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-

All rights reserved by *www.ijirst.org*

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.

All rights reserved by *www.ijirst.org*

1) Advantages
- It is best predictive model as deep analysis of the problem can be done.
- Random forest classifiers is best suited for large amount 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 instable even with a small change in the input.

F. Artificial Neural Network

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

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 - 1
Literature survey for early diabetes detection using various artificial intelligence techniques

Year	*Journal /Conference*	*Author*	*Central Idea*	*Pros*	*Cons*
2014	*International Journal of Computer Trends and Technology*	*Ravi Sanakal Smt T Jayakumari*	– *This study involves the implementation of* – *FCM and SVM and testing it on a set of PIDD.[3]*	*FCM and SVM gives good Classification*	*Better machine learning algorithm should be employed along with them.*
2016	*International Journal of PharmaMedicine*	*Mohammed Imran, Alhanouf M. Al-Abdullatif, Bushra S. Al-Awwad, Mzoon M. Alwalmani, Sarah A.*	– *Detection of Diabetic Retinopathy (DR) Using Extended Fuzzy Logic.* – *Calculation of damage to Retina using OWE.[5]*	*It allows of detection as well calculation of damage caused to retina.*	*Complex and time taking process*
2008	*Expert Systems with Application*	*Humar Kahramanli Novruz Allahverdi*	– *Artificial neural network combined with fuzzy logic is used to detect diabetes.[6]*	*It allows better result as fuzzy accounts for uncertainties also.*	*Extracting rules from existing methods is not very efficient as it takes times.*
2010	*Expert Systems with Applications*	*Hybrid prediction model for Type-2 diabetic patients B.M. Patil R.C. Joshi, Durga Toshniwal*	– *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.[1]*	*Hybrid approach gives better result as compared to single classifiers.*	*Using all the approaches all together is tedious process*
2014	*Computers and Electrical Engineering*	*A computational intelligence approach for a better diagnosis of diabetic patients Kamadi V.S.R.P. Varma a, Allam Appa Rao b, T. Sita*	– *Authors propose a method to minimize the calculation of Gini indices by identifying false split points.* – *Authors have used the Gaussian fuzzy function [8]*	*Gini indices along with fuzzy function gives good result,*	*Accuracy of model can be improved using fuzzy member-ship functions*
2011	*International Journal on Soft Computing*	*Asha-Gowda Karegowda1 , A.S. Manjunath2 , M.A. Jayaram3*	– *This paper integrates Genetic Algorithm and (BPN).* – *GA is used to initialize and optimize the connection weights of BPN.[9]*	*Hybrid GABPN shows elegant accuracy.*	*BPN is prone to lead to troubles as local minimum problem, slow convergence*

All rights reserved by *www.ijirst.org*

2015	International Journal of Computer Applications	Mani Butwall Shraddha Kumar	– Data mining approach to envisage diabetes behaviour is based on Random Forest Classifier. [10]	Random forest classifiers is good approach to handle large data set.	Single classifier approach is not very effective as comapared to hybrid.
2011	Application of a Unified Medical Data Miner	Nawaz Mohamudally1 and Dost Muhammad	– In this study C4.5, Neural Network, Kmeans, Visualization is used to detect diabetes.[2]	It is good approach as hybrid method is used.	prediction, classification, visualisation requires tremendous effort
2016	International Journal of Bio-Science and Bio-Technology	Kwang Baek Kimand Doo Heon Song2	– This paper presents self-diagnosis system of Disease Classification Index(KCD) and Fuzzy ART/inference method.[13]	Inference system can be used for self use immediately.	More investigation is required to make it use for self use
2015	IEEE Recent Advances in Intelligent Computational Systems	Veena Vijayan V. Anjali C.	– 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]	Adaboost gives an edge to yield combined and better results.	Accuracy of classifiers needs to be improved with nn classifiers and other approaches
2010	Proceedings of ICEE 2010	Mostafa Fathi Ganji	– ACO is used to extract a set of rules for diagnosis of diabetes disease with FADD. [15]	FADD is good approach to detect diabetes.	Single approach for deduction needs to be clubbed with other.
2011	International Journal on Soft Computing	E.P.Ephzibah,	– 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]	Genetic algorithm integrated with fuzzy logic is generating better rules.	Better feature selection mechansism can be used along with fuzzy logic
2016	International Journal of Engineering Research in Africa	G. Thippa Reddy a , Neelu Khare2,	– An attempt has been made to develop Firefly-BAT (FFBAT) optimized Rule Based Fuzzy Logic (RBFL) prediction algorithm.[19]	High accuracy,sensitivity is obtained by this new algorithm.	Other optimization techniques can be applied to improve accuracy
2016	Applied Soft Computing	Kamadi V.S.R.P. Varmaa,	– It presents an approach using principal component analysis and modified Gini index based fuzzy SLIQ decision tree algorithm. [20]	Sharp decision boundary can be overcome by fuzzy SLIQ.	Accuracy can be improved further by better fuzzy membership
2007	International Journal of Computer, Electrical, Automation, Control and Information Engineering	Kemal polat	– Combination of fuzzy c-means and svm is used for diabetes prediction on dataset[7]	Fuzzy C-means classifiy 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
2016	Informatics in Medicine Unlocked	YoichiHayashi ShonosukeYukita	– Use of a rule extraction algorithm, ReRX with J48 graft, combined with sampling selection techniques (sampling Re- RX with J48 graft) is done.[21]	High accuracy in terms of rule extraction.	the diagnosis of T2DM remains a complex problem; diagnosis
2015	International Conference on Computer and Knowledge	Kiarash ZahirniaMehdi Teimouri Rohallah Rahmaniand Amin Salaq	– This paper present and compare different cost-sensitive learning methods for diagnosis of type 2 diabetes[22]	Cost sensitive approach is effective for utilizing resources.	Assumptions are used in data sets,matrices to bring out the results

All rights reserved by www.ijirst.org

	Engineering (ICCKE)				
2015	*International Journal of Computer Science & Wireless Security (IJCSWS)*	*B.Saratha1, A.Vinodhini2*	– *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]*	*Innovative application and easy to use for diabetes deduction.*	*Image analysis needs to be more accurate for this*
2016	*Expert Systems With Applications*	*Carlos F. Vázquez-Rodríguez c , Rubén Posada-Gómez a , Armín Trujillo-Mata a*	– *Fuzzy Expert system is developed to detect neuropathy.[4]*	*This FES provides efficient system to detect neuropathy.*	*Missed data of hypertension as a result inaccurate data is obtained*
2011	*Computational vision and robotics*	*R.C.Joshi,Durga Toshinwal*	– *Binning technique is used to convert continuous data to discrete* – *Apriori algorithm is used[27]*	*Binning helps in identifying hidden patterns*	*PIMA data set prunes large data rules.*

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 health care 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

[1] Patil, Bankat M., Ramesh Chandra Joshi, and Durga Toshniwal. "Hybrid prediction model for Type-2 diabetic patients." Expert systems with applications 37.12 (2010): 8102-8108.
[2] Mohamudally, Nawaz, and Dost Muhammad Khan. "Application of a unified medical data miner (umdm) for prediction, classification, interpretation and visualization on medical datasets: The diabetes dataset case." Industrial Conference on Data Mining. Springer Berlin Heidelberg, 2011.
[3] Sanakal, Ravi, and T. Jayakumari. "Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine." Int. J. Comput. Trends Technol.(IJCTT) 11.2 (2014): 94-98.
[4] Meza-Palacios, Ramiro, et al. "Development of a fuzzy expert system for the nephropathy control assessment in patients with type 2 diabetes mellitus." Expert Systems with Applications (2016).
[5] Imran, Mohammed, et al. "Towards Early Detection of Diabetic Retinopathy Using Extended Fuzzy Logic." (2016).
[6] Kahramanli, Humar, and Novruz Allahverdi. "Design of a hybrid system for the diabetes and heart diseases." Expert Systems with Applications 35.1 (2008): 82-89.
[7] Polat, Kemal, and Salih Güneş. "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease." Digital Signal Processing 17.4 (2007): 702-710.
[8] Varma, Kamadi VSRP, et al. "A computational intelligence approach for a better diagnosis of diabetic patients." Computers & Electrical Engineering 40.5 (2014): 1758-1765.
[9] Karegowda, Asha Gowda, A. S. Manjunath, and M. A. Jayaram. "Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes." International Journal on Soft Computing 2.2 (2011): 15-23.
[10] Butwall, Mani, and Shraddha Kumar. "A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier." International Journal of Computer Applications 120.8 (2015).
[11] Nnamoko, Nonso Alex, et al. "Fuzzy Expert System for Type 2 Diabetes Mellitus (T2DM) Management Using Dual Inference Mechanism." AAAI Spring Symposium: Data Driven Wellness. 2013.
[12] Lukmanto, Rian Budi, and E. Irwansyah. "The Early Detection of Diabetes Mellitus (DM) Using Fuzzy Hierarchical Model." Procedia Computer Science 59 (2015): 312-319.

All rights reserved by *www.ijirst.org*

[13] Kim, Kwang Baek, and Doo Heon Song. "Developing an Intelligent Health Pre-Diagnosis System for Korean Traditional Medicine Public User." International Journal of Bio-Science and Bio-Technology 8.2 (2016): 227-236.
[14] Vijayan, V. Veena, and C. Anjali. "Prediction and diagnosis of diabetes mellitus—A machine learning approach." Intelligent Computational Systems (RAICS), 2015 IEEE Recent Advances in. IEEE, 2015.
[15] Ganji, Mostafa Fathi, and Mohammad Saniee Abadeh. "Using fuzzy ant colony optimization for diagnosis of diabetes disease." 2010 18th Iranian Conference on Electrical Engineering. IEEE, 2010.
[16] Basha, S. Saheb, and K. Satya Prasad. "Automatic detection of hard exudates in diabetic retinopathy using morphological segmentation and fuzzy logic." International Journal of Computer Science and Network Security 8.12 (2008): 211-218.
[17] Ranamuka, Nayomi Geethanjali, and Ravinda Gayan N. Meegama. "Detection of hard exudates from diabetic retinopathy images using fuzzy logic." IET image processing 7.2 (2013): 121-130.
[18] Ephzibah, E. P. "Cost effective approach on feature selection using genetic algorithms and fuzzy logic for diabetes diagnosis." arXiv preprint arXiv:1103.0087 (2011).
[19] Reddy, G. Thippa, and Neelu Khare. "FFBAT-Optimized Rule Based Fuzzy Logic Classifier for Diabetes." (2016).
[20] Kamadi, VSRP Varma, Appa Rao Allam, and Sita Mahalakshmi Thummala. "A computational intelligence technique for the effective diagnosis of diabetic patients using principal component analysis (PCA) and modified fuzzy SLIQ decision tree approach." Applied Soft Computing 49 (2016): 137-145.
[21] Hayashi, Yoichi, and Shonosuke Yukita. "Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset." Informatics in Medicine Unlocked 2 (2016): 92-104.
[22] Zahirnia, Kiarash, et al. "Diagnosis of type 2 diabetes using cost-sensitive learning." Computer and Knowledge Engineering (ICCKE), 2015 5th International Conference on. IEEE, 2015.
[23] Saratha, B., and A. Vinodhini. "Identifying Diabetes Using Tongue Images from Smartphone."
[24] Ferrannini, Ele, et al. "Shift to Fatty Substrate Utilization in Response to Sodium–Glucose Cotransporter 2 Inhibition in Subjects without Diabetes and Patients With Type 2 Diabetes." Diabetes 65.5 (2016): 1190-1195.
[25] Wong, Tien Yin, and Neil M. Bressler. "Artificial Intelligence with Deep Learning Technology Looks Into Diabetic Retinopathy Screening." JAMA 316.22 (2016): 2366-2367.
[26] Rai, Onkar Nath, et al. "Diabetic Peripheral Neuropathy and Its Metabolic Determinants in A North Indian Population." National Journal of Integrated Research in Medicine 7.2 (2016): 1-4.
[27] Patil, Bankat Madhavrao, Ramesh C. Joshi, and Durga Toshniwal. "Classification of type-2 diabetic patients by using Apriori and predictive Apriori." International Journal of Computational Vision and Robotics 2.3 (2011): 254-265.

All rights reserved by *www.ijirst.org*