

BREAST CANCER CELLS CLASSIFICATION

***Aishwarya Deep Rastogi¹ and Nupur Prasad²**

¹*MIET-Meerut, U.P. India*

²*Meerut College-Meerut, U.P. India*

**Author for Correspondence*

ABSTRACT

In recent years various computer adaptive methods and algorithms have successfully tried to find their place in tackling various challenges of medical science, henceforth it is imperative to make use of statistical machine learning techniques coupled alongside with digital image processing that could improve earlier detection and diagnosis of disease and in turn help in the treatment stages, in which the time factor is very important to discover the disease in the patient as fast as possible, especially in various cancer tumors such as the breast cancer. The paper proposes the methodology to identify the type of breast cancer cells, classifying the cells into two categories Benign or Malignant. Implementation of research paper has used K Nearest Neighbours Classification algorithm for classification and breast cancer data for initial training is obtained from UCI Machine Learning repository which in turn utilised digitized image of a fine needle aspirate (FNA) of a breast mass to describe characteristics of the cell nuclei present in the image.

Keywords: *Breast Cancer, K Nearest Neighbours Classification algorithm*

INTRODUCTION

Breast cancer cell classification is a process where patient's breast mass sample data is classified either into malignant or benign tumor. For that the digitised image of a fine needle aspirate (FNA) of breast mass is analysed taking into consideration the characteristics of the cell nuclei present in the image. After the characteristic attributes such as Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses are recorded the further process can be divided into three major steps, pre-processing, applying of algorithms to the pre-processed data to classify the information *i.e.*, cell classification label and finally prediction of test data labels

The k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression). In both cases, the input consists of the k closest training examples in the feature space. When we say a technique is non-parametric, it means that it does not make any assumptions on the underlying data distribution more exact, all (or most) the training data is needed during the testing phase (Eapen, 2004 and Abbass 2002; and Priebe, No Date).

KNN Algorithm is based on feature similarity: How closely out-of-sample features resemble our training set determines how we classify a given data point.

The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbour.

MATERIALS AND METHODS

Proposed System

The paper proposes an algorithm that classifies breast cancer cells into either of the two categories Benign or Malignant based on the sample data. The cell nuclei characteristics obtained from the digitised image of a fine needle aspirate is used as the training data to be fed into the classifier for training.

Research Article

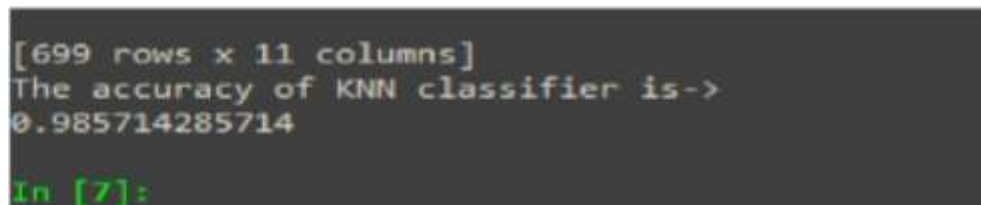
To do away with the inconsistencies of the data present, firstly data scrubbing is done to remove or amend data that is incorrect, incomplete, improperly formatted or duplicated. Then dimensionality reduction is performed using feature selection technique so that only the features that may contribute towards effective classification are only taken into account. The proposed data is then split into 80% to 20% ratio of train and test dataset respectively. The K-nearest classification algorithm is then trained using the 80% train data sample and the rest 20% data is used in obtaining the classifier score i.e. the accuracy of the classifier on unseen data.

Proposed Algorithm

STEP 1	The breast mass sample data is organized using data from UCI Machine Learning repository and from sample data history of patients. The data contains the following attributes as Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli and Mitoses along with the class label i.e. either benign or malignant labels which are used to train the classifier such that when the machine learned model is fed with new and unseen data it is able to predict the type of tumour cell.
STEP 2	Then preprocessing is done on the data set. First data scrubbing is done followed by dimensionality reduction using feature selection technique.
STEP 3	The entire dataset is then split into 4:1 ratio into training and test dataset respectively. The train dataset is used to train the KNN classifier while the test data checks the accuracy of the classifier on unknown dataset. The value for K can be found by algorithm tuning. It is a good idea to try many different values for K (e.g. values from 1 to 21) and see what works best for our situation.
STEP 4	Next step is making the prediction. Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output. To determine which of the K instances in the training dataset are most similar to a new input variable for those K instances Euclidean distance is used. Euclidean Distance $(x, x_i) = \sqrt{\sum (x_j - x_{ij})^2}$ The Euclidean distance between two points in either the plane or 3-dimensional space measures the length of a segment connecting the two points. It is the most obvious way of representing distance between two points. The Pythagorean Theorem can be used to calculate the distance between two points, as shown in the figure below. If the points (x_1, y_1) and (x_2, y_2) are in 2-dimensional space, then the Euclidean distance between them is $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$.

RESULTS

The paper aims at classifying the breast sample data into Malignant or Benign categories with an accuracy of 98% when implemented in python [Python libraries (No Date)].



```
[699 rows x 11 columns]
The accuracy of KNN classifier is->
0.985714285714
In [7]:
```

Figure 1: Results

FUTURE WORK

The computational complexity of KNN increases with the size of the training dataset. For very large training sets, KNN can be made stochastic by taking a sample from the training dataset from which to calculate the K-most similar instances.

There are many other distance measures that can be used, such as Tanimoto, Jaccard, Mahalanobis and cosine distance. We can choose the best distance metric based on the properties of our data. If we are unsure, we can experiment with different distance metrics and different values of K together and see which mix results in the most accurate models.

Moreover, in lieu of single model, a hybrid approach such as a combination of Artificial Neural Network or Bayesian network can be used to obtain a good estimation of prognosis (Choi *et al.*, 2009; and Rani, 2010).

REFERENCES

- Adam H Cannon, Lenore J Cowen and Carey E. Priebe (No date).** Approximate Distance Classification. Department of Mathematical Sciences, The Johns Hopkins University. Available at: <http://rexa.info/paper/095d7064837557bdfbca12fb9c12dbaaeb3a8b0d>
- Arun George Eapen (2004).** Application of Data Mining in Medical Applications, *Master thesis* Waterloo, Ontario, Canada.
- Choi JP, Han TH and Park RW (2009).** A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis. *Journal of Korean Society of Medical Informatics* **15**(1) 49-57.
- Hussein A. Abbass (2002).** An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artificial Intelligence in Medicine*. 25 (3) 265-281.
- Python libraries (No date).** scikit-learn, numpy, scipy [Available at: <https://scikit-learn.org/stable/index.html>]
- Rani KU (2010).** Parallel Approach for Diagnosis of Breast Cancer using Neural Network Technique. *International Journal of Computer Application*, **10**(3) 0975 – 8887.