Early Diabetes Detection using Machine Learning: A Review

Sakshi Gujral
Department of Computer Science & Engineering
IGDTUW, Delhi, India

Abstract

Machine learning is one of the aspect of artificial intelligence that allows the development of computer systems that have the ability to learn from experiences without being the need of programming it for every instance. Machine learning is dire need of today’s scenario to eliminate human effort as well as come up with higher automation with less errors. This paper focuses on the review of Early Diabetes detection using machine learning techniques and detection of the frequently occurred disorders with it - mainly Diabetic Retinopathy and diabetic neuropathy. The data set employed in most of the concerned literature is Pima Indian Diabetic Data Set. Early diabetes detection is significant as it helps to reduce the fatal effects of the diabetes. Various machine learning techniques like artificial neural network, principal component, decision trees, genetic algorithms, Fuzzy logic etc. have been discussed and compared. This paper first introduces the basic notions of diabetes and then describes the various techniques used to detect it. An extensive literature survey is then presented with relevant conclusion and future scopes with analysis have been discussed.

Keywords: Machine Learning, Fuzzy Logic, Fuzzy C-Means, SVM, GA, PCA, ANN

I. INTRODUCTION

Diabetes is one of the diseases that are spreading like epidemics in the entire world. It is seen that every generation ranging from children, adolescents, young people and old age are suffering from it. Pro-long effect can cause worse effects in terms of failure of organs like liver, kidneys, heart, stomach and can lead to death. It is frequently associated with the disorders - Retinopathy and Neuropathy. Diabetes is mainly of two types - type 1 and type 2. [12]

A. Type -1 Diabetes

It is the situation in which liver does not produce insulin at all. Insulin is an hormone that is required to absorb glucose from the blood to utilize this glucose for body building. However, absence of insulin in the body will increase blood sugar and it will lead to Type-1 Diabetes. It is commonly found in children and adolescents. It mainly occurs because of the genetic disorders. It is often known as juvenile disorder. Its common symptoms are frequent urination. Weight loss, increases thirst, blurs vision, nerves problems. This can be treated by insulin therapy.

B. Type -2 Diabetes [24],[11]

It is long term metabolic disorder generally occurs in the adults over age of 40 years. It is evident by high blood sugar, insulin resistance and high insulin. The major cause is obesity and lack of exercise. This bad lifestyle can cause glucose to get store in the blood and develop diabetes.90% of people affected by type-2 diabetes only. To treat insulin resistance metformin is given to ensure this can be treated.

C. Diabetic Neuropathy [26]

These are the nerve disorders developed in diabetic patients with the passage of time. They often occur in foot and hands. The common symptoms are pain, numbness, tingling, loss of feeling in hand, foot, arms etc.

D. Diabetic Retinopathy [16],[17],[25]

It is the diabetic disorder that leads to permanent eye blindness. Initially there is no significant symptom, gradually symptoms are seen. In the second stage, blood vessels are developed at the back of the eyes that could lead to bleeding on bursting as they are quite agile.

II. TECHNIQUES USED FOR DIABETES DETECTION

For Diabetes Detection various artificial intelligence techniques are used as the can mine the data as well as learn from the data set to provide better result. Some of the frequently used techniques used in the various research papers are-
A. **Support Vector Machine (SVM)**

This is supervised learning technique that means data set is trained in such a way that it may give pre-determined output. It represents data set as points of cloud in the space. The aim here is to construct a hyper-plane that divides the data sets into various categories. The hyper-plane divides the data set into the categories so that data mining and classification can be done effectively. This hyper-plane should be at maximum margin from the different categories. However, if categories in which data set have to be classified are large then sophisticated technique is used known as kernel configuration.

1) **Advantages**
   - SVM is used to classify diabetes data set effectively by assigning data set into various categories with the help of the hyper-plane.
   - It removes over fit nature of the samples.

2) **Disadvantages**
   - SVM cannot be used for large data sets.
   - SVM is slow in its execution.

B. **Fuzzy C-means**

It is an extension of K-means clustering algorithm that means it aims at forming the clusters, then finding out the centroids of the clusters, the incoming data set is assigned to that cluster that has minimum distance from it’s centroid. However, it may happen that sometimes very less margin is there so that new data set can be fall for more than one cluster. This was avoided by fuzzy C-means clustering algorithm as it employs fuzzy partition that accounts for the membership function. Hence, results produce are more accurate.

1) **Advantages**
   - The involvement of fuzzy logic here that account of the membership function helps in giving better result for the classification.
   - It is unsupervised learning technique so results are more real time.

2) **Disadvantages**
   - It takes long computational time.
   - It is more susceptible to wrong guesses at initial stages.

C. **Principal Component Analysis**

PCA is a statistical model that is used to classify data set in such a way that the maximum co-relation can be found in the data set. It aims at construction to orthogonal plane so that data can be classified along with this plane, another plane is perpendicular on it, that is known for second co-relation among data set. It helps in feature extraction and makes use of Eigen values and Eigen vectors to calculate the principal component.

1) **Advantages**
   - It helps in reducing dimension thereby preserving the randomness among data sets.
   - It helps in reducing noise as maximum variation data set is chosen.

2) **Disadvantages**
   - There is difficulty to calculate Eigen values and covariance matrix.
   - For diabetes detection alone PCA does not give great performance.

D. **Naives Bayes Classifier**

It is supervised learning technique based on Bayes’ theorem. It is family of algorithms, it assumes that value of one particular feature is independent (naïve) of another feature. It accounts for the conditional probability that it determines the likelihood of an event to take place provided that some of the events have already taken place. It is used for diabetes detection as well as detection of diabetic retinopathy.

1) **Advantages**
   - It helps in reducing noise because values are averaged.
   - Higher value of probability gives more accurate result.

2) **Disadvantages**
   - It makes very strong assumption about the shape of the data distribution.
   - While making continuous features to discrete, data is lost.

E. **Decision Trees**

To support decision making, Decision trees support very sophisticated tools. A tree or graph like structure is constructed on the basis of parameters like cost, classification categories, and effort. The decision is taken by traversing from root to leaf till the criteria is met. The split of node is determined by Gini indices. The inclusion of Gini indices helps in better node splitting. Collection of random collection of decision trees also gives the notion of random forest classifier. These classifier also help us in determine the diabetes detection.
1) **Advantages**
   - It is the best predictive model as deep analysis of the problem can be done.
   - Random forest classifiers are best suited for large amounts of data as well as missing data.

2) **Disadvantages**
   - Random forest is fast to train but slow to create predictions once trained.
   - Decision trees are unstable even with a small change in the input.

### F. Artificial Neural Network

This technique imitates like human minds, just like humans have neurons in the brain to convey messages. Similarly, artificial neural networks have learning capabilities to learn from input and predict the output. When many layers are present, it is called deep neural networks.

1) **Advantages**
   - ANN with back-propagation is used in diabetes detection for feature extraction.
   - When combined with fuzzy logic it can handle uncertainties.

2) **Disadvantages**
   - Large effort is required for training.
   - It is difficult to ensure whether all the inputs have been trained or not.

### III. Exhaustive Literature Survey

Table 1 depicts the exhaustive literature survey that has been carried out for early diabetes detection using various artificial intelligence techniques.

<table>
<thead>
<tr>
<th>Year</th>
<th>Journal /Conference</th>
<th>Author</th>
<th>Central Idea</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>International Journal of Computer Trends and Technology</td>
<td>Ravi Sanakal, Smt T Jayakumari</td>
<td>- This study involves the implementation of FCM and SVM and testing it on a set of PIDD.</td>
<td>FCM and SVM gives good Classification</td>
<td>Better machine learning algorithm should be employed along with them.</td>
</tr>
<tr>
<td>2016</td>
<td>International Journal of PharmaMedicine</td>
<td>Mohammed Imran, Alhanoof M. Al-Abdallatif, Bushra S. Al-Awad, Mzoon M. Alwalmani, Sarah A.</td>
<td>- Detection of Diabetic Retinopathy (DR) Using Extended Fuzzy Logic. - Calculation of damage to Retina using OWE.</td>
<td>It allows detection as well as calculation of damage caused to retina.</td>
<td>Complex and time taking process</td>
</tr>
<tr>
<td>2008</td>
<td>Expert Systems with Application</td>
<td>Humar Kahramanli Novraz Allahverdi</td>
<td>- Artificial neural network combined with fuzzy logic is used to detect diabetes.</td>
<td>It allows better result as fuzzy accounts for uncertainties also.</td>
<td>Extracting rules from existing methods is not very efficient as it takes times.</td>
</tr>
<tr>
<td>2010</td>
<td>Expert Systems with Applications</td>
<td>Hybrid prediction model for Type-2 diabetic patients B.M. Patil R.C. Joshi, Durga Toshniwal</td>
<td>- This study proposes Hybrid Prediction Model which uses Simple K-means clustering algorithm - Subsequently applying the classification algorithm to the result set. C4.5 algorithm is used to build the final classifier.</td>
<td>Hybrid approach gives better result as compared to single classifiers.</td>
<td>Using all the approaches all together is tedious process</td>
</tr>
<tr>
<td>2014</td>
<td>Computers and Electrical Engineering</td>
<td>A computational intelligence approach for a better diagnosis of diabetic patients Kamadi V.S.R.P. Varma a, Allam Appa Rao b, T. Sita</td>
<td>- Authors propose a method to minimize the calculation of Gini indices by identifying false split points. - Authors have used the Gaussian fuzzy function</td>
<td>Gini indices along with fuzzy function gives good result.</td>
<td>Accuracy of model can be improved using fuzzy membership functions</td>
</tr>
<tr>
<td>2011</td>
<td>International Journal on Soft Computing</td>
<td>Asha-Gowda Karegowda1, A.S. Manjunath2, M.A. Jayaram3</td>
<td>- This paper integrates Genetic Algorithm and (BPN). - GA is used to initialize and optimize the connection weights of BPN.</td>
<td>Hybrid GABPN shows elegant accuracy.</td>
<td>BPN is prone to lead to troubles as local minimum problem, slow convergence</td>
</tr>
<tr>
<td>Year</td>
<td>Conference/Journal</td>
<td>Authors</td>
<td>Abstract</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------------------</td>
<td>---------</td>
<td>----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>International Journal of Computer Applications</td>
<td>Mani Butwall, Shraddha Kumar</td>
<td>Data mining approach to envisage diabetes behaviour is based on Random Forest Classifier. [10]</td>
<td>Random forest classifiers is good approach to handle large data set. Single classifier approach is not very effective as compared to hybrid.</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Application of a Unified Medical Data Miner</td>
<td>Nawaz Mohamadally, Dost Muhammad</td>
<td>In this study C4.5, Neural Network, Kmeans, Visualization is used to detect diabetes. [2]</td>
<td>It is good approach as hybrid method is used.</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>International Journal of Bio-Science and Bio-Technology</td>
<td>Kwang Baek Kim, Doo Heon Song</td>
<td>This paper presents self-diagnosis system of Disease Classification Index(KCD) and Fuzzy ART/inference method. [13]</td>
<td>Inference system can be used for self use immediately. More investigation is required to make it use for self use.</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>IEEE Recent Advances in Intelligent Computational Systems</td>
<td>Veena Vijayan V. Anjali C.</td>
<td>Decision support system is proposed that uses AdaBoost algorithm with Decision Stump as base classifier for classification. Support Vector Machine, NaiveBayes and Decision Tree are also implemented as base classifiers. [14]</td>
<td>Adaboost gives an edge to yield combined and better results. Accuracy of classifiers needs to be improved with nn classifiers and other approaches.</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>Proceedings of ICEE 2010</td>
<td>Mostafa Fathi Ganji</td>
<td>ACO is used to extract a set of rules for diagnosis of diabetes disease with FADD. [15]</td>
<td>FADD is good approach to detect diabetes. Single approach for deduction needs to be clubbed with other.</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>International Journal on Soft Computing</td>
<td>E.P. Ephzibah</td>
<td>It is a task of identifying and selecting a useful subset of pattern-representing features from larger set of features. Using fuzzy rule-based classification system. [18]</td>
<td>Genetic algorithm integrated with fuzzy logic is generating better rules. Better feature selection mechanism can be used along with fuzzy logic.</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>International Journal of Engineering Research in Africa</td>
<td>G. Thippa Reddy, Neelu Khare</td>
<td>An attempt has been made to develop Firefly-BAT (FFBAT) optimized Rule Based Fuzzy Logic (RBFL) prediction algorithm. [19]</td>
<td>High accuracy, sensitivity is obtained by this new algorithm. Other optimization techniques can be applied to improve accuracy.</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>Applied Soft Computing</td>
<td>Kamadi V.S.R.P., Varmaa</td>
<td>It presents an approach using principal component analysis and modified Gini index based fuzzy SLIQ decision tree algorithm. [20]</td>
<td>Sharp decision boundary can be overcome by fuzzy SLIQ. Accuracy can be improved further by better fuzzy membership.</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>International Journal of Computer, Electrical, Automation, Control and Information Engineering</td>
<td>Kemal polat</td>
<td>Combination of fuzzy c-means and svm is used for diabetes prediction on dataset[7]</td>
<td>Fuzzy C-means classify data set in better way as it involves membership function Real time data is noisy so effort is required to make it useable for processing.</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>Informatics in Medicine Unlocked</td>
<td>Yoichi Hayashi, Shonosuke Yukita</td>
<td>Use of a rule extraction algorithm, ReRX with J48 graft, combined with sampling selection techniques (sampling Re-RX with J48 graft) is done. [21]</td>
<td>High accuracy in terms of rule extraction. the diagnosis of T2DM remains a complex problem; diagnosis.</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>International Conference on Computer and Knowledge</td>
<td>Kiarash Zamiri, Mehdi Teimouri Rohallah Rahmani and Amin Salaq</td>
<td>This paper present and compare different cost-sensitive learning methods for diagnosis of type 2 diabetes. [22]</td>
<td>Cost sensitive approach is effective for utilizing resources. Assumptions are used in data sets, matrices to bring out the results.</td>
<td></td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Year</th>
<th>Conference/Platform</th>
<th>Authors</th>
<th>Description</th>
<th>Algorithmic Innovations</th>
<th>Literature Survey</th>
<th>Conclusion &amp; Future Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>Journal of Computer Science &amp; Wireless Security (IJCSWS)</td>
<td>B.Saritha, A.Vinodhini</td>
<td>- The tongue images are then individually processed and then texture and color analysis are done. - The differences between these two images (before and after food) are then comparatively analysed. [23]</td>
<td>Innovative application and easy to use for diabetes deduction.</td>
<td>Missed data of hypertension as a result inaccurate data is obtained</td>
<td>The hybrid approaches yield better results than single classifiers. Moreover, some of the techniques when integrated with fuzzy logic gives better results. Not only this, Diabetic retinopathy and Diabetic neuropathy can also be analyzed with fuzzy logic integrated with image analysis. These techniques can be combined with real time data with the help of “Internet of Things” to make real time devices for the healthcare applications. Hence, IOT with intelligence would be acquired. These devices will eliminate need of human involvement at larger pace and will give inculcate the better results with less errors. Data set so acquired or real time data contain noisy data that needs to be mined from proper knowledge discovery. Hence classifiers like SVM and PCA should be used along with more refined techniques for proper feature extraction. Artificial neural network accounts for the drawback of unreliability of learning of input nodes, this needs to be work upon. For Principal component analysis, selection of Eigen values criteria should be more work upon. Random Forest Classifiers require monitoring for the time complexity. Hence, diabetes deduction can be effective with these techniques.</td>
</tr>
<tr>
<td>2016</td>
<td>Expert Systems with Applications</td>
<td>Carlos F. Vázquez-Rodríguez, Rubén Posada-Gómez, Armin Trujillo-Mata</td>
<td>Fuzzy Expert system is developed to detect neuropathy. [4]</td>
<td>This FES provides efficient system to detect neuropathy.</td>
<td>Literature Survey of Diabetes Deduction shows that single approach to detect diabetes is not very sophisticated approach for early diabetes deduction. Hybrid approach with classifiers like Support vector machine, principal component analysis along with Genetic algorithms, Artificial neural network would give better results. As these techniques will give help in reducing noise from data set by feature extracting and then applying learning methodology to detect hidden patterns and give more accurate results. Random forest will give better results than decision trees. However, best combination is integration of machine learning with fuzzy logic as it will account for the uncertainties also. The analysis also shows some of the cost effective approach also for diabetes deduction.</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>Computational vision and robotics</td>
<td>R.C. Joshi, Durga Toshniwal</td>
<td>Binning technique is used to convert continuous data to discrete - Apriori algorithm is used [27]</td>
<td>Binning helps in identifying hidden patterns</td>
<td>PIMA data set prunes large data rules.</td>
<td></td>
</tr>
</tbody>
</table>

G. Analysis of the Literature Survey

Literature Survey of Diabetes Deduction shows that single approach to detect diabetes is not very sophisticated approach for early diabetes deduction. Hybrid approach with classifiers like Support vector machine, principal component analysis along with Genetic algorithms, Artificial neural network would give better results. As these techniques will give help in reducing noise from data set by feature extracting and then applying learning methodology to detect hidden patterns and give more accurate results. Random forest will give better results than decision trees. However, best combination is integration of machine learning with fuzzy logic as it will account for the uncertainties also. The analysis also shows some of the cost effective approach also for diabetes deduction.

IV. Conclusion & Future Scope

The hybrid approaches yield better results than single classifiers. Moreover, some of the techniques when integrated with fuzzy logic gives better results. Not only this, Diabetic retinopathy and Diabetic neuropathy can also be analyzed with fuzzy logic integrated with image analysis. These techniques can be combined with real time data with the help of “Internet of Things” to make real time devices for the healthcare applications. Hence, IOT with intelligence would be acquired. These devices will eliminate need of human involvement at larger pace and will give inculcate the better results with less errors. Data set so acquired or real time data contain noisy data that needs to be mined from proper knowledge discovery. Hence classifiers like SVM and PCA should be used along with more refined techniques for proper feature extraction. Artificial neural network accounts for the drawback of unreliability of learning of input nodes, this needs to be work upon. For Principal component analysis, selection of Eigen values criteria should be more work upon. Random Forest Classifiers require monitoring for the time complexity. Hence, diabetes deduction can be effective with these techniques.

REFERENCES

Early Diabetes Detection using Machine Learning: A Review


