Novel Outlier Detection In Diabetics Classification Using Data Mining Techniques

K.Vembandasamy  
Research Scholar & Assistant Professor, Department of Computer Science, 
PSG College of Arts and Science, Coimbatore-14, Tamil Nadu, India.  
Email: vembupsgphd@gmail.com

T.Karthikeyan  
Associate Professor, Department of Computer Science, 
PSG College of Arts and Science, Coimbatore-14, Tamil Nadu, India.  
Email: t.karthikeyan.gasc@gmail.com

Abstract
Less production of insulin or produced insulin cannot be used by the body leads diabetes. If there is lack of evidence, then it is difficult to understand types of diabetes. Normally several tests are done which includes classification or clustering of large scale data. Several tests will make difficult to achieve the classification process. To solve the problem data mining techniques can be used. So, this works aims to provide the solution for classifying the diabetes found in the data through analyzing the classification by using fuzzy clustering with outlier detection method. This research initially performs the fuzzy c-means clustering algorithm. Then, determination of small clusters are done and considered as outlier detection. The evaluation is carried using PIMA dataset and from the results it is clear that the proposed technique achieves better results.

Keywords- Diabetics Classification, Data mining, Fuzzy c means, Clustering, Outlier detection.

Introduction
Diabetes Mellitus (DM) is a set of diseases where the body unable to normalize the sugar amount in the blood. It is the metabolic disease group where the person will have high blood sugar due to the body is unable to create sufficient insulin or the cells are not responding to the insulin which are produced. So, this will lead to produce the symptoms of polydipsia, polyuria and polyphagia. There are three types of diabetes mellitus such as Type 1, the DM outcome from the body failure to create insulin and presently it needs the person to wear an insulin pump or to inject insulin. Previously, this type is referred as “Insulin Dependent Diabetes Mellitus (IDDM) or “juvenile diabetes”. Type 2, DM outcome from the resistance of insulin, a condition where the cells are failed to utilize the insulin properly, at times it will combine with deficiency of absolute insulin. This type is referred as “Non Insulin Dependent Diabetes Mellitus (NIDDM)” or adult onset diabetes. The third type of diabetes is gestational diabetes which happens mainly in pregnant women without diagnosis of previous diabetes and will increase a higher level of blood glucose.

“Type 2” is the most frequent type of diabetes in the worldwide. It was reported as of 2000, 171 million people are suffered globally from diabetes. To rectify this disease, early detection and treatment is necessary. So, diabetic patients should be screened regularly and database should be maintained for analysis. This process will naturally leads to the knowledge discovery application and data mining techniques for discovering the patterns which exist in the data. Generally, Data mining will extract knowledge from the large amount of data. The aim of this data mining is to extract the unsupervised data from the large amount in particular domain. Classification will maps the data into predefined groups. This classification algorithm will requires the classes which are defined based on the values of data attribute. Even though, knowledge discovery has successful in many domains but for health sectors it is found that they are relatively few in comparison.

So, this work proposed a new Fuzzy Clustering Means (FCM) with outlier detection method for diabetes classification. Clustering the data is a process of grouping the similar data into groups. Clustering algorithms are broadly used not only for categorizing and organizing data but also for construction of models and compression of data. This research uses fuzzy clustering to find the diabetes effects and evaluate them in efficient manner. At first, fuzzy clustering is executed and small clusters are resolute which are considered as the outlier clusters. By removing a point temporary form the data set, the rest of outliers (if any) are identified in the left over clusters and then objective function are re calculated. In the Objective Function (OF), is a noticeable change happened then the point is considered as outlier.

This research is ordered as follows: section 2 reviews the previous research done for diabetes classification, section 3 provides the brief explanation of the proposed research methodology, section 4 evaluates the methodology and section 6 concludes this research.

Literature Survey
The dominance of diabetes mellitus was investigated by [6] for pancreatic diseases. As introduced by the American Diabetes Association, the patient data were studied carefully, and according to the diabetes classification patients are reclassified. Genetic Programming (GP) is used in [2] for diabetes classification. There are three stages in the proposed method: using t-score feature selection is performed, then Kullback-Leibler divergence test, F-score selection, Kolmogorov-Smirnov test and GP. The occurrence of
treatment emergent DM is examined in [7] which contains clozapine dose, weight gain, age, treatment with valproate and lipid abnormalities. Autoimmune diabetes such as diagnostic criteria, epidemiology, clinical classification and etiopathogenesis is discussed in [4]. From the EHR, 4 years data are extracted by [10] of multisite, large, multispecialty ambulatory practice serving 700,000 patients. To determine serum C – peptide, the study was conducted by nation wide cohort in [11] at adolescents and 2734 children diagnosis. A comparative study for different classification techniques are presented by [13] using three tools of data mining which are WEKA, TANAGRA and MATLAB. By using these tools, various classification techniques are analyzed for large data set. [14] aimed for mining the diabetes data relationship for proficient classification. In health care organization, data mining plays an important role for detecting new trends which aids to help all the parties which are associated within this field. A brief introduction is presented by [16] of these techniques and their benefits and limitations. Predictive analysis of diabetic treatment is researched by [1] using regression based data mining techniques. A software mining toll namely Oracle Data Miner (ODM) was introduced for predicting the diabetes treating modes and for analysis support vector machine was used.

The goal of [3] was to utilize the analysis of k-means clustering for identifying the regional peel plantar distribution of pressure in 819 diabetic group feet. [5] characterized and compared the online discussion forums for breast cancer, fibromyalgia and type 1 diabetes. The analysis of cluster facilitated the increased understanding of different features of patient experience which includes significant emotional and temporal features for experience of illness. For characterize the facial metagenome [8] used sequencing of shot gun of 145 European women with impaired, normal or diabetic control of glucose. To develop a protocol [12] carried out analysis on gut microbial content for MetaGenome Wide Association Study (MGWAS) in patient having type 2 diabetes and undertaken a two stage MGWAS. The population based stepwise screening programme effect is assessed by [15] on mortality.

Proposed Research Methodology
Detection of diabetes from symptoms or varied factors becomes an issue and it is not free from false detection accompanied by variable effects. With this condition, adjustment of data mining can be done in ability analysis from data. This research uses fuzzy clustering with outlier detection method in the data mining which admittance information from patient medical records. In data mining, outlier detection and clustering are the most important issues. Outlier detection based on the clustering approach will provide the new good results. First, the input from the PIMA dataset is taken for the process. The input data will contain several noisy and unwanted data. To remove the unwanted data and inconsistencies associated with data, data preprocessing technique is used discussed in section 3.2. The preprocessed data are clustered using FCM clustering method for detecting the outliers given in section 3.3. The outlier detected data’s are classified using MPSO-LS-SVM technique for efficient classification diabetes patients explained in section 3.4.

A. PIMA Dataset
The training dataset used for data mining classification was the Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases. The dataset contains 768 record samples, each having 8 attributes. We used this dataset for our classification exercise, as the data is complete.

B. Preprocessing
The data is not complete in the real world and especially incase of medical data. To remove the unwanted and noise data, the preprocessing is carried out. Preprocessing is an important phase in this work because it will affect the clustering and classification result. This preprocessing method will process in two phases. Initially, unwanted and noisy data will remove from the dataset. These removed data will undergo accumulation and normalization and which is treated as outliers. This normalization and accumulation is carried out using FCM clustering method. Simply, outlier detection is the procedure which determines the objects which are inconsistent or unrelated to other data’s. So, this preprocessing will provide the better and effective result by removing the unwanted data and outlier detection.

C. FCM with outlier detection
This section explains the proposed method for outlier detection. An objective function is produced initially by executing c-means clustering algorithm. Then, determination of small clusters is done which are considered as outlier clusters. Small clusters can be defined as the cluster which will have low points than the average points in the clusters. For detecting the outliers in the remaining clusters (if any) the point from the dataset is removed temporarily and c-means algorithm is re-executed. If there is noticeable decrease in the value of objective function by removing the point, then point is considered as outlier or else it is not considered as outliers. The objective function will represent the sum i.e. Euclidian of the squared distances among the centers of cluster and points which belongs to these clusters will times the values of membership of each cluster which are produced by c means clustering algorithm. By dividing the objective function with N, here N represents the data point’s number in the set and the Average Distance (AD) is produced among each point and the clusters which belong to it. Then multiplying the AD value by the given threshold value, the outliers are determined. The objective function is denoted as OF produced by the FCM algorithm. The objective function after removing the point pi produced by the FCM algorithm is denoted as OFi. Then subtracting the produced OF and OFi will provide the difference among the two values which are expressed as DOFi = (OF – OFi). The proposed method algorithm is as follows,

**Step 1:** Execute the FCM algorithm to produce a set of k clusters as well as the objective function (OF) and Compute the value of T as discussed above.

**Step 2:** Determine small clusters and consider the points that belong to these clusters as outliers.
Step 3: For the rest of the points (not determined in Step 2) 
Begin 
For each point i 
DO
remove a point, pi, 
re-calculate the objective function(OFi) 
compute DOFi as OF – Ofi 
if (DOFj > T) then classify point pi as an outlier and return it back to the set;  
End DO 
End

The main scope of this work is to develop a DM treatment and analysis classification scheme. Diabetes mellitus (DM) is the most widespread complex metabolic disorder among the world’s population currently affecting around 250 million people globally. A lazy associative classification approach for prediction of heart disease is proposed which is used to support the physicians for getting the assessments perfectly. This work done a research on classification and treatment for diabetes disease and implemented a model for analyzing if the data mining techniques is applied for classification of diabetes data will able to provide the reliable performance in diagnosing.

D. MPSO-LS-SVM with outlier detection
Frequent Pattern Growth algorithm is complicated in the medical application. An interesting area where association rule based classification can be used for early diagnosis. Modified Particle Swarm Optimization with the Least Square Support Vector Machine (MPSO-LSSVM) is developed by [9] with the outlier detection method. The CFP growth algorithm will be used to find the frequent patterns where the Minimum Item Support (MIS) tree and CFP array are constructed, and then frequent patterns are produced from the MIS tree. The frequent items sets are found from this set which will create all the association rules and will have confidence which exceeds the minimum confidence. The outlier degree from the association rules are determined using enhanced outlier detection method for outlier detection. Although this algorithm has certain drawbacks which are curse of dimensionality (i.e.,) the larger the degree of freedom the more similar the distance values for values for all points and standard deviation and mean are very responsive to outliers. So, this work proposed a new algorithm for outlier detection using fuzzy clustering.

i. Experimental Results
The experimental results are carried out for PIMA dataset. Pima Indian Diabetes Data (PIDD) set is available publicly from the machine learning database at UCI. This dataset consist only females at the age of 21 of Pima Indian heritage living near Phoenix, Arizona. This data set is extracted from a larger database originally owned by the National Institute of Diabetes and Digestive and Kidney Diseases.

Our proposed method is evaluated using accuracy and execution time in PIMA dataset. By using the fuzzy clustering with outlier detection method, classification accuracy achieved the better result rather using only classification algorithm. Table 1 shows the accuracy and execution time for the MPSO-LS-SVM and the proposed FCM with outlier method. The proposed method achieves better accuracy with less execution time than the preceding method.

Figure 1 and 2 shows the accuracy and execution time for the proposed method. It is clear from the figure that proposed method provides better accuracy than the preceding method. Execution time is also very low when compared with the MPSO-LS-SVM.

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Accuracy (%)</th>
<th>Execution time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPSO-LS-SVM</td>
<td>84</td>
<td>31</td>
</tr>
<tr>
<td>FCM with outlier</td>
<td>93</td>
<td>19</td>
</tr>
</tbody>
</table>

**Table 1: Accuracy and Execution time**

**Figure 1: Accuracy**

**Figure 2: Execution time**

Conclusion
Everyone must check wherever they have affected by type 2 diabetes or not. This work first of all gets the data collection from Pima Indian Diabetes dataset. Then we can find out the type 2 diabetes. The aim of this research is to provide the best algorithms which will classify the given data in multiple aspects. For automatic classification tools, this algorithm is necessary. This work shows how FCM with outlier detection are used to model actual classification of diabetes for local and systematic treatment and tested with Pima Indian diabetes dataset. The proposed method is effective showed in the experimental results. The performance of the techniques was
investigated for the diabetes classification problem. Experimental results demonstrate the adequacy of the proposed model.

References


