Survey on Machine Learning Algorithm Based Disease Prediction

[1]Prachi Mhatre [2] Shivani Sarkale [3]Poonam Padalkar [4]Bhagyashri Patil Prof-Rohini Patil

Terna Engineering College,Nerul

Computer Engineering

Abstract

The paper intends to give details about various techniques of knowledge abstraction by using data mining methods that are being used in today's research for prediction of disease. Data mining methods namely, Naive Bayes, Neural network, Decision tree algorithm are analyzed on medical data sets using algorithms. There are instances where online medical help or healthcare advice is easier or faster to grasp than real world help.

Keywords- Prediction,DataMining,NaiveBayes,Support Vector Machine,Random Forest.

1.Introduction

The Healthcare industry is among the most information intensive industries. Medical information, knowledge and data keep growing on a daily basis. Clinical Prediction is a rapidly growing field that is concerned with applying Computer Science and Information Technology to medical and health data. With the aging population on the rise in developed countries and the increasing cost of healthcare, governments and large health organizations are becoming very interested in the potential of Clinical Diagnosis to save time, money, and human lives.

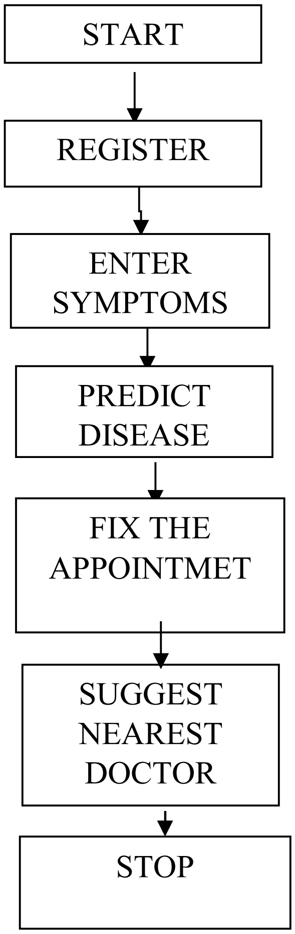
This system allows the users to get analysis on the symptoms they give for predicting the disease they are suffering from. User will be asked to enter the symptoms, then system will processes those symptoms for various illness or disease that user could be a liked with. In this system we use some techniques of data mining to guess the most accurate diseases or illness that could be related with patient’s symptoms. This system tends to replace the existing system for going to the doctor for getting diagnosis on illness you are suffering from to a smart solution where you get instant diagnosis on entering symptoms in the system. The main features of this system will be giving instant diagnosis on the user entered symptoms and suggest specialized doctor near our area. And we take appointment to the doctor. The system will prove helpful in urgent cases where the patient is unable to reach hospital or in cases when there are no doctors available in the area.[7]

Headache disease occurs due to work overload, stressand many reasons behind it. This paper analyzes theheadache diseaseprediction and classification using different classificationalgorithm.Medical data mining techniques like AssociationRule Mining, Clustering, Classification Algorithms such asNaïve Bayes, Decision tree.[3]

2.Literature Survey

AhelamTikotikar, Mallikarjun Kodabagi "A Study of Machine Learning in Healthcare" Conference of Computer Software and Applications 2017 [1]. The papers discussed here reviews about data mining techniques, classification techniques, intelligent techniques and feature selection for prediction of disease. Rohan Bhardwaj, Ankita R. Nambiar, Debojyoti Dutta "A Survey Of Technique For Prediction Of Disease in Medical Data" Conference On Smart Technology for Smart Nation 2017[2]The main focus of this paper is to discuss about decision parameter, attribute, and features used for predicting the disease.. The dataset considered in so many existing techniques . The various data mining techniques are used as classifier, to build a cost effective model for disease prediction. It is well understood by the exhaustive survey that mining the required information from the medical data help us to support well informed diagnosis and decisions. Md. Rajib Hasan, Dr. Md. Shariful Hasan, Fadzilah Siraj **"An Expert System Based HeadacheSolution"** Conference on Computer Applications and Industrial Electronics ,2012[3].This paper investigates the potential developing of an expert system for headache detection. It is to identify the requirements to build an Expert System based Headache Solution (ESHS) for supporting the clinical solution. In ESHS system, doctor can be input the patient's symptoms according to, and ESHS will produce the solution similar to the prescription that is usually made by the doctors. **"Trends and Challenges in Smart Healthcare Research:** A Journey from Data to Wisdom",2017[4].The aim of this article is to identify some of the most relevant trends and research lines that are going to affect the smart healthcare field in the years to come. To do so, the article considers a systematic approach that classifies the identified research trends and problems according to their appearance within the data life cycle, this is, from the data gathering in the physical layer (lowest level) until their final use in the application layer (highest level). By identifying and classifying those research trends and challenges, we help to pose questions that the smart healthcare community will need to address. Consequently, we set a common ground to explore important problems in the field, which will have significant impact in the years to come.

3.Methodology



Following Algorithm are used to Predict Disease:

3.1.Support Vector Machine

SVM is one of the technique from supervised learning based algorithm which is used for classification and regression analysis. This algorithm is used for classification using training dataset Support Vector Machine is a border which best segregates the two classes.

[6]This algorithm is classified into linear dataand non- linear data. Linear classification is implemented using hyperplane. Non-linear classification some kinds of transformation to given training dataset and then after transformation. In Support vector Machine algorithm, it will design each data item set as a point in dimensional space. Then, we achieve classification by finding and constructing the hyper-plane on dataset that divides the dataset into two classes.

3.2.Naïve Bayes

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. Naive Bayes model is easy to build and particularly useful for very large datasets. It learns and predicts very fast and it does not require lots of storage.

Real-time Prediction: As Naive Bayes is super fast, it can be used for making predictions in real time.

Multi-class Prediction: This algorithm can predict the posterior probability of multiple classes of the target variable.

Text classification/ Spam Filtering/ Sentiment Analysis: Naive Bayes classifiers are mostly used in text classification (due to their better results in multi-class problems and independence rule) have a higher success rate as compared to other algorithms.

3.3. Decision Tree

A decision tree is a flowchart structure inwhich each internal node denotes a test on a characteristic,where each branch signifies the result of the test, and each leaf node denotes a class label. The paths from the root toleaf denotes classification rules. In tree the interrelated diagram are used as the analytical, visual and decision supporttool, where the apparent values are calculated.[11]

These algorithms are useful in data exploration. Decision trees implicitly perform feature selection which is very important in predictive analytics. When a decision tree is fit to a training dataset, the nodes at the top on which the decision tree is split, are considered as important variables within a given dataset and feature selection is completed by default.A major drawback of decision tree machine learning algorithms, is that the outcomes may be based on expectations.

3.4.Random Forest

Random Forest algorithm are an ensemble supervised learning method which is used as predictor of data for classification and regression. In the classification process algorithm build a number of decision trees at training time and construct the class that is the mode of the classes output by using each single trees. (Random Forests is introduced by Leo Breiman and Adele Cutler for an ensemble of decision trees). [5] Random Forest algorithm is a grouping of tree predictors where each tree based on the values of a random vector experimented independently with the equal distribution for all trees in the forest. Random Forest is one of the most effective and versatile machine learning algorithm for wide variety of classification and regression tasks, as they are more robust to noise

4.Comparison

Parameter	Support vector machines	Naïve Bayes classifier	Random forest	Decision Tree
Merit	High accuracy, avoid over fitting, flexible selection of kernels for nonlinearity, accuracy and performance are independent of number of features, good generalization ability.	Ability to interpret problem in terms of structural relationship among predictors, takes less computational time for training, no free parameters to be set .	Fast, scalable, robust to noise, does not over fit, offer explanation and visualization of its output without any parameters to worry about.	Dealing with noisy and incomplete data, data classification without much calculations. Handling both continuous and discerte data.
Demerit	Lack of transparency of results.	Requires a very huge number of records for obtaining good results.	Need to choose the number of tree.	Provide less information on the relationship between the predictors and the response.
Application	Text Classificati on.	Document classificati on, medical diagnostic systems.	To find cluster of patients. Classificati on of microarray data, object detection.	Classification and prediction tool.
When to use	To Classify the data.	To calculate the probability of dataset.	When we don't bother much about interpreting the model but want better accuracy.	When we want our model to be simple and explainable.

5. Conclusion

The main focus is on using different algorithm and combination of several targets attributes for different types of disease prediction using data mining. Due to the rapid growth of medical data, it has become indispensable to use data mining techniques to help decision support and predication systems in the field of Healthcare. We observe that Naïve Bayes algorithm performed well with respect to all the factors. Different classification algorithms gives different result on base of accuracy, training time, precision, recall.

6.References

[1]AhelamTikotikar, Mallikarjun Kodabagi" A Study of Machine Learning in Healthcare"2017

[2] Rohan Bhardwaj, Ankita R. Nambiar, Debojyoti Dutta "A Survey Of Technique For Prediction Of Disease in Medical Data"2017

[3] Md. Rajib Hasan, Dr. Md. Shariful Hasan, Fadzilah Siraj "An Expert System Based Headache Solution"2012

[4]AgustiSolanas, Fran Casino UniversitatRoviraiVirgili, Av. Paisos Catalans, 26. 43007 Tarragona, Catalonia, Spain "Trends and Challenges in Smart Healthcare Research"2012

[5]http://www.datasciencecentral.com/profiles/blogs/randomforests-algorithm.

[6]https://www.analyticsvidhya.com/blog/2015

[7]Stegeman CA, Anti-neutrophil cytoplasmic antibody (ANCA) levels directed against proteinase-3 and myeloperoxidase are helpful in predicting disease relapse in ANCA-associated smallvessel vasculitis, Nephrol Dial Transplant 2002, 17:2077-2080.

[8]https://www.analyticsvidhya.com/blog/2015/09/naivebayes-Explained/.v

[9] Md. Tahmid Hossain, Abu Raihan MostofaKamal ,Automated Disease Prediction System (ADPS)"2016

[10] R. Tamilarasi* , Dr. R. Porkodi "A Study and Analysis of Disease Prediction Techniques in Data Mining for Healthcare"2015

[11] Garcia-Chimeno, Y., Garcia-Zapirain, B., M., Fernandez-Ruanova, B., & Garcia-Monco, J. C. " Automatic migraine classification via feature selection committee and machine learning techniques over imaging and questionnaire data. 2017