IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 5 Issue 5, May 2018
ISSN (Online) 2348 – 7968
www.ijiset.com

Breast Cancer Diagnosis Using Machine Learning Techniques

Kriti Jain[1], Megha Saxena[2]andShweta Sharma[3]

[1]Computer Science and Engineering Department, Maharaja Agrasen Institute of Technology Delhi, India

[2]Computer Science and Engineering Department, Maharaja Agrasen Institute of Technology Delhi, India

[3]Computer Science and Engineering Department, Maharaja Agrasen Institute of Technology Delhi, India

Abstract

Breast cancer is an all too common disease in women, successfully predicting and classifying it is a dynamic research issue. Various measurable and machine learning systems have been utilized to create different bosom tumor forecast models. Among them, Support vector machines (SVM) have been appeared to outflank numerous related methods. However, there have been very few studies focused on examining the classification performances of different classification algorithms. Along these lines, the point of this paper is to completely evaluate the expectation execution of SVM, K-Nearest Neighbors, Gaussian Naive Bayes and Classification and Regression trees (CART) on breast cancer dataset. The breast cancer dataset (Wisconsin Breast Cancer (WBC)) was taken from UCI Machine Learning Repository, place for machine learning and insight frameworks. The classification accuracy, precision, F-measure, of various classification algorithms are looked at. The trial comes about demonstrate that SVM classifier can be the better decision for classification, where accuracy of the algorithm is improved by tuning the parameters of the dataset.

Keywords-Machine Learning, CART,Gaussian Naive Bayes, K nearest neighbors, Support Vector Machine

1. INTRODUCTION

Breast cancer has been known to be an important research problem in the healthcare communities. This cancer develops in the woman breast tissue [1]. There are various risk factors for breast cancer including ,bulkiness , being lethargic,, drinking alcohol, hormone replacement therapy during menopause, ionizing radiation, early age at first menstruation, having children late or not at all, and older age.

There are various types of breast cancer, at various stages, levels, and genetic marks. Hence, it would be very useful to have a system that would allow early detection of type of cancer and therefore prevention which would add on to the survival rates for breast cancer.

The paper discusses a number of different statistical and machine learning measures that have been applied to develop breast cancer prediction models, such as Naïve Bayes, CART,, K-nearest neighbor and Support vector machine methods [2–9].

Enhancing the cause, studies have shown that SVM works the best in developing the model.

The achieved dataset for breast cancer prediction is mostly imbalanced, with the minority class containing a small number of patients with cancer and the majority class containing a large number of patients without cancer. This means that prediction accuracy or classification accuracy are insufficient for prediction methods.

The data used in this experiment was obtained from the UCI machine learning repository [11] and described by Dr. William H. Wolberg. The breast cancer data have been used in some research. [14] We discussed the effect of 31 characteristic parameters on the state of breast cancer and the influence of the involved parameter on the performance of the SVM models. Wevisualize the data using density plots to get a sense of the data distribution .At the same time, the comparison between the performance of SVMs and other techniques was performed using these data. The problem is to predict the state of breast cancer. In this database, there are 569 pieces of samples, and every

sample is expressed by 31 characteristic parameters.

Therefore, our research goal is to collate SVM,KNN , CART and Gaussian NB. Their performance will be computed by various evaluation metrics, including the classification accuracy, F-measure, and classifier training time. Hence, the findings of this paper should allow further researchers to easily choose the most capable baseline technique that must provide the optimal prediction performance for future comparison.

2. LITERATURE SURVEY

Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases

Survey of ML applications in breast cancer prediction

A broad pursuit was directed significant to the utilization of ML procedures in disease vulnerability, repeat and survivability expectation. Two electronic databases were gotten to be specific PubMed, Scopus. Because of the tremendous number of articles returned by the inquiry inquiries, assist investigation was required so as to keep up the most pertinent articles. The significance of every production was surveyed in light of the catchphrases of the three prescient assignments found in their titles and modified works. In particular, in the wake of perusing their titles and modified works we just chose those productions that review one of the three foci of tumor forecast and included it in their titles. The greater part of these examinations utilizes diverse sorts of information: genomic, clinical, histological, imaging, statistic, epidemiological information or blend of these. Papers that attention on the forecast of bosom tumor improvement by methods for ordinary factual strategies (e.g. chi-square, Cox relapse) were avoided as were papers that utilization methods for tumor characterization or ID of prescient components. As indicated by [3] and their review in view of ML applications in bosom disease expectation, we noticed a fast increment in papers that have been distributed in the most recent decade. In spite of the fact that it is difficult to accomplish an entire scope of the writing, we trust that a critical number of applicable papers were extricated and are exhibited in this survey. As said above, from the underlying gathering of papers we chose an agent list that takes after an efficient structure. In particular, we chose these examinations that make full use of ML strategies and information from varied sources keeping in mind the end goal to foresee the positive result. We concentrated mainly on things that have been distributed in the most recent years as result to display the best in class in this field.

3. PROPOSED MODEL

Experimental Procedure

The experimental procedure is based on the following steps.

1. First of all, the given dataset is divided into 90% training and 10% testing sets based on the 10-fold cross validation strategy [38].
2. In the second step we visualise the data using density plots to get a sense of the data distribution.
3. Finally, the testing set is fed into the constructed classifiers prior to examination of their classification accuracy, precision, and F-measure rates.
4. Furthermore, the classifier training times are also compared to analyse the computational complexities of training different classifiers.A graph is plotted to compare performance of SVM, CART, KNN and Gaussian Naive Bayes.

We also examine whether performing feature selection to filter out unrepresentative features from the chosen dataset can make the classifiers perform better than the ones without feature selection.

Experimental Setup

The Dataset

In this paper, a breast cancer datasets is used, which is available from the UCI machine learning repository (available at: http://archive.ics.uci.edu/ml/) This a relatively small scale dataset, which is composed of 569 data samples and each data sample has 31 different features.

The Classifier Design

There are four single classifiers, namely, linear SVM, CART, KNN and Gaussian Naive Bayes. In addition, to evaluate the performance of the different SVM classifiers, in addition to the classification accuracy,precision, and the F-measure rate, the time that is spent training each classifier is also compared.

Working of the proposed system

The working of the system is depicted as follows:

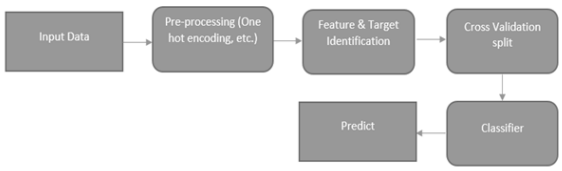

Fig 1.Flowchart of system

4. Implementation of algorithms

CART

CART follows a tree-structured classification scheme where the nodes represent the input variables and the leaves correspond to decision outcomes. DTs are one of the earliest and most prominent ML methods that have been widely applied for classification purposes. Based on the architecture of the DTs, they are simple to interpret and "quick" to learn. When traversing the tree for the classification of a new sample we are able to conjecture about its class. The decisions resulted from their specific architecture allow for adequate reasoning which makes them an appealing technique.

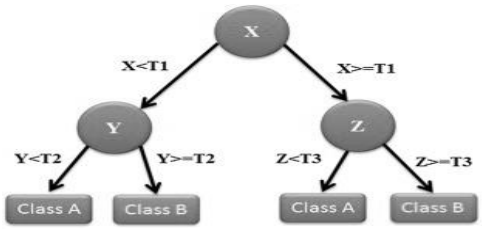

Fig 2. CART

SVM

Support vector machines (SVMs), first introduced by Vapnik [20], have shown their effectiveness in many pattern recognition problems [19], and they can provide better classification performances than many other classification techniques.

An SVM classifier performs binary classification, i.e., it separates a set of training vectors for two different classes (x1, y1), (x2, y2),⋯, (xm, ym), where xi ∈ Rd denotes vectors in a d−dimensional feature space and yi ∈ {−1, +1} is a class label. The SVM model is generated by mapping the input vectors onto a new higher dimensional feature space denoted as Φ: Rd → Hf where d < f. Then, an optimal separating hyperplane in the new feature space is constructed by a kernel function K(xi,xj), which is the product of input vectors xi and xj and where K(xi,xj) = Φ(xi) · Φ(xj).

Fig 3 illustrates this procedure of a linear kernel based SVM, which maps the nonlinear input space into the new linearly separable space. In particular, all vectors lying on one side of the hyperplane are labelled as −1, and all vectors lying on another side are labelled as +1. The training instances that lie closest to the hyperplane in the transformed space are called support vectors. The number of these support vectors is usually small compared to the size of the training set and they determine the margin of the hyperplane, and thus the decision surface.

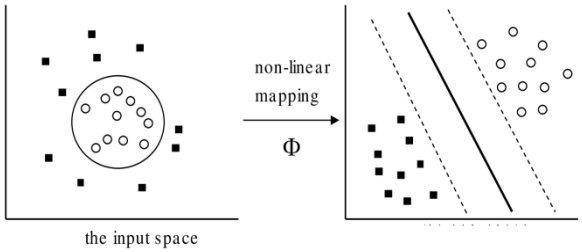

Fig 3.SVM

Gaussian Naive Bayes

A Gaussian Naive Bayes algorithm is a special type of NB algorithm. It's specifically used when the features have continuous values. It's also assumed that all the features are following a gaussian distribution i.e, normal distribution.With real-valued inputs, we can calculate the mean and standard deviation of input values (x) for each class to summarize the distribution.

This means that in addition to the probabilities for each class, we must also store the mean and standard deviations for each input variable for each class.

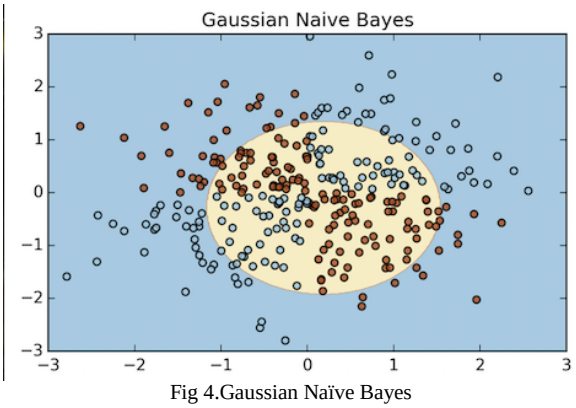

Fig 4.Gaussian Naïve Bayes

K-Nearest Neighbour

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. To evaluate any technique we generally look at 3 important aspects:

1. Ease to interpret output

2 .Calculation time

3. Predictive Power

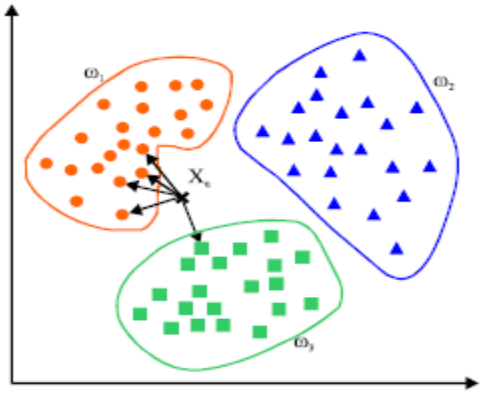
Fig 5.KNN

5. EXPERIMENTAL RESULTS

In this section, the results of the classification are reported. To apply our classifiers and evaluate them, we apply the 10-fold cross validation test which is a technique used in evaluating predictive models that split the original set into a training sample to train the model, and a test set to evaluate it.
After applying the pre-processing and preparation methods, we try to analyse the data visually and figure out the distribution of values in terms of effectiveness and efficiency.

4.1 Density Plots

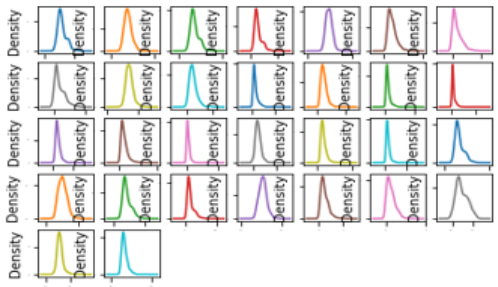
Fig 6 (a). Gaussian Distribution

4.2. Effectiveness

In This section, we evaluate the effectiveness of all classifiers in terms of time to build the model, correctly classified instances, incorrectly classified instances and accuracy. The results are shown in Fig.6(b)

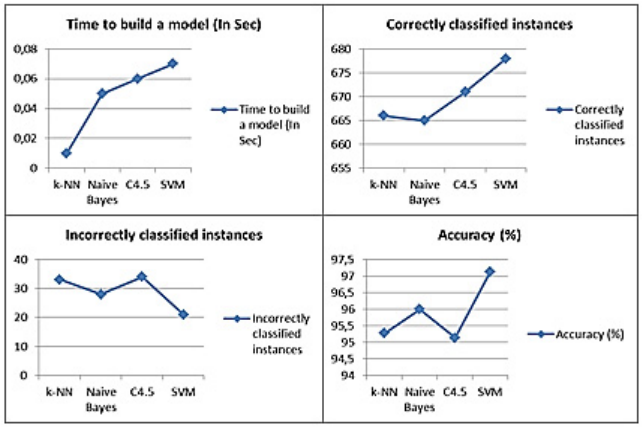

Fig 6 (b) Efficiency comparison

In order to better measure the classification performance of classifiers, we evaluate the accuracy of our classifier in terms of:

- The F measure (F1 score or F score) is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test.
- Precision = (1 – Error) = (TP + TN)/(PP + NP) = Pr(C), the probability of a correct classification
- Sensitivity measures the proportion of positives that are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition).

The results are shown in Fig. 7

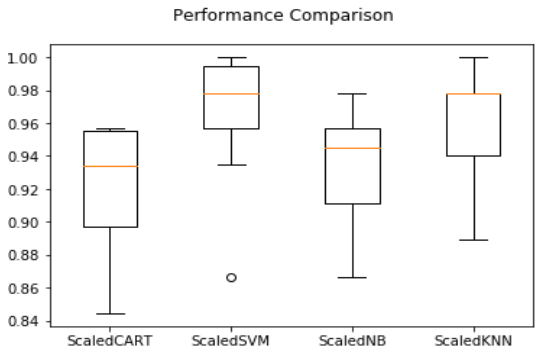

Fig 7 Scaled algorithm comparison

4.2. Efficiency

Once the predictive model is built, we can check how efficient it is. For that, we compare the accuracy measures based on precision, recall, F1- score values for CART, SVM, Gaussian NB and k-NN as shown in Table-3.To better understand efficiency, Fig. 3 presents the classification report of our classifiers that better illustrate the precision of each classifier. It gives a graphical graph that illustrates the performance of different classifiers.

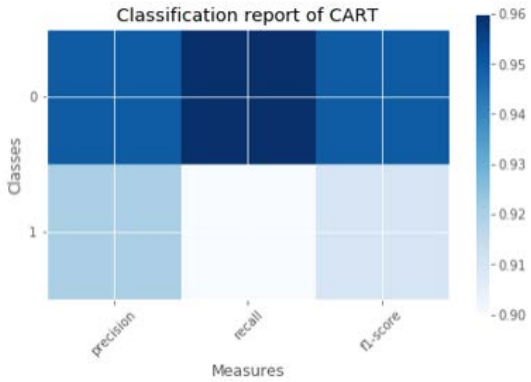

Fig 8.Classification Report of CART

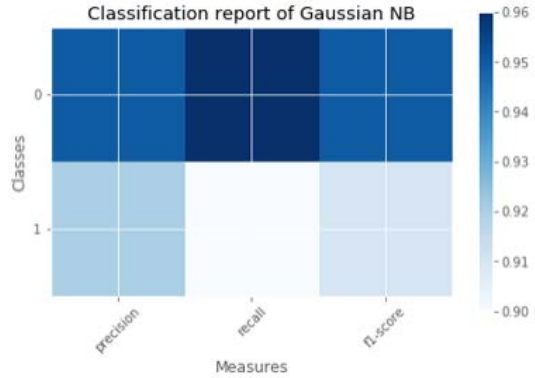

Fig 9.Classification Report of GNB

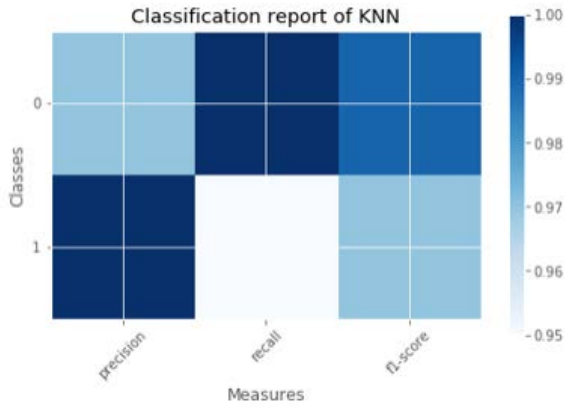

Fig 10.Classification Report of KNN

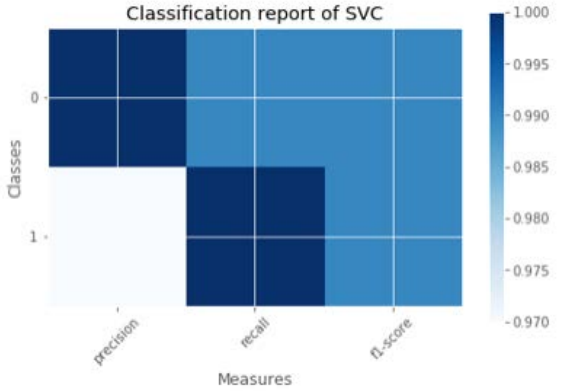

Fig 11.Classification Report of SVM

From the above plots we can easily select optimal models and discard others to best classification. Since Confusion matrices represent a useful way for evaluating classifier, each row of Table 4 represents rates in an actual class while each column shows predictions.

CONCLUSION AND FUTURE SCOPE

To analyze medical data, various data mining and machine learning methods are available. An important task in the field of machine learning is to build accurate and computationally streamlined classifiers for Medical applications. In this study, we employed four main algorithms: SVM, NB, K-NN and CART on the Wisconsin Breast Cancer (original) datasets. We tried to compare efficiency and effectiveness of those algorithms in terms of accuracy, precision, recall and F-measures to find the best classification accuracy. of SVM reaches and accuracy of 97.13% and outperforms, therefore, all other algorithms. In conclusion, SVM algorithm has proven its efficiency and accuracy in breast cancer diagnosis and has achieved the optimum performance in terms of precision and low error rate.

REFERENCES

1. U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based
2. Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2012.
3. Siegel RL, Miller KD, Jemal A. Cancer Statistics , 2016.
4. Noble WS. What is a support vector machine? Nat Biotechnol.
5. Rish I. An empirical study of the naive Bayes classifier. IJCAI Work Empir methods ArtifIntell. 2001.
6. Quinlan JR. C4.5: Programs for Machine Learning.
7. Larose DT. Discovering Knowledge in Data. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2004.
8. X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. Mclachlan, A. Ng, B. Liu, P. S. Yu, Z. Z. Michael, S.
9. David, and J. H. Dan, Top 10 algorithms in data mining. 2008.
10. Datafloq - Top 10 Data Mining Algorithms, Demystified.
11. 44. Accessed December 29, 2015.V. Chaurasia and S. Pal, “Data Mining Techniques : To Predict and Resolve Breast Cancer Survivability,” vol. 3,2014.
12. Djebbari, A., Liu, Z., Phan, S., AND Famili, F. International journal of computational biology and drug design (ijcbdd). 21st Annual Conference on Neural Information Processing Systems (2008).
13. S. Aruna and L. V Nandakishore, “KNOWLEDGE BASED ANALYSIS OF VARIOUS STATISTICAL TOOLS IN DETECTING BREAST,”
14. A. C. Y, “An Empirical Comparison of Data Mining Classification Methods,” vol. 3, no. 2, 2011.
15. A. Pradesh, “Analysis of Feature Selection with Classification : Breast Cancer Datasets,” Indian J. Comput. Sci. Eng., vol. 2, no.5, 2011.
16. Thorsten J. Transductive Inference for Text Classification Using Support Vector Machines. Icml. 1999.
17. L. Ya-qin, W. Cheng, and Z. Lu, “Decision tree based predictive models for breast cancer survivability on imbalanced data,” 2009.
18. D. Delen, G. Walker, and A. Kadam, “Predicting breast cancer survivability: a comparison of three data mining methods,” Artif. Intell. Med., vol. 34, pp. 113–127, 2005. W. Version, “Machine Learning with WEKA,” 2004.
19. “UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set.” [Online].
20. “SUGI 31 Statistics and Data Analysis Receiver Operating Characteristic (ROC) Curves MithatGönen , Memorial Sloan-Kettering Cancer Center SUGI 31 Statistics and Data Analysis FN + FP,”, 2001.

Kriti Jain-Currently she is pursuing bachelors in Computer Science and Engineering from Maharaja Agrasen Institute of Technology, New Delhi, India. Her current research interest includes machine learning and data mining.

Megha Saxena-Currently she is pursuing bachelors in Computer Science and Engineering from Maharaja Agrasen Institute of Technology, New Delhi, India.Her current research interest includes machine learning and data mining.

Shweta Sharma
Currently she is pursuing bachelors in Computer Science and Engineering from Maharaja Agrasen Institute of Technology, New Delhi, India.Her current research interest includes machine learning and data mining.