Analysis of classifier to improve Medical diagnosis for Breast Cancer Detection using Data Mining Techniques

A.subasini[1]
[1]Research Scholar, R.D.Govt college, Sivagangai
Nirase Fathima abubacker[2]
[2]Research Scholar, Putra University Malaysia
Dr.Rekha[3]
[3]Assistant professor, Department of computer science R.D.Govt college, Sivagangai.

ABSTRACT

Many research have been conducted to analyze Breast Cancer Data. Breast cancer is one of the leading cancers for women in developed countries including India. It is the second most common cause of cancer death in women. The high incidence of breast cancer in women has increased significantly in the last years. In this paper we have discussed various data mining approaches that have been utilized for breast cancer diagnosis and prognosis. Breast Cancer Diagnosis is distinguish of benign from malignant breast lumps and Breast Cancer Prognosis predicts when Breast Cancer is to recur in patients that have had their cancers excised .In this work, we explore the applicability of association rule data mining technique to predict the presence of breast cancer. Also it analyzes the performance of conventional supervised learning algorithms viz. C5.0, ID3, APRIORI, C4.5 and Naive Bayes. Experimental results prove that C5.0 serves to be the best one with highest accuracy.

***Keywords*— Data Mining, Breast Cancer, Data mining.**

1. Introduction:

Data mining provides us with a variety of techniques for pattern analysis on large data such as association, clustering, segmentation and classification for better manipulation of data. Constructing fast and accurate classifiers for large data sets is an important task in data mining and knowledge discovery[2]. There is growing evidence that merging classification and association rule mining together can produce more efficient and accurate classification systems than traditional classification techniques.

2. Preprocessing:

Image pre-processing techniques are necessary, in order to find the orientation of the mammogram, to remove the noise and to enhance the quality of the image [10]. Before any image processing algorithm can be applied on mammogram, preprocessing steps are very important in order to limit the search for abnormalities without undue influence from background of the mammogram. Digital mammograms are medical images that are difficult to be interpreted, thus a preparation phase is needed in order to improve the image quality and make the segmentation results more accurate. The main aim of this process is to improve the quality of the image to make it ready to further processing by removing the unrelated and surplus parts in the back ground of the mammogram.

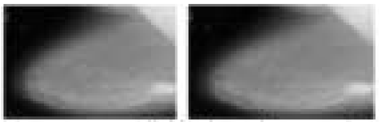


Figure 1: Mammogram digitization Noise Removal Using 2D Median Filtering.

3. Segmentation:

Image segmentation is generally defined as the basic image processing that subdivides a digital image $f\ (x,\ y)$ into its continuous, disconnect and nonempty subset f1 , f2, f 3,…… fn , which provides convenience to extraction of attribute. Practical application of image segmentation range from filtering of noisy images, medical applications

(Locate tumors and other pathologies, Measure tissue volumes, Computer guided surgery, Diagnosis, Treatment planning, study of anatomical structure), Locate objects in satellite images (roads, forests, etc.), Face Recognition, Finger print Recognition, etc. Many segmentation methods have been proposed in the literature.

3.1. Segmentation by Morphological Algorithm:

Mathematical morphology[9] is used as a tool for extracting image components such as boundaries in image segmentation. Since language of mathematical morphology is set theory, this segmentation approach is based on binary image. This algorithm includes two major steps: preprocessing and segmentation. Thresholding is used to convert input image into binary image.Since tumor tissue tends to have maximum intensity in mammograms, normally closed to 1 in gray level, a global threshold could serve as the first cut in the process and convert the image into binary image. Dilation and erosion are two basic morphological operations defined by equation “(1)” and “(2)” respectively.

$$A \oplus B = \left\{ Z \middle/ (\hat{B})_z \cap A \neq \phi \right\} \quad (1)$$

$$A \ \Theta \ B = \{ Z \ / (B)_z \subseteq A \} \ (2)$$

The simplest way to realize boundary extraction of a binary image A is given by equation “(3)”,

$$\beta(A) = A - (A \ \Theta \ B) \quad (3)$$

where B is a suitable structuring element. However, there could be noise present by this method, instead of using original binary image A; dilation of A, that is, $A \oplus B$ could be used. “(4)” is the resulting edge detection formula.

$$\beta(A) = (A \oplus B) - ((A \oplus B)\Theta B)) \ (4)$$

4. Feature extraction:

Feature is used to denote a piece of information which is relevant for solving the computational task related to a certain application. More specifically, features can refer to: The result of a general neighborhood operation (feature extractor or feature detector) applied to the image,

Specific structures in the image itself, ranging from simple structures such as points or edges to more complex structures such as objects.

4.1. Mean:

The mean, m of the pixel values in the defined window, estimates the value in the image in which central clustering occurs. The mean can be calculated using the formula:

$$\mu = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j)$$

Where p(i,j) is the pixel value at point (i,j) of an image of size MxN.

4.2. Standard Deviation

The Standard Deviation, σ is the estimate of the mean square devi-ation of grey pixel value

p(i, j) from its mean value m. Standard deviation describes the dispersion within a local region. It is determined using the formula

$$\sigma = \sqrt{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (p(i,j) - \mu)^2}$$

4.3. Contrast:

Measure of intensity contrast between a pixel and its neighbour.

$$\sum_{i,j=0}^{N-1} (p(i-j)*(i-j)^2)$$

4.4 Entropy :

Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. Entropy, h can also be used to describe the distribution variation in a region. Overall Entropy of the image can be calculated as:

$$h = -\sum_{k=0}^{L-1} Pr_k(log_2 Pr_k)$$

Where, Pr is the probability of the k-th grey level, which can be calculated as Zk /m*n, Zk is the total number of pixels with the kth grey level and L is the total number of grey levels.

4.5. Energy :

Energy returns the sum of squared elements in the Grey Level Co-Occurrence Matrix (GLCM). Energy is also known as uniformity. The range of energy is [0 1]. Energy is 1 for a constant image. The formula for finding energy is given in below equation:

$$E = \sum_{i,j} P(i,j)^2$$

4.6.Variance :

Variance is the square root of standard deviation. The formula for finding Variance is:
Where SD is the Standard Deviation.

$$Var = \sqrt{SD}$$

After extracting the features of segmented mass/tumor, then the dataset has to be constructed in the proper format, so that it can be given to any of the standard classifier tools.

5. ASSOCIATION RULE MINING:

Association rule mining gives the interesting relationship among a large items[1]. Association-rule mining is a data-mining task that is used to discover relationships among items in a transactional database .

Association rule mining gives the interesting relationship among a large items[2]. If the rules satisfy the minimum support and minimum confidence threshold. The association rule mining divides the problems into two parts. First it finds the frequent item sets : Means each item set is frequent if it satisfies minimum support. Second it finds the strong association rules from the frequent item sets: Means these rules should satisfy minimum support and minimum confidence. Many scans are required to find frequent sequence by using association rule mining[5-8].

6. Classification:

Classification algorithm is supervised method that is first trained on a set of sample images (whose classification label is known) called the training set. The performance of the algorithm is then tested on a separate testing set. The extracted features are input to the classifier. Here We Compare the Performance Evaluation Of Five Classifier. By applying various classifiers the dataset is analyzed. The accuracy measures such as Precision and recall are used to evaluate the performance of the classifiers.

6.1) C4.5:

C4.5 was developed by Quinlan Ross which is an extension to ID3[2]. It is mainly used for generating decision tree. The splitting area defined here is gain ratio. C4.5 classification uses entropy and information gain for tree splitting. It is suitable for handling both categorical as well as continuous data. A threshold value is fixed such that all the values above the threshold are not taken into consideration. The initial step is to calculate information gain for each attribute. The attribute with the maximum
gain will be preferred as the root node for the decision tree. Given a set S of cases, C4.5 first grows an initial tree using the divide-and-conquer algorithm as follows:

- If all the cases in S belong to the same class or S is small, the tree is a leaf labeled with the most frequent class in S.
- Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets S1, S2,... according to the outcome for each case, and apply the same procedure recursively to each subset.

6.2) Iterative Dichotomizer (ID3):

ID3 is a simple decision learning algorithm, developed by J.Ross Qunilan. It accepts only categorical data for building a model. The basic idea of ID3 is to construct a decision tree by employing a top down greedy search through the given sets of training data to test each attribute at every node. It uses statistical property known as
information gain to select which attribute to test at each node in the tree. Information gain measures how well a given attribute separates the training samples according to
their classification. [4]

$$Info(D) = -\sum_{i=1}^{m} P_i \log_2(P_i)$$

6.3) Naive Bayes:

Naive Bayes Classifier is a probabilistic model based on Baye's theorem. It is defined as a statistical classifier. It is one of the frequently used method for supervised learning. It provides an efficient way of handling any number of attributes or classes which is purely based on probabilistic theory. Bayesian classification provides practical learning algorithms and prior knowledge on observed data.[3]

6.4)C5.0 ALGORITHM:

C5.0 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set S = S1,S2,...Sn of already classified samples. Each sample Si = x1,x2,... xk is a vector where x1,x2,... xk represent attributes or features of the sample. The training data is augmented with a vector C = c1,c2,... cm where c1,c2,... cm represent the class to which each sample belongs.

At each node of the tree, C5.0 chooses one attribute of the data that most effectively splits the set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain, which is the difference in entropy that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C5.0 algorithm then recurses on the smaller sub lists.

6.5)Apriori algorithm: One of the most popular data mining approaches is to find frequent itemsets from a transaction dataset and derive association rules. Finding frequent itemsets (itemsets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. Apriori is a seminal algorithm for finding frequent itemsets using candidate generation. It is characterized as a level-wise complete search algorithm using anti-monotonicity of itemsets, "if an itemset is not frequent, any of its superset is never frequent". By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order. Let the set of frequent itemsets of size k be Fk and their candidates be Ck . Apriori first scans the database and searches for frequent itemsets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement. It then iterates on the following three steps and extracts all the frequent itemsets.

1. Generate Ck+1, candidates of frequent itemsets of size k +1, from the frequent itemsets of size k.
2. Scan the database and calculate the support of each candidate of frequent itemsets.
3. Add those itemsets that satisfies the minimum support requirement to Fk+1.

7. Accuracy Measures:

Accuracy measure represents how far the set of tuples are being classified correctly.TP refers to positive tuples and TN refers to negative tuples classified by the basic classifiers. Similarly FP refers to positive tuples and FN refers to negative tuples which is being incorrectly classified by the classifiers. The accuracy measures used here are precision and recall.

8. Confusion matrix*:*

The confusion matrix contains four classification performance indices: true positive, false positive, false negative, and true negative as shown in Table 1. These four indices are also usually used to evaluate the performance the two-class classification problem.

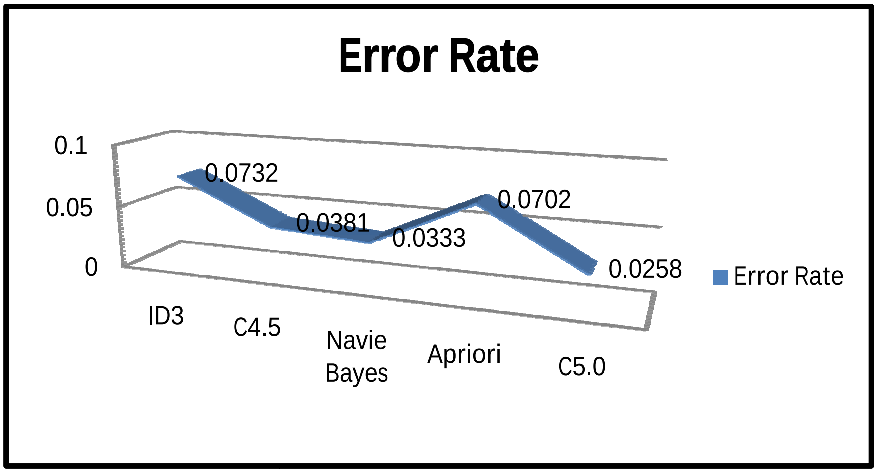
The four classification performance indices included in the confusion matrix:

Table1: Confusion Matrix

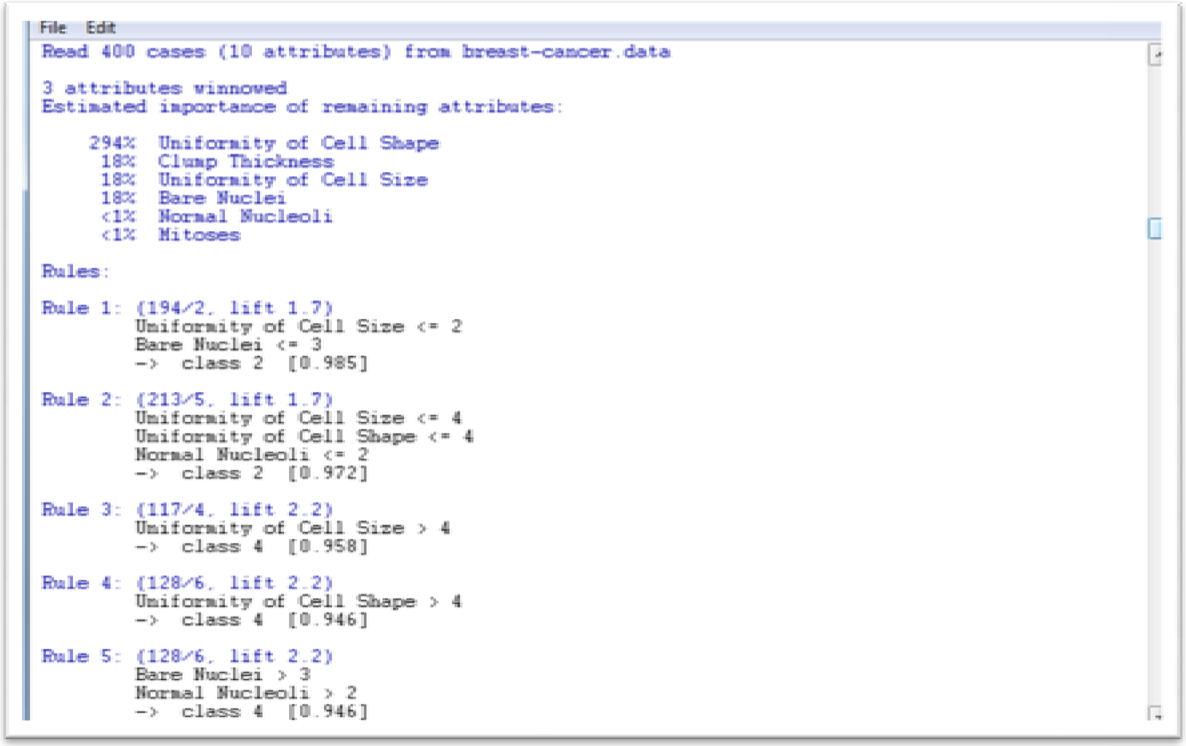
Actual Class	predicted Class	
	Positive	Negative
Positive	True Positive(TP)	False Negative(FN)
Negative	False Positive(FP)	True Negative(TN)

ERROR RATE

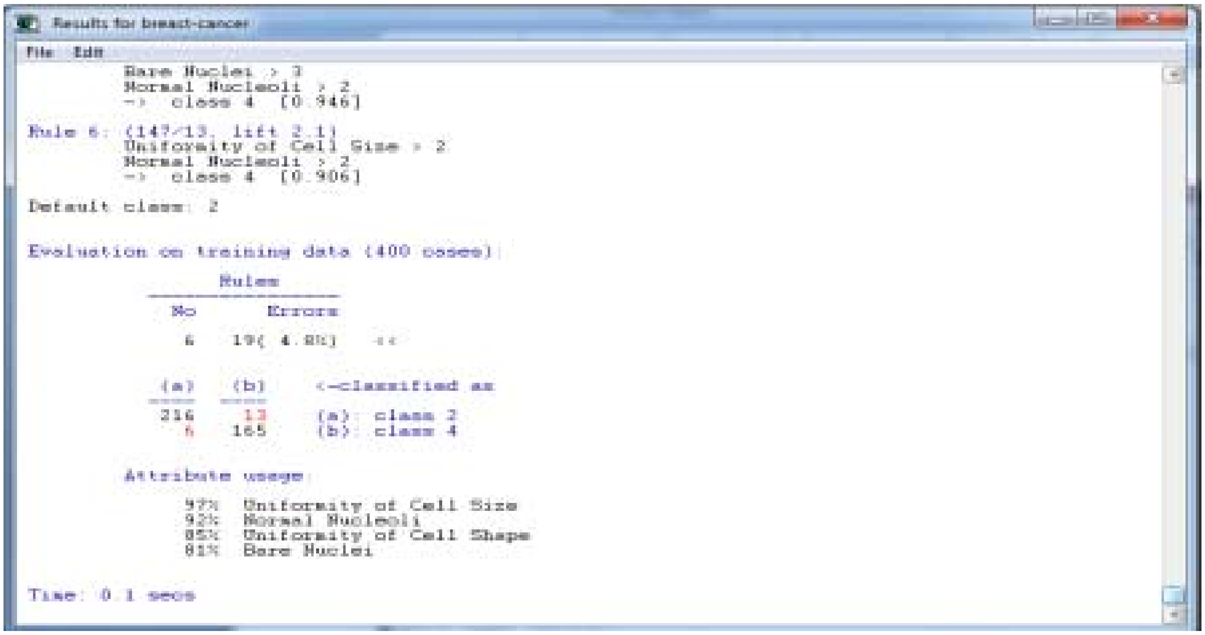
The error rates of various classifiers are compared.



SAMPLE OUTPUT:



```
Read 400 cases (10 attributes) from breast-cancer.data

3 attributes winnowed
Estimated importance of remaining attributes:

    294%  Uniformity of Cell Shape
     18%  Clump Thickness
     18%  Uniformity of Cell Size
     18%  Bare Nuclei
     <1%  Normal Nucleoli
     <1%  Mitoses

Rules:

Rule 1: (194/2, lift 1.7)
        Uniformity of Cell Size <= 2
        Bare Nuclei <= 3
        ->  class 2  [0.985]

Rule 2: (213/5, lift 1.7)
        Uniformity of Cell Size <= 4
        Uniformity of Cell Shape <= 4
        Normal Nucleoli <= 2
        ->  class 2  [0.972]

Rule 3: (117/4, lift 2.2)
        Uniformity of Cell Size > 4
        ->  class 4  [0.958]

Rule 4: (128/6, lift 2.2)
        Uniformity of Cell Shape > 4
        ->  class 4  [0.946]

Rule 5: (128/6, lift 2.2)
        Bare Nuclei > 3
        Normal Nucleoli > 2
        ->  class 4  [0.946]
```



```
        Bare Nuclei > 3
        Normal Nucleoli > 2
        ->  class 4  [0.946]

Rule 6: (147/13, lift 2.1)
        Uniformity of Cell Size > 2
        Normal Nucleoli > 2
        ->  class 4  [0.906]

Default class: 2

Evaluation on training data (400 cases):

                 Rules
            ----------------
             No      Errors

              6   19( 4.8%)   <<

            (a)   (b)    <-classified as
            ----  ----
            216    13    (a): class 2
              6   165    (b): class 4

        Attribute usage:

              97%  Uniformity of Cell Size
              92%  Normal Nucleoli
              85%  Uniformity of Cell Shape
              81%  Bare Nuclei

Time: 0.1 secs
```

Conclusion:

In this research various algorithms are compared to predict the best classifier. It is found that among various classification techniques c5 outperforms of all other algorithms with highest accuracy rate. Therefore an efficient classifier is identified to determine the nature of the disease which is highly essential in a clinical investigation of life threatening disease like breast cancer.

Reference:

[1]. Kiran Amin" Sequential Sequence Mining Technique in Mammographic Information Analysis Database", ISSN 2250-2459, Volume 2, Issue 5, May 2012.

[2]. Harveen Buttar" Association Technique in Data Mining and Its Applications" *International Journal of Computer Trends and Technology (IJCTT) - volume4Issue4 –April 2013.*

[3]. AUTOMATIC CLASSIFICATION OF MAMMOGRAM MRI USING DENDOGRAMS" Asian Journal Of Computer Science And Information Technology 2: 4 (2012) 78– 81".

[4]. XindongWu," Top 10 algorithms in data mining", Knowl Inf Syst (2008) 14:1–37
DOI 10.1007/s10115-007-0114-2.

[5]. Herwanto" Association Technique based on Classification for Classifying Microcalcification and Mass in Mammogram", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No 2, January 2013.

[6]. Shweta Kharya" USING DATA MINING TECHNIQUES FOR DIAGNOSIS AND PROGNOSIS OF CANCER DISEASE", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.

[7]. DeepaS.Deshpande" ASSOCIATION RULE MINING BASED ON IMAGE CONTENT",*International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 143-146.*

[8]. Qiankun Zhao" Association Rule Mining: A Survey".

[9]. Yao Yao" SEGMENTATION OF BREAST CANCER MASS IN MAMMOGRAMS AND DETECTION USING MAGNETIC RESONANCE IMAGING"

[10]. D.Narain Ponraj," A Survey on the Preprocessing Techniques of Mammogram for the Detection of Breast Cancer", VOL. 2, NO. 12, December 2011 ISSN 2079-8407, Journal of Emerging Trends in Computing and Information Sciences.