An Analytical Study on Early Diagnosis and Classification of Diabetes Mellitus

Dr. T. Karthikeyan, Department of computer science, PSG college of arts and science, Coimbatore, 9443820196, t.karthikeyan.gasc@gmail.com.

K. Vembandadsamy, Department of computer science, PSG college of arts and science, Coimbatore, 9942844799, vembupsgphd@gmail.com.

ABSTRACT

Diabetes mellitus (DM) is a chronic, general, life-threatening syndrome occurring all around the world. It is characterized by hyperglycemia occurring due to abnormalities in insulin secretion which would in turn result in irregular raise of glucose level. In recent years, the impact of Diabetes mellitus has increased to a great extent especially in developing countries like India. This is mainly due to the irregularities in the food habits of several IT professionals. Thus, early diagnosis and classification of this deadly disease has become an active area of research in the last decade. A number of techniques have been developed to deal with his disease. Numerous clustering and classifications techniques are available in the literature to visualize temporal data to identifying trends for controlling diabetes mellitus. This survey presents an analytical study of several algorithms which diagnosis and classifies Diabetes mellitus data effectively. The existing algorithms are analyzed thoroughly to identify their advantages and limitations. The performance evaluation of the existing algorithms is carried out to determine the best approach. A best approach among the existing approach is determined and a solution is also suggested to improve the overall performance of diagnosis process.

Key Words: Data mining, Diabetes mellitus, Clustering, Classification.

INTRODUCTION

The occurrence of diabetes mellitus with certain complications has been utilized to determine the diagnostic cut-points for diabetes, especially using data from epidemiological investigations which have resulted in several other disorders [1,2].

The major symptoms and effects of diabetes mellitus include long-term injury, dysfunction, and functional abnormalities in eyes, kidneys, nerves, heart, and blood vessels [3]. Thus, it is considered as one of the most important health issue and economic costs in various countries like United States, India, etc [4].

Diabetes occurrence and predominance in the United States is well understood at the circumstances [5, 6]. Recognizing a type of diabetes to an individual is mostly based on the situation at the time of diagnosis. It is observed that most of the diabetic patients do not just fix into a single category [7]. The three major categories of diabetes mellitus includes

- **Type 1 Diabetes:** This type of abnormality occurs due to the malfunction of the body to generate insulin.
- **Type 2 diabetes:** This type of class results from insulin resistance, where the cells could not employ insulin in proper proportion.
• **Gestational diabetes:** This occurs when pregnant women, without diabetes but having high blood glucose level during pregnancy. It may possibly lead to the development of type 2 DM.

In recent years, the number of diabetic patients has increased drastically mainly due to the aging population and irregular western food habits. Genetic inheritance is the main reason for the cause type 1 and 2 diabetes classes. The main aim of treating diabetes is to control the sensitive complications of diabetes, and to eliminate the chronic complications of diabetes. For effective diagnosis of diabetes, the main factor that has to be considered is the risk of diabetic complications, early and accurately.

Recently, a novel classification and diagnostic criteria have been presented by the American Diabetes Association (ADA), World Health Organization (WHO) and Japan Diabetes Society (JDS). This survey illustrates the major points of the report of the Committee on the classification and diagnostic criteria of diabetes mellitus.

The investigation has strongly suggested that the new and efficient diagnostic and classification techniques have to be adapted for early and efficient diagnosis of diabetes mellitus. The new diagnostic techniques should have all the validated evaluations.

Numerous techniques have been developed for the diagnosis of diabetic mellitus. Most of the techniques used clustering and classification for the effective diagnosis of the diabetic mellitus disease. But, there is always a scope for improvement and still several techniques are being developed to overcome the limitations of the existing techniques.

This paper presents an analytical study on the existing techniques available for diabetes mellitus. The characteristic features of the approaches are investigated to develop a better approach for the early and efficient diagnosis of the disease.

**LITERATURE SURVEY**

This section discusses about the existing techniques and algorithms used for the diagnosis of diabetes mellitus. Each and every algorithm used for the diagnosis of diabetes mellitus has its own limitations and advantages. This section presents an analytical study on the features of the existing techniques.

**THE K-NEAREST NEIGHBOR ALGORITHM:**

The idea in k-Nearest Neighbor methods [21] is to identify ‘k’ samples in the training set whose independent variables 'm' are similar to 'n', and to use these 'k' samples to classify this new sample into a class, v. Assume that ‘f’ is a smooth function, an idea is to look for samples in our training data that are near it and then to compute ‘v’ from the values of ‘y’ for these samples.

When we talk about neighbors we are implying that there is a distance or dissimilarity measure that we can compute between samples based on the independent variables. For the moment we will concern ourselves to the most popular measure of distance is say, Euclidean distance.

The initial training stage for kNN [22] consists of storing all known occurrences and their class labels. Either a tabular representation or a specialized design such as a kd-tree can be used. If we want to adjust the value of 'k', an alternate method of n-fold cross-validation on the training data set can be used.

The k-NN algorithm for continuous-valued functions
Calculate the mean value of the k nearest neighbors
Distance-weighted nearest neighbor algorithm
• Weight the contribution of k neighbors according to their distance to the query point $x_q$
  – giving greater weight to closer neighbors
  $$w = \frac{1}{d(x_q, x)^2}$$
  – Similarly, for real-valued target functions.

CLUSTERING AND HYBRID APPROACHES FOR DIAGNOSIS DIABETES MELLITUS

A novel hybrid integration of Iterative Learning Control (ILC) and Model Predictive Control (MPC) forms a Model Predictive Iterative Learning Control (MPILC) for glycemic control in type 1 diabetes mellitus as discussed by Youqing Wang et al [12]. MPILC uses two key factors such as frequent glucose readings and the repetitive feature of glucose-meal-insulin dynamics with a 24-h cycle. This algorithm formulates based on individual's lifestyle, facilitating the control performance to be improved.

Hemant and Pushpavathi [14] presented a novel approach for diagnosing and classifying the diabetes mellitus disease. Initially K-means clustering approach is used to group the disease related data into clusters and assigns classes to clusters. Then, various classification approaches are trained on the result set to generate the final classifier framework based on K-fold cross validation. This approach is validated using 768 raw diabetes data. The proposed approach helps clinical professionals in their diagnosis judgments and also in their treatment procedures for different class of diabetes mellitus.

Nirmala Devi et al [13] presented a fusion model that integrates k-means clustering and k-Nearest Neighbor (KNN) with multi-sep preprocessing. It is observed that KNN algorithm provides significant performance on various data sets. In this fusion model, the quality of the data is improved through eliminating noisy data thus improving the accuracy and efficiency of the KNN algorithm. K-means clustering is determines and avoids incorrectly classified instances. An efficient classification is carried out through KNN by taking the correctly clustered samples with preprocessed subset as inputs for the KNN. The finest choice of k is based on the data. The main goal of this work is to identify the value of k for PIID for better classification accuracy using fusion based KNN. It is observed from the results that this fusion work based on KNN along with preprocessing provides best result for different k values. If k value is more, the classification accuracy of the proposed fusion framework is 97.4%. The results are also compared with simple KNN and cascaded K-MEANS and KNN for the same k values. A hybrid framework which integrates clustering and classification to obtain higher accuracy result is presented. In the earlier approaches, hybrid classification approaches have been developed to predict different types of diabetes through analyzing the patients’ profiles [15]. This information is useful in diagnosing a diabetic patient. Hybrid Prediction Model (HPM) using K-means clustering algorithm integrated with a C4.5 classifier is presented by Patil et al [16]. This framework is trained on patients’ characteristics and measurements to inquire on the diabetic diagnosis. K-fold cross validation approach has been used in Decision Tree C4.5 Classifier and it is found that the accuracy is increased to 92.4% when compared with the C4.5 algorithm. The recent hybrid framework includes K-Means clustering and Decision Tree C4.5 classifier. This framework clearly classifies the patients with high risk of diabetic mellitus and the people with lesser probability in getting diabetic mellitus [16].
Using the same dataset, which consists of 392 cases with no missing values, the proposed cascaded model obtained the classification accuracy of 93.3% when compared to accuracy of 73.6% using C4.5 classifier alone.

Norul Hidayah Ibrahim et al [17] presented a novel hybrid framework by exploiting Agglomerative Hierarchical Clustering and Decision Tree Classifier on Pima Indians Diabetes dataset. This work presents a comparative analysis of the performance accuracy of the Decision Tree Classifier against the same classifier augmented with Hierarchical clustering. It is observed that the hybrid model achieved higher accuracy with 80.8% when compared to 76.9% of the conventional model. The results showed the potential significance of the hierarchical clustering in a rule-based classifier.

Table 1: Summary of the existing Approaches

<table>
<thead>
<tr>
<th>Various Approaches</th>
<th>Results Obtained</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Nearest Neighbor Algorithm</td>
<td>Provides significant results with accuracy (78%)</td>
<td>It requires more time to process, Large values of k.</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>Higher classification accuracy</td>
<td>SVMs do not directly provide probability estimates, so these must be calculated using indirect techniques</td>
</tr>
<tr>
<td>Hybrid Approaches</td>
<td>Very high classification accuracy and Prediction Accuracy</td>
<td>It is very to simple to implement, Robust to noisy training data etc</td>
</tr>
</tbody>
</table>

COMPARATIVE ANALYSIS OF THE EXISTING APPROACHES

The performance evaluation of the above discussed approaches is based on certain performance metrics. The diagnostic model of diabetes mellitus of the data is evaluated using three metrics like classification accuracy and processing time.
Table 2: Methods to diagnostic the diabetes mellitus

<table>
<thead>
<tr>
<th>APPROACHES</th>
<th>CLASSIFICATION ACCURACY</th>
<th>PROCESSING TIME IN (SEC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Nearest Neighbor Algorithm</td>
<td>78</td>
<td>0.964</td>
</tr>
<tr>
<td>NAHLA H BARAKAT, ET AL., (2010)</td>
<td>94</td>
<td>0.856</td>
</tr>
<tr>
<td>NIRMALA DEVI ET AL., (2013)</td>
<td>97.4</td>
<td>0.753</td>
</tr>
<tr>
<td>NORUL HIDAYAH IBRAHIM ET AL., (2013)</td>
<td>81</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Fig 1: Comparison of Methods to diagnostic the diabetes mellitus in terms of Classification accuracy
PERFORMANCE COMPARISON

The performance of the various algorithms is evaluated based on the parameters like
- Convergence behavior
- Processing Time
- Classification Accuracy

Convergence Behavior

Figure 3 shows the comparison of the convergence behavior of the Decision trees, neural networks and Clustering Techniques. It is observed from the figure that the clustering Approaches and hybrid approaches converge in lesser iterations when compared with the other techniques. For instance, the decision trees approach takes 90 iterations for convergence; neural network Algorithms approach takes 70 iterations where as the clustering approaches takes 40 iterations for convergence, hybrid approaches takes 30 iterations to convergence. Thus the hybrid approaches are very
significant when compared with the other optimization approaches taken for consideration.

**Processing Time**

Table 3 shows the performance comparison of the techniques such as C4.5 Classifier, Neural networks and Clustering Techniques.

**Classification Accuracy**

Table 3: Performance Comparison of the various Techniques using Classification Accuracy

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Classification Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4.5 Classifier [18]</td>
<td>92.38</td>
</tr>
<tr>
<td>K-NN [8]</td>
<td>80.88</td>
</tr>
<tr>
<td>Clustering Techniques [3]</td>
<td>93.67</td>
</tr>
<tr>
<td>Hybrid Approach [15]</td>
<td>96.68</td>
</tr>
</tbody>
</table>

![Fig 4: Comparison of classification accuracy](image)

It is observed from the figure 4, that the clustering Approaches performs better when compared with other techniques.
CONCLUSION

Diabetes mellitus is a deadly disease which is one of the chief public health challenges around the world. It is a fact that 80% of type 2 diabetes complications can be prevented by early identification of people at risk. Numerous machine learning techniques have been used for the diagnosis and classification of diabetes mellitus. Early detection of this disease has become an essential issue to improve the overall clinical efficiency of the diagnosis process. This paper clearly presents an analytical study of numerous algorithms which includes clustering, classification, vector machines and neural networks. An analytical result has been validated for the approaches such as clustering, neural network, vector machines and hybrid approaches. It is observed that the hybrid approaches are observed to produce significant results in terms of the classification accuracy, processing time, etc. This survey would help clinical assistance to doctors for earlier diagnosis of the dreadful disease.

REFERENCES


