IJCSN - International Journal of Computer Science and Network, Volume 8, Issue 3, June 2019
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

Optimized Medical Disease Analysis Using Autoencoder and Multilayer Perceptron

[1] Juby Mary Abraham; [2] Kavitha V K; [3] Dr. Radhakrishnan B

[1] CSE Department, BMCE
APJ Abdul Kalam Technological University, Sasthamcotta, Kerala, India.

[2] CSE Department, BMCE
APJ Abdul Kalam Technological University, Sasthamcotta, Kerala, India.

[3] CSE Department, BMCE
APJ Abdul Kalam Technological University, Sasthamcotta, Kerala, India.

Abstract-The machine learning and health care combination are sharply related. Machine Learning can play a crucial role in predicting the presence or absence of kidney disease. An effective method for kidney disease prediction is discussed in this work. The proposed system consists of autoencoder combined with a multilayer perceptron for a classification problem. An autoencoder is an artificial neural network that trains a model to extracting useful features. We used kidney disease analysis as a case study for simulating the proposed system and its efficiency is evaluated against the current approaches. An autoencoder is able to integrate into an optimal representation which is then classified by the MLP network to derive the final output. The proposed system clearly gives a better result than the traditional ones. This learning method has a good effect on the classification of disease prediction and guidance for the diagnosis of disease in medical.

Keywords- *Machine learning, artificial neural network, Autoencoder, Multilayer Perceptron*

1. Introduction

Kidney failure can also be called as chronic kidney disease. Ten in hundred people worldwide are suffering from kidney disease. 15% of the population worldwide experience from chronic kidney disease. The present work focuses on predicting weather a person is miserable from CKD or non-chronic disease using machine learning. The precise evaluation of medical data assets quick disease identification, patient concern, and community services due to the data mining advancement in healthcare and biomedical communities. In the medical field, diagnosis of chronic diseases is very essential as these diseases persist for long time. When the standard of medical data is incomplete the analysis accuracy is lowered. Different regions exhibit unique features of certain divisional diseases, which may weaken the prediction of disease outbreaks. Machine learning is playing an important role in different infrastructures which can make better decisions on patient's diagnosis and lead to overall improvement of healthcare services due to the improvement of the Internet, big data, cloud computing and artificial intelligence.

There are different categories in machine learning algorithms for classification of kidney disease. Nearest Neighbours, Clustering, regression algorithm, Random forest, neural network, Bayesian, support vector machine and more. Some proposed algorithms need to be improved with the rapid development of computer level, which is restricted to bottlenecks of the amount of computing capacity at the time. Artificial neural network prediction of the diagnosis of kidney disease includes easily-obtained diagnostic methods and of low cost. The classification of the diagnosis of the patient is studied by combining the knowledge of machine learning and data mining. The accuracy and error rate of the classification are compared to judge the performance superiority of the algorithm. The application of machine learning in the medical data mining will doubtlessly bring a new approach to medical diagnosis.

The rest of this paper is organized as follows. A brief literature review of existing works on kidney disease prediction is given in Section 2. The proposed method is explained in Section 3. The experimental results are discussed in Section 4. Finally Section 5 concludes the paper.

2. Related Work

Dr. S. Vijayarani, Mr.s.dhayanand, [1] have used two machine learning techniques namely SVM and ANN to predict kidney diseases. To handle discovery of hidden patterns and relationships and provides a remedy by utilizing advanced data mining techniques. Based on its accuracy and execution warrant, compare the functioning of these two algorithms is the main purpose of this work. The SVM catch the hyperplane wielding support vectors and margins. The biggest limitation of the support vector approach lies in choice of the kernel and greater execution time. ANNs have the ability to learn and model non-linear and complex relationships. ANN does not impose any restrictions on the input variables ANN algorithm is enhanced to minimize the execution time. ANN outperforms best in classification process compared to SVM algorithm.

Sujata Drall, Gurdeep Singh Drall, Sugandha Singh, Bharat Bhushan Naib, [2] performed a Chronic kidney disease prediction using a new machine learning approach. First the dataset is preprocessed and select subset of relevant attributes from the total given attributes. This stage helps in reducing the dimensionality and making the model simplest and easy to use. KNN scans through all the previous experiences known as data points and looks up the closest experience to find a solution for finding a class for a new data point. The data of previous data points is maintained and the class of a new data point is determined by the majority of nearest data points. This algorithm is fast and easy. The results show that Chronic Kidney Disease can be better predicted by using K-Nearest Neigbour algorithm with higher accuracy compared with naïve Bayes classification.

Manish Kumar, [3] applied a machine learning algorithm called random forest for the prediction of chronic kidney disease. A random forest is a union of tree predictors so that values of a random vector examined sovereignly and with the matching dispersion for all trees in the forest. Draw n tree bootstrap using original samples data. For every bootstrap samples, an unpruned classification tree is produced. At each node, arbitrarily sample m try of the predictors and choose the best split among that variables. By aggregating the predictions of the n trees using majority votes for classification to predict new data. An error rate can be found on estimation, based on the training data. In this study, the experiments were conducted for the prediction task of Chronic kidney using different machine learning algorithms namely Random Forest (RF) classifiers, Naïve Basis and SVM. The results obtained show that the RF classifier performs well and greater accuracy compared to other classifiers.

M. Praveena, N. Bhavana, [4] performs a prediction of chronic kidney disease using C4.5 algorithm. Machine learning is employed by buildup a decision tree using the C4.5 algorithm and predicted while the person is suffering or normal from kidney trouble. Data entered is processed and a decision tree will be generated by computing entropy and information gain values as per the rules of c4.5 algorithm. From root node to the leaf node, prioritizing of the node to be placed depends on the homogeneity of the node and also the calculated values. In decision tree, the prediction usually occurs at the leaf node. The c4.5 learning algorithm to predict patients with chronic kidney failure (ckd) disease and patients who do not (notckd) suffer from the disease.

Himanshu Kriplani, Bhumi Patel and Sudipta Roy, [5] have used a deep artificial neural network technique for the prediction of chronic kidney diseases. A varied range of parameters is used for training and it outcome is the more accurate predictions. Our proposed method is based on deep neural network which predicts the presence or absence of chronic kidney disease with highest accuracy compared to other available algorithms like SVM and Random Forest. Among all these classifiers, deep artificial neural Network classifier results good accuracy.
Pinar Yildirim, [6] performed a chronic kidney disease prediction on imbalanced data by multilayer perceptron. The most beneficial technique to handle imbalanced data is sampling. It adjusts the data set by replicating the examples of minority class. To analyze the data set three sampling algorithms were used and their performance was evaluated by multilayer perceptron by changing learning rate. The results were evaluated on accuracy metrics and execution time. Among the different sampling algorithms, Resample method has better accuracy results on the data set than the others. Sampling algorithms can improve the performance of multilayer perceptron with optimum learning rate parameter for learning process.

3. Proposed System

The proposed architecture is shown in Fig 1. After data preprocessing and standardization, the data to be passed to a neural network called an autoencoder is used for optimized useful representation of the input. Autoencoder tries to minimize the reconstruction error. It learns some important features present in the input to reduce the error.

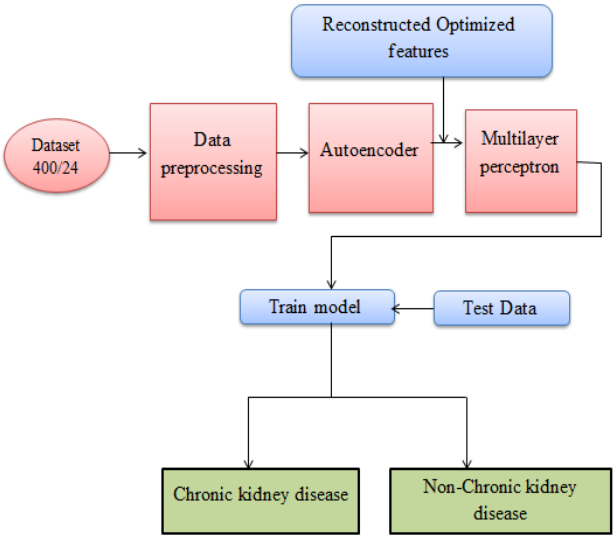

Fig.1 Proposed Architecture

Encoder generates a new set of features which is a combination of original features. Encoding in autoencoders helps to identify the latent features present in the input data. After encoding and decoding process, that is after attaining the input and output almost similar trash the decoder and takes the code layer as output that contains compressed useful features. Once the useful input will be extracted using an autoencoder the reconstructed useful features will be provided as input into a multilayer perceptron which will be trained to classify the result. The test data's are given to the trained model and predict whether a patient has CKD or NCKD. of autoencoder encodes the input vector received from input layer into code (features). By increasing or reducing the number of hidden layer neurons with respect to the input layer neurons, an autoencoder is trained. During the training phase, the input vector is mapped to the features. Autoencoder tries to represent the input vector into features which are useful for data classification process. First take the input, encode it to identify latent feature representation. Decode the latent feature representation to recreate the input. Calculate the loss by comparing the input and output. To reduce the reconstruction error we back propagate and update the weights. Weight is updated based on how much they are responsible for the error.

3.1 Dataset

The kidney disease data's for training is taken from UCI Machine Learning Repository which is publicly available. It includes 400 patients with 24 attributes collected from each of the patients for the training to prediction algorithms. A patient has chronic kidney disease or not is predicted as output.

3.2 Data Preprocessing

Remove rows that have missing data and replace to binary strings (yes/no, present/not present) with 1's and 0's. Then standardize the dataset. Data standardization is the critical process of bringing data into a common format.

$$Z=(x_1-\mu_1)/\sigma_1 \qquad (1)$$

μ_1 is the mean of the population andσ is the standard deviation.

3.3 Autoencoder

AE is used for extracting useful features and reconstructing the input. An autoencoder has different layers namely input layer, code layers and output layer. Autoencoders have the equal number of neurons in the input layer and output layer, as it train itself to reconstruct the given input. The hidden layer

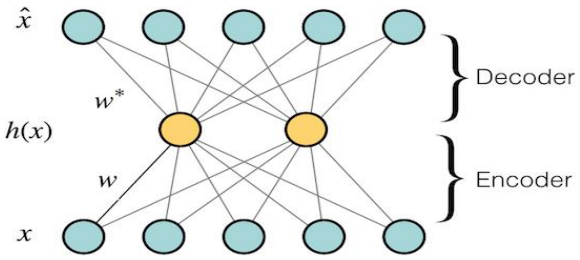

Fig. 2 Architecture of Autoencoder

An auto-encoder is a type of neural network that main objective is to find a new representation of input without too much loss of information and the input can be reconstructed. An auto-encoder has one input layer, one output layer and multiple hidden layers.

The encoder part takes the input X∈ P^d and maps it to the h∈ P^L where

$$h = \sigma(Wx+b) \quad (2)$$

Take the first row bought in an array as the input. Encode the input into another vector h. *h* is a lower dimension vector than the input. Decode the vector h to recreate the input. This h is generally referred to as code layer or latent representation. Here σ, W and b are the activation function, input weight matrix, and bias vector respectively. The decoder part maps the latent representation h∈ P^L to the output $\hat{X} \in P^m$. Here

$$\hat{X} = \gamma\,(W_0 h + b_0) \quad (3)$$

Where $\hat{X}$ is the reconstruction of the input and γ,W_0, b_0 are the activation function, output weight matrix, and bias vector respectively. The network comprises rectified linear unit (ReLU) hidden layers to implement the learning algorithm. It is an activation function. The differentiation of the function is defined as

$$F(x) = \max(0, x).$$
$$\frac{df}{dx} = \{x; x > 0, 0: x < 0\} \quad (4)$$

Calculate the reconstruction error Re. Reconstruction error is the difference between the input and output vector. Our goal is to minimize the reconstruction error so that output is similar to the input vector. Back propagate the error from output layer to the input layer to update the weights.

In most cases, auto-encoder outperforms Principal Component Analysis in processing high dimensional complex datasets because auto-encoder performs both linear and non-linear projections, while PCA performs only linear projection. Auto-encoders have been successfully used to efficiently extract meaningful features in disease diagnosis based on kidney disease dataset.

3.4 Multilayer Perceptron

An autoencoder is able to represent to optimal representation which is then classified using multilayer perceptron to derive the final result. A multilayer perceptron (MLP) is a type of feed forward artificial neural network. An MLP consists of minimum three layers of nodes that is input layer, hidden layer and an output layer. MLP uses a supervised learning technique named back propagation for training. In multilayer perceptron, a node can be modeled as a neuron which calculates the weighted sum of the inputs at the existence of the bias and passes this sum through the activation function.

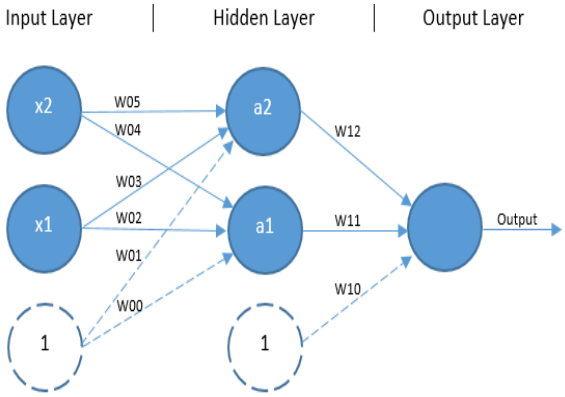

Fig. 3 Architecture of Multilayer Perceptron

The total process is defined as follows:
Input data is provided to input layer for processing, which produces a predicted output

$$u_j = b_j + \sum_{i=1}^{q} w_{ij} X_i \quad (5)$$
$$z_j = f(u_j) \quad (6)$$

Where u_j is the linear combination of inputs X_1, $X_2, \ldots X_q$, b_j is the bias, w_{ij} is the connection weight between the input *Xi* and the neuron *j*, and $f(u_j)$ is the activation function of the neuron. The sigmoid function is a popular choice of the activation function and it is defined as

$$f(u_j) = \frac{1}{1+e^{-u_j}} \quad (7)$$

and y_j is the output. There is a cost involved in calculations is the difference between the real and predicted values. Thus we need to reduce the cost of the model that is increase the accuracy of the predictions. The predicted output is subtracted from actual output and error value is calculated

$$E = \frac{1}{2}\sum(t - y_j)^2 \quad (8)$$

After the computation it is possible that estimated output is far away from actual output resulting high error. To cope up with this problem we go back and change the weights to get the least error possible. Such a process is known as backward propagation. A round of forward and backward propagation is termed as epoch. The network then uses a Back propagation algorithm which adjusts the weights. For weights adjusting it starts from weights

between output layer nodes and last hidden layer nodes and works backwards through network. The needed change for each weight can be represented as

$$\Delta w_{ij}(n_1) = - \mu \frac{\partial \varepsilon(n_1)}{\partial v_j(n)} y_i(n_1) \quad (9)$$

Where the output of the previous neuron is denoted as y and the learning rate is represented by μ.

Update the weights with

$$w(t_1+1) = w(t_1) + \Delta w(t_1) \quad (10)$$

Where t_1 is the iteration number, W is the connection weight. When back propagation is finished, the forwarding process starts again. The process is repeated until the error between predicted and actual output is minimized. After that test the data with this network and an output is obtained. Then obtained output and trained output is compared and choose the minimum distance and a class is obtained.

4. Experimental Result and Discussion

This study is carried to predict weather a patient is suffering from Chronic Kidney Disease or not. For the implementation of the proposed algorithm, this prediction model is created in Python programming language. The proposed approach used an autoencoders for optimized representation and multilayer perceptron as classification algorithm is used as a perfect model to examine the kidney disease prediction. The results are presented in Table 1. We can see that chronic kidney disease prediction with multilayer perceptron gives more accurate result than the SVM, KNN and Random Forest The proposed model gives 96.475% accuracy. The main advantage of this method is the low computational complexity, cheap implementation cost and accurate result.

Table 1: Prediction Accuracy of different methods

Method	Accuracy
SVM	63. 3237
KNN	66.8479
Random Forest	55.3202
Proposed method	96.475

5. Conclusion

An effective method for kidney disease prediction is discussed in this paper. In this work, we proposed a framework for chronic kidney disease classification of data using autoencoders and MLP. Chronic kidney disease dataset is taken for training process of the model. This model is compared with several neural network approaches and other state-of art approaches in the literature. From the results it is evident that the proposed model outperforms other model with an accuracy of 96.475% which is quite excellent for the ideal classification model. Based on the experiments and observations it is concluded that the proposed framework for kidney disease prediction can be used as powerful tool for the disease diagnosis process. This model helps in predicting the chronic kidney disease of a patient with better which are important in the medical world. In the future, this system can be improved using DNN based techniques to enrich the input features.

References

[1] Dr. S. Vijayarani1, mr.s.dhayanand, "Kidney disease prediction using svm and Ann algorithms", International Journal of Computing and Business Research (IJCBR).

[2] Sujata Drall, Gurdeep Singh Drall, Sugandha Singh, Bharat Bhushan Naib, "Chronic Kidney Disease Prediction Using Machine Learning: A New Approach", International Journal of Management, Technology And Engineering.

[3] Manish Kumar, "Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm", IJCSMC, Vol. 5, Issue. 2, February 2016.

[4] M. Praveena, N. Bhavana, "Prediction of Chronic Kidney Disease using c4.5 Algorithm", International Journal of Recent Technology and Engineering(IJRTE).

[5] Himanshu Kriplani, Bhumi Patel and Sudipta Roy, "Prediction of Chronic Kidney Diseases using Deep artificial neural network Technique", Researchgate.

[6] Pinar Yildirim, "Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron", 2017 IEEE 41st Annual Computer Software and Applications Conference.

Author Profile

Juby Mary Abraham received her B.Tech (CSE) degree from University of Kerala in 2016. She is currently pursuing her Masters in Computer Science & Engineering from KTU. Her research interests are data mining, machine learning and image processing.

Dr. Radhakrishnan B is working as the Head of CSE department. He has more than 14 years' experience in teaching and has published papers on data mining and image processing. His research interests include image processing, data mining, image mining.

Kavitha V K is working as Assistant Professor in computer science and engineering Department. She has more than 10 years' experience in teaching. Her research interests focus data mining and machine learning. She has published several papers on data mining.