Examining the Accuracy of Machine Learning Disease Prediction Algorithms

SMRUTI RANJAN SWAIN
ASSISTANT PROFESSOR, Mtech,
Department of CSE
Gandhi Institute for Technology,Bhubaneswar.

Abstract

Machine learning algorithms use a broad variety of quantitative, probabilistic, and efficient strategies to learn from past knowledge and discover meaningful examples from huge, organized, unstructured, and complex datasets. Data mining and machine learning help in this manner with the diagnosis and treatment of many different diseases. Medical professionals may be able to better anticipate a patient's illness progression and personalize their treatment plans with the use of predictive analysis, provided that the study makes use of various machine learning algorithms that have been shown to be successful. By using machine learning methods, we can now make timely and precise predictions about the spread of disease. Using pathological data, this research aims to compare the performance of multiple supervised machine learning algorithms for disease prediction. As a last step in comparing the efficacy of various machine learning algorithms, we gather data on disease symptoms and visualize the results.

Introduction

Different symptoms are associated with each and every disease which is difficult to diagnose quickly and to keep the patient better healthy. In traditional practice, to foresee the existence of a disease and for assisting the diagnosis process of a disease, each symptom is identified with some weights, and whichever symptom is having a high impact on the disease will be considered primarily. Data mining provides excess and substitutes knowledge for making decisions on predicting the disease by healthcare professionals. This process extracts the Hidden and unknown patterns, relationship and knowledge where statistical traditional methods difficult to process. The effective treatment is constantly credited by privilege and exact conclusion. The chronicled clinical information needs systematic strategies to analyze the information and to extract the potential data from it for which the data mining ideas go accommodating to find the unrevealed figures, connections from database and machine learning ways to deal with further analyze a patient. There are various algorithms of machine learning which are used to predict diverse types of diseases and for making good protocols for better health avoiding the high cost, late recovery, and error treatment. Machine learning algorithms are found to be gratefully successful in predicting disease and many more techniques are yet to explore. Machine learning is a huge strategy for information analysis that iteratively gains from accessible information with the guide of learning algorithms.

In this paper, we proposed a GUI application utilized by machine learning methods for disease prediction system, applying mining on the disease symptoms

Copyright @ 2021 Authors

and finally detecting the disease on comparing the performance or accuracy in finding the disease with various techniques. To achieve using the classification methods of machine learning like Naive Bays, Decision tree, Support Vector Machines and Random Forest that could give precise results for every new insert. Evaluate and analyze statistical and visualized results, which find the standard pattern for all regiments. We further, predict the outcome of a patient whether the patient is likely to suffer from a disease or not using a GUI application. The proposed framework intends to determine the issue precise prediction of disease for a patient by right off gathering suitable information and components identified with a different disease, distinguishing essential characteristics that could assist us with verifying the normal examples in a selection of the model and parameters. The presentation is briefly separated as: initially we review basic strategies of machine learning identified with anticipating various diseases at that point examine the methodology including the information and techniques utilized. Next, we investigate the outcomes lastly present ends.

Related Work

The proposed optimization hybrid approach method by Youngest Khourdifi [1] increased the predictive accuracy of medical data sets. These methods on comparison with the supervised procedures depend upon existing datasets of classification accuracy estimation which are utilized to assess the performance. KNN (99.65%) and RF (99.6%) are the precision score achieved by the proposed model using FCBF, PSO and ACO. H.Benjamin Fredrick David [2] proposed an algorithm which gave maximum precision with the classification carried of a typical and anomalous individual. From UCI machine learning repository heart disease dataset is used in his methodology for the assessment of the performance of the algorithms. The conclusion is 8% exactness traced by the calculations using Random forest algorithm when contrasted with different calculations for coronary disease expectation. Meghan Shah [3] recommended data mining techniques for Heart Disease Prediction System. In the proposed strategy WEKA programming is utilized for programmed determination of sickness by providing characteristics of administrations in healthcare place. They used various methods like Support Vector Machine, Naive Bayesian, KNN, Association rule, ANN, and Decision Tree. They suggested Support Vector Machine is reasonable providing dense exactness as differentiated and various data mining procedures. A few investigations are led to clinical datasets utilizing numerous classifiers and features selection as suggested by M. Fatima [5]. The proposed method uses the classification of the heart disease dataset where good classification accuracy is traced. The hybrid effective algorithm is suggested by Malay [6] which is used to predict coronary disease. The popular clustering algorithm K means and ANN are used to extort anonymous information about coronary disease. An accuracy of 97% is given by the proposed methods. In this, a hybrid approach that includes merging various procedures like FCBF (Fast Correlation-Based Feature Selection) strategy to filter redundant features so as to improve the nature of heart disease classification.

Proposed Work

Methodology

Copyright @ 2021 Authors

The fundamental commitments of this work are to extract the classified precision valuable for predicting various diseases, Although Machine learning models have been broadly contemplated and seen as extraordinarily fruitful, the disease prediction is a confounded issue and there are as yet numerous upgrades to be made and techniques to investigate. We assemble predictive models and look at them utilizing explicit execution measurements including Accuracy, Precision, Recall and F-score, recognizing the best ones that could be utilized for foreseeing different diseases. Programmed algorithms are used by machine learning to find out and upgrade the activities by breaking down info information to make forecasts inside a worthy choice. In concern with the recent information, these techniques will in general, build progressively exact expectations.

Dataset

Datasets from the effectively accessible repositories can be utilized for training the machine. Data pre-preparing or Data cleaning is one the significant angle to be completed before executing machine learning calculations for mining purposes. The dataset used in the proposed system contains 4988 rows and 133 columns. This dataset is downloaded from an open-source (https://github.com/feat7/symptom-to-disease-prediction/tree/ master/data/clean). The training dataset has 4920 rows and 133 columns and the testing dataset has 68 row and 133 columns. Some of the attributes in the training dataset are shivering, skin rash, acidity, itching, obesity, mild fever, high fever, knee pain, puss-filled pimples, puffy face and eyes, blurred and distorted vision, phlegm, throat irritation, redness of eyes, sinus pressure, runny nose, congestion, chest pain, weakness in limbs, fast heart rate, pain during bowel movements, pain in the anal region, bloody stool, red spots over body, belly pain, abnormal menstruation, diachronic patches, watering from eyes, increased appetite, polyuria, skin peeling, silver-like dusting, small dents in nails, inflammatory nails, blister, red soar around nose, yellow crust ooze, itching, headache, cough, high fever, fatigue, weight loss.

Architecture of the proposed system

In the proposed system, a relative study and performance evaluation between four different Machine learning classification algorithms is made using performance metrics such as accuracy, precision, recall and F-score. In this work, the various machine learning algorithms like used are Naïve Bayesian, Decision tree method, Random Forest method and Support Vector Machine method for calculating the accuracy in predicting disease. In machine learning, the most repeatedly used technique is classification. The disease prediction is carried out by the technique which is depicted in Figure1which decided the investigation approach for developing a classification model necessary for the forecast of the patient disease. In order to make forecasts, a classifier should be prepared with the reports and then produce a classification model which is taken care of with another obscure report and the anticipation is made. The setup of this exploration incorporates the Performance Evaluation of the four algorithms used for classification.

Copyright @ 2021 Authors

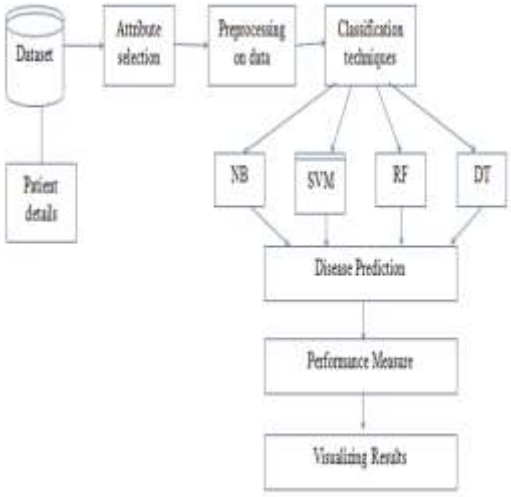

Figure1. Architecture of the proposed system

Data preprocessing

Data preprocessing is a very important innovate mining method particularly in health data. Data collection strategies square measure typically loosely controlled, leading to out of vary, inapplicable information mixtures, missing values etc. Data preprocessing cannot be done carefully screened for disease prediction which gives results that mislead the diagnosis. Data arrangement and sifting steps will took a lot of handling session. Data preprocessing incorporates cleaning, Instance choice, standardization, change, include extraction and determination. The result of information preprocessing is the last training set.

Classification techniques

The dataset is spitted into two parts in terms of percentage, 80% of the dataset is training dataset and the remaining 20% of the dataset is test dataset. In this work at the training phase, a classification model is built on the dataset which is trained by the four classification methods like Naïve Bays, Decision Tree, Support Vector Machine, and Random Forest used in the training phase of the proposed method. Each of the four methods is depicted below.

Naïve Bays Method

The most important aspects of machine learning are classification and prediction where the world brimming with AI and machine learning consciousness encompassing, nearly everything around. Naïve Bays is a basic yet surprisingly incredible algorithm for predictive examination. It is a classification strategy dependent on Bays the hypothesis with suspicion of freedom among predictors. It involves two sections which are Naïve and Bays, in straightforward terms In the Naïve Bayesian method, the classifier will accept when the nearness of a precise feature of a class is random with presence of another attribute. Regardless of whether these features rely upon one another or upon the presence of different features these properties autonomously add to the probability that is the reason named after seeing that Naïve.

For huge datasets especially we can use Naïve Bays model which is easy to build. In probability theory and measurements, Bays hypothesis which is then again known as the Bays law or the Bays rule depicts the likelihood of an event dependent on earlier information on the conditions that may be identified with the event. Bays theorem is a technique to build out the conditional probability. By considering the prior knowledge of a condition that relates to the event, the condition probability of the event occurrence can be calculated by using Bays theorem. The conditional likelihood is the likelihood of an event happening given that it has some relationship to at least one different event. Bays hypothesis is marginally more nuanced more or less; it gives the real likelihood of an event. For the given information

Copyright @ 2021 Authors

Naïve Bays method is defined as given a speculation H and the proof E, Bayesian theorem expresses the connection involving the likelihood of the theory before getting the proof P (H) and the likelihood of the speculation in the wake of getting the proof is given by

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)}$$

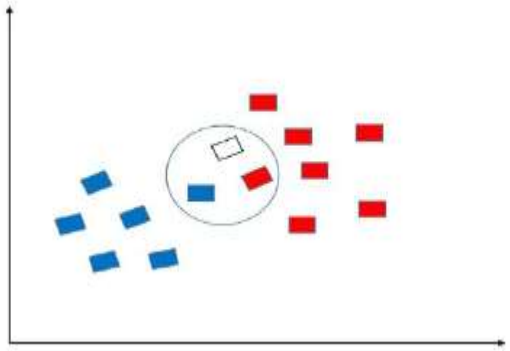

Figure2. Illustration of Naïve Bays method

This relates the probability of theory before cutting the proof which is P (H) to the likelihood of the speculation in the wake of getting the proof which is P (H|E) thus P (H) is known as the prior probability while P (H|E) is known as the posterior probability and the factor P (E|H) that narrate both is known as the likelihood ratio. Figure 2 shows the delineation of the Naïve Bays method. Now using this term Bays theorem specifies differently as the procedure probability equals the prior probability times the likelihood ratio.

Support Vector Machine method

The various applications of SVM that are generally used with it are detecting face, normal text in hypertext classification, image classification and bio-information. The support vector machine is specific to supervised learning machine learning model learns from the past input data and makes future predictions as output. SVM is a method that looks at data and sorts into one of two categories. In the larger picture of the machine learning model and under supervised learning we can see that the support vector fits in under classification deciding what yes-and-no is and there is also a regression version but it is primarily used for classification. SVM exists in both linear and non-linear forms. There are two data sets like train dataset and test dataset which is involved in SVM.

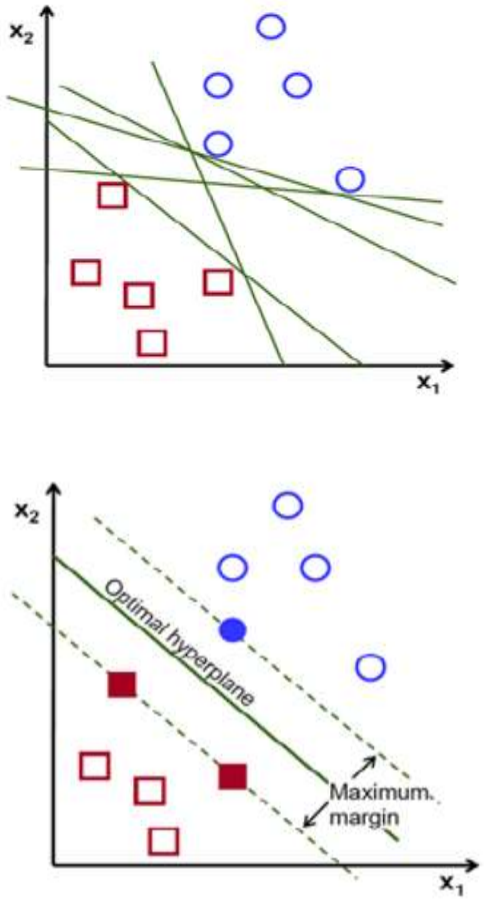

Figure3. Illustration of Support vector machine

In general, if two classes are there, it splits the two classes perfectly which is called linearly separable.

Copyright @ 2021 Authors

On the other side, a set of lines can be used to split the dataset; among them one separating line is chosen as the best line. The ideal line is selected in such a way that the distance should be maximum to the closest purposes of the two classes in the training dataset. The separation between the support vector and the hyper plane ought to be beyond what many would consider possible and this is the place the support vectors are the outrageous focuses in the dataset.

Random Forest method

Random forest algorithm which is a popular supervised machine learning technique used in regression and classification. It is called random forest because the forest has trees and a tree in the machine learning world means a decision tree. The random forest it constructs is based on an ensemble of decision trees using the bagging method. This bagging method combines various learning methods to increase the overall results. The random forest is built on various decision trees and combines all to form accurate and stable predictions.

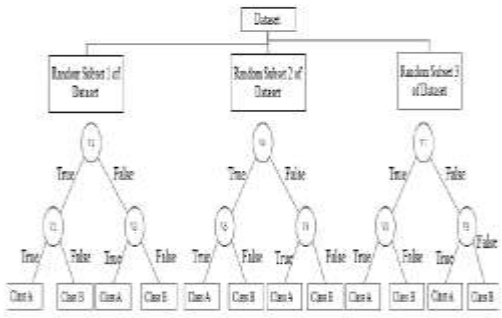

Figure3. Illustration of Random forest method

Random forests are a blend of tree indicators utilizing a decision tree to such an extent that each tree relies upon the estimations of an irregular vector examined autonomously and with a comparative conveyance for all trees in the forest. The theory mistake of a forest of tree classifiers relies on the idea of the individual trees in the forest and the association between's them. They are continuously enthusiastic regarding commotion. It is a supervised classification algorithm utilized for the desire and it is considered as the ideal because of its enormous number of trees in the forest giving improved precision than decision trees. Normally, the trees are trained autonomously and the desires for the trees are merged through averaging. Random forest algorithm can utilize both for classification and the regression dependent to the difficult space.

Initially, the k features are removed from absolute m features. Figure 4 shows the illustration of Random forest method. In the following phase, every tree haphazardly chooses k features so as to discover the root node by utilizing the best split loom. The following phase on the disease dataset includes computing the daughter nodes utilizing a similar best split methodology. Likewise, the tree is shaped from the top hub, that is the root and until all the leaf hubs are produced from the features. A random forest is created by using the tree formed by randomly which is being used for making disease prediction.

Decision Tree Method

A decision tree is simple and used as one of the machine learning implementations of classification methods, which are represented in the form of a hierarchical structure. The decision tree algorithms can handle both numerical and categorical which is a supervised learning algorithm. In a decision tree, the data items are being mapped based on certain predictive conclusions. The classification in the decision tree categorizes the input data into an outcome. To clearly specify the accuracy of the algorithm is directly depended upon the features it

Copyright @ 2021 Authors

selects for training model. Figure 5 shows the illustration of Decision tree method. The nodes of the decision tree contain different levels where the highest node is represented as the root node. Nodes which are having at least one child, all internal nodes signify tests on input factors or properties. Contingent upon the test result, the classification algorithm branches towards the fitting child node where the procedure of test and fanning rehashes until it arrives at the leaf hub. The leaf or terminal hubs compares with choice results.

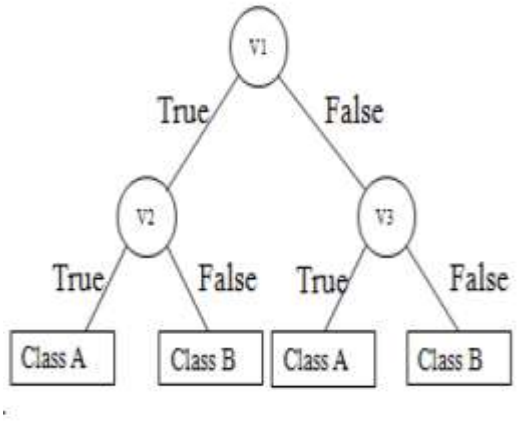

Figure5. Illustration of Decision tree method

Evaluating classification algorithms performance metrics

In this paper, the various performance metrics used to assess and compare classification algorithms are Accuracy, Precision, Recall and F-score. These are measured on the basis of a two by two matrix called Confusion matrix. This confusion matrix holds the models predicted values to the actual class values. All the measures are focused on the values present in the four quadrants of the Confusion matrix.

Table1. Confusion matrix

Class-Actual values	Class-Predicted values	
	P	N
P	TP	FN
N	FP	TN

In the above table 1 the terminology is given as

- Observation positive is given by P
- Observation negative is given by N
- Positive observation and negatively predicted is given by TP

Accuracy

Accuracy or success rate is defined as the proportion of correctly classified test instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

Precision or positive predicted value is the proportion of the total number of correctly classified positives with the total number of predicted positive instances.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall

Recall or True positive rate is characterized as the proportion of the absolute number of accurately grouped positive cases to the all out number of positive cases. High Recall demonstrates effectively perceived (for example few FN).

Results and Discussions

Figure 6 shows the output with a GUI for a disease dataset. From the dataset the input symptoms are chosen as blackheads, puffy_face_and_eyes,

Copyright @ 2021 Authors

puss_filled_pimples and irritability. The predicted disease for the given symptoms by all the classification algorithms like Support vector machine, Random Forest, Naïve Bays and Decision tree is Acne.

Figure 6: Disease predictor-Acne

Figure 7 shows the results for the attributes like Accuracy, Precision, Recall and F-score for the classification algorithms Support vector machine, Random Forest, Naïve Bays and Decision tree.

The results obtained after running the code are compared and analyzed. Decision Tree algorithm gives an accuracy of 0.9527, Random Forest gives an accuracy of 0.9504, Naive Bayesian algorithm gives an accuracy of 0.9693 and Support vector machine algorithm gives an accuracy of 0.9504 .

From the above results it is clear that Naïve Bays algorithm gives the highest accuracy with the given training and testing datasets when compared to other algorithms like Support vector machine, Random Forest and Decision tree.

```
Training set :  (4920, 133)
Testing set :  (331, 133)
--------DECISION TREE--------
ACCURACY OF DECISION TREE ALGORITHM IS:  0.9527436477372293
RECALL  0.9795585785462744
PRECISION  0.9759001161440187
F1 SCORE  0.9742369361909198
--------RANDOM FOREST--------
ACCURACY OF RANDOM FOREST ALGORITHM IS :  0.9504805065138826
RECALL  0.986775419702249
PRECISION  0.9651567944250871
F1 SCORE  0.969393526233399
--------NAIVE BAYES--------
ACCURACY OF NAIVE BAYES ALGORITHM IS:  0.9693080632105023
RECALL  0.9890983000739096
PRECISION  0.9814169570267132
F1 SCORE  0.982808160379740
--------SUPPORT VECTOR MACHINE--------
ACCURACY OF SVM ALGORITHM IS:  0.9504805065138826
RECALL  0.986775419702249
PRECISION  0.9651567944250871
F1 SCORE  0.969393526233399
```

Figure 7: Calculations of the metrics for the disease Acne

Figure 8 compares the values of the obtained attributes from all the supervised learning methods.

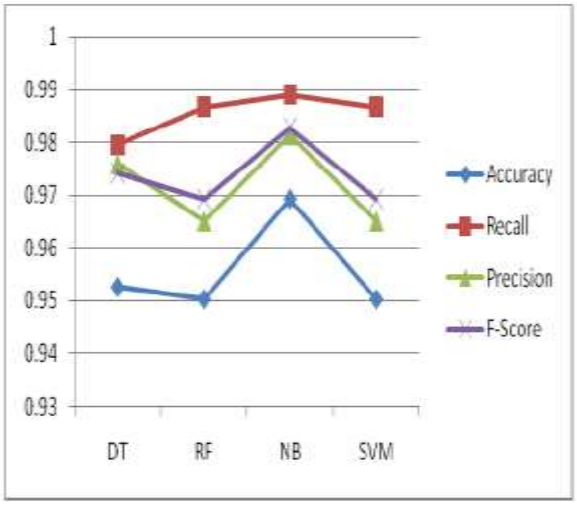

Figure 8: Graphical representation of comparative analysis

Conclusion and Future work

The primary objective of this article is to use Machine Learning techniques to better accurately predict the occurrence of the disease. Decision Tree,

Copyright @ 2021 Authors

Random Forest, Support Vector Machine, and Naive Bays were some of the methods we used. These algorithms take symptoms into account as input variables, allowing for an evaluation of Accuracy, Precision, Recall, and F-score on a disease dataset. The Nave Bays algorithm provides the greatest accuracy in illness prediction among all the compared algorithms. Because this new condition may emerge at some point in the future, the system may be improved further. Adding additional symptoms to the data set allows for the prediction of these emerging illnesses. The GUI is made more user-friendly by using more potent machine learning supervised algorithms, which allows for the provision of more specific information about the condition to patients, and the inclusion of a component prescribing medication to the patient in the event of an emergency.

References

[1] Youngest Khourdifi, Mohamed, "Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization", International Journal of Intelligent Engineering and Systems, 12(1), (2019), pp 242-252.

[2] H. Benjamin Fredrick David, "Heart Disease Prediction using Data Mining Techniques", ICTACT Journal on Soft Computing, October (2018), 09(01), pp 1817-1823.

[3] Megha Shahi, Kaur Gurm, "Heart Disease Prediction System using Data Mining Techniques", Orient J. Computer Science Technology, vol.6 (2017), pp.457-466.

[4] Sanjay Kumar."Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms", International Journal of Engineering and Computer Science, ISSN: 2319-7242, 6 (6) June 2017, Page No. 21623-21631,

[5] Fatima and Pasha, "Survey of machine learning algorithms for disease diagnostic. "Journal of Intelligent Learning Systems and Applications", Vol.9, No.01, pp.1, (2017).

[6] A. Malay, Adam, and Kamet, "Prediction of heart disease using k-means and artificial neural network as a hybrid approach to improve accuracy", International Journal of Engineering and Technology, Vol.9, No.4, (2017).

[7] Katie and Sumac, "Efficient Heart Disease Prediction System using K-Nearest Neighbor Classification Technique". In: Proc into Conf Big Data Internet Things. (2017).

[8] J. Sumatra, "Performance Evaluation of Machine Learning Algorithms in the Classification of Parkinson Disease Using Voice Attributes:, International Journal of Applied Engineering Research Vol 12, Number 21 (2017) pp. 10669-10675

[9] Gomati, Dared. Shanmuga Priyaa, "Multi Disease Prediction using Data Mining Techniques", International Journal of System and Software Engineering, December (2016), pp.12-14.

[10] Ashwini Shetty, Chandra Naik, Different Data Mining Approaches for Predicting Heart Disease, International Journal of Innovative in Science Engineering and Technology, Vol.5, May (2016), pp.277-281.

Copyright @ 2021 Authors