An Efficient Disease Prediction in Big Data using Neuralnetwork based Optimization Method

K.Tamilselvi\(^a\) and Dr. K.Ramesh Kumar\(^b\)
\(^a\)Research Scholar, Research and Development, Bharathiar University, Coimbatore - India
\(^b\)Research Guide, Bharathiar University, Coimbatore – India

**Article History:** Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

**Abstract:** Generally, with Big Data concept large complex data are handled easily and in healthcare it helps to access data rapidly. Previously, Big Data in healthcare reached a complexity level but sometimes it is impossible to obtain the required data for use; thereby the growth in healthcare sector is slow. Big Data is much interesting when used to analyze healthcare data. Classifiers with cost-sensitive factors increases the stability of classification and reduces its computational costs when dealing with large scale, imbalanced and redundant datasets like medical data. Moreover, the nature and growth of disease are unknown; thus, making prediction complex. This work predicts various diseases where the parameters of neural network are optimized by using big data which includes data from social media. The efficiency of the TRAPezoidal Neural Network (TRAP-NN) and Improved Whale Optimization Algorithm (WOA) learning models named as (IWOA-TRAP-NN) were compared with the two standard methods such as Genetic Algorithm optimized Convolutional Neural Network (GA-CNN) and Ant Colony Optimized Convolutional Neural Network (ACO-CNN). The proposed TRAP-NN achieves 86% accuracy, 74% of F1 score and 56.2% of kappa static. The results show that the TRAP-NN performs better than GA-CNN and ACO-CNN.

**Keywords:** Optimization, neural network, disease prediction, bigdata, healthcare.

1. Introduction

Several kinds of pressures come in life like aging and accelerating pace of life. Every factor increases the frequency of diseases like diabetes and cancer year by year. Hence, preventing measures and providing treatment to cure the diseases is of demand in medical as well as healthcare environment. Simultaneously, as medical informatization has been advanced continuously, public health sector of China has gathered a wealth of data resources which is similar to big data. The enormous resources of medical data generally contain more valuable information which includes diagnosed results of patients and the rules to provide treatment. As a distinctive continuing epidemic, the frequency of diabetes is high recently which is an increasing trend [1].

At present, models related to machine learning are commonly employed for predicting the diseases [2]. Several research works have been carried out in diagnosing diabetes and providing treatment where some results were obtained. When 16 patients were monitored, statistical methods were used for analyzing the risks of diabetes [3]. Simultaneously, few research works employed decision tree and multilayer perceptron approaches for comparative analysis. Machine learning techniques like random forest (RF), Support Vector Machine (SVM), and naive Bayes (NB) were used to observe the early symptoms for predicting diabetes. From the results, it was proved that decision tree (DT) and RF methods achieved better prediction results while dealing with diabetes data [4]. In the meanwhile, deep convolutional neural network (DCNN) was used in several applications due to its powerful ability of extracting features. It possibly extracted deeper features from a huge training data due to the hierarchical network structure which was not able to obtain using conventional classifiers[5]. Thus, this approach was extensively employed in applications like image recognition, speech recognition, text detection etc. [6]. It is well known that medical dataset has rich characteristics and features which helps in discovering potential laws and valuable information from when DCNN was applied to medical data [7]. Practically, it has several significant potential and social value [8]. Most of the existing works tried to predict disease using only the search query data via Internet. But there aroused a demand to consider different big data and environmental factors while predicting diseases [9]. Moreover, with models using deep learning approaches, performance of prediction rate was improved by optimizing the parameters of these approaches [10]. In view of this, the main objective is to classify the risk level of the patient by predicting the disease using efficient optimization technique which hence reduce the massive challenges.

The remaining part of this paper is presented as follows: Section 2 discusses the works in the literature relevant to the system proposed. Section 3 elaborates the designed disease prediction model with optimization adopted at various stages. The results are presented in section 4. At last, section 5 concludes the work with closing remarks and future directions.

2. Literature survey

In [11], generic structure-based assessment method was predicted which included non-imaging and imaging data termed as Graphic Convolutional Networks (GCNs) and was employed in application dealing with
brain. In [12] a framework was designed to examine the efficiency of various classifiers and classifiers developed based on Ensemble Classifier (EC). In [13], a novel methodological regime was formulated for diabetes and the classification was based on fuzzy rule to provide treatment for the disease. In [14], heart disease prediction approach was developed using Machine Learning techniques like Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) techniques. In [15], five different classifiers were examined which included Artificial Neural Network (ANN), SVM, RF, DT, K-Nearest Neighbor (KNN) and observed that RF classifier produced the highest accuracy among all. In [16], Decision tree based Neural Fuzzy System (DTNFS) approach was introduced for analyzing and predicting different sort of heart diseases. This work developed a cost-effective intelligent system which outperformed the existing systems. Particularly, data mining approaches were employed for improving the prediction of heart disease. It was found that SVM and NN produced higher prediction rate for heart disease. Yet, data mining techniques are not suitable for the prediction of heart disease. In [17], for heart disease prediction employed Genetic algorithm and generated optimum set of attributes which as useful for prediction.

3. Proposed Methodology

The major objective of this proposed model is to classify the risk level of the patient by predicting the disease. Fig. 1 demonstrates the general architecture for the proposed framework. Once the symptoms are identified in a patient efficiently, an optimal approach is applied to the model to improve the performance of the system thereby determining an optimum series of diseases. This prediction model is designed using TRAP-NN as higher accuracy rates for prediction can be produced.

![Figure-1 System architecture for disease prediction](image)

4. Dataset description

UCI archive data is used which consists of 270 cases with complete characteristics. UCI is one of the open-source online dataset with huge several sicknesses, area speculations and information inventors utilized by specialists.

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Items</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured</td>
<td>Patient’s Demographics</td>
<td>Details like Gender, height, weight, age etc.,</td>
</tr>
<tr>
<td></td>
<td>Living habits</td>
<td>Smoking or not</td>
</tr>
<tr>
<td></td>
<td>Examination items and results</td>
<td>Physical check</td>
</tr>
<tr>
<td></td>
<td>Diseases</td>
<td>Cerebral infraction and other heavy disease</td>
</tr>
<tr>
<td>Unstructured</td>
<td>Readme illness of the patient</td>
<td>Health history (Illness or not)</td>
</tr>
<tr>
<td></td>
<td>Medical records maintained by doctors</td>
<td>Integration records</td>
</tr>
</tbody>
</table>

5. Construction of network

To classify various diseases automatically, a TRAPezoidal Neural Network (TRAP-NN) was designed and implemented with unifying 3D bounding box estimation and in-region where features and weights are shared among various tasks. Basically, TRAP-NN uses the principle of FPN, where various feature map levels are combined for promoting discriminative feature extraction. The responsibility of the encoder is to focus on the 3D bounding box estimation. Decoder classifies VoI in-region of shared features with various levels of pyramid network. Convolutional 3D kernels are more expensive than 2D variants. Further, 3D framework has numerous trainable parameters, where every layer of the model adds

\[ C_{i} = \prod_{i=1}^{n} k(x,y,z) \]

Cl indicates the number of feature maps in layer l, and k \( \{ x,y,z \} \) represent the kernel size related to the spatial dimension. Due to this, the network is increasingly inclined to overfitting thereby drastically increasing the GPU memory.
Improvised Whale Optimization Algorithm (IWOA)

There are three main stages in the whale optimization algorithm, which are searching prey, encircling prey and Logarithmic spiral prey. Among them, searching for prey is the exploration stage of the algorithm. Moreover, encircling prey and Logarithmic spiral prey is the exploitation stage of the algorithm.

Step 1: Obtaining the objective function value of all points, and getting the best point as \( X_g \), and then getting the second-best point as \( X_b \), supposing that \( X_s \) is the one should be substituted. \( f(X)_s \), \( f(X)_b \) and \( f(X)_g \) show the objective function values.

Step 2: Obtaining the objective function value of the middle point \( X_c \) between point \( X_g \) and point \( X_b \)

\[ X_c = \frac{X_g + X_b}{2} \]

Step 3: Obtaining the reflection point \( X_r \) by using the following formula. \( \alpha \) is the reflection coefficient, which is set as 1:

\[ X_s = X_c + \alpha (X_r - X_c) \]

Step 4: If \( f(X_r) < f(X_g) \) getting the expansion point from formula

\[ X_e = X_c + \mu (X_r - X_c) \]

where the expansion coefficient is \( \mu \), which usually set to 2. If \( f(X_e) < f(X_g) \), \( X_s \) will be replaced by \( X_e \), otherwise, \( X_s \) will substitute \( X_r \).

Step 5: If \( f(X) < f(X_r) \), the compression point can be acquired

\[ X_c = \lambda (X_s - X_c) \]

Step 6: If \( f(X_g) < f(X_r) < f(X_s) \), shrink point is \( X_w \) and the shrink coefficient is \( \Omega \)

\[ X_w = X_c - \alpha (X_s - X_c) \]

IWOA IN TRAP-NN

The representation between slices is learnt using slice-wise NN by using cuboid kernels of size 1×1×n. The aim is to wrap the 3D input into a 2D feature map using 1D slice-wise convolution without considering the channels and this operation enables the network to concentrate on slices. Therefore, initially, 2D convolutional kernels are set to 1, where the third one n, depends on the number of expected sets, a convolution can stack before deriving a 2D feature map,

\[ n = \frac{D-1}{2} + 1 \]

here D denotes the third dimension. The feature maps of various scales obtained for the same view are combined using element-wise addition for strengthening the unique patterns

\[ y(v) = f(x(w_i)1 + Fv(x, w_i)2 + Fv(x(w_i)3) \]

\( Fv(x, \{W_i\}) \) is a function which has to be learned for transforming input \( x \) to different feature maps, where \( j \in \{1, 2, 3\} \) represent various scales.

IWOA IN TRAP-NN

Before optimizing perceptron, initially, neural network structure is obtained. The neurons of the input as well as output layer is adopted using classified datasets, while the neurons of the hidden layer is decided by the Kolmogorov theorem

\[ \text{Hidden} = 1 + \text{Input} \times 2 \]
When IWOA is adopted to obtain the values of connection weights and bias, D represents the dimension of the candidate solution

\[ D = (\text{input} \times \text{hidden}) + (\text{hidden} \times \text{output}) + \text{hidden} \{ \text{bias} \} + \text{output} \{ \text{bias} \} \]

where the number of neurons of the input, hidden and output layers are expressed by Input, Hidden and Output. For the hidden and output layer, number of bias is shown by Hidden \{ bias \} and Output \{ bias \}

The mathematical models of the IWOA algorithm are denoted by vectors, so vectors \( N \) is used to represent agents in the population. Each agent in population is denoted, \( \{ X_1, X_2, X_3, \ldots \} \) \( \{ iw, hw, hb, ob \} \) by \( X \) \( \{ \text{by} iy \} \) \( \{ \text{and} \text{hw} \} \). Generally, after agents are defined by vectors, the objective function of NN defines the fitness function of TSWOA algorithm. Hence, the fitness function is considered as the difference between the theoretical and actual output in the neural network

\[ \text{MSE} = \sum_{i=0}^{n}(O_k - D_k) \]

where the total number of output is shown by \( m \), the desired output and the actual output of \( i^{th} \) input are exhibited by \( k_d \) and \( k_o \)

**Algorithm: Improved Whale Optimized Algorithm TRAPEzoidal Neural Network (IWOA-TRAP-NN)**

**Input:** Dataset (D), weight matrix \( W \), hidden and visible layer element bias \( B \) and \( A \) respectively

**Output:** Categorized disease

**Start**

\[ \theta = \{ W, A, B \} \]

**Initialize global search**

Input parameters- \( \text{max}_\text{iter}, s, n2, \eta \)

**Construct the population** \( M \)

\[ M_i = \{ M_{i1}, M_{i2}, \ldots, M_{iD} \} \]

**Evaluate the fitness** \( f(t) \)

\[ \tau_i \{ t \} \leftarrow D \]

\[ D+1 \leftarrow \{ \tau_i \{ t-1 \} \} \]

**Update spiral** \( (p) \), \( \rho = \tau_i \{ t \} \)

**Update the position**

\[ X_i(t) = x_i(t-1) + R_w(\tau_i x_i(t-1)) \]

**Concurrent factor recusrcion in network by**

\[ W = \{ wij \in R(n*m) \} \]

**If** \( w < D \) then

**Initialize the bias threshold**

**Else**

**Go for global search**

Hidden layer \( (H) = \{ H_1, H_2, H_3, \ldots, H_n \} \)

**If** \( n < n_k \)

\[ P(x < m), ni(n_l, n_2, n_3, \ldots, n_m) \]

**MSE** = \[ \sum_{i=0}^{n}(O_k - D_k) \]

**N_m \leftarrow MS**

**6. Performance analysis**

For experiment, standard datasets including 20 types of disease are used which can be accessed from secondary database. The data available in the datasets were captured from various scenarios in health sector environment, which broadly evaluates the proposed TRAPEzoidal Neural Network (TRAP-NN) and Improved Whale Optimization Algorithm (IWOA) learning models which is named as IWOA-TRAP-NN) were compared with two standard methods such as Genetic Algorithm optimized Convolutional Neural Network (GA-CNN) and Ant Colony Optimized Convolutional Neural Network (ACO-CNN) The tool used for obtaining the result is PHYTHON.

**Accuracy** is the ability of prediction obtained by proposed deep learning model. True positive (TP) and true negative (TN) are the predictions made by the model indicating the presence and absence of attack. False positive (FP) and false negative (FN) are the false predictions made.

\[ \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \]

Table 2 shows the comparison of accuracy between existing GA-CNN, ACO-CNN and proposed IWOA-TRAP-NN.

<table>
<thead>
<tr>
<th>Number of Datasets</th>
<th>GA-CNN</th>
<th>ACO-CNN</th>
<th>IWOA-TRAP-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>45</td>
<td>78</td>
<td>81</td>
</tr>
<tr>
<td>200</td>
<td>59</td>
<td>80</td>
<td>83</td>
</tr>
<tr>
<td>300</td>
<td>61</td>
<td>84</td>
<td>85</td>
</tr>
</tbody>
</table>
The figure 4 shows the comparison of accuracy between existing GA-CNN, ACO-CNN, and proposed IWOA-TRAP-NN methods. X axis represents the number of datasets used for analysis and Y axis indicates the obtained accuracy values in percentage. When compared, existing method achieves 81% and 80% while the proposed method achieves 4% better than GA-CNN and 2.1 better than ACO-CNN.

**Precision and sensitivity** give the success of the attack and classification model accordingly. Precision describes the positive predictions made by the classifier in the presence of disease. It is given by:

\[
\text{Precision} (P) = \frac{TP}{TP + FP}
\]

Specificity gives the negative prediction of the classifier in the absence of the disease and is estimated by:

\[
\text{Specificity} (S) = \frac{TP}{TP + FN}
\]

Table 3 shows the comparison of specificity between existing GA-CNN, ACO-CNN, and proposed IWOA-TRAP-NN methods where X axis represents the specificity while Y axis represents the sensitivity in percentage.

**F1-Score** is utilized to determine the prediction performance. It is the harmonic mean (or weighted average) of both the precision as well as recall. If the score is 1, the model is said to be the best else if 0 it is worst. F1-Score is estimated by:

\[
F1 - \text{Score} = \frac{2 * P * R}{P + R}
\]

Table 4 compares the F1-score between existing GA-CNN, ACO-CNN, and proposed IWOA-TRAP-NN methods where X axis represents the F1-score while Y axis represents the number of datasets used for analysis.
Table 4 Comparison for f1-score

<table>
<thead>
<tr>
<th>Number of Datasets</th>
<th>GA-CNN</th>
<th>ACO-CNN</th>
<th>IWOA-TRAP-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>65</td>
<td>68</td>
<td>71</td>
</tr>
<tr>
<td>200</td>
<td>69</td>
<td>70</td>
<td>73</td>
</tr>
<tr>
<td>300</td>
<td>71</td>
<td>74</td>
<td>75</td>
</tr>
<tr>
<td>400</td>
<td>72</td>
<td>76</td>
<td>79</td>
</tr>
<tr>
<td>500</td>
<td>75</td>
<td>79</td>
<td>84</td>
</tr>
</tbody>
</table>

Figure 6 Comparison of f1-score

The figure 6 illustrates the comparison of f1-score between existing GA-CNN, ACO-CNN, and proposed IWOA-TRAP-NN whereas X axis shows the number of datasets and Y axis shows f1-score in percentage.

The kappa static generally test the interrater reliability whose importance represents the correctness of the data collected.

Table 5 compares the kappa static between existing GA-CNN, ACO-CNN, and proposed IWOA-TRAP-NN methods.

Table 5 Comparison for kappa static

<table>
<thead>
<tr>
<th>Number of Datasets</th>
<th>GA-CNN</th>
<th>ACO-CNN</th>
<th>IWOA-TRAP-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>60</td>
<td>64</td>
<td>68</td>
</tr>
<tr>
<td>200</td>
<td>64</td>
<td>68</td>
<td>70</td>
</tr>
<tr>
<td>300</td>
<td>69</td>
<td>70</td>
<td>73</td>
</tr>
<tr>
<td>400</td>
<td>70</td>
<td>72</td>
<td>75</td>
</tr>
<tr>
<td>500</td>
<td>72</td>
<td>75</td>
<td>83</td>
</tr>
</tbody>
</table>

Figure 7 Comparison of kappa static

The figure 7 shows the comparison of kappa static between existing GA-CNN, ACO-CNN, and proposed IWOA-TRAP-NN methods. X axis represents the number of datasets and Y axis provides the obtained kappa static values in percentage.

Table 8 Overall comparative analysis
Figure-8 shows the comparison of various parameters between existing GA-CNN, ACO-CNN, and proposed IWOA-TRAP-NN. As a result the proposed method achieves 86% accuracy, 74% of f1 score and 56.2% of kappa static.

7. Conclusion

In this paper, a Improved Whale Optimized Algorithm TRAPezoidal Neural Network (IWOA-TRAP-NN) that is based on the optimization process is proposed to speed up the entire computational process of hospital sector big data. This proposed method consists of three modules such as, classification, prediction and learning process which uses few characteristics, like persist/cache strategies, fault tolerance, for faster decomposing than the existing algorithms. As a result, several experiments were carried out to test the efficiency of IWOA-TRAP-NN method for medical big data disease prediction and achieved 86% of accuracy, 74% of f1-score and 56.2% of kappa static.

References


16. Akhil Jabbar M , Priti Chandra, Deekshatulu B L. Heart Disease Prediction System using Associative Classification and Genetic Algorithm. Soft Computing inData Analytics, part of AISC volume 78,pp, 735-742, 2018