International Journal of Engineering Sciences & Research Technology

(A Peer Reviewed Online Journal)
Impact Factor: 5.164

IJESRT

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

AN IN DEPTH EXPERIMENTATION WITH CLASSIFIERS FOR PREDICTION OF DIABETES

Sonal Kumari & Vandana Bhattacharjee*
Department Of Computer Science and Engineering BIT Mesra, Ranchi

DOI: https://doi.org/10.29121/ijesrt.v10.i6.2021.4

ABSTRACT

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. Hyperglycaemia, or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels. Therefore the objective of this paper is to analyse different classification algorithm such as SVM, decision tree, KNN, NN to detect diabetes at early stage. Among all the four algorithms used NN gives maximum accuracy of 82.46%.

KEYWORDS: Diabetes, machine learning, support vector machine, decision trees, neural networks.

1. INTRODUCTION

Diabetes is one of the most common diseases worldwide. It is a disease which can occur due to heredity and / or lifestyle factors. If one is living a stressful life or is obese, and carries extra weight in the belly portion of the body, this complicates the functioning of insulin, thus leading to diabetes. The early detection of diabetes can help to improve the general health of people, and thankfully machine learning approaches are providing great help in this regard. Several researchers have applied classification, clustering and other machine learning techniques for health care data. Authors in [1] have proposed the use of Decision Tree, SVM and Naive Bayes for prediction of diabetes. Machine learning algorithms like Decision Tree, Decision Table etc have been used extensively for prediction of this disease by researchers.. It has been proved that these learning algorithms [2-4] work better in diagnosing different diseases. Data Mining and Machine learning algorithms are useful in this regard due to their capability of managing large amounts of data, the ability to integrate data from multiple sources, to pre-process and handle erroneous data, as well as to integrate any domain information seamlessly [5]. In [6] a web application has been proposed by using disease classifiers. Authors in [7] have applied SVM with Radial basis function kernel for classification of Diabetes Disease. Sharief and Sheta in [8] developed a classifier based on Multigene genetic programming. GP mathematical model was built to provide a solution to the diabetic problem. In another research [9], the authors develop a new short-term glucose prediction algorithm based on a neural network which eventually leads to detection of diabetes. Yu et al in [10] take an alternative approach based on SVM and use it to detect persons with diabetes and pre-diabetes. SVM was also used by authors in [14]. The Adaboost and bagging ensemble techniques using J48 (c4.5) decision tree as a base learner along with standalone data mining technique J48 has been used to classify patients with diabetes mellitus using diabetes risk factors [11]. Han et al in [12] use the Rapid-I's RapidMiner tool to develop a decision tree based diabetes prediction model. The paper by Aiswarya et al [13] aims at finding solutions to diagnose the disease by analyzing the patterns found in the data through classification analysis by employing Decision Tree and Naïve Bayes algorithms. Genetic Programming (GP) showed significant advantages on evolving nonlinear model which can be used for prediction and several researchers have developed models based on GP [14-17].

The organization of the rest of the paper is as follows: Section 2 presents the brief overview of classifiers. Section 3 presents the Experimental setup, Section 4 gives the Results and Analysis. Finally Section 5 concludes the paper.

2. BRIEF OVERVIEW OF CLASSIFIERS

2.1 Support vector machine(SVM)

SVM is a supervised learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in an n dimensional space where n is the no of features we have. The value of each feature here is the value of a particular coordinate. Then, we perform the classification by finding the plane(3-D)/hyper plane(n-D) that differentiates the two classes very well. For two-class, separable training data sets, such as shown below, there are lots of possible linear separators. Intuitively, a decision boundary drawn in the middle of the void between data items of two classes seems better than one which approaches very close to examples of one or both classes. In process of finding correct hyper plane, we encounter the points in either direction, called *support vectors* and the width that the boundary could be increased by before hitting a data points is called *margin*. In case of non-linearly separable data, we map the data points to higher dimension or add new feature to data and try to make linearly separable. The forming of new feature with existing features is called *kernel transformation*. There are generally three types of kernel used in SVM -Gaussian, polynomial and linear kernel. Most of the time polynomial kernel works better with our real life related data. But in case of PIMA diabetes dataset, the classification of diabetic and non-diabetic classes is performed better using linear kernel. SVM classifies the classes accurately prior to maximizing margin.

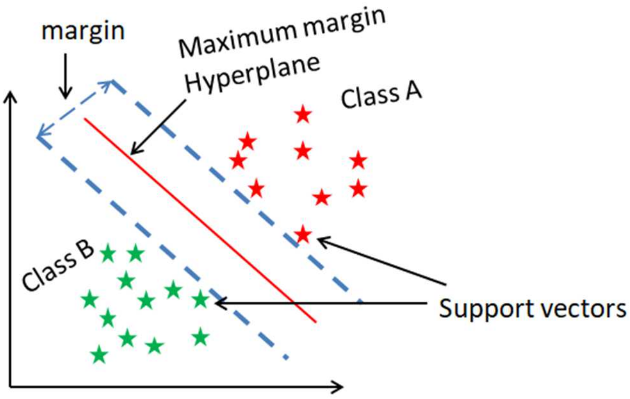

Figure 1 The maximal margin SVM classifier

2.2 K-Nearest Neighbour(KNN)

K-Nearest Neighbour is a lazy learning algorithm and the simplest supervised learning algorithm which is mostly used for classification analysis. K-NN algorithm just stores the available data during training phase and when a new data appears then it classifies on basis of similarity. K-NN can be seen on the basis of below algorithm:

1. Take as input *K*: the number of neighbours, *d*: the data to be classified.
2. Calculate the Euclidean distances of *d* with all the datapoints.
3. Take the K nearest neighbours as per the calculated Euclidean distance.
4. Among these K neighbours, count the number of the data points in each category.
5. Assign the new data point that category for which the number of the neighbour is maximum.
6. The model is ready.

2.3 Decision Tree

Decision Tree is a supervised machine learning technique that is mostly preferred for solving classification problems. It is a tree-structured classifier, where internal nodes represent features of a dataset, branches represent the decision rules and each leaf node represents the outcome. ID3, C4.5, and CART adopt a greedy approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner.

Steps in ID3 algorithm:

i. It begins with the original set S as the root node.
ii. On each iteration of the algorithm, it iterates through the very unused attribute of the set S and calculates Entropy(H) and Information gain(IG) of this attribute.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

iii. It then selects the attribute which has the smallest Entropy or largest information gain.
iv. The set S is then split by the selected attribute to produce a subset of the data.
v. The algorithm continues to recur on each subset. Considering only attributes never selected before.

There are some methods of attribute selection: entropy, information gain, Gini index, Gain ratio, Reduction in variance, chi-square.

$info(D) = -\sum_{i=0}^{n} p_i log_2(p_i)$,

Where, p_i is the non zero probability that an arbitrary tuple in D belongs to class C_i and is estimated by $|C_{i,D}|/|D|$.

2.4 Neural Network

Neural networks, as the name suggests, are networks inspired by the working of neurons in the brain.Neural networks are used to model complex patterns in datasets using multiple hidden layers and non-linear activation functions. A neural network takes an input, passes through multiple layers of hidden neurons and outputs a prediction representing the combined input of all the neurons.

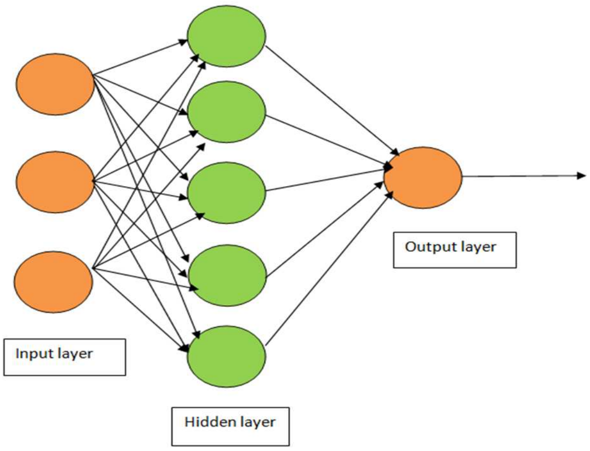

Figure 2 A Neural Network Model with one hidden layer

3. EXPERIMENTAL SETUP

3.1 Datasets

Table 1. Description of dataset

Dataset	Number of attributes	Number of instances	Number of negative classes	Number of positive classes
Pima Indians Diabetes Dataset	9	768	500	268

3.2 Experiments conducted

In this research work, the different classifiers were applied on the dataset with varying parameter values. The first experiment conducted was with the SVM classifier, with the polynomial kernel, Linear kernel, rbf and the default kernel. Table 2 presents the result of applying the SVM classifier with different kernels on the dataset. Table 3 presents the result of applying the KNN classifier with different number of neighbours and distance measure technique. Table 4 presents the result of applying the DT classifier with different depth of tree and attribute selection measures. Next Table 5 presents the various experiment with the Neural network classifier.

4. RESULTS AND DISCUSSION

Table 2. SVM accuracy on different kernels

Types of kernel	Accuracy
Poly(polynomial)	0.7532
Linear	0.8116
rbf(Gaussian /radial basis function)	0.7662
Default	0.7662

Table 3 Accuracy values with KNN

Distance measures	Number of neighbours			
	25	50	75	100
Euclidean	0.7922	0.7662	0.7792	0.7857
Manhattan	0.7857	0.7922	0.7987	0.7922
Default	0.7922	0.7922	0.8051	0.7792

Table 4 Accuracy values with the DT classifier

Attribute selection measure	Max Depth values					
	3	4	5	6	7	8
Gini	0.7597	0.7662	0.7532	0.7597	0.7857	0.7532
Entropy	0.7597	0.7402	0.7402	0.7467	0.7207	0.7142
Default	0.7597	0.7662	0.7532	0.7467	0.7922	0.7402
Gini-Gini index; Entropy - information gain						

Table 5 Accuracy values with different activation function

Number of iterations	Activation function			
	’relu’ (default)	’identity’	’tanh’	’logistic’
100	0.7532	0.7467	0.7012	0.6948
200	0.7532	0.8051	0.7792	0.6948
300	0.7922	0.7987	0.7922	0.6948
400	0.8051	0.8181	0.7987	0.6948
500	0.7987	0.8116	0.8181	0.6948
600	0.7922	0.8311	0.8116	0.6948
700	0.7857	0.8051	0.8116	0.6948

The experiments with the NN classifier were conducted for different solvers, learning rates, variation in learning rates and activation functions. The NN algorithm performs better with accuracy of 82.46% after applying different parameters in the following ways: Activation=’identity’, solver=’adam’, alpha=0.0005, learning_rate=’constant’, early_stopping=False, max_iter=800 . The training time is 0.8061 sec and the prediction time is 0.0030 sec. For the final analysis, we computed the performance measures for all the 4 classification techniques used in the study. The results were based on the values of training time prediction time and Accuracy.

TABLE 6 Comparative results for all four classifiers

Classifier	Training time	Prediction time	Accuracy
Decision tree	0.0077 sec	0.0037sec	0.7922
KNN	0.0078sec	0.0233 sec	0.8116

SVM	8.7507 sec	0.0059 sec	0.8116
NN	0.8061 sec	0.0030 sec	0.8246

5. CONCLUSION

The basic machine learning classifiers have been applied on the diabetes dataset. The KNN algorithm was executed with varying number of neighbours. The SVM classifier was experimented with different kernels. The decision tree was executed with different depths and attribute selection criteria. Finally the neural network classifier was applied with varying activation functions, solvers, learning rates and change in learning rates. The accuracy obtained for KNN and SVM in the best cases is same, 81.16% but the prediction time for KNN algorithm is more. The decision tree algorithm gives the accuracy of 79.22% , and also very less training time and prediction time. The NN classifier gives the best accuracy of 82.46% but the training time is quite high. It is important that we choose the algorithm depending upon different parameters. The machine learning paradigm gives us very good solutions to the real life problems.

6. ACKNOWLEDGEMENTS

The authors thank the reviewers for their valuable comments.

REFERENCES

[1] Deepti Sisodia, Dilip Singh Sisodia, Prediction of Diabetes using Classification Algorithms, Procedia Computer Science 132 (2018) 1578–1585

[2] Aishwarya, R., Gayathri, P., Jaisankar, N., 2013. A Method for Classification Using Machine Learning Technique for Diabetes. International Journal of Engineering and Technology (IJET) 5, 2903–2908.

[3] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., 2017. Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal 15, 104–116. doi:10.1016/j.csbj.2016.12.005.

[4] Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 5–10.

[5] Kumar, P.S., Umatejaswi, V., 2017. Diagnosing Diabetes using Data Mining Techniques. International Journal of Scientific and Research Publications 7, 705–709

[6] Nai-Arun, N., Moungmai, R., 2015. Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Computer Science 69, 132–142. doi:10.1016/j.procs.2015.10.014

[7] Kumari, V.A., Chitra, R., 2013. Classification Of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications (IJERA) www.ijera.com 3, 1797–1801

[8] Sharief, A.A., Sheta, A., 2014. Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. International Journal of Advanced Research in Artificial Intelligence (IJARAI) 3, 54–59. doi:doi:10.14569/IJARAI.2014.031007

[9] C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, and C. Cobelli, "A new neural network approach for short-term glucose prediction using continuous glucose monitoring time-series and meal information," Proceedings of the IEEE International Conference on Engineering Med. Biol. Soc., pp. 5653–6, 2011

[10] Yu, W., Liu, T., Valdez, R., Gwinn, M., Khoury, M.J., 2010. Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making 10. doi:10.1186/1472-6947-10-16, arXiv:arXiv:1011.1669v3

[11] Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K., 2016. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science 82, 115–121. doi:10.1016/j.procs.2016.04.016

[12] Han, J., Rodriguez, J.C., Beheshti, M., 2008. Discovering decision tree based diabetes prediction model, in: International Conference on Advanced Software Engineering and Its Applications, Springer. pp. 99–109

[13] AiswaryaIyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[14] Aishwarya R., Gayathri P., Jaisankar N, "A Method for Classification Using Machine Learning Technique for Diabetes '',International Journal of Engineering and Technology (IJET), 5 (2013), pp. 2903-2908

[15] Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770. doi:10.1007/978-3-319-11933- 5

[16] Sharief, A.A., Sheta, A., 2014. Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. International Journal of Advanced Research in Artificial Intelligence (IJARAI) 3, 54–59

[17] Pradhan, P.M.A., Bamnote, G.R., Tribhuvan, V., Jadhav, K., Chabukswar, V., Dhobale, V., 2012. A Genetic Programming Approach for Detection of Diabetes. International Journal Of Computational Engineering Research 2, 91–94.