European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 07, Issue 03, 2020

A Study Of Breast Cancer Analysis Using K-Nearest Neighbor With Different Distance Measures And Classification Rules Using Machine Learning.

M.D.Bakthavachalam[1], Dr.S .Albert Antony Raj[2]

[1]*Research Scholar ,Department of Computer Science, College of science and humanities, Associate Professor and Head*
[2]*Research supervisor, Department of Computer Application, College of Science and Humanities, SRMIST, Kattankulathur, 603203*

Abstract
Breast Cancer is one of the life threatening disease among females all over the world. This killer disease however when it can be detected in its early stages can be a life saver for many. Radiologists uses the mammography images to detect the presence and absence of Breast Cancer. The field of Bio-informatics leverages the Machine learning techniques for diagnosis of Breast cancer in particular. This research work experiments with the two most popularly used Supervised Machine Learning Algorithms, K-Nearest Neighbour and Naive Bayes. This work predicts Breast Cancer on the The Breast Cancer Data Set (BCD) taken from the UCI Machine Learning Repository. A comparative analysis between the two approaches are made in terms of its performance metrics using CV techniques. The proposed work has achieved a best accuracy of 97.15% by employing the KNN algorithm and a lowest error rate of 96.19% using NB classifier.

Keywords:
Image Processing, Mammography, Machine Learning, Naive Bayes , K-nearest neighbor

1. INTRODUCTION

The life threatening disease that stands next to lung cancer is Breast Cancer. Premature detection of this type of Cancer can save the lives of people. Tumor that signifies the abnormal growth of cells can be treated at the early stage by preventing it from expanded. On a very basic level, a self-test can be done by every women to find the presence of any lumps or abnormal growing cells. The most reliable and popular method of screening breast cancer is to do a mammography screening. Mammography is "a radiology process that examines the human breast by taking X-rays and it can be utilized as a screening tool for diagnosing cancer" [4]. For quite a long time, the X-ray was the only method that was used to detect the breast cancer [1, 2]. Unfortunately, early detection of cancer is not that very

easy as no symptoms or signs are shown at the incipient stage. This challenging nature of the disease enlists cancer detection as one of the most sought research topic in health care sector.

Multitude of researchers have investigated in this health care domain by collecting data through surveys and studies to uncover hidden patterns and associations in the large volume of data. With the rapid growth in the field of Information Technology, decision making tools can assist the clinicians to take better informed decisions for the well-being of the patients. Precisely, Predictive Analytics built softwares are needed to assist the oncologists to detect cancer, choose the best treatment path, prevention of recurrence of the disease and to provide effective medical treatment for the patients.

With the increasing momentum of Machine Learning and Artificial Intelligence, more efficient methods are proposed for the detection of Breast Cancer. Several Clinical parameters like patient age, histopathological variable are crucial for the machine learning algorithms to work. The causes of Breast Cancer are multivariate and involve history of family, hormonal issues, weight gain and obesity and even reproductive factors. Malignant tumors expands to other tissues whereas the benign tumors cannot expand to other tissues.[1][2]. Even though detection of BC may be hard at the beginning of the disease, due to the absence of symptoms, classifying the tumor as Malignant and Benign tumors is the need of the oncologists. [1][2]. A best predictive model should yield low false positive (FP) rate and false negative (FN) rate[3].

2. THE DATASET

The Breast Cancer Dataset (BCD) that we used is taken from the public repository for Machine Learning repository from UCI. There are 11 attributes and the first attribute ID is removed as it is not an important feature. The major criterion that determines whether a tumor is benign or malign are listed below. The last feature is the class label that contains a binary classification value (2-benign tumor, 4-malign tumor). The dataset contains 699 instances and the BCD data contains missing values for 16 samples. Missing values were ignored and Ultimately, the dataset contains 683 instances.

S.No	**Features**
1	radius (mean of distances from center to points on the perimeter)
2	texture (standard deviation of gray-scale values)
3	Perimeter
4	Area
5	Smoothness
6	Compactness
7	Concavity
8	Concave Points
9	Symmetry
10	Fractal Dimension

3. MACHINE LEARNING APPROACHES

Machine learning is a subfield of Artificial Intelligence. Machine Learning methods employ mathematics, statistics, probabilities, conditional logics, Boolean logic, and many optimization strategies to bring insights from the data and uncover the hidden patterns [8]. ML approaches can be classified as Supervised, unsupervised and semi-supervised learning depending on the problem and dataset in hand [9]. This section, presents two supervised learning classifiers.

1) Naïve Bayesian Classifier (NBC)

Naive Bayes is a probabilistic machine learning algorithm that can be used to model decisions. In Naive Bayes, the features are conditionally independent and it is purely based o conditional probabilities. The conditional probability is the measure of the probability of an event given another even has occured. The Bayesian Classifiers are well suited for compound databases .[10] The Bayesian classifiers use Bayes theorem, given by the equation (1)

$$p(h \mid d) = \frac{p(d \mid h)p(h)}{p(d)} \qquad (1)$$

In Eq. 1, P(h) -priori probability
P(d) is the prior probability of the training data.
p (d | h)- Conditional probability
P(h | d) - Conditional probability of h, provided d
P (h | d) -probability of d given h
Using the above said equation, we can determine whether xi belongs to Si,where Si is the class variable [11].

2) Logistic Regression (LR)

Linear Regression is a statistical model used popularly in medical field and it is simplest and most properly used machine learning algorithm, Logistic Regression requires a hypothesis and cost function to be defined. The below equations represents the hypothesis and the cost function.

$$h_{\emptyset}(x) = \frac{1}{1+e^{-\emptyset T_x}} \qquad (2)$$

$$J(\emptyset) = -\frac{1}{m}[\sum_{i=1}^{m} y^i log h_{\emptyset}(x^i) + (1-y^i) log(1-h_{\emptyset}(x^i))] \qquad (3)$$

Where, X- input features
Y-Class Varible
J(Θ) - Cost Function
Repeating the equation (4) several times, the cost function is optimized.

$$\emptyset_j = \emptyset_j + \alpha \frac{d}{d\emptyset_j} J(\emptyset) \qquad (4)$$

3. K-NEAREST NEIGHBORS METHOD

The *kNN* algorithm is one of the most simplest and popularly used supervised machine learning algorithm [19, 15, 6]. It is a parameterized learning method based on samples. The training samples, are associated with a distance function like Euclidean distance or Manhattan distance and the majority class of the k-nearest data points to the data point in question is used to build the model. Its *k*-nearest neighbors are considered, and so choosing the value of k is crucial.

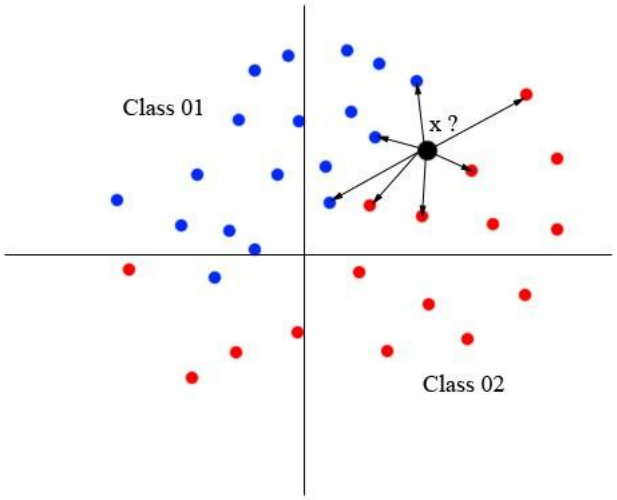

Fig. 1. The K-nearest neighbors method**.**

4. EXPERIMENTAL SETUP

This work is experimented on WBCD (Wisconsin breast cancer database) [10, 3] obtained by the University of Wisconsin. The database comprises of breast cancer data taken from the human breast tissue. There are totally 699 cases, of which 458 cases are benign tumors and 241 are malignant cases.In the WBCD dataset, there are around 16 instances that are having missing values. These records are deleted and we are left with 683 clinical cases.We have applied the k-Nearest Neighbor algorithm investigating the several variants of distance metrics, different K values and different classification rules . This algorithm does not require learning phase. In order to evaluate the performance of the method, the dataset is partitioned into Training Set and Testing Set. Training set with class labels are used to train the model. To further build the model effectively, cross validation method was incorporated with

HoldOut Cross Validation Method. Accordingly, 455 (65, 10%) clinical cases for the training phase and 244 (34, 90%) for the testing phase were experimented. The performance of the algorithm for Breast Cancer Classification was tested for different distance metrics and classification rules. The Classification rules used are enlisted:

The Nearest Rule suggests that a new data point is assigned to the majority classes among the nearest neighbours based on the value of k.

The **random rule** recommends that the new element will be assigned to the majority class among the nearest neighbors and it is done randomly.

For the **consensus rule**, the new query datapoint will only be impacted if all classes are the class for the neighbors and a consensus must be arrived.

5. THE PROPOSED ALGORITHMS

A. Steps in Nearest Neighbors Algorithm with k=3

1- Load the dataset and partition that into Train and Test Split

2- Calculate the distance between various samples using various distance metrics like Euclidean Distance, Manhattan Distance, City Block and Cosine distance. Sort them in ascending order.

3- The class label for the query datapoint in question is the most frequent class f the first k(k=3) training samples.

B. Steps in Naive Bayes classifier

Algorithm Steps:

- Load the dataset
- Divide the dataset separating the Classes (2-benign, 4-malignant)
- Summarize the dataset by calculating the measures of central tendency and dispersion , namely mean and standard deviation.
- Generate the summary of the data based on the Class.
- Calculate the guassian probability density function
- Multiply the individual probabilities of all features.
- Ascertain the target prediction of the query input based on the probability value.

Description:

We start by dividing the dataset into validation and training sets. The training phase consists of separating it into binary classification as presence and absence of tumor. The Class outcome 4 represents malignant tumor and 2 represents benign tumor. The descriptive statistics of the features like mean, standard deviation for each feature from set T and then for each class from set D is calculated. The summary information of each feature is taken into consideration and class that we will use for our prediction is given by the equation (see equation 5).

$$P(T \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(T-\mu)^2}{2\sigma^2}} \quad (5)$$

Our final prediction is done by multiplication the probability of each feature given a class. The probability of each feature is calculated using the density of normal distribution (6).

$$P(D) = \prod P(T \mid \mu, \sigma^2) \quad (6)$$

σ : Mean of Features
μ : Measure of Variance
P (D): Probability of Tumor Class
P (T): Probability of Features

6. RESULTS AND DISCUSSIONS

The classification accuracy of each of the distance measures and by varying the k values are investigated and plotted for visual interpretation.. The high classification accuracy rate achieved is 98.70% for the Euclidean Distance Metrics with k value of 1. When tested with Manhattan Distance with the k value of 1, the accuracy obtained was 98.48%. Apparently, the best accuracy yielded was from K-NN Algorithm with k=1 and distance measure of Euclidean.

Evidently, it is observed that as we keep increasing the value of k, the accuracy decreases and it steadies to a value closse to 50 with the accuracy hovering around 94.40%. The best outcome is yielded only when the distance metric of Euclidean is used. The classification time taken by the algorithms are also recorded for different distance metrics as well. The City block distance and Euclidean distance measure takes lot of computation time. The results are presented in figure 3.

Figure 4 demonstrated the classification Accuracy rate in function k value, being the number of nearest neighbors, using the random rule of classification.

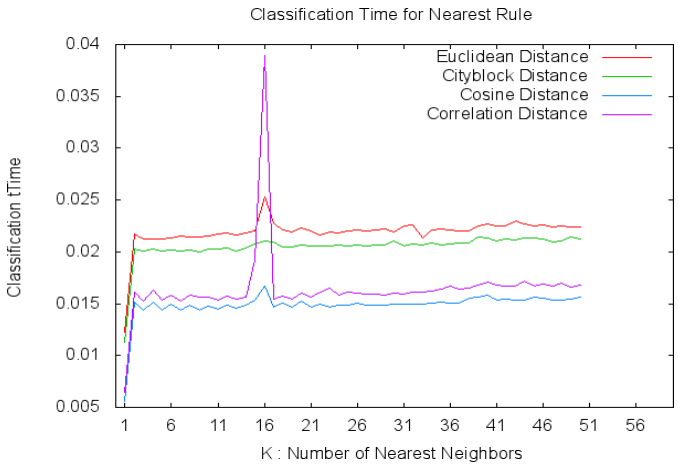

Fig2: Graphical Plot of classification time Vs K value, Nearest Rule based on random rule

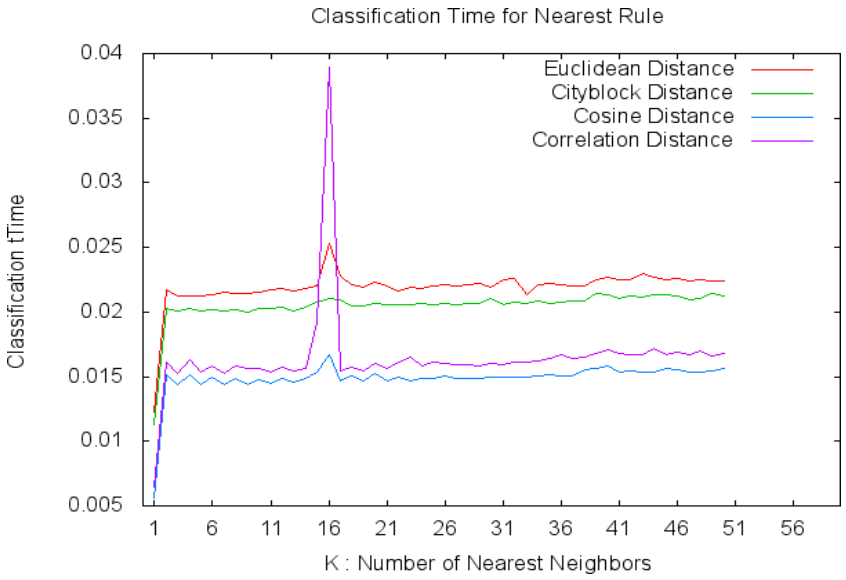

Fig 3 : Graphical Plot of classification time Vs K value, Random Rule

Fig. 5 depicts the cost function with variant values of alpha during the training process.

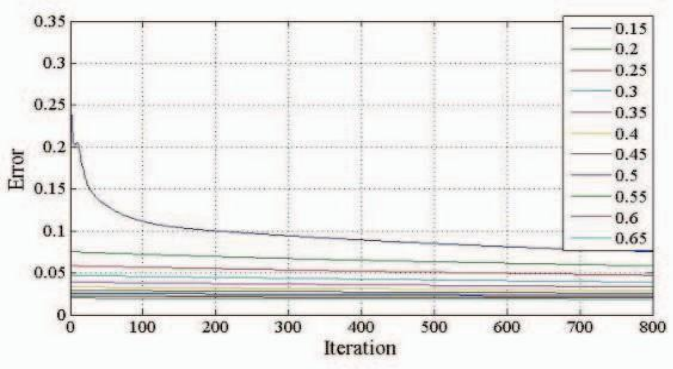

Fig. Error values of the cost function

From Table 1 one can observe that the two approaches are impressive and powerful in the diagnosis of tumor in breast, as they both display a high accuracy rate inspite of the dataset size being small.

Table 1: Comparison between KNN and NB

Approach	Acc Rate	Training Phase	Test Phase	Time Duration
KNN	0.975106	0.000730	0.001741	0.002472
NB	0.961930	0.000757	0.000420	0.001181

The above comparative analysis demonstrates that KNN classifiers are topping first in terms of accuracy and time duration. As a result, KNN is the most reliable and promising classifier for this cancer classification problem. On the flip side if voluminous dataset is used, the amount of time taken for computation will be more and it is time consuming and cumbersome.

7. CONCLUSION

In this paper, a very dreadful disease that reasons to be the fatality among many women all over the world was focused and we proposed a method for the prediction of Breast Cancer. On the WBC datasets, we implemented well-known algorithms such as Naive Bayes and K-Nearest Neighbour for the purpose of Classifying the Breast Cancer. The implementation was investigated on different distance measures and classification rules and the metrics are recorded. Evidently, the best results were obtained as 98.70% employing Euclidean distance and 98, 48% using Manhattan with k = 1. An accurate comparison between the algorithms, it is observed that KNN yields a best accuracy outcome of 97.51%, even though NB has a good accuracy at 96.19 %, But for larger datasets, the running time of the KNN Algorithm can be high. Prediction of malignant tumor in the very early stages can be done directly on the mamography images with high degree of accuracy using deep learning techniques.

REFERENCES

[1]. M. Brown, F. Houn, E. Sickles, and Kessler. “Screening mammograph I Community practice”. Amer.J.Roentgen, vol. 165, 1995.

[2]. M. Alhadidi, M. Al-Gawagzeh , B. Alsaaidah, “Solving A Mammography Problems of Breast Cancer Detection Using Artificial Neural Networks And Image Processing techniques”, Indian Journal of Science and Technology, Vol.5, No.4, 2012.

[3]. E. D. Ubeyli,”Implementing automated diagnostic systems for breast cancer detection”, Elsevier, Expert systems with applications, vol.33, (2007).

[4]. D. Kulkarni,S. Bhagyashree, G. Udupi, ”Texture analysis of mammographic images”,International Journal of Computer Applications, vol.5 , (2010).

[5]. T. Acharya, A. Ray,”Image processing: principles and applications”,Wiley-Interscience,Hoboken new jersey, ISBN 0471719986,(2005).

[6]. A. Hopgood, Intelligent systems for engineers and scientists, Library of Congress Cataloging in Publication Data, (2000).

[7]. H. Zhang, T. Arslan, B. Flynn,”A Single Antenna Based microwave System for Breast Cancer Detection: Experimental Results”, IEEE, (2013).

[8]. M. Brown, F. Houn, E. Sickles, and L. K. P. Bennett and O. L. Mangasarian. Robust linear pro- gramming discrimination of two linearly inseparable sets. *Optimization Methods and Software 1*, 1992.

[9]. E. D. beyli. Implementing automated diagnostic systems for breast cancer detection. *Expert Systems with Applica- tions*, 4(33), 2007.

[10]. D. Bremner, E. Demaine, J. Erickson, J. Iacono, S. Langer- man, P. M., and Godfried. Output-sensitive algorithms for computing nearest-neighbour decision boundaries. *Dis- crete and Computational Geometry*, 33(4), 2005.

[11]. S.Chakraborty, "Bayesian kernel probit model for microarray based cancer classification", Computational Statistics and Data Analysis, Vol. 12, pp. 4198–4209, 2009.

[12]. I. Guyon, J. Weston, S. Barnhill, V. Vapnik. "Gene selection for cancer classification

using support vector machines". Machine Learning, Vol. 46, pp. 389–422, 2002.

[13]. S. Gokhale., "Ultrasound characterization of breast masses", The Indian journal of radiology &imaging, Vol. 19, pp. 242-249, 2009.

[14]. T. Jinshan, R.R., X. Jun, I. El Naqa, Y. Yongyi, "Computer-Aided Detection and Diagnosis of Breast Cancer With Mammography: Recent Advances", Information Technology in Biomedicine. IEEE, Vol. 13, pp. 236-251, 2009.

[15]. A. Jemal, R.S., E. Ward, Y. Hao, J. Xu, T. Murray, M.J.Thun, "Cancer statistics", A Cancer Journal for Clinicians, Vol. 58, pp. 71-96, 2008.

[16]. L. Adi Tarca, V.J.C., X. Chen, R. Romero, S. Drăghici, "Machine Learning and Its Applications to Biology", PLoS Comput Biol,, Vol. 3, pp. 116- 122, 2007.