Analysis of Performance in competitive examinations & pattern analysis using statistical and Data Mining techniques-A case study

Ananth.Y.N
Associate Professor & Ph.D student- Dept of Compute Science
JainUniversity
Bangalore-INDIA,
aninotes@gmail.com

Dr. Narahari.N.S
Professor, Dept of Industrial Engineering & Management
R.V.College of Engineering Autonomous-VTU
Bangalore-INDIA,
naraharins@rvce.edu.in

***Abstract*: Competitive exams are the order of the day, starting from the High school level to the post graduate level. The two aspects which are the corner stones of these processes are the assesses and the assessment process. The assessment process has got its own parameters like the nature of questions, number of questions , strength of the questions and so on. The assesses on his part has got the performance record and his aptitude to perform in the examination. These two are closely related each other. This paper focuses on some data mining and statistical techniques which could be used to draw these relationships and discusses the scope of these tests. Statistical techniques could be used to perform the multifactor analysis of the parameters under concern. Data mining software's like Rapid miner, R and Weka can be used to analyze data in large sets –effectively to find out clusters, correlation between the fields in the data and so on.**

Keywords—data mining,Statitical Techniques,Rapid Miner,Classification, Regression, Correlation Introduction

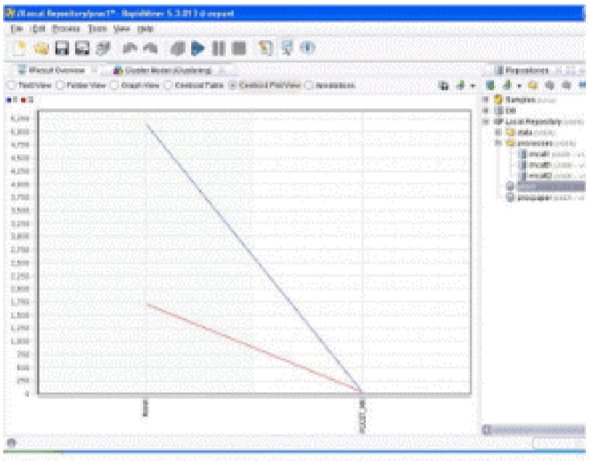
I. INTRODUCTION

Competitive exams are happening at every level of the education scenario – starting from the HIGH school level till post graduation. The extent to which these tests would be effective is a concern for investigation. In order to study this results scored are indicative of the performance of the students, at least to some extent. An analysis of the results of the examination, with a perspective of finding patterns in these scores would be an appropriate step .This paper focuses on some data mining and statistical techniques to study the performance of the students.

Competitive examinations include several types of questionnaire testing the various faculties of the student. Many of them are multiple choice questions with the student having either a correct or a wrong choice among the answers and there is no "average marks "scored by a student. On the other hand, some examinations have descriptive type of questions and answers to be the rule and here the student could score some marks which is not "right" or "wrong" and get the full marks available , but some marks scored indicative of his faculties of writing skills. Some examinations contain a mixture of these two kinds of questions. The NET examination conducted by UGC earlier had this type of pattern. So there can be no one method to analyze and find the patterns of the scores in these exams. Different methods are available and have to be used.

II. DATA MINING TECHNIQUES TO FIND PATTERNS –CLUSTERING AND CLASSIFICATION:

Data mining techniques come to be useful when there are large amounts of data to be analyzed and we can expect patterns of data to be present in the data set. There can be clusters where in the scoring would be following a particular pattern say students scoring x plus or minus some amount of marks. This can be found to include smaller clusters or larger clusters. At a very basic level this kind of analysis would lead to identifying similar or dissimilar patterns of scores.
The resulting patterns so evolved can be used to classify the clusters. In this step, the clusters could be classified based upon the proximity of the scores within the clusters .A further analysis would allow



Us to lead to explore the relationship between the scores and the pattern of the questions in the question papers. Questions in such competitive exams can be having categories of questions like knowledge based questions, skill based questions and memory based questions. Classification techniques would lead us to isolate groups of students who have scored well in a certain category of questions. The following screen shot shows the distribution of clusters in the rank and marks of PGCET –MCA 2012 results, generated using Rapid Miner.

It can be observed that there are overlapping clusters between the two fields taken individually. The relative distribution may It can be observed that there are overlapping clusters between the two fields taken individually. The relative distribution may be generated by taking the correlation between these two fields which is shown in a further section.

Another independent variable which could be studied in this scenario is that of the students' scores in the qualifying exams. For example the relationship between the B.E marks of a student who is taking PGCET, although is not very direct, is certainly present. We can use the classification methods to classify the students' groups based upon this relationship. This relationship is a very strong indicator of the student's aptitude because the students would have put in 4 \years of efforts in studying the core aspects of his chosen area and would have been tuned to answer the questions in a particular way. Although there could be exceptions which defy this relationship, they could be quite less in number and these sets would be treated as outliers in data mining techniques

III. ANALYSIS

The analysis of the student performance would indicate or shed light on several factors-what is the general trends in aptitude of the students- whether the questions are satisfactorily measuring the aptitude ,whether the right types of questionnaire is present, what is the actual distribution of the marks. These could also be studied in order to predict the future behavior and performance of the students, the possible range of scores for the students who fall in a different set. Generalizations in this regard are quite difficult to make. Statistical techniques step in for analysis.

The basics of this analysis include some dependent and independent factors which need to be mentioned. The independent variables are the past performance of the students, aptitude tests whereas the main dependent variable is the performance record.

Some of the basic statistics which could be studied are the mean, median and mode of the scores. The next level is to study the distributions of the scores, which is done by calculating the range of the scores, and statistics like standard deviations could be used here. Under ideal conditions the scores follow a Gaussian distribution, 34.13 % of the scores would fall between mean and mean plus one standard deviation(x *hat* x *hat* +σ).while the other 34.13 percent would fall on the other side of the distribution(x *hat* x *hat* -σ)Thus 68.26 %of the scores would fall between x *hat* plus or minus σ).and another 13.59% of the scores would fall between x *hat* plus or minus two standard deviations This is an observed fact over many situations but it cannot be said that they are applicable to all the scores of all the examinations.

IV. CALCULATION OF SCALED SCORES

An absolute score cannot determine the comparison between a student's performance and that of his peers. In such cases, scaled scores can be used to determine where exactly does a student stand in relation to the entire set of students. Two of the popular scaled scores are the z and T- scores. The z score is equivalent to the number of standard deviations that a raw score either falls below or above the mean. For example if a student scores 20 on an exam with the mean being equal to 50 and a standard deviation of 15, ,the raw score is two standard deviations below the mean, and the z – score will be -2.A raw score of 65 on the same exam would be having a z score of 1. T- scores transform the grade to a scale on which the mean has been adjusted to 50 and the standard deviation has been scaled to exactly 10 points.T scores can be calculated from the raw scores (x*i*), the mean x *hat* and the standard deviation σ) using the following formula

$$T=10 * (x_i - x\ \mathit{hat})/\ S + 50$$

Our student who was two standard deviations below the mean would have a T- score of 30, while the student who was one standard deviation below would have a T score of 60.
The advantage of the scaled scores is that the student knows exactly where he stands in relation to his peers. Some are confident while some others are inherently afraid of failure. If the student is just given his raw score, mean and the range, he cannot understand where exactly he stands in relation to his peers.

V. CORRELATION AND REGRESSION ANALYSIS

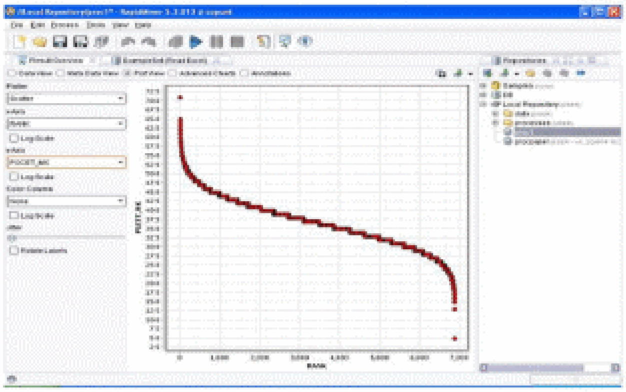
Correlation is the statistical tool which can be used to describe the degree to which one variable could be linearly related to another. Often, correlation analysis is used in conjunction with regression analysis to measure how well the regression line explains the variation of the dependent variable. There are mainly two types of measures to describe the correlation between two variables: the coefficient of determination and the coefficient of correlation. These could be used to study the effect of the questions distributions on the scores that the student is going to score. These two techniques are briefly explained here.

The sample coefficient of determination is developed from the relationship between two kinds of variation. The variation of the dependent variable Y in a data set around
1. The fitted regression line
2. Their own mean.

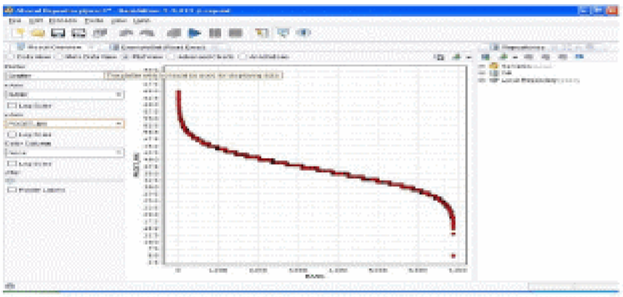
In our analysis, if we make the dependent variable to be the performance record, that is the score of the students, the independent variable to draw the regression line could be the question type , difficulty level, number of questions and so on.

In simple correlation analysis with a single variable, this could be studied with only one variable at a time to be the independent variable. In such a case the statistic used will be the sample of coefficient of determination, denoted as r^2 .The interpretation of r2 is that it measures only the linear relationship between two the variables. We can illustrate the perfect correlation by considering the following.

The following screen shot shows the distributions of the marks scored and the corresponding ranks in PGCET examinations, Karnataka for MCA exams 2012. The number of data items taken is over 6000.



The above chart shows the correlation between of the marks versus the ranks scored by the candidates. It can be observed that there are equal ranges of marks at several values of the ranks which indicate that the same ranks are shared between candidates scoring the same marks. This leads to the conclusion that only PGCET marks is not the exact indicator of the performance of the students. This leads to the inclusion of another parameter which could decide the ranks, namely the scores of the students in their qualifying degree. The same can be inferred from the graph for the same data when the covariance of the two given fields is generated. The graph is shown in the following fig.



Linear regression can also accommodate for multiple variables as independent variables. The dependency of the target variable on more than one independent variable can be studied. The independent variables could be simple variables having a linear relationship with the dependent variable or they themselves can be functions which vary based upon some other variable. For a normal relationship between the dependent and the independent variable, the regression equation would be

Y= a+b1x1+b2x2+b3x3+….where xi(i= 1 to n) denote the independent variables and bi(i= 1 to n)denote the regression coefficients. In our case Y would be the score card and xi would be the factors like number of questions, type of questions, strength of questions, previous performance and so on. When xi are independent variables, then it would be easy to calculate the values of bi and the resulting equation could be used to predict the future trends. In other words, we can establish a definite relationship between the scores of the students and the parameters on which it depends and evolve a predictive model to predict the performance in the coming years. When xi themselves are variants according to some other functions, linear regression methods cannot be used to develop a predictive model, then causal models and probabilistic have to be resorted to.

In a similar manner, we can consider multivariate analysis, where in we can employ the Annova methods. In this case, we can get two different estimates of the variance of the marks of the total population .We can do this by examining the variance among the three sample means of any three factors mentioned above.- and the other estimate being the variation within the individual metrics themselves. Thus we would be correlating many factors and the performance record .By analyzing them separately and using suitable deductions we can narrow down to as less number of independent factors as possible.

VI. ITEM ANALYSIS

The individual questions that are given have to be chosen in a manner that befits the context of the examinations and the student aptitude. In this regard there are two main aspects which determine the ability of an exam to measure the aptitude of a student. These are i) The quality of individual test items and ii) the number of test items. The parameters that could be of interest to study are the proportions of the students who choose a particular answer to the questions and the correlation between the probability of a student choosing his answer and the student's total score in the examination. These factors are grouped under item analysis.

Analysis of the proportions of the students answering a set of questions by selecting a particular choice among the answers would give an indication of the difficulty of the question, as well as the extent to which the questions designed to be distracters would be truly distracters. These data do not necessarily say whether a question is good or bad. They can give an indication as to whether a question designed to be trivial is truly trivial or whether a question designed to be difficult is truly difficult or impossible within the given constraints. It has been suggested that questions which are answered by 85% or more students or 25% or less number of students are of questionable validity.

The correlation between the probability of the student choosing a particular answer to a question and his score in the exam could provide useful information on the ability of that question to identify "strong " or " week" students.[1]In theory a student who answers a question correctly has a better ability to score well than one who answers it wrong. When the correlation coefficient for a correct answer is negative, there is something drastically wrong with the question .Either the wrong answer has been entered into the grading key or the question might be grossly misleading. Conversely if the correlation for a wrong answer should be negative and a positive correlation is disconcerting.

CONCLUSIONS

This paper discusses some of the statistical techniques which could be used in analyzing the performance of students in competitive exams. Although numerous statistical techniques can be used in these situations , the techniques discussed here are the important ones. Particularly regression analysis could be used in a variety of situations, basically to draw relationships between the previous marks and the current patterns generated. The limitation of these techniques is however that they are not causal models and cannot predict the causes behind a particular pattern of scores.

REFERENCES

[1] http://128.148.32.110/courses/cs227/archives/2001/groups/custom/papers/1996-Fayyad.pdf

[2] http://www.refactorthis.net/post/2013/06/30/RapidMiner-tutorial-How-to-explore-correlations-in-your-data-to-discover-the-relevance-of-attributes.aspx

[3] http://www.comp.dit.ie/btierney/oracle11gdoc/datamine.111/b28129/regress.htm