Clustering as Data Mining Technique in Risk factors Analysis of Diabetes, Hypertension and Obesity.

Mohammed G. Ahamad, Mohammed F. Ahmed and Mohammed Y. Uddin

Abstract—This investigation explores data mining using open source software WEKA in health care application. The cluster analysis technique is utilized to study the effects of diabetes, obesity and hypertension from the database obtained from Virginia school of Medicine. The simple k-means cluster techniques are adopted to form ten clusters which are clearly discernible to distinguish the differences among the risk factors such as diabetes, obesity and hypertension. Cluster formation was tried by trial and error method and also kept the SSE as low as possible. The SSE is low when numbers of clusters are more. Less than ten clusters formation unable to yield distinguishable information. In this work each cluster is revealing quit important information about the diabetes, obesity, hypertension and their interrelation. Cluster 0: Diabetes ∩ Obesity ∩ Hypertension = Healthy patient, Cluster 1: Diabetes ∩ Obesity ∩ Hypertension = Healthy patient, Cluster 2: Diabetes ∩ Obesity ∩ Hypertension = Obesity, Cluster 3: Diabetes ∩ Obesity ∩ Hypertension = Patients with Obesity and Hypertension, Cluster 4: Boarder line Diabetes ∪ Obesity ∪ Hypertension = Sever obesity, Cluster 5: Obesity ∩ Hyper tension ∩ Diabetes = Hypertension, Cluster 6: Border line obese ∩ Border line hypertension ∩ Diabetes = No serious complications, Cluster 7: Obesity ∩ Hypertension ∩ Diabetes= Healthy patients, Cluster 8: Obesity ∩ Hypertension ∩ Diabetes= Healthy patients, and Cluster 9: Diabetes ∪ Hyper tension ∪ Obesity = High risk unhealthy patients

Index Terms— Data mining; Diabetes; Hyper Tension; Obesity; Simple k-Means Clusters.

I. INTRODUCTION

The data mining assumes to be a vital technique in the medical diagnosis and health care. In the present study simple k-means clustering technique of WEKA is utilized to identify the risk factors of obesity, hypertension and diabetes from the diabetes data provided by Dr. John Schorling, Department of Medicine, University of Virginia school of Medicine [18]. The clustering algorithm detects the cognitive patterns of various risk indices of obesity, hypertension and CVD. A cluster, defines similar group of objects or indices making meaning full segmentation of data. Two broad categorization of cluster analysis exists. 1) Hierarchical 2) Non Hierarchical cluster techniques. In case of hierarchical clustering two types distance measures are used. One is Euclidean distance another is squared Euclidean distance. In the present case nonhierarchical clustering such as k-means clustering method is adopted. In this method, the number of clusters to be formed from the objects is to be specified initially. The simple k-means algorithm is an unsupervised machine learning cluster algorithm from WEKA tool.

Frank et al, explored the Weka machine learning workbench general purpose environment for automatic classification, regression and feature selection, specifically for data mining problems in bioinformatics [1]. Md Zahidul Islam and Ljiljana Brankovic, in their work, proposed a framework for privacy preservation data mining technique using clustering method[2] Sanker Rajagopal in his business investigation identified the high profit, high value and low risk customers through clustering technique using IBM –miner[3]. Rakesh Kumar Arora, and Dharmendra Badal proposed admission management of quality students using Weka cluster analysis [4]. Murlidher Mourya, Phani Prasad studied an effective execution of diabetes dataset using weka. The various cluster techniques are examined [5]. Samir Kumar Sarangi and Dr. Vivek Jaglan, integrated cluster and classification techniques to analyze the large diabetes data set, to interpret the performance of various cluster algorithms [6]. Suman and Pooja Mittal, compared the performance of different cluster algorithms of weka using the educational data set [7]. T. Soni Madhulatha, in her investigation, discussed the clustering algorithms, benefits, applications and limitations [8]. M. Durairaj, and C. Vijitha while predicting the students’ performance, used weka data mining through clustering, which paved way to strategic management tool [9]. Sanjay Chakraborthy, NK Nagwani, implemented the incremental k-means Clustering algorithm, in performance evaluation of pollution data base [10]. D. S.V.G.K.Kaladhar et al.while making inferences and predictions from colon cancer data, used the weka data mining tool and its cluster techniques [11]. Narendra Sharma et al discussed the comparision of various clustering algorithms of weka tool [12]. Jyoti Agarwal et al, carried out k-means cluster analysis on the crime data set using rapid miner tool [13]. Amitava Karmaker, and Syed M. Rahman, successfully implemented outlier analysis through cluster analysis on the spatial data [14]. Yi-Hong Lu, and Yan Huang, surveyed various clustering algorithms such as, k-means, grid based and regression analysis and compared the performance of these three algorithms [15]. Mohammad

Published on December 30, 2016.
M. G. Ahamad is Professor in College of Computer Engineering, Department of Computer Engineering. Prince Sattam bin Abdulaziz University, Saudi Arabia (e-mail:mg.ahmad@psau.edu.sa).
F. Ahmed, Network Engineer LBNco Ltd, Sydney, Australia (e-mail:faisal.ahmed519@gmail.com).
Y. Uddin Lecturer, Information Systems department Prince Sattam bin Abdulaziz University Al Khair kingdom of Saudi Arabia (e-mail:m.yousuf@psau.edu.sa).
Khubeb Siddiqui and Shams Naahid, analyzed the KDD CUP 99 data set using clustering technique, to detected the protocols and attack types by the hackers [16]. In view of the scattered outcomes of the various studies reported above, an attempt is made for systematic study of the diabetes, CVD, and obesity and the correlation of these indices using cluster technique of data mining using the open software WEKA. Many interesting features of these indices and their interrelations are found.

Simple K-means algorithm: Initially it sorts preset k of the objects, which is a cluster mean. The sorting or clustering mechanism assign the remaining objects to the cluster to which it is most similar, based on cluster mean. The iteration process continues until the set condition is zeroed. The algorithm is described as below [17].

1. Initial cluster seed value is set. This is a temporary means of clusters.
2. The squared Euclidean distance from each instance is computed, and accordingly assigned to nearest cluster.
3. For each cluster, a fresh centroid is generated and each seed value replaced by the respective cluster centroid.
4. The squared Euclidean distance from each instance is generated, and the instance is assigned to the cluster minimum squared Euclidean distance.
5. The cluster centriodes are freshly estimated based on the new membership assignment.
6. Steps 4 and 5 are recycled and the process comes to end with no attribute is added to clusters.

II. RESULTS

The diabetes data set is obtained from courtesy of Dr. John Schorling, Department of Medicine, University of Virginia School of Medicine. This data consists of 19 variables on 403 subjects from 1046 patients. This data set is subjected to study various risk factors using the free open source WEKA data mining tool. The Data is not only consists of diabetes complaints, but also obesity and cardiovascular complaints. The present study is to classify the data to identify the various complaints. The clustering technique is used to categorize the data. The results are reflecting the features of obesity, cardiovascular, and diabetes features. The results are compared with the standard prototype values recorded in table-2 for the purpose of interpretation. In this cluster classification using WEKA, version10, clusters are formed, as shown in Figure-1. The number of clusters below 10 results in merging the other clusters, and the resulted clusters couldn’t give the clear clarity of the nature of clusters. After trying with trial and error methods, it is decided to form 10 clusters. The sum of squared error (SSE) should be minimum for better clustering. In the present case with 10 clusters the SSE resulted to be 278.57. The cluster mode selected here is using training set. In this mode 90 % data is used for experimentation and 10 % is set for training purpose. As a result most of the data is used for experimentation. The time taken to build the model with full training data is 0.17 sec. There are sufficient numbers of output instances in each cluster and this is recorded in the Table 2.

### TABLE I: STANDARD PROTOTYPE VALUES

<table>
<thead>
<tr>
<th>SNO</th>
<th>Attribute</th>
<th>Target values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cholesterol (Chol)</td>
<td>&lt;200</td>
</tr>
<tr>
<td>2</td>
<td>Stable Glucose (Stb.glu)</td>
<td>FBS&lt;110, PBS&lt;140</td>
</tr>
<tr>
<td>3</td>
<td>High Density Lipoproteins (HDL)</td>
<td>&gt;59</td>
</tr>
<tr>
<td>4</td>
<td>(Cholesterol/HDL) Ratio</td>
<td>&lt;4</td>
</tr>
<tr>
<td>5</td>
<td>Glycosylated</td>
<td>4 – 6%</td>
</tr>
<tr>
<td>6</td>
<td>Body Mass Index =BMI</td>
<td>BMI= Weight(Lbs)X703/Height(in)^2 &lt;24.9</td>
</tr>
<tr>
<td>7</td>
<td>Blood Pressure Systolic (Normal)</td>
<td>&lt; 120</td>
</tr>
<tr>
<td>8</td>
<td>Blood Pressure Diastolic (Normal)</td>
<td>&lt; 80</td>
</tr>
<tr>
<td>9</td>
<td>Blood Pressure Systolic Stage1</td>
<td>140-159</td>
</tr>
<tr>
<td>10</td>
<td>Blood Pressure Diastolic Stage1</td>
<td>90-99</td>
</tr>
<tr>
<td>11</td>
<td>Blood Pressure Systolic Stage 2</td>
<td>=&gt;160</td>
</tr>
<tr>
<td>12</td>
<td>Blood Pressure Diastolic Stage2</td>
<td>=&gt;100</td>
</tr>
<tr>
<td>13</td>
<td>Waist</td>
<td>40 (inches) or 102 cm (Men), 35(inches) or 88 cm (Women)</td>
</tr>
<tr>
<td>14</td>
<td>Waist to Height Ratio</td>
<td>0.5</td>
</tr>
<tr>
<td>15</td>
<td>Waist to Hip Ratio</td>
<td>0.90 – 0.95 (Men), 0.80 – 0.086 (Women)</td>
</tr>
<tr>
<td>16</td>
<td>Time.ppm</td>
<td></td>
</tr>
</tbody>
</table>

III. DISCUSSION

The simple k-means cluster analysis of diabetic data presented above depicted quiet interesting results. Not only the individuality of parameters, but also their interdepends are quiet interesting. The individual clusters are described below.
Cluster 0: In this cluster all parameters are normal when compared with the prototype values of table 2. This is a female group with 41 years of age. The Waist to height ratio is 0.47, and waist to hip ratio is 0.88. The BMI is 22.6, and small frame. These four parameters indicate that this cluster group of patient is not obese. The glyhb and stable glucose levels are 4.8 and 196 respectively are indicative of diabetes negativity. The cholesterol, the high density lipoprotein (hdl), and the blood pressure levels are well under control, interpreting negativity of heart alignments. The cluster results also shown in cluster 0 histogram.

\[ \text{Diabetes} \cap \text{Obesity} \cap \text{Hypertension} = \text{Healthy patient} \]

Cluster 1: This is a male group of cluster with the age of 39 years. The waist to height ratio, waist to hip ratio and BMI parameters are 0.5, 0.86 and 23.8 respectively are indicative of non-obese. The cholesterol, hdl and blood pressure levels are under control making this cluster negative to heart diseases and hypertension. The diabetic parameters such as stable glucose and glyhb are under controllable range, which states the negativity to diabetes. This is shown in figure 3 histogram. This is related mathematically.

\[ \text{Diabetes} \cap \text{Obesity} \cap \text{Hypertension} = \text{Healthy patient} \]

Cluster 2: This cluster includes male patients with age raised to 55 years. The cholesterol (208), hdl (43.4), chol/hdl ratio (5.2) and blood pressure values exceeded the safe limits indicated to positivity to heart disease and hypertension. The stable glucose (120) and glyhb (6.2) values exhibit border line diabetes. The BMI (28), waist to height ratio (0.565), waist to hip ratio (0.91). Though the waist to hip ratio is optimum, waist to height ratio is nominal high and BMI is sufficiently high predicting the obesity. Thus this cluster group is prone to heart diseases, obesity, and marginal diabetes and advised to undergo treatment for these alignments simultaneously.

\[ \text{Diabetes} \cup \text{Obesity} \cap \text{Hypertension} = \text{Patients with Obesity and Hypertension} \]
Cluster 4: This cluster presents female patients with the age of 51 years. The BMI (36), waist to height ratio (0.675), and waist to hip ratio (0.91). These elevated anthropometrics quiet evident of grade II obesity leading to health risk. It also indicates greater amount of visceral fat around the vital organs such as hips, buttocks and thighs. The cholesterol (211), the hdl (45), chol/hdl ratio (4.9) and blood pressure parameters (143/84(s1), 152/92(s2)). The blood pressure of stage 2 are used to classify the blood pressure measurement. These values predict the hypertension with diabetes and chronic disease. The stabglu (129) glyhb (6.5) values indicative of diabetes. This cluster group therefore considered to be at increased heart and kidney risk, chronic diabetes and sever obesity.

Border line Diabetes ∪ Obesity ∪ Hypertension = Sever obesity

Cluster 5: This cluster is assigned for females with age 47 years age group. The anthropometric parameters such as BMI (34), waist height ratio (0.63) and waist to hip ratio (0.86). The BMI and waist to height ration seems to be higher and waist to height ratio has reached to maximum limit. These values are considered to be at high risk of obesity. The cholesterol, hdl, chol/hdl ratio are respectively 214, 53 and 4.3. The bp values are 150/92. These parameters are elevated slightly indicating the just onset of hypertension heart and kidney complications. With the regular physical exercise these parameters can brought under control. The diabetic related parameters such as stabglu, glyhb 96 and 5.0 respectively, do not signify the diabetes ailment.

Border line obese ∩ Border line hypertension ∩ Diabetes = No serious complications

Cluster 6: This cluster meant for females with age 41.8 years. The physiological parameters such as BMI, waist to height ratio, waist to hip ratio are 28.9, 0.56, and 0.85 respectively. The BMI is above the target limit and other two values touching the maximum limits. This cluster patients therefore are in boarder line obese. The fat related parameters the cholesterol, hdl, chol/hdl ratio are respectively 214, 53 and 4.3. The bp values are 150/92. These parameters are elevated slightly indicating the just onset of hypertension heart and kidney complications. With the regular physical exercise these parameters can brought under control. The diabetic related parameters such as stabglu, glyhb 96 and 5.0 respectively, do not signify the diabetes ailment.

Border line obese ∩ Border line hypertension ∩ Diabetes = No serious complications

Cluster 7: The age of this cluster patients are 45 years and female gender. The anthropometrics such as BMI, Waist to height, waist to hip are resulted as, 30.22, 0.60, 0.897, respectively. These parameters elevated above the target values. Therefore these factors had the most classical risk indicators for cardio vascular diseases and obesity. The metabolic indices such as stb glu, gly hb are obtained as 92, 5.0 respectively are quiet under the target value predicting free from diabetic. The cardiac risk factors such as cholesterol, chol/hdl ratio and bp measurements are 207, 5.0 and 138/86, 150/92 indicative of pre CVD,and hypertension

Diabetic ∩ Obesity ∪ pre hypertension = Healthy patients
Cluster 8: The age 41 years and female group marked in this cluster. The cross sectional indices BMI = 22.6, WHpR = 0.84, and waist to height ratio (WHR) = 0.49. The indices are within the target values predicting healthy body free from obesity. The CVD risk parameters such as cholesterol, chol/hdl ratio and bp marked as 195, 3.5 and 132/82, 155/94. Except the bp2s which shows the indication of hypertension, other CVD risk parameters are under well control. The diabetic risk factors such as stb glu, glyhb are predicted to be free from diabetes.

Obesity ∩ Hypertension ∩ Diabetes = Healthy patients

Cluster 9: This cluster is unique with female age group 64 years. The body physical indices BMI = 29.9, WHpR = 0.93, and waist to height ratio (WHR) = 0.62. These indices are indicative of obesity. The CVD risk indices such cholesterol = 250, the ratio of chol/hdl = 6.65 and blood pressure measurements are 156/88(1s), 157/89 (2s). These indicators are very beyond the healthy respective targets forecasting the presence of CVD, kidney complaints and hypertension. The diabetic indices stbglu, glyhb obtained as 215 and 10.7 respectively. The high glyhb predicts the prolonged and uncontrolled diabetes. This cluster is highly prone to high obesity, hypertension, high CVD and uncontrollable prolonged diabetes.

Diabetes ∪ Hyper tension ∪ Obesity = High risk unhealthy patients.

IV. CONCLUSION

Conclusion: The WEKA open software of data mining is used in analysis of risk factors of diabetes data supplied by Virginia school of medicine. Using these data ten clusters are formed. These clusters distinctly differentiating among the risk factors like obesity, hypertension and diabetes. Cluster 1 indicates healthy patients, cluster 2, obesity, cluster 3, obesity and hypertension, cluster 4, sever obesity, cluster 5, hypertension, cluster 6 no serious complications, cluster 7&8, healthy patients, and cluster 9 high risk unhealthy patients. The clusters appear to be indicators of risk factors.

REFERENCES

[18] John Schorling, Department of Medicine, University of Virginia school of Medicine, diabetes data set. Available: http://biostat.mc.vanderbilt.edu/DataSets
Dr Mohammad Gulam Ahamad Professor in College of Computer Engineering, department of Computer Engineering, Prince Sattam bin Abdulaziz University, Saudi Arabia. Completed PhD from Osmania University India. He had published many papers in different areas like Data mining, Digital forensics. (e-mail: mg.ahmad@psau.edu.sa)

Mohammed Faisal Ahmed
Network Engineer in LBN Co Sydney, Australia. Completed his Master's in IT from Federation University Australia. He worked in Fuzenet Australia and Senior Associate Wipro Technologies and system administrator in Exatech software solutions Hyderabad. (e-mail:faisalahmed519@gmail.com)

Mohammad Yousuf Uddin
Lecturer in College of Computer Engineering, department of Information Systems. Prince Sattam bin Abdulaziz University, Saudi Arabia. Completed MSc IS from Osmania University India, published few papers in area of data recovery and digital forensics (e-mail: m.yousuf@sau.edu.sa)