ISSN:0975-9646

Prediction of Cancer Diseases Using Navie Bayes Classification

[1]Shruthi S, [2]Mithun Mohan,[3] Akshaya Aravind,

[1,2,3]Department of Computer Science, Amrita Vishwavidyapeetam,
Mysuru campus. Karnataka, India

Abstract : Cancer is a set of diseases in which some cells of the body grow abnormally. These cells then destroy other surrounding cells and their normal functions. Cancer can spread throughout the human body. Since it is a very treacherous disease its diagnosis is very important. In some forms it spreads within days. So the diagnosis of cancer at early stages is very important. The challenge is to first diagnose the main type and then its subtypes. This research uses data mining classification tools to make a decision support system to identify different types of cancer on the Genes dataset. Data mining technology helps in classifying cancer patients and this technique helps to identify potential cancer patients by simply analyzing the data using Navie Bayes classification.

Keywords - Data mining,Classification, Navie Bayes.

I INTRODUCTION

Cancer is normally diagnosed by examining the cells using a microscope. Imaging tests like computerized tomography (CT) or mammography help in indicating the possible presence of cancer by depicting an abnormal growth or mass. Final decision is usually taken by having different kinds of lab tests of the patient and observing closely the cancer cells under study. Another method used by Doctors is called biopsy. Biopsy is done by surgery. Doctors take a sample of the tissue that is understudy. This sample is then examined with the help of a microscope. The appearance of normal cells is uniform; they are organized in order and are of equal size. Cancer cells are different than normal cells. They are in dispersed order, their sizes are different and they are not structured well. The problem with this is that a medical image such as CT scan or MRI cannot show all the patterns and information for a particular type of cancer or subtypes of cancer. Another issue is that a doctor with his/her naked eye and a microscope cannot remember a large number of patterns of the disease. It is frightening for a patient to know that he/she has cancer. A patient can lose all hope after being diagnosed with cancer. Therefore cancer diagnosis is a process that needs proper care and patience on both sides i.e. the patient and doctor/hospital.

Early diagnosis of cancer can help save the life of a patient because Cancer cells cause destruction to other cells and spread to other parts of body very quickly. If it is diagnosed in the early stage, the treatment begins earlier and this can prevent further spread of the disease.

I.I Genes and their importance in Cancer Diagnosis

Genes provide very valuable information which canbe used to study any disease in depth. Study of genesfrom a cancer patient helps us diagnose cancer anddifferentiate between types of cancer. It also helps inseparating the healthy people from the patients.Genes contains infinite patterns that cannot berecorded manually using a microscope. DNA MicroArrays are used to study the information obtainedfrom Genes.

I.II DNA Micro Arrays

DNA microarrays are the latestform of biotechnology. These allow the measurementof genes expression values simultaneously fromhundreds of genes. Some of the application areas ofDNA microarrays are obtaining the genes valuesfrom 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.

II RELATED WORKS

Numerous methodologies for localization in cancer diagnosis have been developed in the last 20 years.Classifications for molecular subtypes of colorectal cancer based on microsatellite instability (MSI), the CpG island methylator phenotype (CIMP), and somatic mutations in BRAFV600E and KRAS have been proposed to reflect distinct pathways of colorectal tumorigenesis. We assessed the relationship between these subtype classifications and colorectal cancer survival in a population-based study. Methods: Tumor-markers were evaluated in incident colorectal cancer cases diagnosed between 1997-2007 in western Washington State. Cases were classified into pathway-based subtypes according to the following tumor-marker combinations: (1) traditional pathway [microsatellite stable / low MSI (MSS/MSI-L), non-CIMP, BRAF-wildtype, KRAS-wildtype, n = 631]; (2) MSI-high serrated (MSI-high, CIMP-positive, BRAF-mutated, KRAS-wildtype, N=100); (3) MSS/MSI-L serrated (MSS/MSI-L, CIMP-positive, BRAF-mutated, KRAS-wildtype, n = 55); (4) alternate pathway (MSS/MSI-L, non-CIMP, BRAF-wildtype, KRAS-mutated, n = 353); and (5) MSI-high familial (MSI-high, non-CIMP, BRAF-wildtype, KRAS-wildtype, n = 50). Multiple-imputation was used to classify tumor-marker status for cases with missing data on one to three markers. Differences in survival across subtypes were assessed through multivariable-adjusted Cox regression. Results: Relative to cases of the predominant traditional pathway subtype, cases with MSS/MSI-L serrated and alternate pathway tumor subtypes experienced statistically significantly worse disease-specific survival [hazard ratio (HR) = 2.25, 95%

confidence interval (CI): 1.50-3.36 and HR = 1.39, 95% CI: 1.12-1.71, respectively]; cases with MSI-high familial tumors had the most favorable disease-specific prognosis (HR = 0.28, 95% CI: 0.12-0.64). With respect to overall survival, associations were similar but slightly attenuated. Conclusions: In this large, population-based study, colorectal cancer subtype classifications based on integrated pathways were associated with marked differences in survival, highlighting the significance of molecular heterogeneity in colorectal cancer.

III NAVIE BAYES ALGORITHEM

Bayesian classifiers are statistical classifiers. They can predictclass membership probabilities, such as the probability that a given tuple belongs toa particular class.Naïve Bayesian classifiers assume that the effect of an attribute value on a given classis independent of the values of the other attributes.

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)/subjectvaluevj)= (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.

IV PROBLEM FORMULATION

Attributes considered for proposed methodology and descretization of values are tabulated as follows

Sl.No	Attributes	Values
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)

Table 1

A **benign tumor** is a mass of cells (tumor) that lacks the ability to invade neighboring tissue or metastasize. These characteristics are required for a tumor to be defined as cancerous and therefore benign tumors are non-cancerous.

A **malignant tumor** or malignant neoplasm commonly called as CANCER, is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body

Here Clump Thickness, Uniformity of Cell size, Uniformity of Cell shape, Marginal adhesion, Single Epithelial cell size, bare nuclei, Bland chromatin, Normal nucleoli and Mitosis are the attributes we need to consider.

Class becomes the subject and it can take either of the two values: 2 or 4.

2 means barely means cancer is not present, whereas 4 means cancer is present.

A. *Performance analysis*

Suppose we want to classify a patient with following values for the attribute,

Sample code number: 1023468
Clump Thickness: 3
Uniformity of Cell Size: 6
Uniformity of Cell Shape: 7
Marginal Adhesion: 5
Single Epithelial Cell Size: 4
Bare Nuclei: 9
Bland Chromatin: 3
Normal Nucleoli: 2
Mitoses: 9

Applying the equation (Eq1) to the above given information,

P(3|2) , P(6|2) , P(7|2) , P(5|2) , P(4|2) , P(9|2) , P(3|2) , P(2|2) , P(9|2) ,

P(3|4) , P(6|4) , P(7|4) , P(5|4) , P(4|4) , P(9|4) , P(3|4) , P(2|4) , P(9|4)

After computation it was known that the above training example is classified into class 4, which indicates the presence of cancer.

B. *Experimental analysis*

In this section, we analyzed our scheme. From the beginning we pointed out that cancer type and sub type prediction as a cancer diagnosis.

In order to evaluate the proposed scheme we considered gene expression value as data set. We have to find different cancer type and its sub type. There will be separate attribute value and id for gene data corresponding to the cancer types. The attribute value of the newly entered patient can refer the older patient gene data. The efficiency of the application depend on the speed and the accurate prediction of the large patient data. In solution module, system generate the out put for the input given by the application user, output is categorizing the cancer type and sub type based on the gene data set. DNA micro arrays are used to study the information obtained from the Genes. These allow the measurement of genes expression values simultaneously from hundreds of genes. This will be automated application for the cancer diagnosis. Here we classified the particular cancer type and sub types by using data mining techniques.

The application which we are developing is going to be used by the hospitals or the stakeholders. This is going to help them in predicting the cancer type and subtype based on the gene dataset for the new patients from plenty of old patients genes dataset. It will takes less time to accomplish a particular task such as predicting the cancer type and its subtype for new patients, generating reports, extracting the genes dataset from server which also reduce time complexity. It reduces the complications when an information has several functionalities thus increases the efficiency. Here the data updates are done automatically without loss of data that already exists.

V. CONCLUSION

According to results Naive Bayesian Classification has the most accurate prediction for dataset samples. Naive Bayesian classified 95% of the samples correctly in their respective classes. It has only error rate of 5%. Naive Bayesian is the best method for classifying DNA Microarray genes expression data. According to results naive Bayesian Classification has the most accurate prediction for dataset samples. Naive Bayesian classified 95% of the samples correctly in their respective classes. It has only error rate of 5%. Naive Bayesian is the best method for classifying DNA Microarray genes expression data.

The results above can be improved by reducing the number of attributes we have in the dataset. This can be done using dimensionality reduction techniques like principle component analysis. Problem with principle component analysis is that we do not have the track of data which is considered redundant by this method. If we can somehow obtain the pattern of data that is redundant and get information about which attributes or values are retained then it would be a great improvement in the classification results of any of the learning algorithm.

REFERENCES

[1] Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.Sandrine Dudoit1, Yee Hwa Yang2, Matthew J. Callow3, and Terence P. Speed2;4Technical report # 578, August 2000.

[2] Glutamic acid analogues used as potent anticancer: A review SatyajitDutta*, 1, Supratim Ray2 and K. Nagarajan3Der PharmaChemica, 2011, 3(2):263-272 (http://derpharmachemica.com/archive.html).

[3] EVALUATING THE MATHEMATICAL FUNCTION OF MRI IMAGES OF NORMAL AND ABNORMAL HUMAN LUNGS Pavan Khandelwal1, Prof. M.P Parsai2.M.E Student [Microwave Engg.], Dept. of ECE, Jabalpur Engineering College, Jabalpur, Madhya Pradesh, India 1Professor, Dept. of ECE, Jabalpur Engineering College, Jabalpur, Madhya Pradesh, India 2

[4] Diagnosis and Risk Assessment of Cancer On Genes Dataset Using Data Mining TechniquesArunkumar Sivaraman1, Dr.M.Lakshmi2, S.Arun Rajesh31Research Scholar, Department of Computer Science and Engineering ManonmaniumSundaranar University, Trinelveli, India.International Journal Of Engineering And Computer Science ISSN:2319-7242Volume 2 Issue 8 August, 2013 Page No. 2430-2433Arunkumar Sivaraman, IJECS Volume 2 Issue 8 August, 2013 Page No.2430-2433 Page 2430Diagnosis.

[5] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02.

[6] .Gwangju, Republic of Korea. Classification and diagnostic prediction of cancers using gene, expression profiling and artificial neural networks, JAVED KHAN, JUN S. WEI, MARKUS RINGNÉR, LAO H. SAAL, MARC LADANY, FRANK WESTERMANN, FRANK BERTHOLD, MANFRED SCHWAB, CRISTINA R. ANTONESCU, CARSTEN PETERSON & PAUL S. MELTZER

.