Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 9, Issue, 4(I), pp. 26126-26130, April, 2018

ISSN: 0976-3031

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

ANALYSIS OF K- NEAREST NEIGHBOR TECHNIQUE FOR BREAST CANCER DISEASE CLASSIFICATION

Arpita Joshi[1] and Ashish Mehta[2]

[1]Department of Computer Science, Kumaun University, S.S.J.Campus, Almora, Uttarakhand
[2]Department of Computer Science, Kumaun University, D.S.B. Campus, Nainital, Uttarakhand

DOI: http://dx.doi.org/10.24327/ijrsr.2018.0904.1997

ARTICLE INFO

Article History:

Received 5th January, 2018
Received in revised form 20th February, 2018
Accepted 8th March, 2018
Published online 28th April, 2018

Key Words:

Machine Learning, Classification, Breast Cancer Dataset, K- Nearest Neighbor, R Tool.

ABSTRACT

Breast Cancer becomes the life-threatening disease in the female. Breast Cancer can start in breast and spread to different parts of the body. Early detection and diagnosis of Breast Cancer have been pointed at as the most reliable approach to reducing the number of deaths. There are different machine learning techniques available that are widely used in various domains such as classification and prediction process. In the present study, we employed one of the most popular used machine learning technique K-Nearest Neighbor(KNN) for Wisconsin Diagnostic Breast Cancer dataset in R environment. The dataset has been taken from the UCI Machine Learning Repository containing 32 attributes, 569 instances, and 2 classes. We analyzed classification results of KNN with and without Dimensionality Reduction Techniques. We employed two most important Dimensionality Reduction techniques namely Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) respectively. The objective of the present study is to analyze the performance of KNN technique in Wisconsin Diagnostic Breast Cancer Dataset based on the confusion matrix. The results show that to classify benign or a malignant using KNN with Linear Discriminant Analysis technique outperforms 97.06% accuracy as compared to KNN and KNN with PCA.

INTRODUCTION

Breast Cancer is one of the most frequent diagnosis diseases occurs in female and main reasons for death among female. Early detection and diagnosis of Breast Cancer disease pose a great challenge to researchers/ doctors[20]. Breast Cancer occurs when the cell tissues of breast become abnormal and divide uncontrollably. These abnormal cell tissues form large lumps, which consequently become a tumor[15]. It can be detected by clinical breast examination. Breast Cancer has been a large topic in research area for the last few decades. A lot of research work has been done in the medical field to detect cancer. Still, the progress in detection and diagnosis of breast cancer remains very time to consume and expensive[6].

According to the report of cancer statistics 2018, In 2018, estimated 1735350 new cancer cases are expected to be diagnosed and 609640 cancer deaths the United States[11]. According to the report available at www.breastcancerindia.net, 144937 women were newly detected with breast cancer, 70218 women died of breast cancer in 2012 in India[22]. According to the report of National Cancer Institute available at https://seer.cancer.gov,estimated 266,120 new breast cancer(female) cases are expected to be diagnosed and estimated 40,920 breast cancer deaths in 2018 in the United States[10].

There are different machine learning techniques available for classification purpose that can differentiate breast cancer into benign (non-cancerous) or a malignant (cancerous). Nowadays Machine Learning has become one of the most promising technology. This exciting technology opens the ways to new possibilities which helps the researchers /doctors working towards detection of diseases and therefore, prevent the spread of disease into the more severe stage. Machine Learning is the most prominent sub-field of Artificial Intelligence that involved self-learning techniques that derived knowledge from data in order to make predictions. Machine Learning provides a more efficient way of capturing the knowledge in data to gradually improve the performance of predictive models and make data-driven decisions[19]. However, Machine Learning techniques have been applied in various domains and it has

Corresponding author:* **Arpita Joshi
Department of Computer Science, Kumaun University, S.S.J.Campus, Almora, Uttarakhand

been proved that their practice is unavoidable in various applications [17].

In the present study, using KNN, an attempt is made towards correctly predicting the class of breast cancer (benign or malignant). We know that KNN stores all the cases and finds similarity between cases. There are many tools available for machine learning techniques execution such as R tool. We have used R tool for implementation .R is an open source tool which is one of the most powerful tools for statistical programming and machine learning.

Related Work

Breast Cancer diagnosis is one of the most serious problems in the medical field**.** Various researches have been carried out by researchers in order to improve the performance and obtain satisfactory outcomes. Machine learning is particularly used in Breast cancer diagnosis. KNN is one of the most popular used technique for cancer classification. Palaniammal V. and Chandrasekaran (2014) experimented on Wisconsin breast cancer data (200 rows and 11 columns) and applied fourfold Cross Validation method for testing in which each fold contains 50 instances. They reached 97% accuracy using KNN technique[18]. Medjahed Ahmed Seyyid, Saadi Ait Tamazouzt and Benyettou Abdelkader (2013) used KNN with different types of distances and classification rules in function of the parameter K. They experimented on Wisconsin Breast Cancer Data (WBCD) obtained by UCI machine repository.They reached 98.70% accuracy for Euclidean distance and 98.48% accuracy for Manhattan distance[14]. Fradet Ben (2014) experimented with Wisconsin Diagnostic Breast Cancer data. He reached 96% of right prediction using K-nearest neighbors (KNN) technique[5]. Damini Priya S., Agilandeeswari L and Prabukumar M(2017) proved that the KNN technique produces less misclassification comparing to Random Forest as well as Decision Tree. They concluded that KNN gives 96% classification accuracy[4]. Ibrahim M.Aidarus, Baharudin Baharum, Said MD Abas and Hashimah N.P.(2016) experimented on the Digital Database for Screening Mammography(DDSM).They obtained the classification results using KNN and SVM(Support Vector Machine) for k-means-max pooling and Bag-of-features.KNN technique performed better than SVM, with a high accuracy of 98.19%[13].

Machine Learning Classification Technique

Classification is one of the most important steps. The main task of classification is to predict a class from some set of inputs**[16].**We give a brief description of the Machine Learning classification technique(KNN) that we employed in the present work as following.

K-Nearest Neighbor (KNN)

The k-Nearest Neighbor technique is a non-parametric method for classification. The neighbors are taken from a set of objects for which the class for classification is known[2,8]. KNN simply stores the training data.It is a simple technique that is capable of handling extremely complex tasks such as identifying cancerous masses**.** In KNN technique, There is no assumption about the statistical data structure [1,12]. We can train K-Nearest Neighbor technique for classification with the help of R Tool. The open source tool R is used to analyze data and it is also used to build the statistical model. R contains a large number of packages that are suitable for various applications. To get started, we could first load packages. Then read in the training data downloaded from UCI machine repository. Now build and train K Nearest Neighbor technique through *CARET*(Classification and Regression Training) package that contains some main features for validation and performance assessment. We used the set.seed () function and specify a unique seed so that the outcomes are reproducible. We used train() function. The train () function first takes the feature or predictor data and then final results variable. The train function can work with a variety of models determined through method argument. The train function provides a smooth way to try a number of tuning parameters as named data frame to the tuneGrid argument The argument trcontrol is used to evaluate different tuning parameters to tell the caret package how to validate and select the best tuning parameters. The predict () function generates a set of predictions of data. After calculating and storing the predicted result, examine their distribution. A more formal evaluation of model overall performance is viable the usage of confusion matrix () function in the caret package**[21].**

Proposed System Architecture

The main objective of the present study is to give an excellent result of Breast cancer disease classification using KNN technique in various aspects. The proposed system consists of several stages:

Input Data

We have taken WDBC breast cancer dataset obtained from UCI machine learning repository used for the classification purpose (benign or Malignant).The dataset contains 569 instances, 32 attributes and 2 classes[**3**]**.**A brief description of the dataset is presented in table 1 below:

Table 1 Breast Cancer Data Set

Dataset	Attributes	Instances	Classes
Breast Cancer	32	569	2

Preprocessing of Data

Data preprocessing is the first and thus very important step. Most computational tools are unable to deal with missing values. To overcome this problem we simply removed the corresponding columns (features) or rows (samples) from the dataset that contains the missing value.

Dimensionality Reduction

Dimensionality Reduction is one of the most important steps in preprocessing. One main and common application of unsupervised learning is Dimensionality Reduction. We are working with high dimensionality data that can pose a great challenge for limited storage space and the computational performance of machine learning techniques. Dimensionality Reduction is the most common technique in preprocessing, which compress the data onto the smaller dimensional subspace. In the present study, we have employed two most important dimensionality reduction techniques i.e., Principal Component Analysis(PCA) and Linear Discriminant Analysis(LDA). Principal Component Analysis is unsupervised linear transformation technique, mostly used across different areas such as feature extraction and dimensionality reduction.

Principal Component Analysis uses an orthogonal transformation to go from the raw data to the Principal Components. In addition to being uncorrelated, the principal components are ordered from the components that explain the most variance to that which explains the least [Linear Discriminant Analysis is a supervised dimensionality reduction technique that can be used to reduce the number of dimensions in a dataset. Linear Discriminant Analysis is used to find the feature subspace that optimizes class separability [19, 21].

Normalize the data, if necessary

To avoid bias in the results, the Normalization process is used. It is the process of rescaling of the columns to a range of [0, 1][19].

Divide the dataset into training and testing part

The dataset is divided into two parts

1. training set (The known data are given to the technique for training) and
2. testing part(unknown data are given to the technique for testing).

Train the proposed system

Plot the proposed system

We employed ROC (Receiver Operating Characteristic) graph to select the proposed system for classification based on the performance. Based on ROC curve, ROC Area Under the Curve(ROC AUC) is computed to characterize the performance of classification model**[19].**

Compute the various performance statistical measures from confusion Matrix

The performance of the proposed system is being evaluated using Confusion Matrix. The confusion matrix is a square matrix that contains four outcomes (TP, TN, FP, FN) produced by any machine learning technique. For given input, machine learning technique produces the result with two class values such as negative or positive. Confusion matrix also predicted as error matrix lay out the performance of the classification technique. In confusion matrix, each row and column contains instances of the matrix. In other words, we can say that confusion matrix is a technique that summarizes the prediction results of classification technique. The structure of confusion matrix is presented in table 2 given below:

Table 2 Format of Confusion Matrix

Predictions	**Outcomes**	
	True Positives(TP)	False Negatives(FN)
	False Positives(FP)	True Negatives(TN)

The brief description of TP, TN, FP, and FN is given below

True Negative (TN): No Possibility of disease, Prediction is false.

True Positive (TP): Possibility of disease, Prediction is true.

False Positive (Type 1 error): They do not have the disease but the prediction is true.

False Negative (Type 2 error): They have the disease, Prediction is true.

There are various performance measures available that can be derived from the confusion matrix to evaluate the performance of the proposed system. Sensitivity also called recall or True Positive Rate. It provides useful information about the fraction of positive instances that are correctly identified out of the total number of positives. Specificity also called as True Negative rate provides useful information about the fraction of negative instances that are correctly identified out of the total number of negatives. Accuracy refers to the proportions of instances that are correctly classified. Detection Rate is the percentage of True Positives. The Detection Prevalence refers to the proportion of instances that can be predicted as positive, whether they actually are or not. Balanced Accuracy can be obtained by calculating the mean of the sensitivity and specificity. Positive Predictive Value also called as Precision provides useful information about the fraction of relevant instances among the retrieved instances [7,9, 21].

By applying the above-mentioned stages using R tool for Breast cancer dataset the machine learning classification technique (KNN) is employed in the present work.

EXPERIMENTAL RESULTS

The classification process is divided into training (known data are given to the technique for training) and testing part (unknown data are given to the technique). Table 3 depicts Confusion matrix obtained from KNN technique using R tool. The experiment was conducted on Wisconsin (Diagnostic) Breast cancer dataset using K-Nearest Neighbor technique Table 4 shows the performance of Wisconsin (Diagnostic) Breast Cancer dataset using KNN, KNN with PCA and KNN with LDA. Figure 1 depicts the comparative analysis of various results obtained from confusion matrix using KNN technique in terms of different parameters, Fig 2,3 and 4 showcases the performance of KNN using ROC graph. The classification result shows that KNN technique with Linear Discriminant Analysis technique (one of the most popular used dimensionality reduction technique) outperforms 97.06% accuracy as compared to KNN(95.29%) and KNN with Principal Component Analysis(95.88%).

Table 3 Confusion Matrix

KNN

	Test outcomes	
Predictions	Benign	Malignant
Benign	105(TP)	6(FN)
Malignant	2(FP)	57(TN)

KNN(LDA)

	Test outcomes	
Predictions	Benign	Malignant
Benign	107(TP)	5(FN)
Malignant	0(FP)	58(TN)

KNN(PCA)

	Test outcomes	
Predictions	Benign	Malignant
Benign	106(TP)	6(FN)
Malignant	1(FP)	57(TN)

Table 4 Performance of dataset

	KNN	KNN(PCA)	KNN(LDA)
Accuracy(%)	95.29	95.88	97.06
Sensitivity(%)	90.48	90.48	92.06
Specificity(%)	98.13	99.07	100
PPV or Precision (%)	96.61	98.28	100
Negative Predicted Value(%)	94.59	94.64	95.54
Prevalence(%)	37.06	37.06	37.06
Detection Rate(%)	33.53	33.53	34.12
Detection Prevalence(%)	34.71	34.12	34.12
Balanced Accuracy	94.30	94.77	96.03

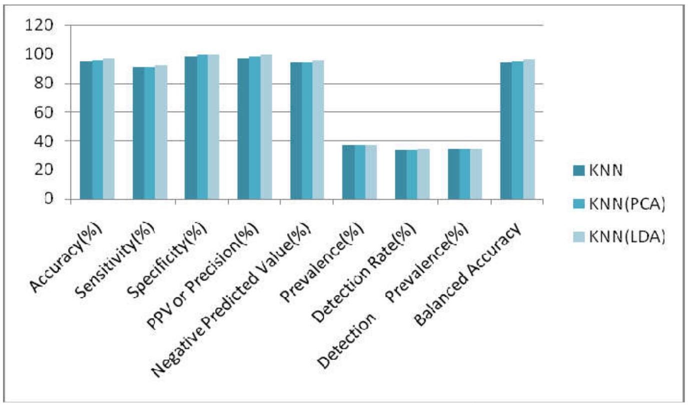

Figure 1

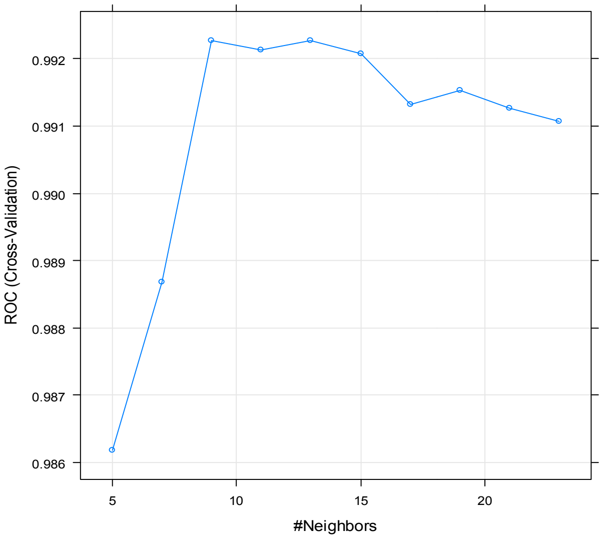

Figure 2 KNN

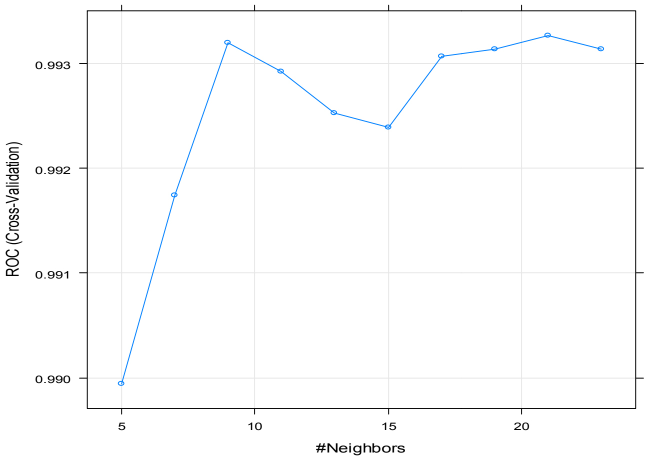

Figure 3 KNN (LDA)

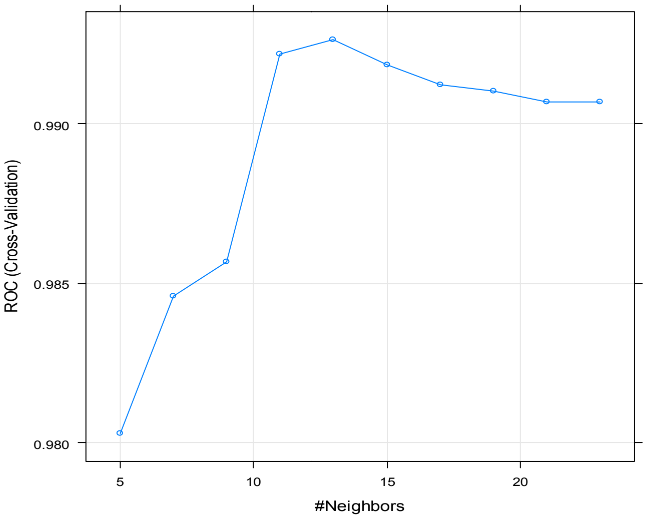

Figure 4 KNN (PCA)

CONCLUSION

Breast cancer is one of the main reason for death in the female. So early detection and diagnosis of breast cancer are very important in reducing life losses. In the present study, the employment of KNN technique on Wisconsin (Diagnostic)Breast Cancer dataset show that KNN technique with Linear Discriminant Analysis technique(Dimensionality Reduction Technique gives better accuracy (97.06%) as compared to KNN without dimensionality reduction technique(95.29%)and KNN with PCA technique(95.88%).

References

1. Abinaya A., Abirami S., Nasrin Faridha N. and Kalaiyarasi R.(2017),"Detection and Classification of Brain Tumor Using SVM And K-NN Based Clustering", SSRG *International Journal of Electronics and Communication Engineering*, Page(s)-70-74.
2. Altman N.S.(1992),"An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression, Volume 46, Issue 3: Page(s)-175-185.
3. Breast Cancer Wisconsin Data[online], http://archive.ics.uci.edu/ml/datsets/Breast + Cancer+ Wisconsin+ (Diagnostic).
4. Damini Priya S., Agilandeeswari L and Prabukumar M(2017),"A Case Study on Effective Approach to Predict the Class of Breast Cancer on Numerical Dataset", *International Journal of Pure and Applied Mathematics,* Volume-117,Page(s)-161-167.
5. Fradet Ben (2014),"Using k-Nearest Neighbors to Diagnose Cancer",benfradet.github.io.
6. Giri Prannoy and Saravanakumar,"Breast Cancer Detection using Image Processing Techniques", Oriental *Journal of computer science and technology*, Volume 10, Issue 2, ISSN:l0974-6471.
7. https://classeval.wordpress.com,"Basic Evaluation measures from the confusion matrix".
8. https://en.m.wikipedia.org," k-nearest neighbors"
9. https://en.m.wikipedia.org/wiki/Precision and recall
10. https://seer.cancer.gov,Cancer Stat Facts: Female Breast Cancer.

11. https://www.cancer.org,Cancer Facts and Figures 2018(cancer statistics 2018, American Cancer Society Journal, CA: A Cancer *Journal for Clinicians*.
12. https://www.packtpub.com,"Summary-Machine Learning".
13. Ibrahim M.Aidarus, Baharudin Baharum, Said Md Abas and Hashimah N.P.(2016),"Classification of Breast Tumor in Mammogram Image Using Unsupervised Feature Learning", *American Journal of Applied Sciences,* Volume 13, Issue 5,Page(s)-552-561.
14. Medjahed Ahmed Seyyid, Saadi Ait Tamazouzt and Benyettou Abdelkader (2013),"Breast Cancer Diagnosis by using K-Nearest Neighbor with Different Distances and Classification Rules", *International Journal of Computer Applications,* Volume 62-No.1
15. Mehdy M.M., Ng Y.P., and Gomes C., "Artificial Neural Networks In Image Processing for Early Detection of Breast Cancer", *Computational and Mathematical Methods in Medicine*.
16. Nithya B. and Ilango V (2017),"Comparative Analysis of Classification Methods in R Environment with two Different Data Sets", *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*. Volume 2,Issue 6,2(6),Page(s)-136-141.
17. Nithya B.(2016),"An Analysis on Applications of Machine Learning Tools, Techniques an Practices in Health Care System", *International Journal of Advanced Research in Computer Science and Software Engineering,* Volume 6, Issue 6,Page(s)-1-8.
18. Palaniammal V.and Chandrasekaran M.R.(2014),"Analysis for breast cancer diagnosis using KNN classification", *International Journal of Applied Engineering Research,* Volume 9 Issue 22,Page(s)-14233-14241.
19. Raschka Sebastian and Mirjalili (2017),"Python Machine Learning",2nd edition, Packt Publishing Ltd., ISBN:978-1-78712-593-3.
20. Swagatika Sushree and Barik Ranjan Smruti (2013),"Breast Cancer Diagnosis and Prognosis in India: A Comparative Study Based on Data Mining Techniques", National Conference on Recent Advances on Business Intelligence and Data Mining.
21. Wiley F. Joshua(2016),"R Deep Learning Essentials", Packt Publishing Ltd, ISBN:978-1-78528-058-0.
22. www.breastcancerindia.net, Statistics of Breast Cancer in India.

How to cite this article:

Arpita Joshi and Ashish Mehta.2018, Analysis of K- Nearest Neighbor Technique for Breast Cancer Disease Classification. *Int J Recent Sci Res.* 9(4), pp. 26126-26130. DOI: http://dx.doi.org/10.24327/ijrsr.2018.0904.1997

****