International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 3, Issue 3, pp 14-17, March-2019

Design and Development of Cancer Prediction using Machine Learning Technology System

[1]Ashish Singh, [2]Vishal Tripathi, [3]Ankit Singh, [4]Shiwani Gupta

[1,2,3]Computer Science and Engineering Department, Thakur College of Engineering and Technology, Mumbai, India
[4]Assistant Professor, Computer Science and Engineering Department, Thakur College of Engineering and Technology, Mumbai, India

***Abstract* - Breast cancer is one of the utmost shared disease in women, classifying and predicting it is a vibrant research issue. Various machine learning system have been utilized to create different cancer models. Among various algorithms, Support Vector Machines and k nearest neighbors have been appeared to outnumber other algorithms. Though there are few studies concentrated on examining the performance of different classification algorithms .The motive of this paper is to evaluate the performance of SVM and KNN on breast cancer dataset. The cancer dataset (Wisconsin Dataset) is taken from UCI machine Repository, place for machine learning and insight Framework. The precision, accuracy F-measures of different classification algorithms are looked at. The outcome shows that SVM classifier can give the better result for classification, while accuracy of the algorithm is improved by modifying the attributes of the dataset.**

Keywords: Breast cancer classification, Machine Learning, Support Vector Machine, K nearest neighbors, classification.

I. INTRODUCTION

Breast cancer is considered to be a crucial problem in the healthcare communities. This cancer takes place in the tissue of the women's breast [1]. There are numerous threat involved in breast cancer including cumbersomeness, feeling languid, consuming alcohol, genetic mutations, reproductive history (having periods before age 12 and menopause after age 55), having offspring late or not at all, and getting older. Breast cancer is a result of mutation in a cell, which can be sealed by system or causes uncontrolled cell division. If the problem remains unfixed after months, masses are shaped from cells having incorrect instructions. Malignant tumours enlarge to surrounding cells which can lead to metastasize, while benign masses cannot enlarge to other tissues, so the expansion is limited to benign mass. There are several types of breast cancer hence it would be much beneficial to have a system that would consent early detection of sort of cancer and thus prevention which would supplement the survival rates of the cancer The paper focuses on deliberating statistical and machine learning measures which are used to develop cancer prediction models such as Support vector machine and K nearest neighbor. The dataset used for predicting breast cancer was attained from the UCI machine learning depository [2] and designated by Dr. William H. Wolberg. In some research breast cancer dataset have been used [3]. We have discussed the effect of 31 characteristic parameters of breast cancer and the performance of the SVM and KNN models based on the influence of these parameters. These dataset contains a total of 569 pieces of samples and each sample is articulated by characteristic parameters. The performance of SVM and KNN will be evaluated based on metrics including the classifier training time, F-measure, and the classification accuracy. Hence, the outcomes of this paper should promote researches to choose the one model that provides the optimal prediction performance for future comparison.

II. LITERATURE SURVEY

Various approaches and algorithms have been approved on classification of breast cancer. Some related works have been discussed here. Analysis of Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection‖ by Borges and Lucas Rodrigues [7]. In this two machine learning algorithms called Bayesian Networks algorithm and J48 are studied. Various experiments were conducted using these algorithms and they concluded out that Bayesian network has a much better performance than the J48 algorithm Analysis of knn clustering approach on the breast cancer Wisconsin dataset by Ashutosh Kumar Dubey et.al [6]. The study was meant to find the effects of k-means clustering algorithm with diverse calculation measures like distance, method, centroid, and iteration and to sensibly consider and discovery out the combination of measures that has accurate clustering accurateness. Classification of Cancer Dataset in Data Mining Algorithms Using R Tool by P.Dhivyapriya and Dr.S.Sivakumar [5]. CART and Support Vector Machine classifiers are being used. Accuracy is equated classifying two different cancer datasets. The results are achieved using Naive Bayes and Support Vector Machine after data processing and adjustment of the classifiers. Classification of Breast cancer

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 3, Issue 3, pp 14-17, March-2019

using Back propagation training algorithm by F.Paulin[4]. Feed Forward Artificial is used to classify the breast cancer. Levenberg algorithm gives the utmost accuracy.[4]The Wincosin dataset was being developed by Dr.wolberg in 1995 thereafter a lot of research has been done by various medical and allied practioners just to simplify the breast cancer detection and identification.1999 Xin Yao has developed a neural network using negative correlation. S. Aruna and L. V Nandakishore, compare the performance of Naïve Bayes, Support Vector Machine (SVM) and K- Nearest Neighbor (K-NN) to find the best classifier in WBC. SVM is one of the most accurate classifier with accuracy of 96.99%. Angeline Christobel. Y and Dr. Sivaprakasam, achieve accuracy using decision tree classifier in breast cancer datasets.[1]The literature survey reveals that work is done on diagnosis and detection of breast cancer. The SVM and KNN have found best technique of machine learning for breast cancer detection. Still there is no such hybrid model to detect has been explored.

III. PROPOSED SYSTEM

The procedure is based on the given steps

1. Leading of all the dataset is divided into 60% training and 40% testing created on the strategy 10 cross validation.
2. The further steps is to visualize the data using density plots to get the precise sense of the data.
3. Lastly the testing set is provide for into the constructed classifiers before the examination of its accuracy f-measures and precision.
4. Furthermore the classifier training times are also matched to analyze the complexities of training classifiers.

IV. IMPLEMENTATION

a) K Nearest Neighbour

KNN (K-Nearest Neighbors) is a supervised learning algorithm which is used widely in machine learning and data mining. It is a classification algorithm where the learning depends upon how similar the data is from its neighbors.KNN steps are:

1. Collect an unclassified data.
2. Measure the Euclidean distance between the new data and the already existing data.
3. Define the K smaller distances. Select the class having the shortest distance from the list of all classes and take the count of the number of times each class appears. Now the class that appeared the most take it as a correct class.

Then finally classify the new data to the class that resulted in step

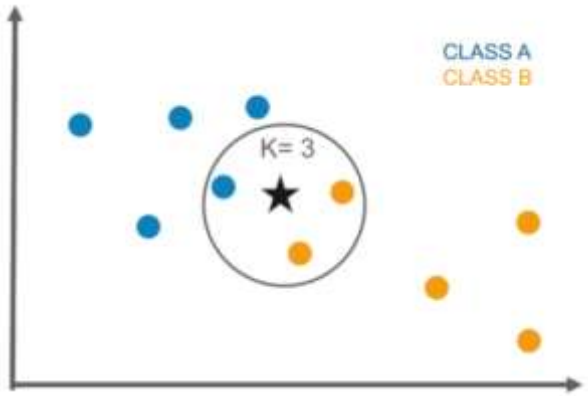

Figure 1: KNN

b) Support Vector Machine

A Support Vector Machine (SVM) is a classifier which is defined formally by separating hyper plane. In simple words it is a supervised learning algorithm with labeled training data which divides a two dimensional plane in two parts.

SVM can solve non-linear and linear problems and work fine for various practical problems. It is a machine learning algorithm that requires data as an input and the resulting output is a line that separates the data into classes. The nonlinear problems are those in which a line cannot divide the data into classes whereas linear problems are those in which a line can divide that data into different classes.

The SVM can be used in many real world applications other than the cancer detection including face detection, text and hypertext categorization, bioinformatics, classification of images, handwriting recognition and many more.

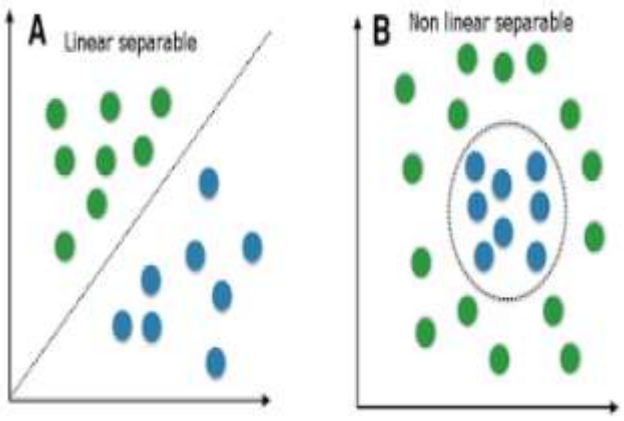

Figure 2: SVM

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 3, Issue 3, pp 14-17, March-2019

V. WORKING OF PROPOSED SYSTEM

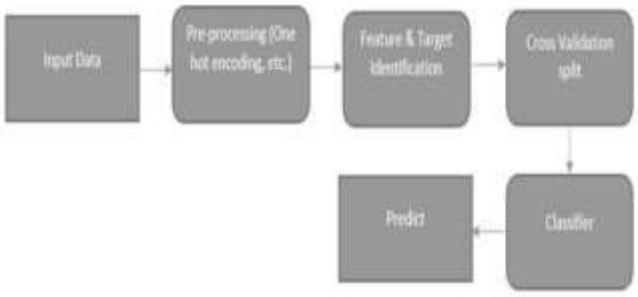

Figure 3: Proposed System

VI. EXPERIMENTAL SETUP

a) The Dataset

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

b) Experimental Result

In this division, the results of the classification are described. 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 are splited the original set into training sample to train the model, and a test set to evaluate it. After applying the pre-processing and preparation methods, we try to analyze the data visually and figure out the distribution of values in terms of effectiveness and efficiency.

c) Effectiveness

In this particular section we will evaluate the effectiveness of all classifiers in terms of period to build the model correctly classified instances and incorrectly classified accuracy and instances In order to improved measure the classification performance of classifiers we will evaluate the accuracy in terms of the following given below measure as:

1. F measure is the measure of a test accuracy and is defined as a weighted harmonic measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test.
2. Precision (1 – Error) = (TP + TN)/(PP + NP) = Pr(C) the probability of the correct classification.
3. Sensitivity events the proportion of the positive that is appropriately identified as the percentage of vile people who are correctly identified as having the condition.
4. Furthermore the classifier training times are also matched to analyses the complexities of training classifiers.

d) Efficiency

When the predictive model is built you can check efficient it is. For this we compare the accuracy measures based on precision, recall, F1score values for SVM and k-NN as shown .To understand efficiency presents the report of our classifiers that better demonstrate the precision of every classifier. It gives a clear understanding of each terms well defined.

	precision	recall	f1-score	support
0	0.95	0.99	0.97	90
1	0.98	0.91	0.94	53
avg / total	0.96	0.96	0.96	143

Figure 4: Classification of KNN

	precision	recall	f1-score	support
0	0.97	0.99	0.98	90
1	0.98	0.94	0.96	53
avg / total	0.97	0.97	0.97	143

Figure 5: Classification of SVM

VII. CONCLUSION

To examine medical data various machine learning and data mining methods are available. The major task in machine learning is to build computationally and accurate classifiers for medical use. In this study we engaged two algorithms SVM and KNN on Wisconsin dataset. We tried compare effectiveness and efficiency of these algorithms in standings of Accuracy, recall, F-measure, precision to find the finest classification of which SVM reaches of 97% and outperforms all the other algorithms. In conclusion SVM algorithm has its efficiency and accuracy in this dataset and has observed optimum in terms of low error rate and precision.

REFERENCES

[1] USA Cancer Statistics Working Group. THE United States Cancer Statistics in 1999–2008 Incidence and Mortality Web-base.

[2] V. Chaurasia and S. Pal Data Minings : To Predict and to Resolve Breast Cancer Survivability vol. 3,2014.

ISSN (online): 2581-3048
Volume 3, Issue 3, pp 14-17, March-2019

[3] A.C.Y, "An Empirical Comparison of Data Mining Classification Methods".
[4] F.Paulin et al. Jan 2011., Classification of Breast cancer by comparing Back propagation training algorithms‖, *International Journal of Computer Sciences and Engineering*, Vol 3, 327 – 332,
[5] P.Dhivyapriya and Dr.S.Sivakumar, Jan-feb 2017Classification of Cancer Dataset in Data Mining Algorithms Using R Tool‖, *International Journal of Computer Science Trends and Technology (IJCST)*, Vol.5, Issue 1.
[6] Dubey, A.K., Gupta, U. & Jain, S, November 2016 Analysis of k-means clustering approach on the breast cancer Wisconsin dataset, *International Journal of Computer Assisted Radiology and Surgery.*
[7] Borges and Lucas Rodrigues, Analysis of Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection‖, *Proceedings Computational*, October, 2015.

AUTHOR'S BIOGRAPHIES

Ashish Singh
Student of computer Engineering,
Thakur college of engineering and technology, Mumbai, India.

Vishal Tripathi
Student of computer Engineering,
Thakur college of engineering and technology, Mumbai, India.

Ankit Singh
Student of computer Engineering,
Thakur college of engineering and technology, Mumbai, India.

Ms. Shiwani Gupta
Assistant Professor at Thakur college of engineering and technology, Mumbai, India.

Citation of this article:

Ashish Singh, Vishal Tripathi, Ankit Singh, Shiwani Gupta, "Design and Development of Cancer Prediction Using Machine Learning Technology System" Published in *International Research Journal of Innovations in Engineering and Technology (IRJIET)*, Volume 3, Issue 3, pp 14-17, March 2019.

******

