IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 7, July 2020

Classification of Pima Indian Diabetes Dataset using Ensemble of Decision Tree, Logistic Regression and Neural Network

Mani Abedini[1], Anita Bijari[2], Touraj Banirostam[2]

Staff Data Scientist, General Electric Aviation, Dubai, UAE[1]

Associate Professor, Dept. of Computer Engineering, Islamic Azad University, Central Tehran Branch, Tehran, Iran[2]

Abstract: This paper proposed an ensemble hierarchical model to combine two or more classifiers which has been trained independently, and then fused them in the next level. This is done in two steps, first we trained a Decision Tree and a Logistic Regression models, step two we feed the output of those models to a Neural Network. The Neural Network is also trained to combine the output of previous classifiers to achieve better overall accuracy. To test our hypothesis, we used PIMA Indian diabetes database as benchmark problem. Our proposed model has achieved classification accuracy above 83% which is better than other states of the art methods in the literature.

Keywords: Data mining, Regression, Neural Network, Decision Tree, Pima Diabetes Data set, Ensemble Learning.

I. INTRODUCTION

Diabetes is the most widespread chronic disease which put a lot of pressure on the public health system. WHO estimates the total death due to diabetes will rise to 50% in the next decade. Diabetes occurs when pancreas does not produce enough insulin or produced insulin is not used effectively in body. According to WHO and ADA, there are four types of diabetes: Type-I, Type-II diabetes, Gestational Diabetes (GDM) and rare specific diabetes. Type-I diabetes is responsible of 5% to 10% of total diabetes, it occurred due to lack of insulin production. Destruction of pancreas organ is the source insulin production loss in human body which leads to insulin-dependent diabetes. Type-II diabetes, however, are more common (90% of the diabetic population) is caused by "insulin resistance" and metabolic disorders, the result is increase in sugar levels in blood [1].

GDM diabetes is s type of II diabetes and related to the pregnancy changes in body. Almost %4 of pregnancy cases will develop GDM at some stage of the pregnancy. To decrease the risk of diabetes in newborn babies, GDM diabetes should be monitored and treated accordingly. The other cause of diabetes is related to genetic and metabolic disorders.

India has the second highest number of diabetes patients in the world. In this paper we propose a novel machine learning approach to predict the risk for diabetes at exceedingly early stage of the disease to help health practitioners control the disease and reducing the impact of it in the Indian population.

This paper is organized as follows: Section II review some basic concepts of machine learnings and Previous work around using machine learning for Diabetes detection using PIMA dataset. Section III describes our proposed the classification process using an ensemble model of logistic regression, decision tree and neural network. The experimental results are also given in this section. Finally, Section V concludes the paper.

II. BACKGROUND

Machine Learning

Machine Learning (ML) is the one of computer science disciplinary, provides systems that learn from data, and improves their behaviour over time by discovering emerge patterns from training datasets automatically [2]. The main objective in ML is to learn from previous experience, it can be supervised or unsupervised learning. In supervised learning the training data contains labelled data (positive and negative examples); thus, the ML algorithm will use training data to predict the labels of new observations. ML is also categorized into three major groups: Classification, Regression, Clustering, and Reinforcement Learning. In classification problems, the objective is predicting the associated labels (categorical values) for the observations, however in regression modelling, the prediction is a continues values rather than nominal values. Clustering methods on the other hands are unsupervised methods which gives some insight about distribution of data and similarity of observations. Reinforcement learning is an area of ML where there is no notion of immediate right or wrong decision, but rather having a strategy to maximize the reward in sequences of actions [3].

DOI 10.17148/IJARCCE.2020.9701

In this study, we utilize supervised learning models for classification for diabetes disk prediction. Some of the most common classification methods are:

- Logistic Regression: that build a statistical model by fitting a linear model to describe the relationship between logit of the features and one or more independent variables. Logistic Regression is a simple approach but popular with the Machine Learning community [4].
- Support Vector Machines (SVM) is another popular ML method that finds a hyperplane which separates positive and negative classes. SVM gives better insight about important features and how they influence the decision boundary [5, 6].
- Random Forest (RF) is an ensemble model, which produce multiple trees during training by randomly selecting features and sample boosting technique. Random Forest is more robust to variance error [7,8].
- Decision Tree (J48) is a simple representation of classifying examples by developing the decision rules through nodes and making final decision in leaves. J48 is the most common decision tree [9].
- Artificial Neural Network (ANN) is a general, parametrized classification method that mimics the underlying computational model in human brain, thought vast network of connection between neurons [10].

Previous work:

There are many publications reporting their methods on Pima Indian Diabetes Dataset by leveraging various machine learning methods. Deng and Kasabov [11] achieved 78.4% accuracy for classification of PIMA data set using a combination of cross validation and Self Organizing Maps (SOM). Yu et al. explore various methods such as Neural Network and combined Quantum Particle Swarm Optimization (QPSO) and Weighted Least Square (WLS) Support Vector Machine (SVM) and achieved the classification accuracy up to 82.18% with the WLS-SVM method. Al Jarullah et al. [9] applied c4.5 algorithm to the PIMA data set and demonstrated the accuracy of 71.1%. Pasi Lukka [13] used a combination of feature selection method based on fuzzy entropy measures and similarity classifier. The proposed model was tested with four medical data sets including the Pima-Indian diabetes; and classification accuracy of 75.29% has been reported in their paper. Seera et al. [14] proposed a hybrid intelligent system, which combines Fuzzy Min–Max neural network, the decision Tree, and the Random Forest model. The hybrid intelligent system provides incrementally learning feature by incorporating the neural network model, ability to explain the decision process by incorporating the decision tree, and also achieve high classification accuracy by leveraging Random Forest model which achieved 71.35% accuracy.

Choubey et al. [15] proposed a combination of Genetic Algorithm (GA) feature selection method and Multilayer Perceptron Neural Network (MLP NN) for the classification task. The proposed model has reported 79.13% accuracy. In another paper Choubey [16] proposed similar approach by using GA for feature selection then Naive Bayes (NB) method for classification. This approach demonstrated a better accuracy of 78.69%. Smith et al. [17] proposed a neural network with Adaptive Learning Routine (ADAP) algorithm and showed an accuracy of 76%. Kumari et al. [6] applied SVM model for PIMA classification and investigated the performance of various kernels in their experiments. The paper reported the classification accuracy of 75.50% using RBF kernel and cross validation method to tune the SVM hyper parameters. Somu et al. [8] introduced RSKHT (Rough Set based K–Helly) feature selection technique and combined it with Random Forest classification method. They have tested their approach on various common classification method: Random Forest, Bayesian Network, Neural Network and Decision Tree. In their experiments they achieved 75.02%, 73.11%, 75.11%, 74.9% accuracy, respectively.

III. EXPERIMENTS

Data set:

The PID database is available from UCI Machine Learning Repository. The data set has taken from 768 women, (500 negative cases and 268 positive cases) from 21 years old and above, and 8 recorded features as follows:

- Number of previous pregnancies
- Plasma glucose concentration at 2-hours in an oral glucose tolerance test
- Diastolic blood pressure (mm Hg)
- Triceps skin fold thickness (mm)
- 2-Hour serum insulin (mu U/ml),
- Body mass index (weight in kg/(height in m) 2)
- Diabetes pedigree function
- Age (years).

One of the biggest challenges in PID data set is the presence of missing data (see Figure I). There are many reasons for lack of complete data, such as death of patients, equipment malfunctions, refusal of respondents to answer certain questions. More than 51% of cases have missing values of one or more features missing.

DOI 10.17148/IJARCCE.2020.9701

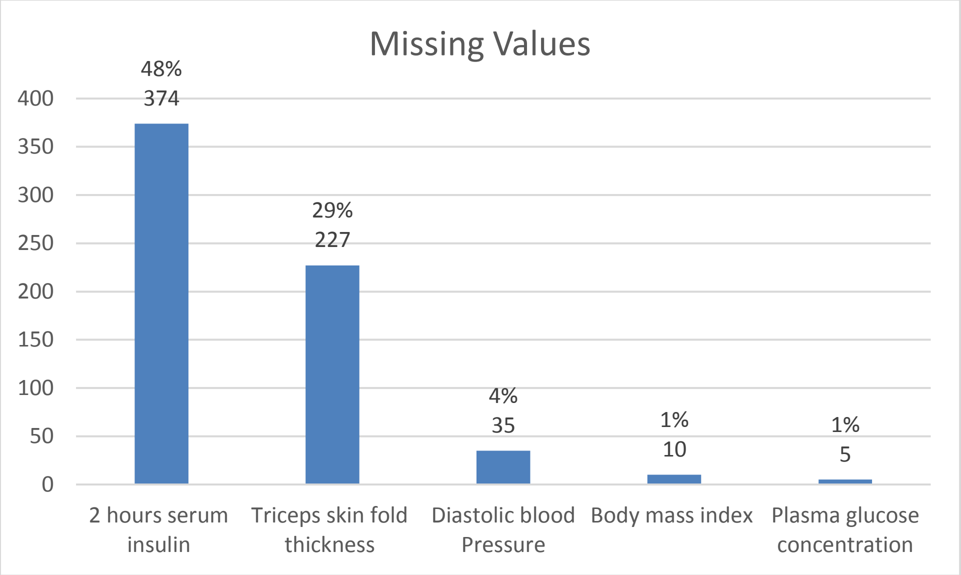

Fig. 1 Number of samples with Missing values per feature in PIMA data set

Proposed model:

In this paper, we proposed a novel hierarchical approach by combining Decision Tree (ID3), Logistic Regression, SVM (RBF Kernel), Random Forest and Neural Networks. In the first step Decision Tree, and Logistic Regression are trained independently using the training set. In the next step the output of the first level are fed to a Neural network for training. The neural network topology consists of three slayers: input layer of 4, middle layer of 10 neurons, and 15 neurons, then 1 output neuron (See Figure 2). We used adaptive learning method to adjust the learning rate according to the output of the model (See Figure 3).

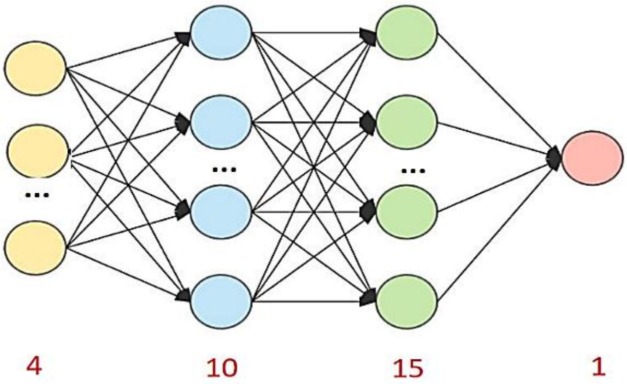

Fig. 2 The Neural Network topology

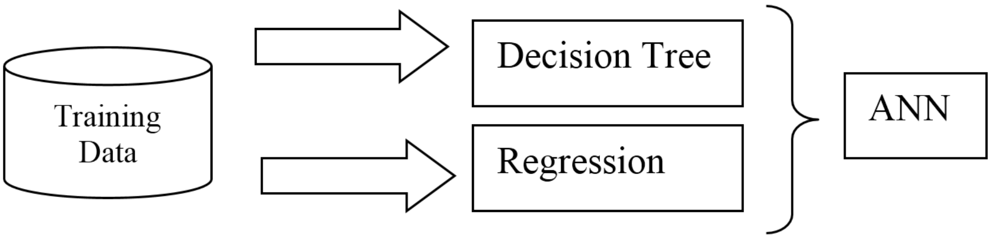

Fig. 3 Proposed model

Results:

In our experiments we have split the PID into %30 test and %70 training portion.

We could achieve %65 accuracy using decision tree; the regression model achieved better accuracy, over %80. However, by combining the output of previous models and trained an Artificial Neural Network (ANN), the overall accuracy jumped to %83. (see Table I).

DOI 10.17148/IJARCCE.2020.9701

Table I Decision tree result

Method	Overall Accuracy	Label	Precision	Recall
Decision Tree (ID3)	65.84%	Positive (Diabetes)	65.84	33.33
		Negative (Normal)	84.7	95.53
Logistic Regression	80.71%	Positive (Diabetes)	51.97	81.48
		Negative (Normal)	94.38	80.51
Our proposed Ensemble model: (Artificial Neural Network + Logistic Regression + Decision Tree)	83.08%	Positive (Diabetes)	25.00	82.35
		Negative (Normal)	98.57	83.13

The experiment results suggest that the proposed model can effectively combine the Decision Tree and Regression model and improve the overall accuracy. In comparison to other methods in the literature, our proposed model also demonstrated improved performance.

IV. CONCLUSION

Diabetes is a medical condition that cause serious complication, and result in unnecessary pressure on the healthcare system. The untreated diabetes can cause blindness, heart disease, kidney failure and nerve damage. In this paper, we proposed a machine learning method to analyze PIMA data set and predict the risk of diabetes based on available features. We proposed an ensemble classification method, comprise of two levels: in the first level we trained a Logistic Regression and a Decision Tree (ID3) independently. At the next level, we combined the outputs of previous level by using an ANN. In our experiments, we demonstrated that, the ensemble model improves the accuracy, in compare to the individual method.

REFERENCES

[1]. https://www.who.int/news-room/fact-sheets/detail/diabetes
[2]. Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996), " Data Mining to Knowledge Discovery in Databases".
[3]. Kaelbling, Leslie P.; Littman, Michael L.; Moore, Andrew W. (1996). "Reinforcement Learning: A Survey". Journal of Artificial Intelligence Research. 4: 237–285
[4]. Alan Agresti Department of Statistics University of Florida, Gainesville, Florida, An Introduction to Categorical Data Analysis 2nd Edition, (2007).
[5]. Cortes C, Vapnik VN. Support-vector networks. Mach Learn. 1995;20(3): 273–97
[6]. Jegan, Chitra. (2013). Classification Of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications. 3. 1797 - 1801.
[7]. Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.
[8]. Somu N, Raman MR, Kirthivasan K, Sriram VS. Hypergraph Based Feature Selection Technique for Medical Diagnosis. J Med Syst. 2016;40(11):239.
[9]. A. Al Jarullah, "Decision tree discovery for the diagnosis of type II diabetes," 2011 International Conference on Innovations in Information Technology, Abu Dhabi, 2011, pp. 303-307
[10]. Bishop, C. M. 2006. Pattern Recognition and Machine Learning. Springer.
[11]. D. Deng and N. Kasabov, " On-line pattern analysis by evolving self- organizing maps", In Proceedings of the fifth biannual conference on artificial neural networks and expert systems (ANNES), 2001, pp. 46-51.
[12]. Yue, et al. " An Intelligent Diagnosis to Type 2 Diabetes Based on QPSO Algorithm and WLSSVM," International Symposium on Intelligent Information Technology Application Workshops, IEEE Computer Society, 2008.
[13]. Luukka, Pasi. (2011) 'Feature selection using fuzzy entropy measures with similarity classifier', Expert Systems with Applications, Elsevier, Vol. 38, pp. 4600–4607.
[14]. Seera, Manjeevan., Lim, Chee Peng. (2014) 'A hybrid intelligent system for medical data classification', Expert Systems with Applications, Elsevier, Vol. 41pp. 2239–2249.
[15]. Choubey, Dilip Kumar., Paul, Sanchita. (2016) 'GA_MLP NN: A Hybrid Intelligent System for Diabetes Disease Diagnosis', International Journal of Intelligent Systems and Applications (IJISA), MECS, ISSN: 2074–904X (Print), ISSN: 2074–9058. (Online), Vol. 8, No. 1, pp.49–59.
[16]. Choubey, D.K., Paul, S., Kumar, S., & Kumar, S. (2016). Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection.
[17]. Smith, JW., J.E.Everhart, et.al.- "Using the ADAP learning algorithm to forecast the onset of diabete mellitus", Proceeding of the Symposium on Computer Applications & Medical Care (Washington, DC). R.A. Greenes. Los Angeles, CA, IEEE Computer Society Press, 1988, pp.261-265.