International Journal of Scientific and Innovative Mathematical Research(IJSIMR)
Volume 8, Issue 2, 2020, PP 29-36
ISSN No. (Online) 2454-9444
DOI: http://dx.doi.org/10.20431/2347-3142.0802004
www.arcjournals.org

Evaluation and Classification of Master Health Checkup Database using Data Mining Techniques

R. LAKSHMI PRIYA[2], MANIMANNAN G[1*]

[1]*Assistant Professor, Department of Mathematics, TMG College of Arts and Science, Chennai*

[2]*Assistant Professor, Department of Statistics, Dr. Ambedkar Govt. Arts College, Vyasarpadi, Chennai.*

Corresponding Author:** ***MANIMANNAN G, *Assistant Professor, Department of Mathematics, TMG College of Arts and Science, Chennai*

Abstract: *The intention of this paper is to explore the possibility of identifying meaningful groups of MHC database that are scaled as the best with respect to their medical observations (parameters) using Self Organizing Map (SOM). Initially, k means clustering is used to identify underlying groups based on 29 medical parameters and cross validate the derived clusters using SOM. The next stage of this research paper is to analyze the MHC database and achieved that only 3 groups could be meaningfully formed for all the data. This indicates that only 3 types of patients existed over the 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 given samples.*

Keywords: *Self Organizing Map, k-Mean Clustering, Classification, Data Mining and Master Health Checkup.*

1. INTRODUCTION

In recent years particularly in India, Master Health Checkup (MHC) is becoming more familiar in almost all government and private hospitals. The MHC is now available in different types of packages. This type of medical checkup is must now days, because of food habits, change in climatic condition, stress, consuming more of junk foods, etc. In this research paper the primary data were collected from a private hospital and analyzed using data mining techniques.

The Master Health Checkup (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 checkup. Master Health Checkup consists of five permanent packages, which are as follows: Master Health Checkup, Executive Health Checkuup, Heart Checkup, Whole Body Checkup and Well Women Checkup (B. Krishan Reddy, G.V.R.K. Acharyulu , 2002).

Non-invasive measurement of blood pressure (BP) using cuff-based methods provides adequate data for many applications in medicine. However, cuff-based methods have some disadvantages, which limit their use in certain clinical situations. First, a continuous measurement of blood pressure is not possible, since a pause of at least 1–2 min between two BP measurements is necessary to avoid errors in measurement. Therefore, short-term changes in BP cannot be detected. Furthermore, the inflation of the cuff may disturb the patient and the consequences of these disturbances are alterations of BP. The blow up of the cuff leads to an arousal going along with an increase in the systemic blood pressure. Hence, in this case, blood pressure measurement may result in false-positive values**.**

2. REVIEW OF LITERATURE

Many research scholars are working in the filed of MHC database. Most patents have explicit desires when they visit their physicians. Identification of patients request and need is starting points of a patient centred approach to care. Number of subjects associated with patient satisfaction was examined for 243 patients with chronic diseases in general medicine clinics of the Department of Veterans Affairs Hospital. An average of 67% was meeting their expectations.

In “A study of patient’s expectation and satisfaction in Singapore hospital”, it is described that in today’s competitive healthcare environment, hospitals increasingly realize the need to focus on service quality

as a means to improve their competitive position. Customer based determination and perception of service quality therefore play an important role when choosing the hospital. An analysis covering 252 patients revealed that there was an overall service quality gap between patients expectation and perception. Thus improvements are required across the entire six dimensions, namely, tangibility, reliability, responsiveness, assurance, empathy, accessibility and affordability (Lim PC and Tang NK, 2000)

There are at least two additional benefits of health checkups that will be important in analysis of demand for these checkups. First, a checkup will likely give an individual a more objective diagnostic health analysis, in addition to his or her own subjective evaluation of health, made under un-certainty. Secondly, health checkups lead to further demand for preventive medical care when necessary. Early medical care often curtails serious illness. In this aspect, demand for health checkups differs from the demand for health. The former is a derived demand, and the latter is a final demand. That is, health checkups appear in demand for health, which in turn appears in individual utility function (Grossman 2000).

In particular, individuals demand more health information as age increases (Kenkel 1990). Time costs are also major determinants of the demand for health checkups, which exhibits larger time-price elasticity than the demand for other medical inputs (Coffey 1983). Income has a positive effect on demand for preventive medical care (Kenkel 1994). A better knowledge of one's own health also increases the demand for preventive medical care (Hsieh and Lin 1997). However, better health gives individuals less incentive to collect health information. Furthermore, lack of knowledge about health leads individuals to adopt unhealthy consumption patterns (Kenkel 1991). Thus, uncertainty plays a vital role in determining the demand for health checkups, as well as the demand for health itself (Arrow 1963). In the present context the problem of MHC patients has been studied, without making any assumptions with regard to the number of groups or any other structural patterns in advance, which reflected the classification of patients based on certain medical observations (G. Manimannan, S. Hari and G.Vijaythiraviyam).

3. 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 460 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.

3.1. Selection of Variables

In this study, 29 medical observations (parameters) were chosen among the many that had been used in MHC case sheets. These 29 medical observations were chosen to assess socio economic and demographic characteristics, Blood Pressure, Fat, Liver and diabetics. Some of them are given below.

Parameters	Description
Blood_Hb	Hemoglobin
Blood_PCV	Packed Cell Volume
Blood_RBC	Red Blood Cells
Blood_ESR	Erythrocyte Sedimentation Rate
Liver_ALT	Alanine Transaminase
FAT_Creatinine	FAT_Creatinine
Liver Albumin	Liver Albumin
Liver_GGT	Gamma-glutamyl Transpeptidase
FAT_Acid	Fatty Acid
FAT_VLDL	Very low density lipoprotein
BP_Dias	Blood Pressure diastolic
Diabetes_Post Prandial	Diabetes Post Prandial
Diabetess_Fasting	Diabetes Fasting
Cholestral	Cholesterol
Blood_TC	Total Count
Liver_AST	Aspartate Aminotransferase

FAT_LDL	Low-density lipoprotein
Blood_MCH	Mean Corpuscular Hemoglobin
Blood_Platelet	Platelet
Blood_MCV	Mean Corpuscular Volume
Diabet_Sugar	Diabetes Sugar
Blood_Albumin	Blood Albumin
BP_Syst	Blood Pressure Systolic
Blood_Urine	Blood Urine
Blood_TSH	Thyroid Stimulating Hormone
FAT_HDL	High Density Cholesterol
Blood_MCHC	Mean Corpuscular Hemoglobin concentration
Liver Total Protein	Liver Total Protein
Liver_SAP	Alkaline Phosphates

The main objective of the study is to find the hidden pattern of signs and symptoms of patients based on their medical examinations using Self Organizing Map (SOM).

3.2. Data Mining Techniques

Data mining has attracted a great deal of attention in information industry and in society as a whole in recent year, due to wide availability of huge amounts of data and imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging market analysis, fraud detection, and customer retention to production control and science exploration.

Data Mining or Knowledge Discovery in Databases (KDD) is the process of discovering previously unknown and potentially useful information from the data in databases. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, pattern analysis, data archaeology, and data dredging.

Many researcher treat data mining as a synonym for another popularly used term, Knowledge Discovery from data, or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery. Knowledge discovery as a process is depicted in figure 1 and consists of iterative sequence of the following steps.

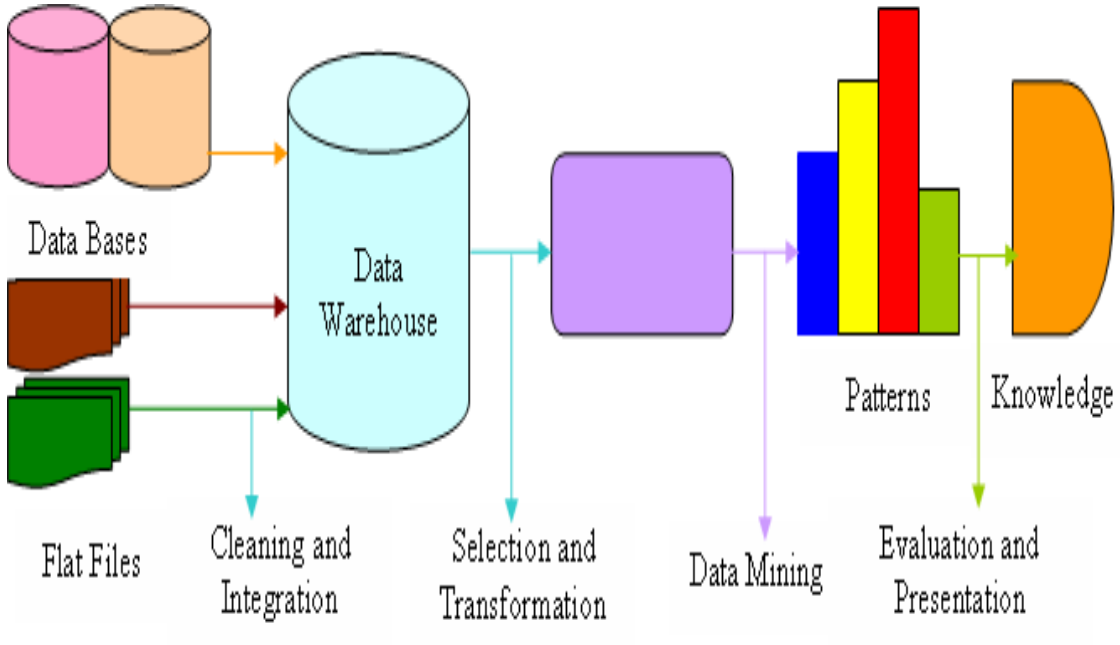

Fig1. *Data mining as a Step in the Process of Knowledge Discovery*

Step 1*: *Data cleaning - to remove noise and inconsistent data.

Step 2: Data integration - where multiple data sources may be combined. A popular trend in information industry is to perform data cleaning and data integration as a preprocessing step, where the resulting data are stored in a data warehouse.

Step 3: Data selection - where data relevant to the analysis task are retrieved from the Database.

Step 4: Data transformation - where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations. Sometimes data transformation and consolidation are preformed before data selection process, particularly in case of data warehousing. Data reduction may also perform to obtain a smaller representation of the original data without sacrificing its integrity.

Step 5: Data mining – an essential process where intelligent methods are applied in order to extract data patterns.

Step 6: Pattern evaluation - to identify the truly interesting patterns representing knowledge based on some interesting measures.

Step 7: Knowledge presentation – where visualizing and knowledge representation techniques are used to present the mined knowledge to the user.

Steps 1 to 4 are different forms of data preprocessing, where the data are prepared for mining. The data mining step may interact with the user or knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base. Note that according to this view, data mining is only one step in the entire process, although an essential one because it uncovers hidden patterns for evaluation (Jiawei Han and M. Kamber, 2006).

Of these above iterative process Steps 5 and 7 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 user, which is the final phase of data mining. In the present context data mining exhibits the patterns by applying two techniques namely, factor analysis and ***k***-mean clustering techniques.

4. SELF-ORGANIZING MAP (SOM)

The nets discussed in this chapter use Kohenen learning approach. In this learning, the units update their weights by forming a new weight vector that is a linear combination of old weight vector and current input vector. The unit whose weight vector is closest to the input vector is allowed to learn.

Like a codebook vector in vector quantization, the model is then usually a certain weighted local average of the given data items in data space. But in addition to that, when models are computed by SOM algorithm, they are more similar at the nearby nodes than between nodes located farther away from each other on the grid. In this way the set of models can be regarded to constitute a similarity graph, and structured 'skeleton' of distribution of the given data items.

SOM was originally developed for visualization of distributions of metric vectors, such as ordered sets of measurement values or statistical attributes, but it can be shown as a SOM-type mapping that can be defined for any data items, the mutual pairwise distances of which can be defined. Examples of non-vectorial data that are feasible for this method are strings of symbols and sequences of segments in organic molecules (Kohonen and Somervuo 2002).

SOM is an unsupervised learning procedure based on artificial neural network. It is very effective and frequently used network popularly known as ***Kohonen's*** neural network. These networks have only two layers, a standard input layer and an output layer known as the ***Competitive*** (Kohonen) layer. Each input neuron is connected to each and every neuron on the competitive layer which are organized as a two dimensional grid. Each MHC medical observations is associated with exactly one neuron whereas each neuron may have one or more medical observations attributed to it. This grid map enables to discover statistical regularities in its input space and develops different modes of behavior to represent different classes of input databases.

4.1. Network Architecture

The network is created from a 2D lattice of 'nodes', each of which is fully connected to the input layer. Figure 2 shows a very small Kohonen network of 4 X 4 nodes connected to the input layer (shown in green) representing a two dimensional vector.

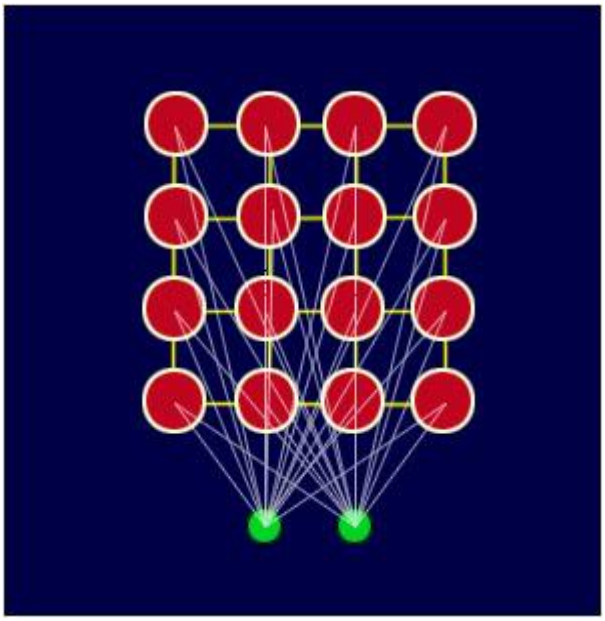

Figure2. *A Simple Kohenen Network*

Each node has a specific topological position (an x, y coordinate in the lattice) and contains a vector of weights of the same dimension as the input vectors. That is to say, if the training data consists of vectors, *V*, of *n* dimensions, then each node will contain a corresponding weight vector *W*, of *n* dimensions: The lines connecting the nodes in Figure 2 are only there to represent adjacency and do not signify a connection as normally indicated when discussing a neural network. There are no lateral connections between nodes with the lattice.

4.2. Training Algorithm

Initially, the weights and learning rate are set. The input vectors to be clustered are presented in the network. Once the input vectors are given, based on the initial weights, the winner unit is calculated either by Euclidean distance method or sum of products method. Based on the winner unit selection, the weights are updated for that particular winner unit using competitive learning rule. An epoch is said to be completed once all the input are presented to the network. By updating the learning rate, several epochs of training may be performed.

Step 1: Set topological neighborhood parameters, set learning rate and initialize weights.

Step 2: While stopping condition is false, do step 3-9

Step 3: For each input vector x, do steps 4-6

Step 4: For each j, compute squared Euclidean distance,

$$D(j) = \sum \left(w_{ij} - x_i\right)^2 \quad i = 1 \, to \, n \text{ and } \quad j = 1 \, to \, m$$

Step 5: Find index j, when $D(j)$ is minimum

Step 6: For all units j, with the specified neighborhood of j, and for all i, update the weights.

$$w_{ij(new)} = w_{ij(old)} + \alpha \left[x_i - w_{ij(old)}\right]$$

Step 7: Update the learning rate.

Step 8: Reduce the radius of topological neighborhood at specified times.

Step 9: Test the stopping condition

The map formation occurs in two phases:

Initial formation of perfect (correct) order

Final Convergence

The second phase takes a longer duration than the first phase and requires a small value of learning rate. The learning rate is a slowly decreasing function of time and the radius of the neighborhood around a cluster unit also decreases as the clustering process goes on. The initial weights are assumed random values. The learning rate is updated by, $\alpha(t+1) = 0.5\,\alpha(t)$ (S. N. Sivanandam et.al.2012).

4.3. Proposed 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).

Step 1: A map of weight vectors with 460 x 11 neurons having hexagonal topology for neighbourhood is obtained using SOMPAK.

Step 2: Factor analysis is initiated to find the structural pattern underlying the data set.

Step 3: K –means analysis is used to partition the data set into k-clusters using the factor scores obtained in Step 2 as input.

Step 4: Construct a SOM using the prototype vector with appropriate hits of the MHC patients in the data set that are assigned group labels in step 3.

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 belongs 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 called Obesity. This suggests 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 3and are called as Under weight.

Inspite of incorporating the results of MHC patients, only the summary statistics are reported in Table 1. The first column in Table 1 provides the groupings done by cluster analysis. .

Table1. *Number of MHC patients in the Clusters*

MHC Patients	k-Means			SOM		
	1	2	3	1	2	3
460	321	68	71	325	64	71

1– Scale N 2 – Scale UW 3 – Scale O

Figure 2 shows 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 MHC patients. Initially, a SOM was trained separately, with the sequential procedure algorithm, using the SOM Toolbox version 2 for MATLAB. In the construction process several maps are initialized and trained by considering the Gaussian neighborhood function and a map with hexagonal topology. Among the prototype vector maps, the best ones in respect of average quantization error are carefully chosen to explore the data in different dimensions. From the study it is found that a weight vector map units of size 295 by 11 neurons is well spanned within the data set as in Figure 3.

In addition, SOM is used efficiently in data visualization due to its ability to represent the input data in two dimensions. Among the various visualization techniques the most widely used method for visualizing the cluster structure of SOM is the distance matrix technique, especially the unified distance matrix (U-Matrix). For the present study, the method of displaying the number of hits in each map unit is well-thought off. Figure 3 show the groupings of MHC patients into 3 clusters over the SOM grid using the visualization method. In the following Figures, each colour (shades) represents classes of MHC patients.

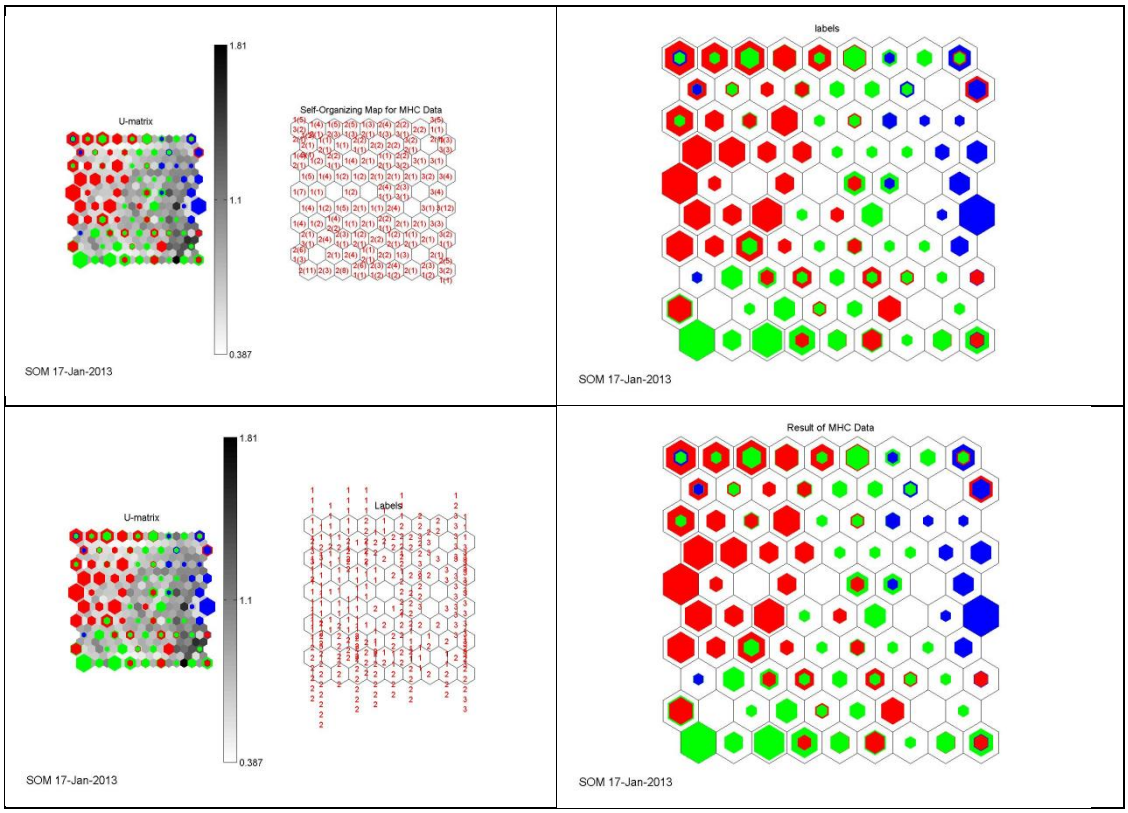

Fig3. *Self-Organizing Map (SOM)*

5. CONCLUSION

The purpose of this paper is to explore the possibility of identifying meaningful groups of MHC patients that are scaled as the best with respect to their medical observations (parameters) using SOM and few related 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.

MHC patient's data is mined by choosing random weight vectors in a map of size 460 x 11 neurons with different training parameters and a final map with the least quantization error is obtained using SOM Toolbox version 2. An attempt is made for Scaling of MHC patients based on certain medical observations by applying the proposed algorithm. The ability of SOM provides several methods of representing the high-dimensional vectors to be projected to a low dimension preserving the inter-sample distances as faithfully as possible and essential topological properties. In the present study, unified distance matrix with corresponding hits of data to prototype vectors are illustrated. This enables us to visualize the clustering of MHC patients grouped into different categories of patients on a two dimensional map unambiguously. Hence, the members of Cluster 1 are labeled as Score N (Normal). Similarly, the Cluster 2 includes patients which performed moderately well being scored as UW (Under Weight) and the Cluster 3 with low-profile MHC patients with score O (Obesity).

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 (Scale N), Under Weight (Scale UW) and Obesity (Scale 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] Arrow (1963), Uncertainty and the welfare economics of medical care. *American Economic Review* 53 (5): 941–73.
[2] Becker (1976), A theory of the allocation of time. In *The economic approach to human behavior,* 89–114. Chicago: University of Chicago Press.
[3] Coffey (1983), The effect of Time price on the demand for medical-care services. *Journal of Human Resources* 18 (3): 407–24.
[4] Grossman (2000)., The human capital model of the demand for health. In *Handbook of health economics*, vol. 14, ed. A. J. Culyer and J. P. Newhouse, 347–408. Amsterdam:
Elsevier Science.
[5] Han J. and M. Kamber,(2006), Data Mining Concepts and Techniques, Second Edition, Academic Press, USA
[6] Hsieh and Lin (1997), Health information and the demand for preventive care among the elderly in Taiwan. *Journal of Human Resources* 32 (2): 308–33.
[7] Joos S.K., Hickam DH (1993), Patient's desires and satisfaction in general Medicine clinics, 108(6), pp 752-59, Public Health Report.
[8] Kenkel (1991), Health behavior, health knowledge, and schooling. *Journal of Political Economy* 99 (2): 287–305.
[9] Krishsan Reddy B., G.V.R.K. Acharyulu (2002), Customer Relationship Management in Health Care Sector – A Case Study on Master Health Check, Journal of Academic of Hospital Administration, India, Vol. 14, No. 1.
[10] Lim PC and Tang NK (2000), A study of patient's expectation and satisfaction in Singapore Hospital", International Journal Health care quality Assurance Inc. Leadership Health Services, 13(6-7): pp-290-99.
[11] Manimannan G., S. Hari and G. Vijaythiraviyam (2013), Data Mining Applications in Master Health Checkup: A Statistical Exploration, International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 2.
[12] Kohonen and Somervuo (2002). How to make large self-organizing maps for nonvectorial data. Neural Networks 15(8-9): 945-952.

Citation: *MANIMANNAN G & R. LAKSHMI PRIYA, Evaluation and Classification of Master Health Checkup Database using Data Mining Techniques, International Journal of Scientific and Innovative Mathematical Research (IJSIMR), vol. 8, no. 2, pp. 29-36, 2020. Available : DOI: http://dx.doi.org/10.20431/2347-3142.0802004*