Data Mining: Exploring the Role of Discriminant Function Analysis

Dr . Ravindranath K . Murthy,
Assistant Professor
Department of Education,
Institute of Advanced Study in Education,
Osmania University,
Hyderabad, India,

***Abstract*— Data mining has recently emerged as interdisciplinary sub branch of computer sciences. It refers to the computational techniques of discovering relations or patterns in large data sets, to extract meaningful information from a data set and converting it or transforming it into comprehensible forms for further exploration. Thus Data mining is a collection of analytical techniques used to uncover new trends and patterns in massive databases. Discriminant Function Analysis is a popular multivariate data analysis technique used in statistics. This article explores how Discriminant funtion analysis serves as an important data mining technique in organizational research.**

Keywords*— *Data mining, Discriminant Function Analysis, Employee effectiveness

I. INTRODUCTION

Data mining has recently emerged as interdisciplinary sub branch of computer sciences. It refers to the computational techniques of discovering relations or patterns in large data sets, to extract meaningful information from a data set and converting it or transforming it into comprehensible forms for further exploration. Thus Data mining is a collection of analytical techniques used to uncover new trends and patterns in massive databases. Data mining is the process of selecting, exploring, and modelling large amounts of data to uncover new trends and patterns in massive databases. These analyses lead to proactive decision making and knowledge discovery in large databases by stressing data exploration to thoroughly understand and study the structure of data and to check validity of the model so designed

Discriminant Function Analysis is a popular multivariate data analysis technique used in inferential statistics. This article explores how Discriminant funtion analysis serves as a important data mining technique in organizational context.

II. DISCRIMINANT FUNCTION ANALYS

Discriminant Analysis (DA) is a multivariate statistical technique commonly used to build a predictive / descriptive model of group discrimination based on observed predictor variables and to classify each observation into one of the groups. In Discriminant function analysis, several metric variables measured atleast in an inerval scale are used to discriminate single classification variable. Discriminant Analysis is to be differentiated from cluster analysis. In cluster analysis prior knowledge of the classes, usually in the form of a sample from each class is required. The basic purpose of discriminant analysis is to estimate the relationship between a single nonmetric (categorical) dependent variable and a set of metric independent variables (Hair, Anderson, Tatham, & Black, 2003). Discriminant analysis determines how best different groups of people can be discriminated from each other in terms of a number of discriminating variables (independent variables). In Discriminant analysis the dependent variable is the categorical variable and the independent variables are the metric variables. The common objectives of Discriminant Analyis are 1) to investigate differences between groups 2) to discriminate groups effectively; 3) to identify important discriminating variables; 4) to perform hypothesis testing on the differences between the expected groupings; and 5) to classify new observations into pre-existing groups. Stepwise, canonical and discriminant function analyses are commonly used DA techniques available in the statistical softwares.

Generally in the organizational database the details of the employees on several variables (fields) are collected and recorded in regular intervals. If the details on employees working in different departments or sections of an organization is available on several variables and the objective is to examine the effectiveness of these several variables in discriminating employees on high job performance employees and low job performance employees then discriminant funtion analysis can be effectively used for this purpose.

There are two methods of computing discriminant analysis:
(1) Simultaneous method, or Direct method, and
(2) Stepwisemethod (Hair Anderson, and Tatham, 1990).

Of the several variables available on the employees in the database, if the objective is to determine, which variables are most efficient in discriminating between the two groups, then stepwise method of discriminant analysis may be selected

There are several different discriminant analysis procedures available in stepwise method. In IBM SPSS software the following different procedures are available 1) Wilk's Lambda, 2)Unexplained variance, 3) Mahalanobis distance, 4) Smallest F ratio, 5) Rao's V. While delineating

about the different methods in stepwise method, Hair, Anderson, and Tatham (1990) observed: "In general, Mahalanobis is the preferred procedure when one is interested in the maximal use of available information. The Mahalanobis procedure in the SPSS package performs a stepwise discriminant analysis that is similar to a stepwise regression analysis. The stepwise procedure is designed to develop the best one-variable model, followed by the best two-variable mode, followed by the best three variable model, and so forth, until no other variables, meet the desired selection rule. The selection rule in this procedure is to maximise Mahalanobis distance D^2 between groups" (p. 94). The step wise discriminant function analysis if performed results in different steps and at each step, the variable that maximizes the Mahalanobis distance between the groups (Low job performance group - High job performance group) is entered or removed.

To further test how well separated are the centroids of the groups, Wilks Lambda is used. Wilks Lambda also called the 'U' statistic is used for testing the equality ofgroup centroids. It is a multivariate analysis of variance test statistic that varies between 0 and 1. Small values indicate that the group centroids differ. Wilks Lambda is the proportion of the total variance in the discriminant scores not explained by differences among the groups. To test the significance of lambda, it is transformed to a variable with an approximate chi-square distribution.

III CONCLUSION

Discriminant function analysis serves as a good data mining technique, which fall under the multivariate dependence technique. In the organizational data base if the company or organization wants to examine the differences between low performing employees and high performing employees, on several variables then discriminant function analysis can be used very effectively

REFERENCES

[1] Glick N (1978) Additive estimators for probabilities of correct classification Pattern Recognition 10: 211-222.

[2] Hora S. C and Wilcox J.B. (1982) Estimation of error rates in several population discriminant analysis. Journal of. Marketing Research. 19 57-61.

[3] Hair, J.F., Black, W.C., Babin, B.J., Anderson, R., Tatham,R.L. (2009). Multivariate Data Analysis, New York: Person Education

[4] Khattree R. and Naik D.N , (1995) Applied Multivariate statistics with SAS software Cary NC . SAS Institute Inc

[5] Lachenbruch P. A and Mickey M.A. (1968) Estimation of error rates in discriminant analysis Technometrics 10 1-10.

[6] SAS Institute Inc. (1999) SAS/STAT Users Guide, Version 8, Cary NC .SAS Institute Inc