Classification Based Neural Network Technique For Breast Cancer Diagnosis

Dr. Puneet Misra,

Assistant Professor
Department of Computer Science,
University of Lucknow, Lucknow,India

Abstract:
Cancer diagnosis is one of the most researched and studied problems in the medical field. Several researchers have focused on improving performance and achieve satisfactory results. Breast Cancer is the most often identified cancer among women and one of the silent killer in the world. However, detecting this cancer in its early stages helps in saving lives. As the diagnosis of this disease manually takes long hours and there is less availability of effective diagnostics systems, so there is need to develop the automatic diagnosis system for earlier detection of the cancer. In artificial intelligence, machine learning is a field, which allows machines to evolve through processes, learnt, and trained to act intelligent. Machine Learning is widely used in the field of bioinformatics and cancer diagnosis**.**

Index Terms - Classification, Diagnosis, Breast Cancer, Machine Learning, K-nearest Neighbors

I. INTRODUCTION

With the increase in the number of breast cancer in India, the fear of cancer is on the rise. Detection of cancer at early stage is necessary for a rapid response and better chances of cure. Unfortunately, early detection of cancer is difficult since symptoms are not visible at an early stage. One cannot prevent breast cancer but can increase survival rates by being informed and choosing the right treatment at the right time.
Breast cancer is most common cancer in women in India. It accounts for 25% to 32% of female cancer in all cities across India. According to National Cancer Institute and American Cancer Society in 2018 over 265.000 new cases of breast cancer will be diagnose each year in women and over 2200 in men. The study released said "For women diagnosed during 2010-2014, five-year survival for breast cancer in now 89.5% in Australia, and 90.2% in the USA but international differences remain very wide, with levels as low as 66.1% in India", according to study titled Global Surveillance of trends in Cancer Survival (2000-2014) (CONCORD-3).
For many years, the X-ray was the only method used to detect breast cancer. However, many other methods have been generated and proposed for detecting process that are more efficient than x-ray procedure such as, neural networks, artificial intelligence, and data mining.[3] There is a self-test every woman can do it monthly using her hand to check for any abnormal growing cells, another way is going to a specialist doctor for mammography test. Mammography is "the process of using low-dose X-rays to examine the human breast and is used as a diagnostic as well as a screening tool"

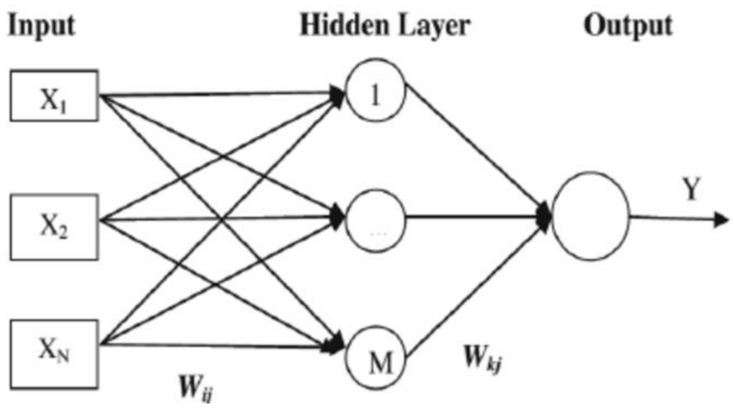


Figure 1: Basic Neural Network Architecture

To develop tools for cancer detection, machine learning and ANN (Artificial Neural Network) methods and medical factors such as patient age and histopathological variables form the basis for daily decision-making are used.
Artificial Neural network is one of the fields of Machine learning that simulates human body neural network that comprises of group of neurons in different layers as shown in Figure 1.These groups of neurons function differently at same time. In neural networks we have a training stage or learning stage in which weights are adjusted to achieve the desired output i.e., machines are trained accordingly to give

required results then in a testing stage these neural networks are tested for their accuracy and efficiency in the detecting process.

Machine Learning, with its advancements in detection of critical features from complex datasets is largely acknowledged as the method in the prediction of breast cancer. Machine Learning techniques can be used for the classification of benign and malignant tumors. This process of classifying benign and malignant tumors can be done by the application of classification techniques of machine learning. Lot of research is done in this area by the application of various machine learning and data mining techniques for many different datasets on Breast Cancer. Most of them show that classification techniques give good accuracy in forecasting the type of tumor.

II. Classification Algorithm

Classification has been an age-old problem. Early in the 4th century BC, Aristotle tried to group organisms into two classes depending on whether they are beneficial or harmful to a human. He also introduced the concept of classifying all forms of life for organizing the rich diversity in living organisms. Classification is defined as the process of finding a set of models (or functions) that describe and distinguish data classes and concepts, with the goal being to use the model to predict the classes of objects whose class labels are unknown. Thus, classification is a supervised learning problem where the task is to predict the value of a discrete output variable given a set of training examples and a test sample where each training example is a pair consisting of the input object and the desired class. Generally, data classification is a two-step process. In the first step, a classifier is built describing a pre-determined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or learning from a training set. In the second step, the model is used for classification.

K-Nearest Neighbor Classification Algorithm

The K-NN is a supervised learning algorithm[1]. The K-Nearest Neighbor is one of the introductory supervised classifier algorithms i.e., supervised classification leaning algorithm. Fix & Hodges proposed the K-nearest neighbor (K-NN) classifier algorithm in the year 1951 for performing pattern classification task. K-NNaddresses pattern recognition problems and the best choices for addressing some of the classification related tasks [2]. The simple version of the K-nearest neighbor classifier algorithms is to predict the target label by finding the nearest neighbor class. The closest class will be identified using distance measures like Euclidean distance. It is a learning method based on instances that does not require a learning phase. The training sample, associated with a distance function and the choice function of the class based on the classes of nearest neighbors is the model developed. Before classifying a new element, we must compare it to other elements using a similarity measure. Its k-nearest neighbors are then considered, the class that appears most among the neighbors is assigned to the element to be classified. The neighbors are weighed by the distance that separates it to the new elements to classify.In the current problem we have two target classes white and orange circles and 26 training samples as shown in Figure 2. The prediction for the target class for blue circle is to be done.

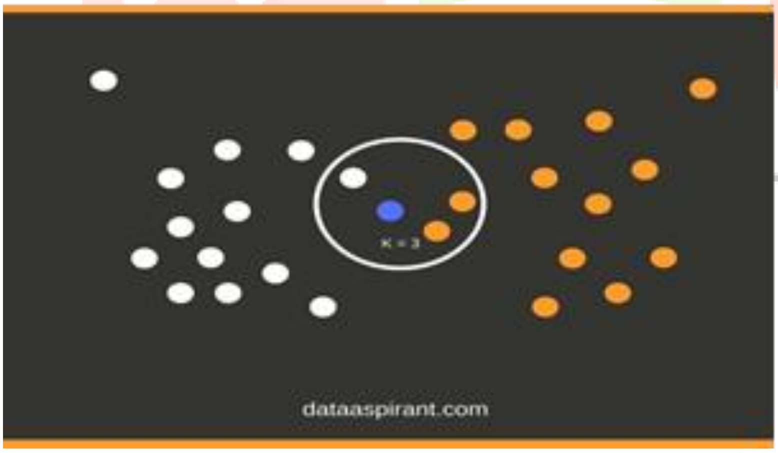


Figure 2: K- Nearest Neighbor Method

The working principle behind KNNpresumes that alike data points lie in same surroundings. It reduces the burden of building a model, adapting several parameters, or building furthermore assumptions. It catches the idea of proximity based on mathematical formula called as Euclidean distance, calculation of distance between two points in a plane. Suppose the two points in a plane are A (x0, y0) and B (x1, y1) then the Euclidean distance between them is calculated as follows.

$$\sqrt{(x0 - x1)^2 + (y0 - y1)^2}$$

An object to be classified is allotted to the respective class, which represents the greater number of its nearest neighbors. If *k* takes the value as 1, then the data point is classified into the category that contains only one nearest neighbor. Given a new input data point, the distances between those points to all the data points in the training dataset are computed. Based on the distances, the training set data points with shorter distances from the test data point are considered as the nearest neighbors of our test data. Finally, the test data point is classified to one of the classes of its nearest neighbor. Thus, the classification of the test data point hinges on the classification of its nearest neighbors. Choosing the value of K is the crucial step in the implementation of the KNN algorithm. The value of K is not fixed and it varies for every dataset, depending on the type of the dataset. If the value of K is less the stability of the prediction is less. In the same manner if we increase its value the ambiguity is reduced, leads to smoother boundaries, and increases stability. In KNN, assigning a new data point to a category entirely depends upon K's value. K represents the number of nearest training data points in the proximity of a given test data point and then the test data point is allotted to the class containing the highest number of nearest neighbors(i.e. class with high frequency).

Algorithm:

a) Calculate "$d(x, x_i)$" i =1, 2, ….., **n**; where **d** denotes the Euclidean distance between the points.
b) Arrange the calculated **n** Euclidean distances in non-decreasing order.
c) Let **k** be a positive integer, take the first **k** distances from this sorted list.
d) Find those **k**-points corresponding to these **k**-distances.
e) Let $\mathbf{k}_i$ denotes the number of points belonging to the i^{th} class among **k** points i.e. $k \geq 0$
f) If $k_i > k_j \; \forall \; i \neq j$, then put x in class i.

III. HOW TO CHOOSE THE VALUSE OF K?

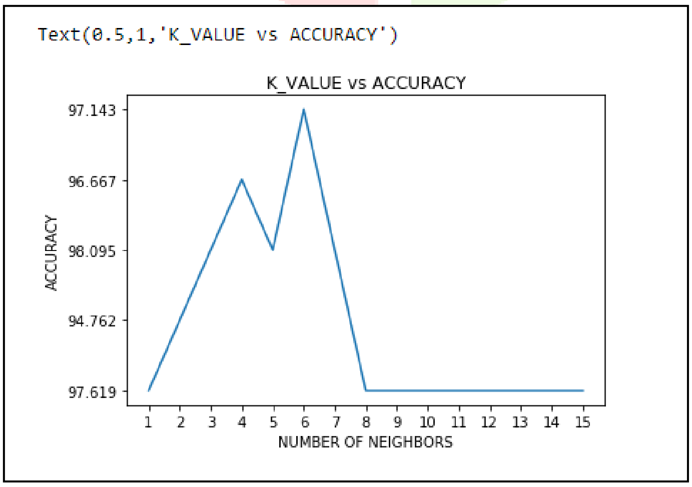
Some value of k means that noise will have a higher influence on the result i.e., probability of overfitting is very high. A large value of k make it computationally expensive and defeats basic idea behind k-NN that points that are near might have similar classes.

To select k, k=n^ (1/2), where n is total training samples. To optimize the results we can use Cross Validation technique (test k-NN algorithm with different values of k).

IV. EXPERIMENTATION

For the classification of benign and malignant tumor, we have used k-Nearest Neighbor classification algorithm of machine learning in which machine is learnt from the past dataset and can predict the category of new input. This experiment is conducted on the Wisconsin Breast Cancer database (WBCD) obtained from UCI repository. Dr. William H. Wolberg (physician), University of Wisconsin Hospitals, USA, collected the Wisconsin Breast Cancer Database. This dataset consists of 10 continuous attributes and 1 target class attributes. Class attribute shows the observation result, whether the patient is suffering from the benign tumor or malignant tumor. Benign tumors do not spread to other parts while the malignant tumor is cancerous. The dataset was collected & openly distributed to find out some patterns from this data [4]. This data set corresponds to 699 clinical cases where 458 are benign cases and 241 are malignant cases.

The evaluation of performance of learning methods required the database to be divided into two parts: the training dataset that represents the initial base for which the class of different clinical cases are known, and the testing set for predictive analysis. We have used Cross validation technique, which is used to divide the database randomly between training data set and testing dataset and we have obtained: 460 clinical cases for the training phase and 237 for the testing phase.



```
Text(0.5,1,'K_VALUE vs ACCURACY')
```

Figure 3: k- value vs classification accuracy graph

V. Results and Discussion

In this study different values of k ranging from 1 to 15 are used.

The high classification accuracy rate, 98.095%, is recorded by the algorithm that uses the Euclidean distance with a value of k = 3 and k=5.The training dataset accuracy is recorded as 96.933% and testing dataset accuracy is recorded as 97.619%.

Figure 3 shows that when k increases over the classification accuracy rate decreases and stabilizes at a value close to 8, It shows that we are getting 98.09% accuracy on K = 3, 5. Choosing a large value of K will lead to greater amount of execution time & under-fitting. Selecting the small value of K will lead to overfitting.

VI. CONCLUSION

In this paper, we have mainly focused on the advancement of the already available methods for detecting and diagnosing the most dangerous death causing disease among females. We have used k-Nearest Neighbour algorithm for classification and predicting the accuracy of the given dataset using the different values of k using Euclidian Distance. K-NN is time consuming while testing the data set, nevertheless no training is required if the new training pattern is added to the training set. The algorithm gives the best results when k= 3 and 5 (98.09%) and values are not significantly affected after it.

REFERENCES

[1] Peterson, Leif E. "K-nearest neighbor." Scholarpedia 4(2), 1883, 2009.

[2] Keller, J. M., Gray, M.R. and Givens, J.A.”A fuzzy K-nearest neighbor algorithm”. Systems, Man and Cybernetics, IEEE Transactions 15(4), 580 – 585, 1985.

[3] De Oliveira, J.P.S., Conci, A., Prez, M.G., Andaluz, V.H.: Segmentation Of Infrared Images: A New Technology For Early Detection Of Breast Diseases. In: 2015 Ieee International Conference On Industrial Technology (Icit), Pp. 1765–1771 (2015).

[4] http://breastcancerindia.net/