International Engineering Journal For Research & Development
Vol.5
Issue -1
E-ISSN NO:-2349-0721
Impact factor : 6.03

A SURVEY ON PREDICTION OF DISEASES THROUGH DATA MINING

Alok Nagargoje, Anand Bohara, Vijay Kakade, Mrunal Kale, Prof P.S.Hanwate

NBN SINHGAD SCHOOL OF ENGINNERING, AMBEGAON(BK), PUNE

nagargojealok@gmail.com,aanandbohara@gmail.com,vijaykakade.vk@gmail.com,
mrunalkale2277@gmail.com, ajayshanwate@gmail.com

Abstract:

A vast amount of data is bring out in the fields of health-care and diagnostics, doctors have to make an in-person contact with patience to determine the wounds, diseases and injuries by which the patient is affected. Wrong clinical decisions taken by medical practitioners can cause any harm and result in serious loss of life of patience, which is hard to afford by any hospital. To acquire a precise and cost effective treatment, technology based data mining system can be considered to make worth decision. This survey paper analyses different data related to symptoms and diseases which can be used for predicting different types of diseases. The main focal point of this project is the application of classifying and predicting a precise disease by achieving the operations on medical data generated in the field of health-care and medical. In this project an affected multi-class Naive Bayes, Decision Tree and Random Forest algorithm is used for prediction of particular disease by training it on a set of data before implementation. Data mining plays a crucial role of predicting diseases in thealth-care. In order to determine a particular disease, numerous different tests are needed to be done. This makes the whole process tedious and it can be reduced with the help of symptoms of the patients in data mining.

Keywords: - Decision Tree (DT), Naive Bayes (NB), Random Forest (RF), K Nearest Neighbour (KNN), Support Vector Machine (SVM), Neural Network (NN), Iterative Dichotomiser 3 (ID3), Internal Health Disease Prediction System (IHDPS), Frequent Pattern Growth (FP Growth), Multilayer Perception (MLP), Waikato Environment For Knowledge Analysis (WEKA).

I. INTRODUCTION

Data is collection of facts, figures and statistics together which can be used for reference or analysis purposes. Data mining consist of selecting data, modelling data and discovering huge amounts of data. Various different hidden patterns that are valuable, are been discovered from large data sets. The problems occurred in the process of data mining are often dealt by computer sciences, such as machine learning, soft computing and data visualization and include classification and regression techniques. Some of the research works are accomplished in this side, but all of them are concentrated on a few approaches of analysis, prediction or diagnosis of this disease by using different techniques and tools. The planned system can clarify problematic queries for recognize of a specific disease and also can help medical professionals to take smart decisions that the conventional systems were incompetent to perform. The decisions seized by medical practitioners with the

help of technology can result in powerful and low expense treatments. Data mining is an act of turning material to use as study of obtaining previously undetected patterns from an selected data set.

II. MOTIVATION

This paper is completely based upon developing a cost efficient and fully functional disease prediction system. It can be used in rural areas, where people cannot afford huge amount of hospital fees. It can be used in urgent basis for quick disease prediction. We are going to design disease prediction using data mining for health-care system. This will help multiple patients to predict their diseases from their symptoms. The most difficult area in data mining is to construct precise and efficient classifier in medical sectors. In this project an affected multi-class Naive Bayes, Decision Tree and Random Forest algorithm is used for prediction of particular disease by training it on a set of data before implementation. An important challenge in this system development is it requires huge amount of precise medical data and the system also needs support area with internet connection.

I. OVERVIEW

Existing Paper	Method Used
Human Heart Disease Prediction system using Data Mining Techniques Theresa Princy R 2016	The paper has proposed a framework to precisely forsee heart disease
Review of heart disease prediction system using data mining and hybrid intelligent techniques R Chitra, V Seenivasagam 2013	The paper has developed a computer based heart disease predicting system that helps the physician as a tool for diagnosis
Prediction of Heart Diseases using Data Mining Techniques K. Manimekalai 2016	The paper demonstrates Random Forest as best classifier for disease categorization of WEKA tool for large data sets
Analysis of Data Mining Techniques for Diagnosing Heart Disease Jyoti Rohilla, Preeti Gulia 2015	The paper proposes discretization and IQR filters to enhance the efficiency of Hidden Naive Bayes
Kidney Disease Prediction using SVM and ANN algorithms Dr. S. Vijayarani, S. Dhayanand 2015	The paper has uncovered that the Neural Networks shows significant results over all other data mining techniques
Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease Vivekanandan T. 2017	The paper has proposed the challenging tasks of selecting results from the enormous set of features and diagnosis of heart disease
Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and K-nn based weighting preprocessing Polat K., S. Sahan, Gunes S. 2007	The paper has proposed automatic detection of heart disease using AIRS and K-NN

II. PROPOSE SYSTEM

The system proposes is an advance and efficient perdition system for multiple diseases. This system is based on the training of data set obtained by survey through multiple hospitals and symptoms of each individual

disease those patients gains. The strategy of the system is to analyze and test data mining algorithms (Decision Tree, Naive Bayes and Random Forest) and implement the algorithm whose outcome has high test degree of accuracy. The data set is created for implementation purpose and it contains symptoms of diseases. The algorithms which are used for implementation purpose in this project are Naive Bayes, Decision Tree and Random Forest which are selected by evaluating the prediction accuracy and latency analysis results.

III. SYSTEM ARCHITECTURE

The below diagram is the system's architecture:

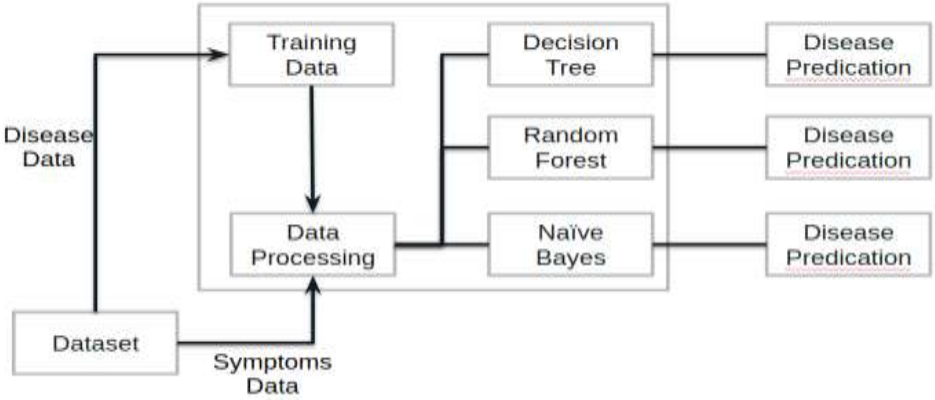

Figure 1: system architecture

In system architecture we are using three types of algorithms i.e. Decision Tree, Naive Bayes and Random Forest. By using these three algorithms, the perdition of disease are done. Data set contains the data of symptoms and diseases. The symptoms are taken from the patient and send for data processing. The data processing is a combination of trained data and untrained. Based on the symptoms these algorithms (Decision Tree, Naive Bayes and Random Forest) predict the disease of the patients. The algorithm with high degree of accuracy is selected and based on the patient symptoms respective output of disease is shown to patient.

VI. METHODOLOGIES

In this Paper we are considering the three algorithms which are having more accuracy rate than other algorithms such as ID3, J48, BayesNet and KNN. We have elaborated DT, NB and RF in detail with their steps and algorithm.

Decision Tree:

Decision tree is supervised learning algorithm that is used for solving regression and classification problems too. The main intention to use decision tree is to develop training model, so that it can be used to forecast the class or value of target variable from training data. The decision tree is so easy to understand than any other classification algorithm. It tries to provide the solution to the problem by using tree representation, in which each internal nodes of the tree associated to an attribute as well as each leaf node associate with a class label.

STEPS:

1. The best attribute from the data-set is selected as root of the tree.

2. The training set is divided into subsets. Each and every subset should contain data with the same value for an attribute.
3. Repeat step 1 and step 2 on separate subset until you find leaf nodes in all the branches of the tree.

In decision tree to predict the class label for an particular record we have to start from the root node of the tree. We compare the values of the root node's attribute with record's attribute and based on that we follow the branch associated with the given record's attribute value and jump to the next node. We continue the same process of comparing with other internal nodes of the tree until we reach to the leaf node with predicted class value.

Naive Bayes:

Naive Bayes algorithm is supervised learning algorithm. It encompasses a family of simple "probabilistic classifiers" settled on applying Bayes theorem with strong straightforward independence expectation between the features. They are among the simplest Bayesian network models. It is not a single algorithm, rather it consists of a same family, and having a common principle "Every pair of features being classified is independent of each other".

The Naive Bayes algorithm divides the data-set into 2 different parts:

1) Feature Matrix - It contains all the rows that appear in the data-set. It contains the value of dependent features.
2) Response Vector - It contains the value of prediction or output (class variable) for each row of feature matrix.

Naive Bayes Pseudo code:

INPUT - Training data-set T,
F= (f1, f2, f3… fn) // Predictor variable
OUTPUT - A class of training data-set

STEPS:

1. Read the training data-set T.
2. The mean and standard deviation of the predictor variable in each class is calculated.
3. Repeat. Calculate the probability of "fi" using the gauss density of equation in each class, until the probability of all predictor variable has been calculated.
4. Calculate the likelihood for each class.
5. Get the greatest likelihood.

Random Forest:

Random forest algorithm is a supervised learning model. In case of large data sets, the decision tree occurs with the problem of over-fitting. To overcome this problem, random forest operates by constructing multiple decision trees at the training data set. It uses ensemble learning that solves a particular problem by creating multiple models for classification, regression and other tasks.

The random forest model uses 2 key concepts while constructing a tree:

1. The sampling of training data points while building trees are selected at random.
2. Than random subsets of the data is considered while splitting the nodes.

There is selection of random samples from the data sets to generate a training data sample. Various different training data sets are made by replacement of samples. It is called as bootstrap data set. There can be repetition of some samples in same tree. For each bootstrap data set, a decision tree is created predicting new nodes. Predictions are made by averaging the prediction of each decision tree. This concept of learning from different bootstrap data set and then averaging the prediction is called bagging.

The bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias.

Bias - The bias-error is an error from erroneous (incorrect) assumptions in learning algorithm. High bias can cause under-fitting.

Variance - The variance is an error from sensitivity (quality) to small fluctuations in training set. High variance can cause over-fitting.

The prediction of a single tree may lead to high noise; rather it can be decreased by using large number of different decision trees. Training of different trees on same data set may lead to strongly correlated trees. With the help of numerous different data set and their decision trees, it is easy to classify and predict results. It helps to reduce the chances of over-fitting, which occurs while working on large data sets. The same random forest algorithm may be used for both classification and regression. It identifies the most important features out of the available features from the training data-set to provide the best. The random forest outperforms the single decision tree.

STEPS:

1 It selects “k” features randomly from total “m” features, Where, k << m.
2 Among the “k” features, calculate the node “d” using the best spit point.
3 Split the node into child node using best split.
4 Repeat 1 to 3 steps until “l” number of nodes have been reached.
5 Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.
These randomly created trees form the random forest.

Steps for prediction of Random Forest:

1) The test features and rules of each randomly created decision tree is used to predict the outcome and stores the predicted outcome.
2) Calculate the votes for each predicted target.
3) The highest voted predicted target is selected as the final prediction from the random forest algorithm.

VIII. CONCLUSION

This paper enlightens all the issue, challenges and issue faced by medical practitioners while determining the disease of patient without considering the medical data of patient. By using the multiple machine learning algorithms based on data set user can predict the disease. The system becomes more accurate because of multiple machine learning algorithms. The patient is able to find out disease based on his symptoms, hence reducing the time and cost of direct treatment from hospital. The proposed system features a low computation cost and confidentiality of the training set and prediction.

IX. REFERENCE

1) Theresa Princy R, "Human Heart Disease Prediction System using Data Mining Techniques", International Conference on Circuit, Power and Computing Technologies [ICCPCT], IEEE (2016)
2) Chitra R., Seenivasagam V., "Review of heart disease prediction system using data mining and hybrid intelligent techniques", ICTACT journal on soft computing, volume: 03, issue: 04, 2013.
3) K.Manimekalai, ―Prediction of Heart Diseases using Data Mining Techniques‖, International Journal of Innovative Research in Computer andCommunication Engineering, Vol. 4, Issue 2, ISSN(Online):2320-9801, ISSN (Print):2320- 9798, February 2016.
4) Jyoti Rohilla, Preeti Gulia, ―Analysis of Data Mining Techniques for Diagnosing Heart Disease", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 7, ISSN: 2277 128X, July 2015.
5) Dr. S. Vijayarani, Mr.S.Dhayanand, "Kidney Disease Prediction using SVM and ANN algorithms", International Journal of Computing and Business Research (IJCBR), Volume 6, Issue 2, ISSN (Online):2229-6166, March 2015.
6) Vivekanandan T et al., "Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease", www.elsevier.com/locate/compbiomed, https://doi.org/10.1016/j.compbiomed, Pages: 125-136 (2017)
7) Polat K., Sahan S., Gunes S., "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing", ScienceDirect, 2007.
8) Parvathi I, Siddharth Rautaray, ―Survey on Data Mining Techniques for the Diagnosis of Diseases in Medical Domain, International Journal of Computer Science and Information Technologies, Vol. 5 (1), 838-846, ISSN: 0975- 9646, 2014
9) Dhanya P Varghese, Tintu P B, ―A Survey on Health Data using Data Mining Techniques‖, International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 07, e-ISSN: 2395-0056, p-ISSN: 2395-0072, Oct2015.
10) Vahid Rafe, Roghayeh Hashemi Farhoud, ―A Survey on Data Mining Approaches in Medicine‖, International Research Journal of Applied and Basic Sciences, Vol 4 (1), ISSN 2251-838X, 2013.
11) T. Revathi, S. Jeevitha, ―Comparative Study on Heart Disease Prediction System Using Data Mining Techniques, Volume 4 Issue 7, ISSN (Online): 2319-7064, July 2015.
12) Devendra Ratnaparkhi, Tushar Mahajan, Vishal Jadhav, ―Heart Disease Prediction System Using Data Mining Technique, International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 08, e-ISSN: 2395 -0056, p-ISSN: 2395-0072, Nov-2015.