Unsupervised Study of Cell Behaviour for Breast Cancer

Supratim Bhattacharya,Jayanta Poray and Sampurna Mandal
Department of Computer Science & Engineering, Techno India University,
Saltlake, kolkata - 700091, India
E-mail:bhattacharya.supratim@gmail.com, jayanta.poray@gmail.com,piu.sampurna@gmail.com

August 13, 2016

Abstract

Women's most threatened diseases is breast cancer.It is the leading cause of death for women today and it is the most common cancer in developed countries.Recently, the collection of biological data has been increasing at an ex-plosive rate. In this domain, Data Mining plays an important role.

1 Introduction

In India the event of breast cancer cases are increasing day by day. A new global study estimates that by 2030, the number of new cases of breast cancer in India will increase from the current 115,000 to around 200,000 per year.Cancer treatment and early successful diagnosis of the patients is a challenge since so many years. Doctors and Researchers have been working every day to find new ways to treat cancer.Data mining for cancer treatment can become a great support tool for doctors and physicians for decision making and estimation purpose. The need for biological data mining is that there is too much data but they are mostly unstructured.Data mining and machine learning depend on classification which is the most essential and important task. Many experiments are performed on medical datasets using multiple classifiers and feature selection techniques. A good amount of research on breast cancer datasets is found in literature. Many of them show good classification accuracy.

Data mining approaches in medical domains is increasing rapidly due to the improvement effectiveness of these approaches to classification and prediction systems, especially in helping medical practitioners in their decision making. In addition to its importance in finding ways to improve patient outcomes, reduce the cost of medicine, and help in enhancing clinical studies. Although there was a great deal of public education and scientific research, Breast cancer considered the most common invasive cancer in women, with more than one million cases and nearly 600,000 deaths occurring worldwide annually.Data mining and machine learning depend on classification which is the most essential and important task. Many experiments are performed on medical datasets using multiple classifiers and feature selection techniques. A good amount of research on breast cancer datasets is found in literature. Many of them show good classification accuracy.

We have applied WEKA machine learning technique to analyze the risk factors so that early diagnosis and proper preventive measures can be taken. We have used the Wisconsin Diagnostic Breast Cancer Dataset from UC Irvine for our study. The classification is based on benign or malignant.

2 Problem Definition

The dataset used in this experiment were obtained from Wisconsin Diagnostic Breast Cancer Dataset from UC Irvine machine learning repository and described by Dr. William H. Wolberg. The breast cancer data have been used in some other research. We study the effect of nine characteristics parameter on the state of Breast cancer and the influence of the involved parameters on the performance of the SVM model. We have used WEKA Tool as a classifier in our experiment. Our aim is to predict the various state, behaviour and characteristics of breast cancer.

In this dataset, there are 698 samples taken from different women and every sample is expressed by nine characteristic parameter. The nine parameter are as follows:- Clump thickness, Uniformity of cell size, Uniformity of cell shape, Marginal adhesion, single epithelia cell size, Bare Nuclei, Bland chromatic, Normal Nucleoli, Mitoses. According to the properties of these nine parameter, the brest cancer is classified into benign & malignant. Every single parameter is given a range between 1 to 10 and the resultant class is expressed by 2 for benign and 4 for malignant. Among the 698 samples in the dataset there are 16 samples with missing or incomplete data. So we have used remaining 682 samples in this machine learning.

3 Proposed Model

Fig 1 shows the functional block diagram of our proposed model. It consists of four steps: (a) Acquisition, (b) Preprocess, (c) Feature Extraction and (d) Feature Selection.

In acquisition step, feature selection & feature extraction is accomplished in order to determine the input vector and based on either feature selection or feature extraction, dimensionality reduction is accomplished. In the preprocessing phase, filtering is done to clear the noise & map the entire data into lower dimension. Also less important and redundant information are ignored. In the classification step different classifier is used to get the best result out of it. We have also applied clustering method & Association Rule Mining to obtain more decisions & to predict more accurately

4 Methodology

The dataset's attributes are found listed in Table 2.

Table2: WISCONSIN BREAST CANCER DATASET ATTRIBUTES

	Attribute	Domain
1	Sample Code No	id no
2	Clumb Thickness	1-10
3	Uniformity(Cell Size)	1-10
4	Uniformity(Cell Shape)	1-10
5	Marginal Adhesion	1-10
6	sgl Epithelial(cell size)	1-10
7	Bare Nucleoli	1-10
8	Normal Nucleoli	1-10
9	Mitoses	1-10
10	Class	2 or 4

In the Clump thickness benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in multilayered. While in the Uniformity of cell size/shape the cancer cells tend to vary in size and shape. That is why these parameters are valuable in determining whether the cells are cancerous or not. In the case of Marginal adhesion the normal cells tend to stick together, where cancer cells tend to lose this ability. So loss of adhesion is a sign of malignancy. In the Single epithelial cell size the size is related to the uniformity mentioned above. Epithelial cells that are significantly enlarged may be a malignant cell. The Bare nuclei is a term used for nuclei that is not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumors. The Bland Chromatin describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells the chromatin tends to be coarser. The Normal nucleoli are small structures seen in the nucleus. In normal cells the nucleolus is usually very small if visible. In cancer cells the nucleoli become more prominent, and sometimes there are more of them. Finally, Mitoses is nuclear division plus cytokines and produce two identical daughter cells during prophase. It is the process in which the cell divides and replicates. Pathologists can determine the grade of cancer by counting the number of mitoses.

We next present our algorithm and further describe the dataset on which we have evaluated. our first step is to discretize the dataset into three major groups.

1. Low
2. Mid
3. High

Algorithm 1 An algorithm for discretization of dataset

Input : Dataset in excel format with 9 parameters.
Output: : Dataset in csv file(space delimiter) format in discrete format with all 9 parameters.
Algorithmic Steps:

1. Obtain the ranges of high, middle and low.
2. collect every cell value for computation for every parameter.
3. **For** parameter 1 to 9 do
4. **If** Cell value¿= high value **Then Put** new Cell value= 'H' **Else If** Cell value¿= middle value **Then Put** new Cell value= 'M' **Else Put** new Cell value= 'L'
5. **End if**
6. **Next**
7. **For** 10^th parameter
8. **If**Cell value=2 **Then**
9. **Put** new Cell value = "Benign"
10. **Else If**Cell value=4 then
11. **Put**new Cell value = "Malignant"
12. **End If**
13. Construct another excel file based on this discrete value.
14. Convert the excel file into csv(space delimiter) file.

As the parameter of the dataset ranges from 1 to 10 we made this discretization based on different ranges like:-

1. low:- 1 to 1 mid:- 2 to 6 high:- 7 to 10
2. low:- 1 to 1 mid:- 2 to 7 high:- 8 to 10
3. low:- 1 to 2 mid:- 3 to 7 high:- 8 to 10
4. low:- 1 to 3 mid:- 4 to 6 high:- 7 to 10
5. low:- 1 to 4 mid:- 5 to 6 high:- 7 to 10
6. low:- 1 to 4 mid:- 5 to 7 high:- 8 to 10

5 Methods

6 Analytical framework

Cancer cell description:The breast cell attributes behavior can be helpful in determining whether the cell is normal or cancerous.Some cells can be visualized like in given figure-2.In the figure it is seen that the nucleoli in normal cell is approximately invisible but the cancerous cell has an enlarged one.In the mitosis attribute the cell division is so fast and uncontrolled then the cell size and shape varies a lot in the cancerous cell.Then according to the bare nuclei attribute, a cancerous cell is comparatively dry. Decision tree-Decision tree is a popular classification method.Decision

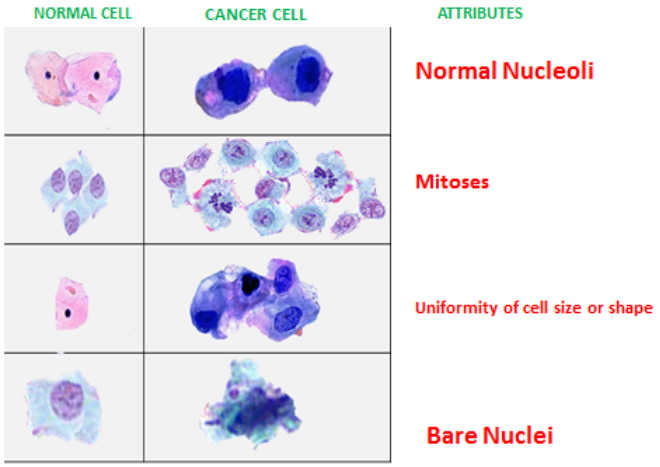

Figure 1: normal vs cancerous cell attributes'behavior

tree is used as a predictive model which maps observations about an item to conclusions about the item's target value.Rules produced by decision tree induction are easy to interpret and understand and hence can help greatly in appreciating the underlying mechanism that separate samples in different classes.One of the decision tree algorithms is c4.5.C4.5 builds decision trees from a set of training data using the concept of information entropy.It uses the information gain ratio criterion to determine the most discriminatory feature at each step of its decision tree induction process.Pruning helps to reduce the size of decision trees by removing sections of the tree that provide little power to classify instances.Pruning reduces the complexity of the final classifier and hence improves predictive accuracy by the reduction of over fitting. Algorithm: Step 1.The leaf is labelled with the same class if the instances belong to the same class. Step 2: For every attribute, the potential information will be calculated and the gain in information will be taken from the test on the attribute. Step 3: Finally the best attribute will be selected based on the current selection parameter.

FORMULa Association Rules Mining Since its introduction in 1993[1] the task of association rules mining has received a great deal of attention.Association Rules mining is the datamining process of finding the rules that may govern associations and causal objects between sets of items.It is used to find out association rules that satisfy the predefined minimum support and confidence from a given database.The problem is usually decomposed into two subproblems.One is to find those item sets whose occurrences exceed a predefined threshold in the database;those item sets are frequent or large item sets.The second problem is to generate association rules from those large item sets with the constriants of minimal confidence. One of the popular algorithms is the aprioiri

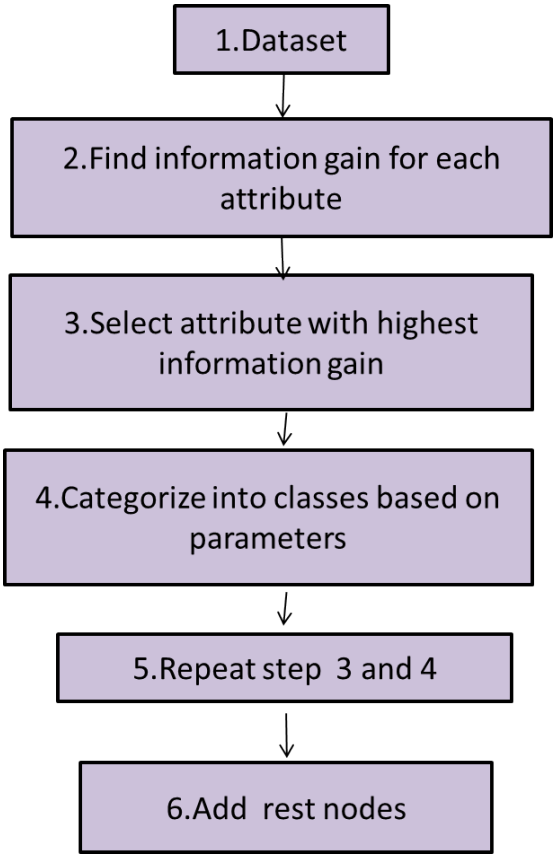

Figure 2: flowchart of c4.5

algorithm.Its iteravely reduces the minimum support until it finds the required number of rules with the given minimum confidence. The algorithm has an option to mine class association rules. Mathematical formulae-

7 Discussion

8 Results

Based on these cell's behavior the dataset's attributes values can largely help in determination and diagnosis.On discretisation and preprocessing,each attribute shows its own graphical result divided into 3 bins based on high,medium and low values. Fig-3 shows normal nucleoli attribute and the no. of patient instances classified into benign and malignant group. The figure shows that out of 509 instances with low or small value has 30

Similarly all attributes can be visualised at a time shown in figure below.

malignant and 70benign then out of 75 instances with medium value 90are Malignant and 10Benign then out of 98 instances with high value 99are Malignant and 1benign.According to the observation the higher the value of the attributes the greater the tendency towards malignancy. On observing the J48 pruned tree and considering only the malignant category we can see that a combination of two medium values or one high and another medium value or only high values can help in determining malignancy. Selection of the most useful attribute is necessary.Information

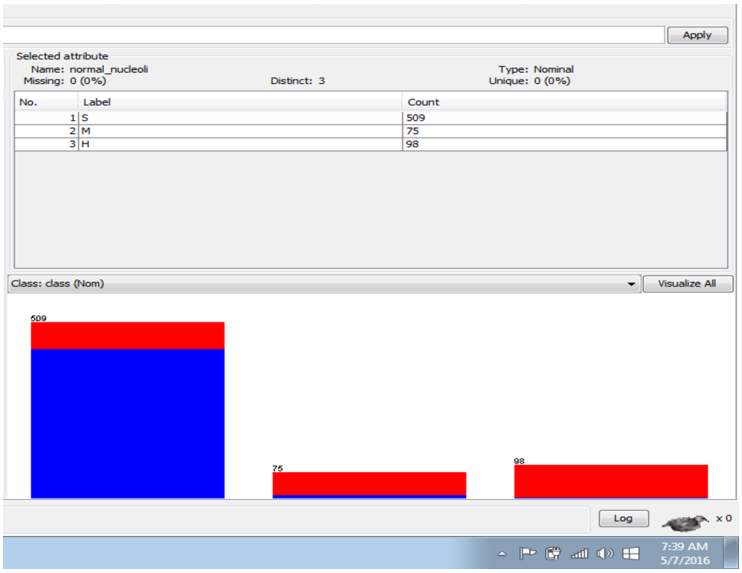

Figure 3: normal nucleoli

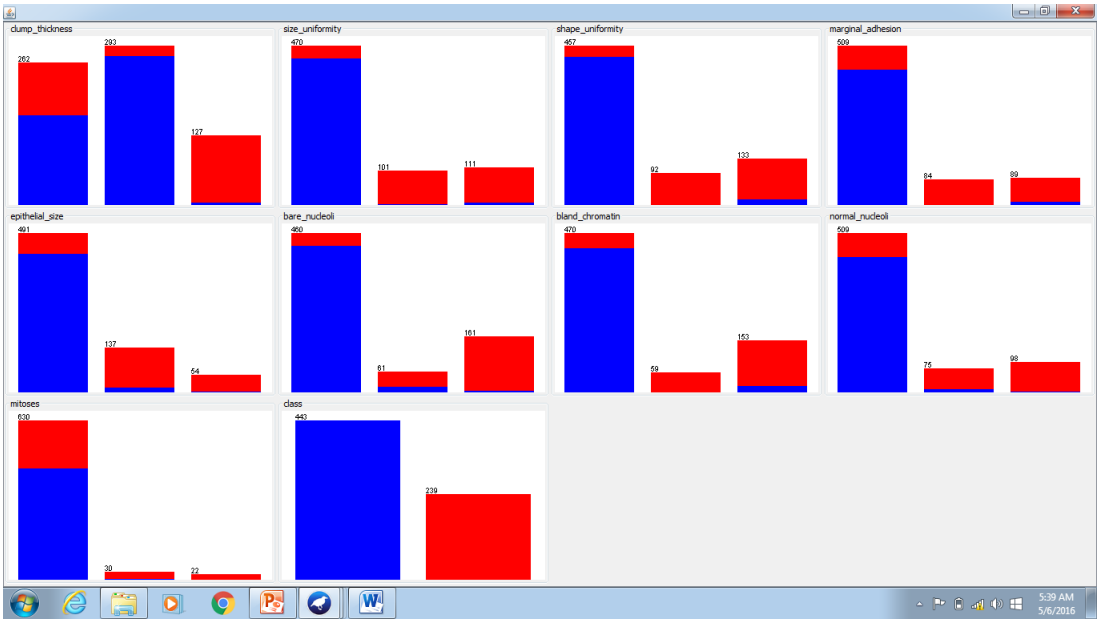

Figure 4: All attributes

gain evaluator can help to find the attribute highest priority.The output presented below shows information gain values of each and every attribute and their ranking.FIGURE Size uniformity is selected as the best one.Then shape uniformity and so on.On selecting a set of attributes based on the ranking we get one or different kinds of rules.Among them the best and of utmost importance are selected. the table below shows different statistical measures.Kappa statistics, mean absolute errorand others have their own significance and importance.Studies that measure the argument between two or more observers that takes into account the fact that observers will sometimes agree or disagree simply by chance.The calculation is based on the difference between how much agreement is acyually present (observed agreement)compared to how much agreement would be

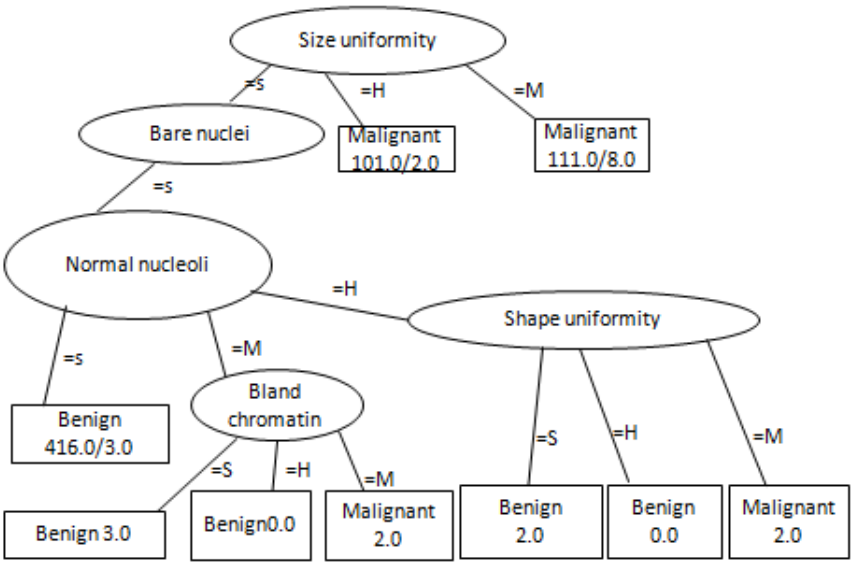

Figure 5: Decision tree

expected to be present by chance alone (expected agreement).The MAE and RMSE can be used together to diagnose the variation in the errors in a set of forecasts.The RMSE will always be larger or equal to the MAE;The greater the difference between them the greater the variance in the individual errors in the sample.If RMSE=MAE then all the errors are of the same magnitude.

RANKING OF ATTRIBUTES

INFO_GAIN	ATTRIBUTE NO.	ATTRIBUTE NAME
0.5788	2	size_uniformity
0.57	3	shape_uniformity
0.5411	6	bare_nucleoli
0.5034	7	bland_chromatin
0.4084	5	epithelial_size
0.3914	4	marginal_adhesion
0.3867	8	normal_nucleoli
0.3828	1	clump_thickness
0.0944	9	mitoses

Figure 6: ranking

Correctly Classified Instances	*654*	*95.8944%*
Incorrectly Classified Instances	*28*	*4.1056 %*

Kappa statistic	**0.9105**
Mean absolute error	**0.0636**
Root mean squared error	**0.1952**
Relative absolute error	**13.9707 %**
Root relative squared error	**40.9069 %**
Total Number of Instances	**682**

Figure 7: classification results

APPENDIX

Pearson correlation coefficient between two expression vectors X and Y is computed as,

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{[\sum_{i=1}^{N}(x_i - \overline{x})^2][\sum_{i=1}^{N}(y_i - \overline{y})^2]}}. \quad (1)$$

Suppose, an event is observed n times out of total N observations and given the evidence that the event originally occurs e times out of E total cases. Then, the p-value is computed assuming a hypergeometric distribution as given follows.

$$p - value = \sum_{i=n}^{N} \frac{\binom{e}{i}\binom{E-e}{N-i}}{\binom{E}{N}}. \quad (2)$$

Supplementary Link: http://www.isical.ac.in/∼malay_r/Supplementary.html.