Data Mining Applications in Master Health Checkup: A Statistical Exploration

G. Manimannan* S. Hari** and G. Vijaythiraviyam**
* Assistant Professor
** Senior Post Graduate Student
Department of Statistics, Madras Christian College, Tambaram, Chennai.

Abstract

Data mining has been used exhaustively and widely by many organizations. In healthcare, data mining is becoming more and more popular, if not increasingly essential. Data mining applications can significantly benefit all parties involved in the healthcare industry. The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making.

This study explores data mining applications in healthcare. In particular, this area was chosen as a model to study Master Health Checkup (MHC). The data were collected from secondary source containing 295 patients in St. Johns Hospital, Bangalore. The case sheet deals with socio demographic characteristic, Blood Pressure, Fat, Liver and diabetic related parameters.

The salient feature of this study is the application of *Factor Analysis*, K-*means clustering* and *Multivariate Discriminant Analysis (**MDA**)* as data mining tools to develop the hidden structure present in the data. The scores from extracted factors are used to find initial groups by K-means clustering algorithm. A few outlier health care profiles, which could not be classified to any of the larger groups, are discarded as some of the parameters possessed higher values. Finally, DA is applied and the groups are identified as MHC patients belonging to **O**-Class (Obesity), **N**-Class **(**Normal**)** and **UW**-Class (Under Weight) in that order. The results of the study indicate that ***DA*** classification maps can be a feasible tool for the health care analysis of large amounts of master health checkup data.

Key Words: *Master Health Check (MCH), Data mining, Factor Analysis, K-means Clustering and Discriminant Analysis (DA)*

1. Introduction

The Master Health Check-up (MHC) is offered by various hospitals and medical research institute is a programme that attempts to reduce health care costs by prevention and early diagnosis. A variety of chronic diseases afflict us, most of which take their toll after the fifth decade of life. Diabetes, hypertension, heart attacks, stroke and cancer are some of the more common examples.

Almost all of these problems first go through a long quiescent phase where they produce no symptoms. This period can be as long as 10 - 20 years. It makes sense, therefore, that a programme that attempts to detect and correct these problems during this silent phase will decrease the ultimate morbidity from these diseases. In the early days of preventive health check-ups, every conceivable test and technology was ordered in the hope that some would be abnormal and provide an avenue of approach. A handful of items, mostly simple, appear to provide the greatest value.

The MHC offered at various hospitals and institutes is a carefully constructed programme that offers a panel of tests that are proven to be valuable. As an incentive to those who have taken the efforts to control their health problems, the programme also includes two or more follow up visits within a year of the MHC and the physician in charge of the checkup.

Good health is by itself of great value. It enhances market earnings by increasing the number of healthy days an individual has available for work (Grossman 1972) and increases non-market productivity, allowing more time for household production (Becker 1976). Health checkups help to secure and maintain good health.

The Master Health Check (MHC) is a series of tests to screen each functional area closely to detect even the smallest symptom of a major illness. It also helps to identify the reason for minor ailments, which are

constant. MHC is considered to be the most comprehensive prevention check. Master Health Check consists of five permanent packages, which are as follows: Master Health Check, Executive Health Check, Heart Check, Whole Body Check and Well Women Check (B. Krishan Reddy, G.V.R.K. Acharyulu, 2002). The main objective of this paper is to investigate whether;

(i) To identify the hidden patterns using Factor analysis for original MHC patients.

(ii) Data mining paradigms together with well known *unsupervised* learning classification model *K-mean clustering method* can be used to exhibit the classification of MHC patients and to cross validate the original classification using MDA.

The rest of the paper is organized as follows. Section 2 describes the methodology we have used, the database and the choice of MHC parameters. Section 3 presents the proposed algorithm which is used as a benchmark to achieve the objective on applying one of the well known statistical classification model MDA and Section 4 presents the empirical results. The conclusions of our study are presented in Section 5.

2 Methodologies and Database

This section brings out the discussion of the database, the MHC (Master Health Checkup) parameters selected and the Data Mining Techniques. The MHC data were collected from secondary source of OPD (Out Patients Department) containing 295 patients in St. Johns Hospital, Bangalore was considered as the database. The data mainly consists of five major categories, such as socio economic and demographic characteristic, Blood Pressure, Fat, Liver and diabetic related parameters. Among the listed patients, number of patients varied over the study period owing to removal of those patients for which the required data are not available or outliers.

2.1 Selection of Variables

In this study, 28 medical observations (parameters) were chosen among the many that had been used in MHC case sheets. These 28 medical observations were chosen to assess socio economic and demographic characteristic, Blood Pressure, Fat, Liver and diabetic. Some of them are given below.

Table 1 MHC Medical observations during the study period

Parameters	Description
BP_Syst	Blood Pressure Systolic
BP_Dias	Blood Pressure diastolic
Blood_Hb	haemoglobin
Blood_PCV	Packed Cell Volume
Blood_TC	Total Count
Diabetess_Fasting	Diabetes Fasting
Diabetes_Post Pran	Diabetes Post Prandial
Cholesterol	Cholesterol
FAT_VLD	High-density lipoprotein
FAT_LDL	Low-density lipoprotein
Liver_SAP	Alkaline phosphate
Liver_ALT	Alanine transaminase

3 Data Mining Techniques

Data Mining or Knowledge Discovery in Databases (KDD) is the process of discovering previously unknown and potentially useful information from the data in databases. In the present context data mining exhibits the patterns by applying few techniques namely, factor analysis, **k**-means clustering and Multivariate Discriminant Analysis (MDA)

As such KDD is an iterative process, which mainly consist of the following steps on the data collected;

Step 1: Data cleaning

Step 2**:** Data Integration

Step 3: Data selection and transformation

Step 4: Data Mining

Step 5: Knowledge representation

Of these above iterative process Steps 4 and 5 are most important. If suitable techniques are applied in Step 5, it provides potentially useful information that explains the hidden structure. This structure discovers knowledge that is represented visually to the user, which is the final phase of data mining.

3.1 Factor Analysis

Factor analysis provides the tools for analyzing the structure of the interrelationships (correlations) among the large number of variables by defining sets of variables known as factors. In the present study, factor analysis is initiated to uncover the patterns underlying MHC medical observations. Orthogonal rotations such as Varimax and Quartimax rotations are used to measure the similarity of a variable with a factor by its factor loading.

3.2 k-Means Clustering Methods

McQueen (1967) suggests the term k-means for describing an algorithm of his that assigns item to the cluster having the nearest centroid (mean). Generally this technique uses Euclidean distances measures computed by variables. Since the group labels are unknown for the data set, k-means clustering is one such technique in applied statistics that discovers acceptable meaningful classes.

3.3 Discriminant Analysis

Multivariate Discriminant Analysis is a multivariate technique using several variables simultaneously to classify an observation into one of several a priori groups. In the present study, discriminant analysis is used to exhibit groups graphically and judge the nature of overall performance of the MHC patients.

3.4 Algorithms

A brief step-by-step algorithm to classify the MHC patients during the study period based on their overall MHC is described below:For the pruned data set the following algorithm is proposed to scale the MHC patients and visualize them on a two-dimensional map during each of the study period based on their overall medical observations (*Table 1*). A brief step-by-step algorithm to classify the MHC patients during the study period based on their overall MHC is described below:

Step 1: Factor analysis is initiated to find the structural pattern underlying the data set.

Step 2: K –means analysis is used to partition the data set into k-clusters using the factor scores obtained in Step 1 as input.

Step 3: Discriminant analysis is then performed with the original MHC patients by considering the groups formed by the k-means algorithm.

4 Results and Discussion

Factor analysis is extended with the techniques of Varimax and Quartimax criterion for orthogonal rotation. Even though the results obtained by both the criterions were very similar, the varimax rotation provided relatively better clustering of MHC medical obervations.

Consequently, only the results of varimax rotation are reported here. We have decided to retain 65 percent of total variation in the data, and thus accounted consistently ten factors for MHC medical observations with eigen values little less than or equal to unity.

Table 2 shows variance accounted for each factors. In the following table we observe that the total variances explained by the extracted factors are over 66 percent, which are relatively higher (that is indicated by *).

Table 2 Percentage of variance explained by factors

Factors	Variance Explained
1	12.149
2	7.256
3	6.592
4	6.168
5	5.962
6	5.776
7	5.760
8	5.304
9	5.050
10	4.599
Total	65.616

Table 3 Financial Ratios in Rotated Factors

Initials	Measures	1	2	3	4
Blood_Hb	Blood-I	*			
Blood_PCV		*			
Blood_RBC		*			
Blood_ESR		*			
FAT_Creatinine		*			
Liver_AST	Liver		*		
Liver_ALT			*		
Liver_GGT			*		
Diabetes_Fasting	Diabetes			*	
Diabetes_Post Prandial				*	
FAT_VDL				*	
Blood_Albumin	Blood-II				*
Liver_SAP					*
Blood_Urine					*

Initials	Measures	5	6	7	8	9	10
BP_Syst	Blood Pressure	*					
BP_Dias		*					
Cholesterol	Cholesterol		*				
FAT_LDL			*				
FAT_VLD			*				
Blood_Platelet	Blood Count			*			
Blood_TC				*			
FAT_Acid				*			
Liver Tot. Protein	Protein and Hemoglobin				*		
Blood_MCHC					*		
Liver_Albumin					*		
Blood_MCV	Mean cell Volume					*	
Blood_MCH	Thyroid and hemoglobin						*
Blood_TSH							*

After performing factor analysis, the next stage is to assign initial group labels to MHC patients. Step 3 of the algorithm is explored with factor score extracted by Step 2, by conventional **k**-means clustering analysis. Formations of clusters are explored by considering 2-clusters, 3-clusters, 4-cluster and so on. Isolated groups with some MHC patients are discarded from the analysis as outliers. A few MHC for these outlier patients are comparatively high or low to those excelled in the analysis.

Out of all the possible trials, 3-cluster exhibited meaningful interpretation than two, four and higher clusters. Having decided to consider only 3 clusters, it is possible to classify MHC patients as Cluster **N**, Cluster **UW** or Cluster **O** depending on whether the MHC patients belonged to Cluster 1, Cluster 2 or Cluster 3 respectively.

Cluster 1 (Cluster **N**) is a group of MHC patients that have high values for the MHC parameters, indicating that these patients are normal. The **O** with lower values for the MHC medical observations are grouped into Cluster 3 (Cluster O). This suggested that Cluster 3 is a group of patients with low-profile. Cluster 2 (Cluster **UW**) are those patients which perform moderately well as compared to the Cluster 1 and Cluster 3.

Inspite of incorporating the results for MHC patients, only the summary statistics are reported in *Table 4*. The first column in *Table 4* provides the groupings done by cluster analysis. .

Table 4 Number of MHC Patients in the Clusters

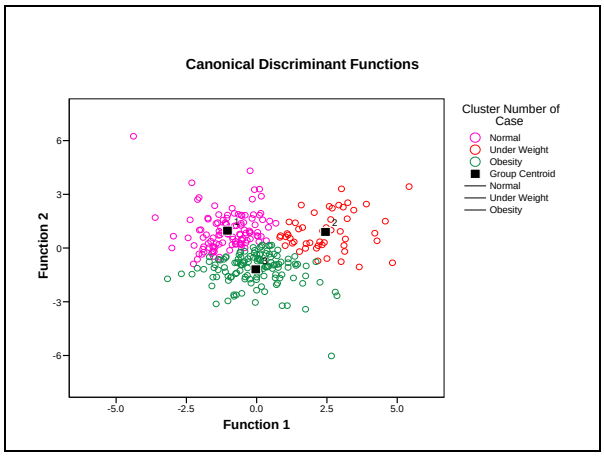
MHC Patients	k-Means			MDA		
	1	2	3	1	2	3
295	115	50	130	113	51	131

1– Cluster **N** 2 – Cluster **UW** 3 – Cluster **O**

Figures 2 show the groupings of MHC patients into 3 clusters for the study period. Patients in cluster 1 tend to be normal, cluster 2 tends to be under weight and cluster 3 tends to be obesity. We classify the members in the first cluster as Cluster **N**, the second as Cluster **UW** and the third as Cluster **O** in terms of MHC medical parameters.

The pruned data set is then subjected to the main algorithm as in *Section 3.4* to assign appropriate classes to the MHC patients. Initially, a MDA was trained separately, with the sequential procedure algorithm, using the SPSS. *Figures* 1 show the groupings of MHC patients into 3 clusters over the MDA map using the visualization method. In the following *Figures*, each colour (shades) represents classes of MHC patients.

Figure 1 Multivariate Discriminant Analysis (MDA)



5 Conclusion

The purpose of this paper is to explore the possibility to identify the meaningful groups of MHC patients that are scaled as the best with respect to their medical observations (parameters) using factor analysis, k-means and MDA classification techniques. Initially, factor analysis is used to identify the underlying structure based on 29 medical observations. The factor scores are used to partition the MHC patients into different clusters by using **k**-means clustering algorithm.

The present analysis has shown that only 3 groups could be meaningfully formed for all the data. This indicates that only 3 types of patients existed over a study period. Further, the MHC patients find themselves classified into *Normal* (Cluster **N**), *Under Weight* (Cluster **UW**) and *Obesity* (Cluster **O**) categories depending on certain medical observations. A generalization of the results is under investigation to obtain an incorporated class of 3 groups of MHC patients for any study period.

References

[1] Becker, Gary S., ed. 1976. A theory of the allocation of time. In *The economic approach to human behavior, 89–114. Chicago: University of Chicago Press.*

[2] Grossman, Michael. 1972. On the concept of health capital and the demand for health. *Journal of Political Economy 80 (2): 223–55.*

[3] Hian Chye Koh and Gerald Tan , Data Mining Application in Health Care *Journal of Healthcare Information Management — Vol. 19, No. 2*

[4] B.Krishan Reddy, G.V.R.K. Acharyulu, (2002) Customer Relationship Management (CRM) in Health Care Sector - A Case Study on Master Health Check, *Journal of the Academy of Hospital Administration,* Vol. 14, No. 1 . (2002-01 - 2002-06).

[5] J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297.

[6] G. Manimannan, S. Hari and G. Vijay Thiraviyam (2012), Data Mining Applications in Master Health Checkup: a Statistical Exploration, Paper presented at *National Conference on Statistics for Twenty First Century-2012 (NCSTC-2012)*, Department of Statistics, University of Kerala, Trivandrum, Kerala, India.

[7] Richard A Johnson and Dean W Wichern (1992), *Applied Multivariate Statistical Analysis,* 3/ed, Prentice-Hall of India Private Limited, New Delhi.

Author Profile

G. Manimannan received his M. Sc. M. Phil. Ph. D in Statistics from University of Madras, Chennai, India during period 1997, 1999 and 2006. He received PGDCA (Post Graduate Diploma in Computer Application) during period 2001 – 2003 from Pondicherry University, Pondicherry, India. Now he is working as Assistant Professor, Department of Statistics, Madras Christian College, Chennai, India from 2006 until now. He gets good experience by working for many Project Guidance and consultation work in application of Statistics. He was published more than fifteen research papers in various national and International journals. He has good experts in many programming languages like, FoxPro, HTML, COBOL, C, C++, VB, RDBMS, SPSS, SYSSTAT, STATISTICA, MINITAB, MATLAB and working knowledge in SAS and R.

S. Hari received his B. Sc. in Statistics from University of Madras during the period of 2008-2011. Now he is a candidate on M.Sc. degree in Statistics from University of Madras from June 2011 until now. He has good experts in many programming languages like C, C++, Visual Basic, DBMS, SPSS and working knowledge in MATLAB.

G. Vijaythiraviayam received his B. Sc. in Statistics from University of Madras during the period of 2008-2011. Now he is a candidate on M.Sc. degree in Statistics from University of Madras from June 2011 until now.

He has good experts in many programming languages like C, C++, Visual Basic, DBMS, SPSS and working knowledge in MATLAB.