Chronic Diabetes Disease Prediction Using Classification

K-Nearest Neighbour (Knn) Algorithm

M Kavitha[1], Dr S Subbaiah[2]

[1]Assistant Professor, PG & Department of Computer Science and Applications, Vivekanandha College of Arts and Science for Women, Tiruchengodu, Tamilnadu, India
[2]Assistant Professor, Department of Computer Science, Shri Krishna Arts and Science College, Coimbatore, Tamilnadu, India
sankavis@gmail.com, subbunaren@yahoo.com

ABSTRACT

Machine learning algorithm is one of the popular algorithms for predicting the disorders. Developing programs that access data and processing by making decisions based on their previous learning. It is branch of Artificial Intelligence (AI). It automatically learn the system and improve it from beforehand experience. Without human assistance the system would work. In this paper, k-nearest neighbour algorithm has been used to predict the diabetes disease with help of template dataset.

KEYWORDS: Chronic Diabetes Disease (CDD), Machine Learning (ML), Classification

INTRODUCTION

Machine learning aim is to provide systems to learn systems automatically from previous experience or observations. For example, categorizing the fruits, system can be trained based on their color, shape. After trained the system, it can able to find out the given fruit is which category.

Now a day, machine learning algorithms are used in different applications. The applications are particularly for predicting chronic diseases, intrusion detection, fraud detection, social media prediction and market risk modelling etc. For these types of applications various algorithms are used.

In machine learning the algorithm can be categorized as supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. In supervised learning, the system can be predicting the future values based on their known datasets or previous learning methods. Unsupervised learning algorithm is used for unlabelled data that is not classified or categorized. Semi-supervised learning algorithm is a combination of supervised learning and unsupervised learning. It uses the small amount of labeled data and large amount of unlabelled data. The reinforcement learning interacts with the environment and discovers the results and error.

In this paper, section 2 explains about literature survey. Section 3 describes k-nearest neighbour (knn) algorithm and in section 4 specifies the experimental results. Section 5 specifies the conclusion and future work.

LITERATURE STUDY

Murat Pojan [1] predicts the student performance using machine learning algorithms. The author used three algorithms. They are linear regression, decision tree and naïve Bayes classification. To improve the prediction value, he used feature engineering. He concluded that naïve Bayes classification gave best results comparing with other algorithms. He got 98% accuracy in naïve Bayes algorithms to predict student performance.

Ahmed L G et al. [2] used three machine learning algorithms for predicting breast cancer recurrence. They applied techniques to build the models for predicting breast cancer recurrence in patients. They used Decision tree (DT) (C4.5), Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithms. They show that accuracy of different algorithms like DT, ANN and SVM are 0.936, 0.947 and 0.957 respectively. So SVM Model predicts the breast cancer recurrence with highest accuracy and minimum error rate [3, 4].

B Dhomse Kanchan et al. [5] used supervised machine learning algorithms to predict heart disease. The major cause of death is heart disease. If the recent styles continue means 23.6 million people will die due to heart disease said by healthcare.

K-NEAREST NEIGHBOUR(KNN) ALGORITHM

This algorithm is the one of the classification algorithms. It is used for predicting class value for given data sets. This modelling is used for mapping the function from the input variables to discrete variables. It belongs to supervised learning because the target value is given in the input data itself. This algorithm is used in different types of applications like spam prediction, diseases prediction [6,7,8].

EXPERIMENTAL RESULTS

The Prediction of diabetes disease is done by using R tool. The R has number of machine learning packages to perform the different algorithms. In this paper, k-nearest neighbour algorithm is used to perform the prediction.

The dataset is selected from Pima Indians diabetes data set. The dataset contains 768 data and 9 attributes. First the dataset is loaded to r tool and then performed normalization function to the datasets for normalizing the values. Then the data set is divided into two categories. One is train data set and another one is test data set. Next, we applied the knn algorithm to the dataset. We got the result with accuracy as 78%.

Confusion Matrix and Statistics

Reference
Prediction no yes
no 100 8
yes 28 32

Accuracy : 0.7857
95% CI : (0.7159, 0.8452)
No Information Rate : 0.7619
P-Value [Acc > NIR] : 0.266221

Kappa : 0.496

Mcnemar's Test P-Value : 0.001542

Sensitivity : 0.7812

Specificity : 0.8000

Pos Pred Value : 0.9259

Neg Pred Value : 0.5333

Prevalence : 0.7619

Detection Rate : 0.5952

Detection Prevalence : 0.6429

Balanced Accuracy : 0.7906

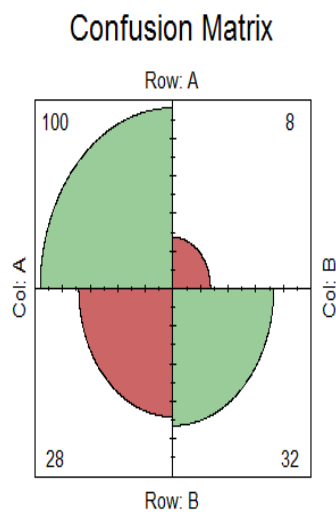**'Positive' Class : no**

Fig2. Confusion matrix plot

CONCLUSION

The machine learning algorithm is used now as days in various applications. Machine learning algorithms are widely applicable for medical diagnosis. Diabetics is one of the growing non-curable disease. K-nearest neighbour algorithm helps to predict the chronic diabetics diseases and it gives accurate results to predict in advance.

REFERENCES

1. Murat Pojon (2017), Using Machine Learning to Predict Student Performance

2. Ahmad (2013), Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence, Journal of Health & Medical Informatics, ISSN: 2157-7420,4(2)

3. B. Dhomse Kanchan and M. Mahale Kishor (2017), Study of machine learning algorithms for special disease prediction using principal of component analysis, ISBN: 978-1-5090-0467-6

4. Mohammad R. Mohebian (2017), A Hybrid Computer-aided-diagnosis System for prediction of Breast Cancer Recurrence (HPBCR) using optimized Ensemble Learning, Computational and Structural biotechnology Journal

5. Abdullah A. Aljumah (2013), Application of data mining: Diabetes health care in young and old patients, Journal of King Saud University-Computer and Information Sciences

6. Loannis kavakiotis,olga Tsave (2017), Machine learning and Data mining methods in Diabetes Research, Computational And Structural biotechnology Journal

7. Neesha Jothi and Nur Aini Abdul Rashid (2015), Data Mining in health care- A Review, Elsevier

8. Kumari Deepika and S Seema (2016), Predictive Analytics to prevent and control chronic Diseases, International conference on Applied and theoretical Computing and Communication and communication Technology