ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 5, May 2016

Cancer Diagnosis Using Naïve Bayes Algorithm

[1] Rajath A N [2] Parashiva Murthy B M
[1][2] Assistant Professor, Department of CSE
GSSS Institute of Engineering & Technology for Women, Mysore

Abstract: **---- The paper "Cancer Diagnosis Using Naive Bayes Algorithm" deals with the diagnosis of cancer using Naïve Bayes Algorithm using Gene Data Set values of previous patients. The genes expression values will be extracted using DNA micro arrays. Gene data from the cancer patients will be stored in the storage server and for the new patient; we do the necessary tests and will get the genes expression values, based on these values system will categorize the type of the cancer. Data mining technology helps in classifying cancer patients and this technique helps to identify potential cancer patients by simply analyzing the data. The need is to automate this process to make the cancer diagnosis efficient and fast with the use of state of the art technology. In recent times Computer Science has been extensively used in the field of medicine. The use of Neutral Networks and Artificial Intelligence can be seen in the diagnosis and prognosis of Cancer. Digital Image processing is used for the diagnosis of Cancer when the images of cancer cells are available. The concepts of Data Mining are used in the diagnosis of different type of diseases. Since Data Mining is mainly based on obtaining results using previously collected values, mining huge amount of data gives accurate results.**

***Keywords*—Data Mining, Classification Rule, Naïve Bayes Algorithm, Cancer Type and Sub Type Diagnosis.**

I. INTRODUCTION

Cancer is known as malignant tumor or malignant neoplasm, is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body. Not all tumors are cancerous; benign tumors do not spread to other parts of the body. Possible signs and symptoms include: a new lump, abnormal bleeding, unexplained weight loss, and a change in bowel movements, among others. While these symptoms may indicate cancer, they may also occur due to other issues. There are over 100 different known cancers that affect humans hold messages indefinitely. Typically many genetic changes are required before cancer develops. Approximately 5–10% of cancers are due to genetic defects inherited from a person's parents. Cancer can be detected by certain signs and symptoms or screening tests. It is then typically further investigated by medical imaging and confirmed by biopsy. The benefits of screening in breast cancer are controversial. Cancer is often treated with some combination of radiation therapy, surgery, chemotherapy, and targeted therapy. Pain and symptom management are an important part of care. Palliative care is particularly important in those with advanced disease. The chance of survival depends on the type of cancer and extent of disease at the start of treatment.

Existing System-

Currently cancer diagnosis system in hospitals is manual. For example when a patient is registered he/she has to go through radiology test process i.e. X-rays, CT or MRI. Radiologist gives his remarks on the test report. After this process an expert doctor reviews the X-rays/CT/MRI and gives his remarks. In some types of cancer the diagnosis is based on the final decision by the doctors e.g. breast and lung cancer, but in other types of cancer like carcinoma some other tests are also required like biopsy. In a manual system the radiologist and the doctor diagnose cancer. Endoscopy is another technique which is used in diagnosis of cancer. Some types of endoscopes are used to look for cancer in people who have no symptoms. For example, colonoscopy and sigmoidoscopy are used to screen for colon and rectal cancer. These procedures can also help prevent cancer because they let doctors find and remove polyps (growths) that might become cancer if left alone. Endoscopes can be used to take out or destroy small cancers. Small instruments passed through an endoscope can be used to cut out small growths. Doctors also can use tools like a cautery or laser through the tips of some endoscopes to burn or vaporize growths.

Limitations-

Manual process is slow as after the radiologist's review the doctor has to review also and give his/her remarks and finally tell if the cancer is present or not. There is a need to automate this process to make the cancer diagnosis efficient and fast with the use of state of the art technology. Scanned images can't show all patterns and

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 5, May 2016

information. It is difficult to recognize and remember large number of patterns. The results are inefficient and lack user satisfaction.

II PROBLEM STATEMENT

Study of genes from a cancer patient helps us diagnose cancer and differentiate between types of cancer. The proposed system predicts the type of the cancer based on the genes dataset. The genes expression values will be extracted using DNA micro arrays. Gene data from the cancer patients will be stored in the storage server and for the new patient; we do the necessary tests and will get the genes expression values, based on these values system will categorize the subtype of the cancer. Data mining technology helps in classifying cancer patients and this technique helps to identify potential cancer patients by simply analyzing the data.

***Importance of genes in Cancer Diagnosis*:** Genes provide very valuable information which can be used to study any disease in depth. Study of genes from a cancer patient helps us diagnose cancer and differentiate between types of cancer. It also helps in separating the healthy people from the patients. Genes contains infinite patterns that cannot be recorded manually using a microscope. DNA Micro Arrays are used to study the information obtained from Genes.

***DNA Micro Arrays*:** DNA microarrays are the latest form of biotechnology. These allow the measurement of genes expression values simultaneously from hundreds of genes. Some of the application areas of DNA microarrays are obtaining the genes values from yeast in various ecological conditions and studying the gene expression values in cancer patients for different cancer types. DNA Microarrays have huge potential scientifically as they can be useful in the study of genes interactions and genes regulations. Other application areas of DNA microarrays are clinical research and pharmaceutical industry.

III RELATED WORK

Data mining is an essential step in the process of knowledge discovery in databases in which intelligent methods are applied in order to extract patterns. Breast cancer has become the primary reason of death in women in developed countries. The most effective way to reduce breast cancer deaths is to detect it earlier. Early diagnosis needs an accurate and reliable diagnosis procedure that can be used by physicians to distinguish benign breast tumors from malignant ones without going for surgical biopsy. The objective of these predictions is to assign patients to one of the two group either a “benign” that is noncancerous or a “malignant” that is cancerous. The prognosis problem is the long-term care for the disease for patients whose cancer has been surgically removed. Predicting the outcome of a disease is one of the most interesting and challenging tasks where to develop data mining applications. The use of computers with automated tools, large volumes of medical data are being collected and made available to the medical research groups. As a result, data mining techniques has become a popular research tool for medical researchers to identify and exploit patterns and relationships among large number of variables, and made them able to predict the outcome of a disease using the historical datasets. The current research is being carried out on various breast cancer datasets using the data mining techniques to enhance the breast cancer diagnosis and prognosis. [1] In one of the approaches used the research helped in classifying cancer patients and the technique used helped to identify potential cancer patients by simply analyzing the data. [2] Sometimes we briefly examine the potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volume of healthcare data. [3] In some of the cases Histogram Equalization is used for preprocessing of the images and feature extraction process and neural network classifier to check the state of a patient in its early stage whether it is normal or abnormal. After that we predict the survival rate of a patient by extracted features. [4]

The potential for using microwaves for detecting breast tumors is based on the concept of tissue-dependent microwave scattering and absorption in the breast to exploit the contrast in the dielectric properties of malignant and normal breast tissues. There are many approaches in which microwaves can be utilized in the imaging tools. It has been widely assumed that normal breast tissue is largely transparent to microwaves because they are featured with a low relative permittivity and conductivity at the microwave frequency bands, whereas lesions, which contain more water and blood are characterized by a high relative permittivity and conductivity at the microwave frequencies and hence they cause a significant backscatter. Microwave imaging systems are being designed to detect the presence of a small object inside a breast causing a considerably larger backscatter than the surrounding medium. [5] Artificial Neural Network is a branch of Artificial

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 5, May 2016

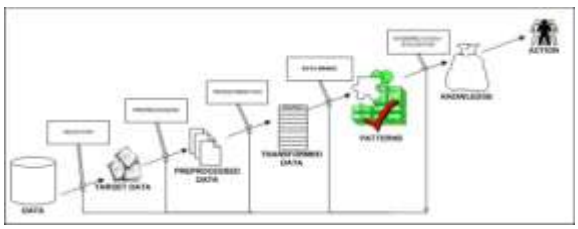
Intelligence has been accepted as a new technology in Computer science. A systematic review that was conducted to assess the benefit of artificial neural networks (ANNs) as decision making tools in the field of cancer. In carcinogenesis, artificial neural networks have been successfully applied to the problems in both pre-clinical and post-clinical diagnosis. The main aim of research in medical diagnostics is to develop more cost-effective and easy–to-use systems, procedures and methods for supporting clinicians. It has been used to analyze demographic data from lung cancer patients with a view to developing diagnostic algorithms that might improve triage practices in the emergency department. For the lung cancer diagnosis problem, the concise rules extracted from the network achieve a high accuracy rate of on the training data set and on the test data set. [6]

IV DATA MINING

Data Mining

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to search large databases in order to find novel and useful patterns that might otherwise remain unknown. In other words, Data mining is a process of analyzing the data from different perspectives and summarizing it into useful information an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. The goal is to extract patterns and knowledge from large amount of data, not the extraction of data itself and is frequently applied to any form of large-scale data or information processing (collection,extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence and business intelligence.

Data Mining Phases:



Classification Rule:

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. An example would be assigning a given email into "spam" or "non-spam" classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient. An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to mathematical function, implemented by a classification algorithm that maps input data to a category.

Naïve Bayes Algorithm:

Naive Bayes algorithm is one of the most effective methods in the field of text classification, but only in the large training sample set can it get a more accurate result. The requirement of a large number of samples not only brings heavy work for previous manual classification, but also puts forward a higher request for storage and computing resources during the computer post-processing. Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature.

Step 1: Scan the dataset (storage servers)
Step 2: Calculate the probability of each attribute value. [n, n_c, m, p]
Step 3: Apply the formulae
P(attributevalue(ai)/subjectvalue(vj))=(n_c + mp)/(n+m)
Where:

- n = the number of training examples for which v = vj
- nc = number of examples for which v = vj and a = ai
- p = a priori estimate for P(aijvj)
- m = the equivalent sample size

Step 4: Multiply the probabilities by p
Step 5: Compare the values and classify the attribute values to one of the predefined set of class.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 5, May 2016

V RELATED DATA

These are the samples from acute breast cancer patients which are derived in prognostic mode. The table consists of the attributes and the range of their respective values.

Sl. No	Attributes	Domain
1	Sample code number	id number
2	Clump Thickness	1 - 10
3	Uniformity of Cell Size	1 - 10
4	Uniformity of Cell Shape	1 - 10
5	Marginal Adhesion	1 - 10
6	Single Epithelial Cell Size	1 - 10
7	Bare Nuclei	1 - 10
8	Bland Chromatin	1 - 10
9	Normal Nucleoli	1 - 10
10	Mitoses	1 - 10
11	class	(2 for benign, 4 for malignant)

REFERENCES

[1] Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease by Shweta Kharya.

[2] Cancer Diagnosis Using Data Mining Technology by Muhammad Shahbaz, Shoaib Faruq, Muhammad Shaheen, Syed Ather Masood.

[3] Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques by V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra.

[4] Early Detection and Prediction Of Lung Cancer Survival Using Neural Network Classifier by Ada, Rajneet Kaur.

[5] Breast Cancer Diagnosis Using Microwave And Hybrid Imaging Methods by Younis M. Abbosh

[6] Application of Neural Networks in Diagnosing Cancer Disease Using Demographic Data by N. Ganesan, K. Venkatesh, M. A. Rama, A. Malathi Palani

[7] Dr. William H. Wolberg, General Surgery Dept. University of Wisconsin, Clinical Sciences Center Madison, WI 53792. wolberg '@' eagle.surgery.wisc.edu

[8] W. Nick Street, Computer Sciences Dept. University of Wisconsin 1210 West Dayton St., Madison, WI 53706. street '@' cs.wisc.edu 608-262-6619

[9] Olvi L. Mangasarian, Computer Sciences Dept., University of Wisconsin 1210 West Dayton St., Madison, WI 53706. olvi '@' cs.wisc.edu