JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

AN EXPERIMENTAL STUDY ON MACHINE LEARNING TECHNIQUES FOR DIABETIC DISEASE PREDICTION

[1]A.Mahalakshmi, [2]G.S.Nandhini,

[1]Assistant Professor, [2]PG student
[1,2]Depratment of Computer Science,
[1,2]Sri Shakthi Institute of Engineering and Technology, Coimbatore, India

Abstract : Diabetics is a considered as complex and wide spread disease as it is rapidly evolving with many peoples. Currently, duration and cost of the treatment process is long and very high due to its high recurrence. Diabetes is a group of metabolic diseases characterized by hyperglycaemia resulting from defects in insulin secretion and insulin action. The chronic hyperglycaemia of diabetes is associated with long-term damage, dysfunction, and failure of different organs, especially the eyes, kidneys, nerves, heart, and blood vessels Accurate early diagnosis prediction of diabetics are become more essential to enhance the patient's treatment procedure. Using machine learning techniques early detection of the disease are made possible. Enabling automated detection and classification of the disease can be carried out using machine learning techniques with low cost and early diagnosis of the disease. In this paper, experimental study on machine learning technique for diabetic's classification has been carried on basis of defining the disease, diagnosis of the disease, classification of the disease on basis of feature processing. Machine learning model is capable of learning the features of the disease extracted from feature extraction and feature selection model. Classification of the patterns has been represented into types. Classification results are highly discriminant with enhanced classification rate on the dynamic characteristics of the dataset. Evaluation of the technique is estimated using PIMA datasets. The evaluation of the classification technique has been done in accordance with the feature extraction and feature selection methods. Finally the performance analysis of the technique has performed with respect to classification accuracy and execution time to attain the effective results on the cross fold validation of the dataset using confusion matrix on basis of precision, recall and f measures.

IndexTerms **- Machine Learning, Diabetics, Classification, Diabetic Disease Types, Feature Selection. Feature Extraction**

I. INTRODUCTION

According to the American Diabetics Society, Diabetics is characterised by occurs mostly in peoples over 40 years of age. Diabetics can be primarily found among obesity people. Diabetes is a group of metabolic diseases characterized by hyperglycaemia resulting from defects in insulin secretion and insulin action Diabetics can be diagnosed using computer vision technique to determine the disease. Further processing is required to analyse the disease to classify it into several types on basis of the features of the disease presence.

Machine learning algorithm has been employed to classify the diabetics in addition to feature extraction and feature selection. In this paper, experimental study on machine learning technique for diabetics classification has been carried using machine learning technique to classify the types [1]. Analysis of the model is carried out on the highly discriminant features extracted using feature extraction and feature selection model. In addition, capability of the learning model is computed using the complex features of the disease. Further performance of the model is evaluated using PIMA dataset.

The evaluation of the classification technique has been done in accordance to partitioning of the dataset into the training and testing data along the validation set. In addition, pre-processing model has been analysed to compute efficiency of the model on removing the noisy, containing some irrelevant or redundant information through pre-processing techniques [2]. The performance analysis of the technique has performed with respect to classification accuracy and execution time to attain the effective results on the cross fold validation of the dataset using confusion matrix to compute the precision, recall and f measure to determine the accuracy and scalability of the models.

The rest of paper is structured as follows, section 2 describes the definition of the disease and types of the disease on basis of analysis, where section 3 describes the PIMA dataset, while section 4 presents the machine learning techniques employed to the diabetic's dataset. In section 5, review of literature on machine learning techniques has been analysed with its advantages and limitation in depth. Finally Section 6 concludes the work

II. DEFINITION OF IMPORTANT TERMS

In this section, definition of the diabetic disease has been provided with types of the disease on basis of clinical trials and pathology analysis as preliminary step of the disease classification and prediction of the dataset. They are

A. Disease Definition: Diabetes is a group of metabolic diseases characterized by hyperglycaemia resulting from defects in insulin secretion and insulin action. Deficient insulin action results from inadequate insulin secretion and/or diminished tissue responses to insulin at one or more points in the complex pathways of hormone action. During this asymptomatic period, it is possible to demonstrate an abnormality in carbohydrate metabolism by measurement of plasma glucose in the fasting state or after a challenge with an oral glucose load.

B. Symptoms of the Diabetics : Polyuria which leads to frequent urination, polydipsia will produces extreme hunger and increased thirst, weight loss, fatigue and irritability, increased appetite, slow healing on wounds, sometimes with polyphagia and blurred vision are symptoms associated with diabetic diseases. Impairment of growth and susceptibility to certain infections may also accompany chronic hyperglycaemia. The severity of the metabolic abnormality can progress, regress, or stay the same. Thus, the degree of hyperglycaemia reflects the severity of the underlying metabolic process. .

C. Type 1 Diabetics – β-cell destruction: Type 1 Diabetics occurs due to results from a cellular-mediated autoimmune destruction of the β-cells of the pancreas β in pathogens usually leading to absolute insulin deficiency [3].

D. Type 2 Diabetics – Insulin Resistance and Deficiency : A type 2 diabetic occurs due to insulin resistance and insulin.

III. DATASET DESCRIPTION

The dataset used for diabetic detection is PIMA Dataset [4]. Its attributes list has been provided in the table 1.

Table 4.1: Pima Diabetic Dataset

Attributes	Use	Value range
Pregnancy count	No of Pregnancy	0-17
Glucose	Glucose level	0-199
Blood pressure	Diastolic blood pressure	0-122
Skin thickness	Triceps skin fold thickness	0-99
Insulin	Insulin level (muU/ml)	0-846
Body mass index	Weight	0-67.1
Diabetes pedigree function	Diabetes information	0.08-2.42
Age	Age	05-81

IV. PMACHINE LEARNING TECHNIQUE FOR DIABETICS PREDICTION & CLASSIIFICATION

Diabetic Prediction is carried out using the machine learning technique. In addition to processing of the dataset, pre-processing technique has to be employed to enhance the data quality and accuracy.

4.1. Analysis of Pre-Processing of dataset

Pre-processing of the dataset is carried out to fill the missing values and normalize the dataset for effective classification. Feature extraction is employed to extract the feature of the dataset for the classification [5]. Linear discriminant analysis and principle component analysis is employed to the extract the feature of the dataset.

4.2. Analysis of processing techniques

Diabetic detection is carried out on the extracted feature vectors containing the patent details using following machine learning technique which is described as follows

4.2.1. Decision Tree

A decision tree is a composed of classification and regression process as CART (Classification and Regression tree). A decision tree is a structured decision analysis. It is easy to learn and interpret. Decision tree indirectly performs selection of features. It works on both numerical and categorical data towards classification and prediction. Decision tree algorithms uses, Giri index, chi-square, information gain and reduction in variance of the features[6].

Drawbacks of the model

- Discovering the relationships among multiple format dataset is complex and difficult

4.2.2. Random Forest

Random forest is simple algorithm in machine learning for disease classification. Random forest can be used for classification and regression tasks of medical data. It works almost similarly as decision tree, it uses bagging method .Bagging is the combination of creating models and improve the output results[7]. Random forest combines two or more decision trees to predict the stages results on oral cancer disease. Therefore Random forest works by splitting a node as random subsets of the features.

Drawbacks of the model

- Complication in determining the global minima

4.2.3. Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the simple machines learning algorithm which produces accuracy with less computational power for cancer stage classification .SVM can be used for both classification and regression process on its main objective is to create classification models. It can be carried by identifying hyperplane in n number of features which classifies the data points. There can be many hyper-planes to differentiate data points. Hyper-planes are also called as decision boundaries. The objects margins can be maximized using support vectors, by eliminating the support vectors the position and distances changes from the boundaries [8].

Drawbacks of the model

- It fails determine the long dependency sequences
- Slow converge

4.2.4. K-Nearest Neighbor (KNN)

KNN is simplest classification algorithm used in machine learning; it is suitable for both large and small datasets. It produces accurate results for more complex problems. KNN is used for classification and regression predictive models, which is mostly used for medical data classification. The commonly used distance measure in KNN is Euclidean distance method. The distance measures are arranged in order to get the top most k-value and frequent class and then results in prediction output. KNN algorithm is also used for regression tasks by calculating averages of nearest objects in a class rather than calculating the mean object in a class [9].

Drawback of the model

- It can't be explored to automated prognoses on large no of missing value in high dimensional dataset and due to unstructured data.

4.2.5. Logistic Regression

Logistic regression is the machine learning algorithm for classification and prediction purpose. Logistic regression is classified into three types namely, Binary logistic regression, Multinomial logistic regression and Ordinal logistic regression. To predict the disease class on stages is set based on threshold value by estimated probability. Logistic regression is a straight technique; for binary/multivariate classification tasks[10].

Drawback of the model

C. It could not find the similarities on slight chronic changes of the similar patients due to elimination of cluster adaption model

CONCLUSION

An experimental study on machine learning algorithm for Diabetic classification has been carried out on PIMA dataset. Especially classification approaches analysed in this study is capable of the classifying the types of the disease and predicting the other disease association with the patient. On analysis, it came to conclusion that types of the disease and prediction has been carried out with high accuracy scalability and it proved that classification model can be used to predict disease efficiently. Finally experimental analysis of machine learning technique is carried out to identify the effectiveness and robustness of the model

REFERENCES

[1] Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K.,2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003

[2] Bhargavi V.R., Senapati R.K., Curvelet fusion enhacement based evaluation of diabetic retinopathy by the identification of exudates in optic color fundus images ,2016, Biomedical Engineering - Applications, Basis and Communications, Vol: 28, Issue: 6, ISSN 10162372

[3] Emerging Risk Factors Collaboration and other, "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies," The Lancet, vol. 375, no. 9733, pp. 2215-2222, Jul. 2010.

[4] Harikumar Rajaguru and Sunil Kumar Prabhakar, Performance Comparison of Oral Cancer Classification with Gaussian MixtureMeasures and Multi Layer Perceptron, The 16th International Conference on Biomedical Engineering p(2017) 123-129

[5] M. Maniruzzaman, M. J. Rahman, M. A. M. Hasan, H. S. Suri, M. M. Abedin, A. El-Baz, and J. S. Suri, "Accurate diabetes risk stratification using machine learning: role of missing value and outliers," Journal of Medical Systems, vol. 42, no. 5, pp. 92, May 2018.

[6] B. Thomas, V. Kumar, and S. Saini, ''Texture analysis based segmentation and classification of oral cancer lesions in color images using ANN,'' in Proc. IEEE Int. Conf. Signal Process., Comput. Control (ISPCC), Sep. 2013, pp. 1–5.

[7] G. I. Webb, J. R. Boughton, and Zhihai Wang, "Not So Naive Bayes: Aggregating one-dependence estimators," Machine learning, vol. 58, no. 1, pp. 5-24, Jan. 2005..

[8I. Jenhani, N. B. Amor, and Z. Elouedi, "Decision trees as possibilistic classifiers," International Journal of Approximate Reasoning, vol. 48, no. 3, pp. 784-807, Aug. 2008.

[9] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdi- nov, "Dropout: a simple way to prevent neural networks from overfitting," The journal of machine learning research, vol. 15, no. 1, pp. 1929-1958, Jan. 2014.

[10] .L.Li, "Diagnosis of Diabetes Using a Weight-Adjusted Voting Approach," in Proc. IEEE International Conference on Bioinformatics and Bioengineering, Nov. 2014, pp. 320-324