Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS

ISSN 2454-6348

Original Research Article

DOI: 10.26479/2021.0702.03

ANALYSING THE EFFICIENCY OF HRBFNN AND DCKSVM ALGORITHMS FOR PREDICTING DIABETICS

S. Nathiya, J. Sumitha*

Department of Computer Science, Dr. SNS Rajalakshmi College of Arts and Science (Autonomous) Coimbatore -49, India.

ABSTRACT: In the healthcare industry data mining and knowledge discovery plays a vital role for disease prediction and diagnosis. Data in the healthcare industry contains patient records and disease related information. Data mining algorithms are used with these databases to predict the diseases. The objective is to analyse the efficiency of Radial Basis Function Neural Network (RBFNN) and Divide and Conquer Kernal Support Vector Machine learning (DCKSVM) algorithms for predicting diabetes among the population. In this paper the pre-processing is done by KNN algorithm for filling missing attribute values in the dataset. The result reveals that the Hybrid Radial Basis Function Neural Network algorithm gives better performance when compared with Divide and Conquer Solver for Kernel Support Vector Machine.

Keywords: Data Mining, KNN, Neural Network, Support Vector machine.

Article History: Received: February 10, 2021; Revised: February 23, 2021; Accepted: March 12, 2021.

Corresponding Author: Dr. J. Sumitha* Ph.D.

Department of Computer Science, Dr. SNS Rajalakshmi College of Arts and Science (Autonomous) Coimbatore -49, India.

1.INTRODUCTION

Diabetes is a metabolic disease which occurs in the human body when the blood sugar or glucose level is very high. Blood glucose is the main source of energy for humans, insulin is a hormone which is made by pancreas and it helps the glucose from food eaten to get into your cells to be used for energy. Sometimes the human body cannot create enough insulin or any insulin or cannot use the created insulin in a proper way [3]. Having very high glucose in the body can cause many health problems. The diabetes can be classified as Type 1, Type 2 and Gestational diabetes. Type 1 diabetes means, where the body cannot produce insulin; type 2 is the created insulin is not used in a proper manner and the Gestational diabetes occurs only for the women in their pregnancy period. In the year of 2019 approximately 400 million peoples were living with diabetes, in future by 2045 this will rise to 700 million people. India is the second most diabetes affected country across the world. Higher Glucose level or Diabetes can create very serious damages to the heart, blood vessels, eyes, kidneys and also nerves in the body [1]. Risk factors for getting diabetes include overweight or obese, lifestyle, amount of food intake, amount of insulin intake, smoking and alcohol intake and a family history of diabetes [1, 2]. Maintaining the healthy diet, doing a regular physical exercise or walking, maintaining a normal body weight and avoiding a tobacco use will be helpful to prevent the type 2 diabetes in human. Hence there is a need to develop software for predicting disease like diabetes. Data mining techniques are used to classify, predict and cluster the data in various fields to make decisions accurately [5]. Machine learning algorithms in data mining are used to analyse medical datasets efficiently. Data mining algorithms are designed to use with logical, probability, statistical and control theory problems to analyse the data in that and retrieve the knowledge from the past experiences. So, the main objective of this paper is to analyse the efficiency of machine learning algorithms for predicting diabetes. The Divide and Conquer Kernal Support Vector Machine (DCKSVM) and Radial Basis Function Neural Network (RBFNN) algorithms are applied over the diabetes dataset which is taken from the UCI machine learning depository for finding performance. The parameters used in this performance measure are accuracy, precision and recall. In this it is proved that neural network gives better accuracy when compared the accuracy of SVM but it is based on the chosen dataset. The other parameters such as precision, recall also gives better performance for neural network compares with Support Vector Machine. This paper is organized as follows: Section 2 is the materials and methods used, Section 3 presents the proposed methods to solve the task of predicting diabetes, and Section 4 contains results obtained and discussions. Finally, Section 5 explains about the conclusion part of this research.

2. MATERIALS AND METHODS

The methods used for this research is DCKSCM, HRBFNN and KNN used for pre-processing the data.

2.1 Dataset Description

In this research the diabetes data set which is named as Pima Indian Diabetes data set is taken from UCI Machine learning depository and it is donated by Vincent Sigillito. Data for this data set is collected from the population of women who were at least in the age of 21 years and they are tested according to the criteria of World Health Organization for diabetes. This data set contains 768 examples and it was collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. Some restrictions were employed on the selection of these instances from a larger database. Table 1 refers the attributes such as Pregnancy, Plasma, Press, Skin, Insu, Mass, Pedi, Image, Var, description of that variables and the value measurements that are used for that attributes.

Table 1: Attribute Information of Diabetes Dataset

S.No	Attribute name	Description	Value
1	Pregnancy	Number of Pregnancies	0-17(number)
2	Plasma	Plasma glucose concentration in an oral glucose tolerance test	(0-199)(mg/dl)
3	Press	Diastolic blood Pressure	(0-122[mm/hg)
4	Skin	Triceps skin fold thickness	(0-99) mm
5	Insu	2-hour serum Insulin	0-846 mu U/ml
6	Mass	Body mass index	[0-67)(weight in kg height in m)^2
7	Pedi	Diabetes Pedigree function	(0-2.45)
8	Image	Age of Individual	21-81(years)
9	Var	Class Variable	0 or 1(numeric)

2.2 KNN Algorithm

The data in the real world contains incomplete or inconsistent data; to handle this part some pre-processing technique is necessary to change the data into an effective manner. Data mining comprises many Pre-processing techniques such as removing of missing values, standard scalar, and MinMax Scalar methods which is applied to the dataset for the effective use of data with the classifiers. The K-Nearest Neighbor is a data mining technique which is used as a pre-processing tool in this research that is depicted in figure 1. The KNN is a simple machine learning algorithm which is used for classifying the data based on the closest example [11,22]. This algorithm will assume the similarity between the new item and the available items and put the new item into the category that is most similar to the available category.

1. Load the preferred data for processing and select the value of k that is the nearest data points.
2. To get the class which we need to predict, repeat the steps starting from 1 to the total number of training points.
3. Compute the distance between the data points whose class is to be predicted and all the training data point using Eculidean distance.
4. Arrange the samples in the ascending order based on the distance calculated.
5. Select the best k Nearest Neighbor, among these neighbors count the number of data points in each category.
6. Allocate the new data to the class for which the number of neighbors is maximum.

Figure 1: Steps of K-Nearest Neighbor Algorithm

2.3 Divide and Conquer Kernal Support Vector Machine Algorithm (DCKSVM)

Support Vector Machine Algorithm is a machine learning algorithm that is based on the concept of decision planes that define decision boundaries. The SVM algorithms are used to classify the data which are both linear and nonlinear [10,22]. The decision plane splits among a set of objects having different class memberships. Support vector machine algorithm is used when data has accurately two classes. Here the diabetes database is having two classes either 0 for negative or 1 for positive. SVM classifies the data by finding the best hyper plane that separates all data points of one class from other class [15] that is describe in figure 2 that is taken from [17]. Support Vector Machines map the training data into kernel space [14]. There are differently used Kernal spaces, in this research the Divide and Conquer kernel space is used andthe algorithm which is named as Divide and Conquer Kernal Support Vector Machine (DCKSVM is analyzed for predicting diabetes. The Kernel SVM problem is partitioned into sub problems in the division step, so that each problem is then being solved independently and efficiently and the answers of these sub problems are nearly combined to provide the solution [16,19].

1. Obtain the input vector of the features and labels
2. Multiply the input vector with another feature space to determine kernel, which is a dot product between the space vectors.
3. Select a subset of data points which are known as Support Vectors.
4. Hyper Plane is recognized with unknown parameters.
5. Distance functions are calculated between the hyper plane and support vectors.
6. The distance across hyper plane and support vectors are called as margin
7. Agree on hyper plane which maximizes distance between support vectors and the hyper plan.

Figure 2: Steps of Divide and Conquer Kernal Support Vector Machine Algorithm

Radial Basis Function Neural Network Algorithm (RBFNN)

The Radial Basis Function Neural Network is a machine learning algorithm which is generally used for approximation problems. In this algorithm there are three layers named as input layer, hidden layer and output layer [13, 15]. The input layer which contains the set of source nodes, the hidden layer which contains the hidden neurons and also the activation of these neurons as Gaussian function and the output layer which gives the reply of the network. The transformation between these three layers might be linear or nonlinear. The transformation that is from the input to the hidden space is non-linear [9] and it provides the complex relationship between these two layers. But the transformation from the hidden-unit to the output space is linear [9,24]. In the training phase of the algorithm, the correct class of each record is known output nodes are assigned the correct value as 1 to the correct class and 0 for others [21,23]. The important feature of a neural network is an iterative learning process in which data cases are presented to the network, one at a time and the weight which is assigned to the input values are adjusted each time [10, 12]. After all cases are presented to the network, this process will start again. During the time of this learning phase, the network learns by adjusting the weights each time to predict the correct class label of input samples[25]. Figure 3 refers the steps that are performed by the HRBFNN algorithm for predicting the disease that is taken from [17, 18].

1. Read Input and Output
2. Initiate the weight and biases with random values
3. Estimate the hidden layer input
4. Do / carry out the non-linear transformation on hidden linear input values
5. Do / carry out the linear and non-linear transformation of hidden layer activations for output layer
6. Gradient of Error (E) will be calculated for the output layer.
7. Calculate the slope for output and hidden layer
8. Calculate the delta for output layer.
9. Compute the error of hidden layer
10. Update the Weight's and Biases at both layers of the Output and Hidden layers

    Train the model multiple times until it comes very closer to the actual outcome.

Figure 3: Steps of Hybrid Radial Basis Function Neural Network Algorithm

3. RESULTS AND DISCUSSION

This research work is implemented using the MATLAB tool which is a numerical computing environment and it is developed by MathWorks. This tool is used for the implementation of many functions such as matrix manipulation, plotting of functions and data and it is also used for the implementation of algorithms. Hence, the algorithms RBFNN and DCKSVM are tested with Pima Indian Diabetes Dataset using this tool with the usage of confusion matrix. Confusion matrix will analyse the performance of algorithms using the parameters Accuracy, Precision and Recall. The information about the actual and predicted value done by a classification system is available in the confusion matrix. The class variable has two values positive or negative. The parameters are calculated using the values True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The True Positive is the value that the positive cases that is correctly classified as positive, False Positive is the value that the negative cases that is incorrectly identified as positive, True negative is the value that the negative cases that is correctly classified as negative and the False Negative is the value that the positive cases that is incorrectly identified as negative. Accuracy is the percentage of predictions that are correct and it is calculated using the summation of true positive and true negative values which is divided by the total value. Precision is the measure of accuracy provided that a specific class has been predicted and it is calculated using true positive value divided by the summation of true positive and false positive value. Recall is the percentage of positive labelled instances that were predicted as positive and it is calculated by true positive divided by the summation of true positive and false negative values

Table 2: Performance of classification algorithms

Algorithm / **Parameters**	**DCKSVM(%)**	**RBFNN(%)**
Accuracy	79	**98**
Precision	72	**78**
Recall	76	**81**

Table 2 shows the performance of classification algorithms using various parameters. DCKSVM shows the accuracy values of 79, 72, and 76 for Accuracy, Precision and recall which is lesser than the value of RBFNN as 98, 78 and 81 for Accuracy, Precision and Recall. So RBFNN gives better performance than DCKSVM for diabetes dataset.

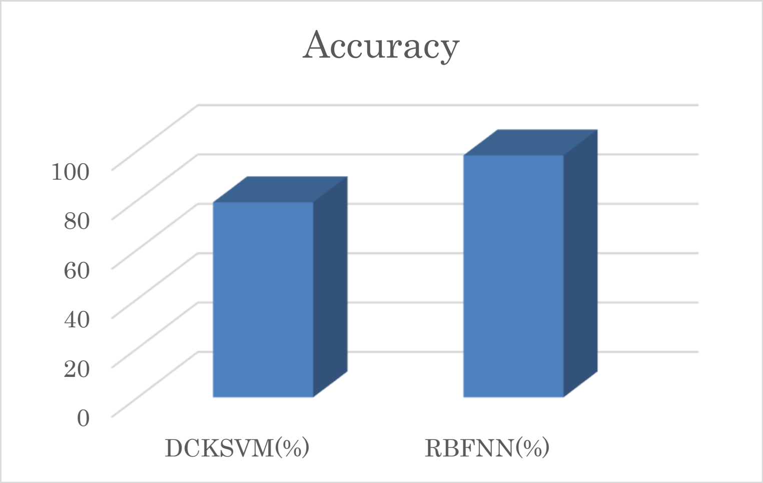

Figure 4: Predicted Accuracy

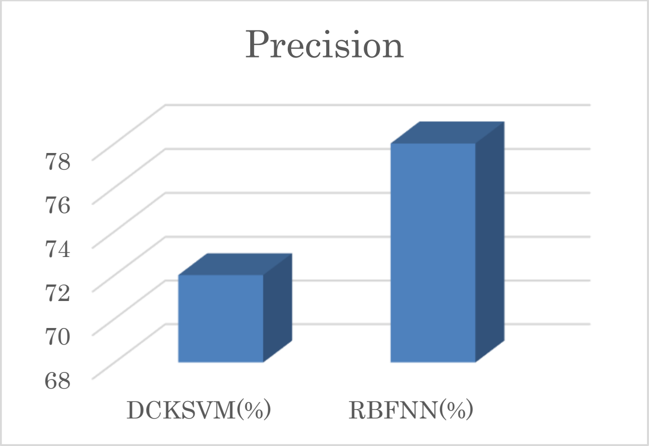

Figure 5: Predicted Precision

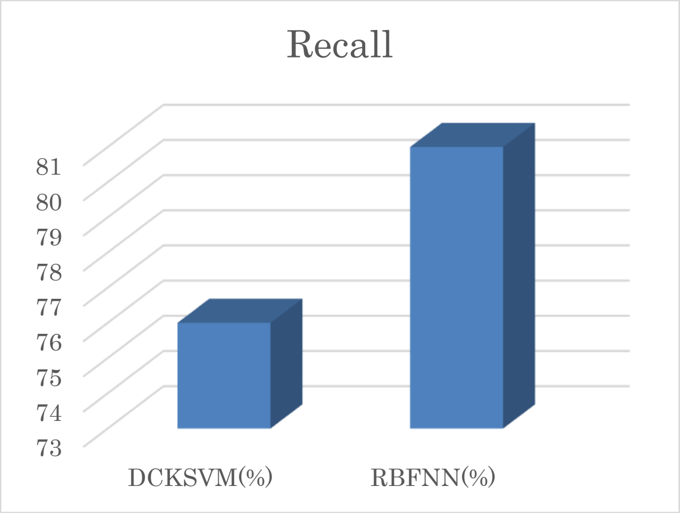

Figure 6: Predicted Recall

The figures 4, 5, and 6 show the graphical representation of the performance. DCKSVM use efficiency in terms accuracy of value of 79, precision of value 72, and recall of value 76 which is lower than RBFNN shows the accuracy of value 98, precision of value 78 and Recall of value 81. So, it is proved that the RBFNN gives better performance for predicting diabetes than the existing algorithm.

4. CONCLUSION

In this research work the performance of two classification algorithms is examined for predicting diabetes with Pima Indian diabetes dataset. It is proved that the HRBFNN algorithm is effective for predicting the disease. HRBFNN algorithm gives the highest percentage of values for all the three parameters accuracy, Precision and Recall. In future different classification algorithms are developed for predicting diabetes and other diseases with different databases.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are base of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The author confirms that the data supporting the findings of this research are available within the article.

FUNDING

None.

ACKNOWLEDGEMENT

None.

CONFLICT OF INTEREST

Authors have no conflict of interest.

REFERENCES

1. Aiswarya Mujumdar, Vaidhei V, Diabetes Prediction Using Machine Learning Algorithms, Science Direct-Procedia Computer Science, Vol. 165:292-299.
2. David Alberts, Lena Mamykina, Data-Driven Blood Glucose Pattern Classification and Anomalies Detection: Machine-Learning Applications in Type 1 Diabetes, JMIR Publications, 2019, Vol.21: 31-49.
3. DeeptiSisodia, Dilip Singh Sisodia, Prediction of Diabetes using Classification Algorithm, International Journal on Computational Intelligence and Data science, 2018, Vol.132: 15787-1585.
4. SumithaJ, BRCA Gene Expression Level Analysis for Identification of Breast Cancer using computer assisted Algorithm's, International journal of Pharma and Biosciences, 2018, Vol. 9: 60-64.
5. ChalaBeyene, Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques, International Journal of Pure and Applied Mathematics, 2018, Vol.118:165-174.
6. Chieh-ChenWu, Wen-ChunYeh, Prediction of fatty liver disease using machine learning algorithms, Computer Methods and Programs in Biomedicine, 2019, Vol.170:23-29.
7. Sumitha J, Analysis of Gene Expression Value Using Bat Algorithm with Multifactor Non-Negative Matrix Factorization, Pakistan Journal of Biotechnology, 2019, Vol. 16 (2):101-104.
8. Elham Nikookar, Embrahim Naderi, Hybrid Ensemble Framework for Heart Disease Detection and Prediction, International Journal of Advanced Computer Science and Applications, 2018,Vol.9:243-248.

9. Divya Jain, Vijendra Singh, Feature selection and classification systems for chronic disease prediction: A review, Egyptian Informatics journal, 2018, Vol .19: 179-189.
10. El-Houssainy A, Radya , Ayman S. Anwarb, Prediction of kidney disease stages using data mining algorithms, Informatics in medicine unlocked ,2019,Vol.15:100178.
11. Muhammad Yusril,HelmiSetyawan, RollyMaulanaAwangga, K-Nearest neighbor algorithm on implicit feedback to determine SOP,Research Gate,2019, Article · June 2019 , Vol .17: 1425-1431.
12. Mucahid Mustafa Saritas, Ali Yasar, Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification, International Journal of Intelligent Systems and Applications in Engineering, 2019, Vol 7: 88-91.
13. Hossein Moayedi 1,2, Dieu Tien Bui, The Feasibility of Three Prediction Techniques of the Artificial Neural Network, Adaptive Neuro-Fuzzy Inference System, and Hybrid Particle Swarm Optimization for Assessing the Safety Factor of Cohesive Slopes, International Journal of Geo-Information, ISPRS Int. J. Geo-Inf. 2019, Vol.8:391-400.
14. Gyeongcheol Cho, JinyeongYim, Review of Machine Learning Algorithms for Diagnosing Mental Illness, Psychiatry Investing. 2019, Vol.16:262–269.
15. Sumitha J, Devi T, Breast Cancer Diagnosis in Analysis of BRCA Gene Using Machine Learning Algorithms,2016, Pakistan Journal of Biotechnology, 2016, Vol. 13: 231-235.
16. Sumitha J, Analysis of Gene Expression Value Using Bio-Inspired Algorithms, Pakistan Journal of Biotechnology, 2019, Vol .16 :115-118.
17. Sumitha J, Devi T.Ravi D, Comparative Study on Gene Expression for Detecting Diseases Using Optimized Algorithm, International Journal of Human Genetics, 2017,Vol. 17:38-42.
18. Sumitha J, Devi T, Analysis of Expression Level of Breast Cancer Gene Using Machine Learning Algorithms for Diagnosis of Breast Cancer, International Journal of Parmanand Bioscience, 2017, Vol.8: 79 – 85.
19. Sumitha J,Mallika R, Cancer Classification in Microarray Data Using Gene Expression with SVM OAA and SVM OAO, International Journal of Advanced Research in Computer Science, 2011,Vol .2 :1-4.
20. Naveen Kishore G, Rajesh V, Vamsi Akki Reddy, Prediction of Diabetes Using Machine Learning Classification Algorithms, 2020, International Journal of Scientific & Technology Research,Vol.9:1805-1808.
21. Sahhadat Uddin, Arif Khan, Md Ekramul Hossain, Mohhammad Ali Moni, comparing different machine learning algorithms for disease prediction, 2019, BMC Medical Informatics and Decision Making, Vol 19:1-16
22. Kesav Srivastava, Dilip Kumar Choubey, Heart Disease Prediction using Machine Learning and Data Mining, 2020, International Journal of Recent Technology and Engineering, Vol 9:216-219

23. Rudra A.Godse, Smita S.Gunjal, Karn A.Jagtap, Neha S.Mahamuni, Multiple Disease Preiction Using Different Machine Learning Algorithms Comparatively, 2019, International Journal of Advanced Research in Computer and Communication Engineering, Vol 8:50-52
24. Shabaz Ali.N, Divya.G, Prediction of Disease in Smart Health Care System using Machine Learning, 2020, International Journal of Recent Technology and Engineering, Vol 8: 2534-2537
25. Anavarapu Naga Prathyusha, Navasinga Rao, Diabetic Prediction Using Kernal Based Support Vector Machine, 2020, International Journal of Advanced Trends in Computer Science and Engineering, Vol 9:1178-1183