work with numerical (floating-point), integer, yes/no (binary), date, and discrete (categorical or string) variables. Using this data and PolyAnalyst's suite of analytical algorithms, relationships in the data can be discovered, predictions made, and the data classified and organized. Data Mining can perform tasks beyond the scope of statistical analysis software.

3.1 The PolyAnalyst Workspace

3.1.1 Projects:

The first step in running an analysis with PolyAnalyst is to create a new project. The PolyAnalyst project file is a single package that will contain all the nodes-datasets, rules, charts, graphs, and reports produced throughout the analysis in one convenient place. The PolyAnalyst interface is shown in figure 1.

3.1.2 Reports*:*

When an exploration engine completes its task, the primary output is a report. The report gives all the details of the solution the exploration engine found, usually including a text report listing several significance measures, a list of terms used, and a symbolic rule (if applicable) as seen in section 5.

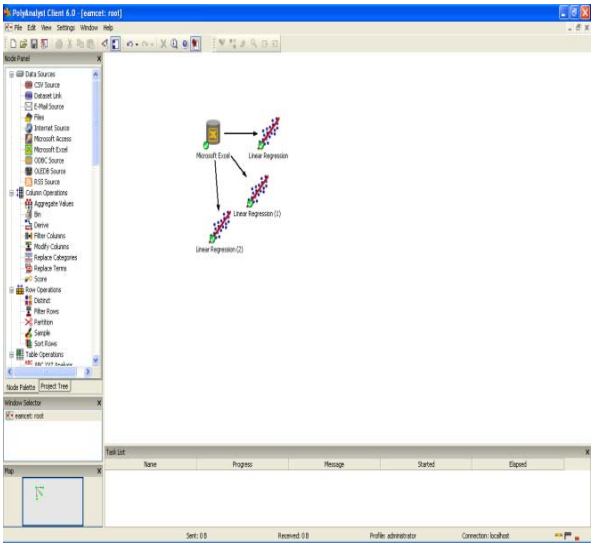

Figure 1: polyanalyst's workspace

4. Materials and Methods

A competitive exam will have the results in the form of ranks or marks. An institution's performance intuitively depends on the ranks that find admittance into it. Analysis of these previous ranks can speak of the trend observed and leads the decision maker to make some important decisions.

In Andhra Pradesh, the entrance into the Engineering and Medical colleges is through the common entrance examination- EAMCET (Engineering And Medical Common Entrance Test).

4.1 Samples taken

The study here considers an engineering organization - Gokaraju Rangaraju Institute of Engineering and Technology. A typical dataset that is used here is of ".xls" format. It contains last 12 years minimum EAMCET ranks[12] that found admittance into the considered college. The data was categorized into Open category, BC reservation, SC reservation for boys and girls accordingly as per the respective branch. These form the rows whereas the years form the columns. For instance consider the years from 1997 to 2008. An example of the dataset prepared for this study is in the figure 2.

The database schema is illustrated in the table 1.

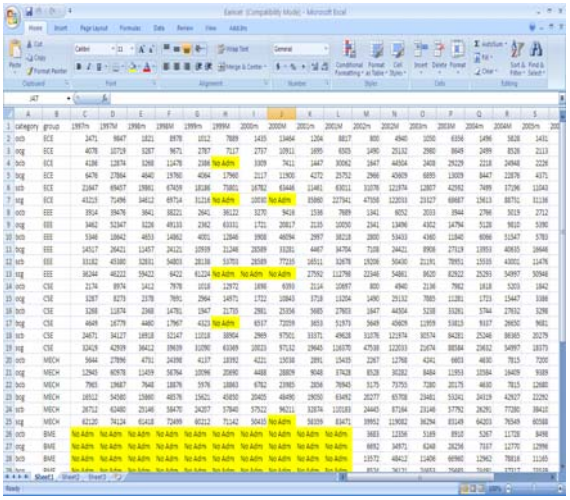

Figure 2: sample dataset

Table I:database schema

Field Name	Source of Data	Data Type	Field Size
Category	Administrative data	String	10
Group	Administrative data	String	10
1997	Statistics	Numerical	Long Integer
1998	Statistics	Numerical	Long Integer
1999	Statistics	Numerical	Long Integer
2000	Statistics	Numerical	Long Integer
2001	Statistics	Numerical	Long Integer
2002	Statistics	Numerical	Long Integer
2003	Statistics	Numerical	Long Integer
2004	Statistics	Numerical	Long Integer
2005	Statistics	Numerical	Long Integer
2006	Statistics	Numerical	Long Integer
2007	Statistics	Numerical	Long Integer
2008	Statistics	Numerical	Long Integer

4.2 Problem Formulation

4.2.1 Study objective

The analysis of standard of students admitted into an institution helps the decision makers- the management and students to draw some important conclusions and make influential decisions. Here an experimental attempt is made for analyzing the variations in these trends.

4.2.2 Approach

a) Preprocess the data

A decision maker first analyses the historical data of the past, obtained from the institution's records. This stage comprises identifying, collecting, filtering, aggregating data into a format required by the data models. Here the missing values are filled with string-"No Adm" ,meaning no admission.

b) Perform mining

Data mining technique- Linear regression is applied here to predict and analyze the variations observed in admittance of students into an institution.

Linear Regression is one of the oldest and most well known methods of statistical prediction – it is the process of creating a line through a space such that the sum of the squares of the distance between the line and each point is minimized. Linear Regression is a valuable tool as it is very fast and produces easily readable and interpretable results. PolyAnalyst's stepwise linear regression can work with any number of attributes, and automatically determines which attributes give the best linear prediction rule.

The data is configured into the 'Microsoft Excel' node. When linear Regression node is selected, it prompts to select which attributes to include in the exploration and which attribute to target for prediction purposes. The target attribute is the one on which the prediction is made. Here it is the 2008 year while all the previous year's attributes until 2007 form the exploration attributes.

c) Analyze patterns

Interesting patterns are analyzed to obtain the useful results as seen in next section.

5. Results and Discussion

5.1 Reports

Reports are the main asset of this study. The linear regression report begins with the prediction rule. The prediction rule is formed based on the most influential attributes. After the prediction rule, a variety of measures of accuracy and statistical significance such as standard deviation,R-Square value etc., are included.

The Prediction rule obtained is in the form of the following equation:

[2008] = -1.15151e-006*[1997] +1.80697e-006*[2000] -4.35346e-006*[2002] +0.991377*[2004] +2.62594e-006*[2007]

Based on this predicition rule, predicted values of the target attribute are obtained. The variations of these predicted and real values are shown in the form of PolyAnalyst's linear regression graphs as in figure 3 and figure 4 .

5.2 Linear regression graphs

Linear regression analyzes the relation ship between Predicted vs Real and also for the Predicted and Real vs Counter in following sub section discussed about the relation as said above

5.2.1 Predicted vs. Real

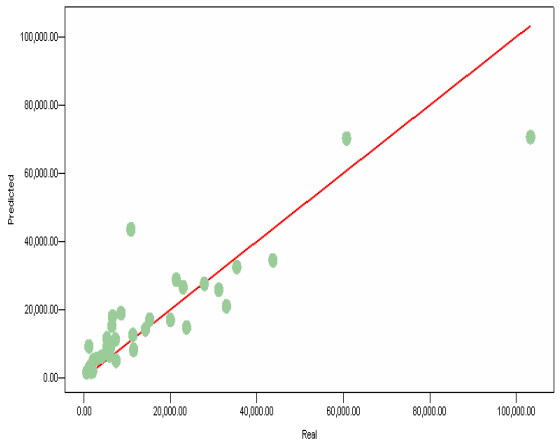

Figure 3

The Predicted vs. Real graph shows all the data points in the actual dataset being explored along with where these data points would have been predicted to fall by the model produced. This allows to get a quick look at the accuracy and predictive power of the model. The real

location of the data point is shown along the x-axis, while the predicted location is shown along the y-axis.

5.2.2 Predicted and Real vs. Counter

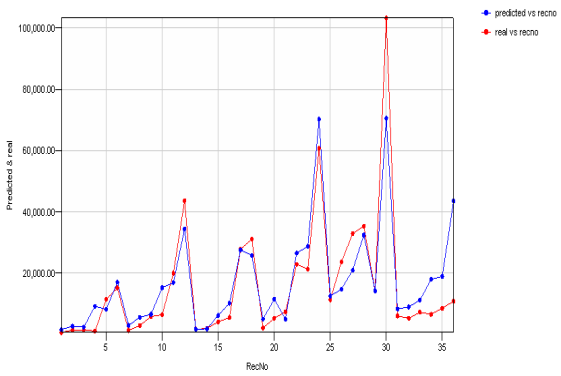

Figure 4

The Predicted and Real vs. Counter graph allows to see how closely the PolyAnalyst prediction follows the actual value of the attribute over the range of the dataset. The blue line represents the PolyAnalyst prediction, while the red line represents the real values. Record number is plotted along the x-axis while the predicted and real locations is shown along the y-axis.

5.3 Prediction Accuracy

The Accuracy of the predicted values can be determined based on the R-Square value[9] (the coefficient of determination). If its value approaches '1' then the prediction is accurate. The R-Square value obtained for the above prediction is 0.82. The variations exist to some extent because of the influence of several other factors besides college previous ranks.

6. Conclusion

The variations observed in the graph shown in figure helps both the students and management to take better decisions. Based on this the students can opt for a good institution from the available ones. This analysis of variations in students who got admitted into the institution can be used by the management to take some beneficial measures. These may include improving the facilities, infrastructure, faculty, ultimately the standards.

The similar approach can be applied to any other institution to determine the best ones based on what kind of students (based on ranks) are getting admitted into that Institution.

While further statistical analysis should provide more quantitative evaluation of the relationships between various variables, the visual analysis clearly shows the variations. The specific application of PolyAnalyst gave a clear scope for evaluation and comparison of predicted and real values.

References

[1] R. Agrawal, A. Arning, T. Bollinger, M. Mehta, J. Shafer, R. Srikant, *The Quest Data Mining System*, proceedings of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, August, 1996.

[2] Jiawei Han, Micheline Kamber , *Data Mining:concepts and techniques*, 2nd ed,Morgan Kaufmann publishers,2008.

[3] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, *Introduction to Data Mining.*

[4] D. J. Hand, Heikki Mannila, Padhraic Smyth, *Principles of Data Mining.*

[5] Pieter Adriaans, Dolf Zantinge, *Data Mining*,3rd ed,2006 .

[6] Arun K. Pujari , *Data Mining Techniques*, 3rd ed,University press,2007.

[7] (2007) The IEEE website. [Online]. Available: http://www.ieee.org/

[8] Daniel T. Larose, *Data Mining Methods and Models.*

[9] The Megaputer website.[Online].Available:http://www.megaputer.com

[10] [online].Available:http://www.indiaedu.com/education-india/

[11] The amazon website.[online]. Available:www.amazon.com/Data-Mining-Techniques

[12] Available:http://www.aicte.ernet.in/regulation_1992.htm

[13] [online]http://www.inae.org/news.htm