ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (802-807), Month: April - June 2015, Available at: www.researchpublish.com

Predicting Breast Cancer Survivability Using Naïve Baysien and C5.0 Algorithm

[1]Mr. D R Umesh, [2]Thilak C R

[1]Asst. Professor (dept of CSE), [2]PG Student (CSE), P.E.S. College of Engineering, Mandya, India
An Autonomous Institution under Visvesvaraya Technological University, Belgaum

Abstract: **Breast cancer affects many people at the present time. The factors that cause this disease are many and cannot be easily determined. Additionally, the diagnosis process which determines whether the cancer is benign or malignant also requires a great deal of effort from a doctors and physicians. When several tests are involved in the diagnosis of breast cancer, such as clump thickness, uniformity of cell size, uniformity of cell shape,…etc, the ultimate result may be difficult to obtain, even for medical experts. This has given a rise in the last few years to the use of machine learning and Artificial Intelligence in general as diagnostic tools. In this paper, we analyses the performance of Naïve Baysien Classifier and C5.0 algorithm in predicting the survivable rate of breast cancer patients. The data set used for analysing is the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia database to evaluate the proposed system performances. These techniques helps the physician to take decisions on prognosis of breast cancer patients. At the end of analysis, C5.0 proves better performance than Naïve Baysien Classifier.**

Keywords: **Breast cancer, Naïve Baysien, C5.0 algorithm.**

1. INTRODUCTION

Breast cancer has become the most hazardous types of cancer among women in the world. The occurrence of breast cancer is increasing globally. Breast cancer begins in the cells of the lobules or the ducts 5-10% of cancers are due to an abnormality which is inherited from the parents and about 90% of breast cancers are due to genetic abnormalities that happen as a result of the aging process. According to the statistical reports of WHO, the incidence of breast cancer is the number one form of cancer among women . In the United States, approximately one in eight women have a risk of developing breast cancer. An analysis of the most recent data has shown that the survival rate is 88% after 5 years of diagnosis and 80% after 10 years of diagnosis. Hence, it can be seen from the study that an early diagnosis improves the survival rate. In 2007, it was reported that 202,964 women in the United States were diagnosed with breast cancer and 40,598 women in the United States died because of breast cancer.

A comparison of breast cancer in India with US obtained from Globocon data, shows that the incidence of cancer is 1 in 30. However, the actual number of cases reported in 2008 were comparable; about 1,82,000 breast cancer cases in the US and 1,15,000 in India. A study at the cancer Institute, Chennai shows that breast cancer is the second most common cancer among women in Madras and southern India after cervix cancer.

Early detection of breast cancer is essential in reducing life losses. However earlier treatment requires the ability to detect breast cancer in early stages. Early diagnosis requires an accurate and reliable diagnosis procedure that allows physicians to distinguish benign breast tumors from malignant ones without going for surgical biopsy.

Data mining techniques have been extensively applied for breast cancer diagnosis. Diagnosis is used to predict the presence of cancer and differentiate between the malignant and benign cases. In this paper, we have attempted to classify breast cancer data using C5.0 algorithm.

2. LITERATURE SURVEY

Delen et al[6], in their work, have developed models for predicting the survivability of diagnosed cases using breast cancer dataset Two algorithms artificial neural network (ANN) and C4.5 decision tree were used to develop prediction models. C4.5 gave an accuracy of 68.6% while ANN gave an accuracy of 66.2%. and the diagnosis was carried out based on nine chosen attributes.

Bellachia et al [2] uses the SEER data to compare three prediction models for detecting breast cancer. They have reported that C4.5 algorithm gave the best performance of 70.7% accuracy.

Endo et al [3] implemented common machine learning algorithms to predict survival rate of breast cancer patient. This study is based upon data of the SEER program with high rate of positive examples (18.5 %). Logistic regression had the highest accuracy, artificial neural network showed the highest specificity and J48 decision trees model had the best sensitivity.

Liu Ya-Qin et al proposed predictive models for breast cancer survivability using SEER data [7]. C4.5 decision tree algorithm was first used on the imbalanced data and then under sampling was applied to the models to overcome the disadvantage of imbalanced data. Bagging algorithm was then used to increase the performance of the classification for predicting breast cancer survivability. The results obtained showed an accuracy of 0.6978.

3. EXISTING SYSTEM

In existing system common machine learning algorithms like ID3 and C4.5 algorithms are used to predict survival rate of breast cancer patients.

Problems of Current System are:

- In this system Preprocessing functions like replacing ,missing values, normalizing numeric attributes and converting discrete attributes to nominal type is not considered.
- Handling continuous attributes.
- Choosing an appropriate attribute selection measure.
- Handling training data with missing attribute values.
- Handing attributes with differing costs.
- Improve computational efficiency.
- Prediction quality using these algorithms retained was low(70.7%).

4. PROPOSED SYSTEM

The processing steps applied to dataset are given in Figure 1.

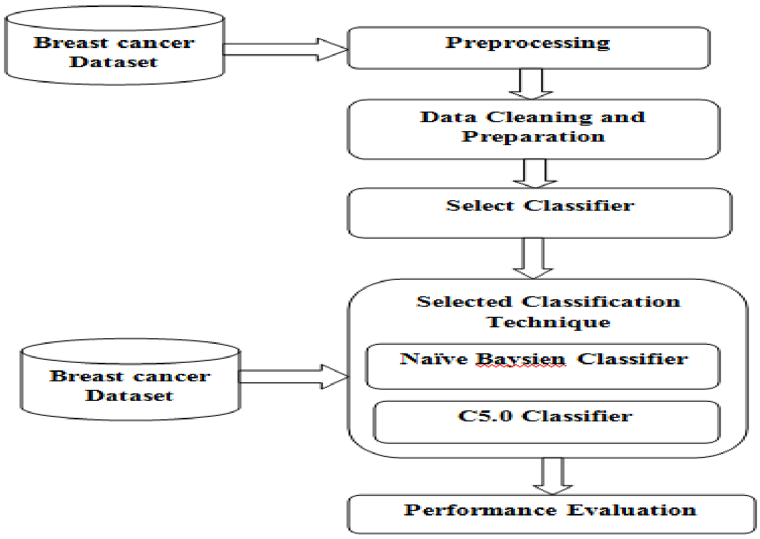

Fig.1 Processing Steps

Breast Cancer Dataset:

The data used in this study are provided by the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. My special thanks go to M. Zwitter and M. Soklic for providing the data for this research work. The data set has 10 attributes and total 286 rows, we restricted testing to these same attributes (see Table 1) and contain the following variables.

1. Age: patient's age at the time of diagnosis.

2. Menopause: menopause status of the patient at the time of diagnosis.

3. Tumor size: tumor size (in mm).

4. Inv-nodes: range 0 - 39 of axillary lymph nodes showing breast cancer at the time of histological examination.

5. Node caps: penetration of the tumor in the lymph node capsule or not.

6. Degree of malignancy: range 1-3 the histological grade of the tumor. That are

grade: 1 predominantly that consist of cancer cells,

grade: 2 neoplastic that consist of usual characteristics of cancer cells,

grade: 3 predominately that consist of cells that are highly affected.

7. Breast: breast cancer may occur in either breast.

8. Breast quadrant: if the nipple consider as a central point the breast may be divided into four quadrants.

9. Irradiation: patient's radiation (x-rays) therapy history.

10. Class: no-recurrence or recurrence depending reappearing symptoms of breast cancer in the patients after treatment.

Table 1: Breast Cancer Dataset

Attributes	Values
Age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
menopause	lt40, ge40, premeno
tumor-size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44,45-49, 50-54, 55-59
inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,27-29, 30-32,33-35, 36-39
node-caps	yes, no
deg-malig	1, 2, 3
breast	left, right
breast-quad	left-up, left-low, right-up, right-low, central
irradiation	Yes, no
class	no-recurrence-events, recurrence-events

Preprocessing:

Before the dataset is used, it needs to be properly preprocessed and a complete relevancy analysis needs to be completed. Preprocessing functions like replacing ,missing values, normalizing numeric attributes and converting discrete attributes to nominal type. Feature selection involves selecting the attributes that are most relevant to the classification problem. The method used in relevancy analysis is information gain ranker. Data pre-processing was applied to given dataset to prepare the raw data. Pre-processing is an important step that is used to transform the raw data into a format that makes it possible to apply data mining techniques and also to improve the quality of data. It can be noted from the related work, that attribute selection plays an important role in identifying parameters that are important and significant for proper breast cancer diagnosis. It was also found that the prediction quality was retained even with a small number of non-redundant attributes. As a first step, non cancer related parameters were identified and removed. For example, parameters relating to race, ethnicity etc. was discarded. The number of attributes removed in this process was 5 and the total number of attributes was reduced from 10 to 5. Next the attributes having missing values and noisy values are replaced. Then after data cleaning and feature selection, records were selected for further processing.

Table 2: Dataset Attributes after Preprocessing

S. No	Attribute	Description
1	Age	Patient's age at the time of diagnosis.
2	Menopause	Menopause status of the patient at the time of diagnosis.
3	Tumor- size	Tumor size (in mm).
4	Inv-nodes	Range 0 - 39 of auxillary lymph nodes showing breast cancer at the time of histological examination.
5	Node-caps	Penetration of the tumor in the lymph node capsule or not.

5. CLASSIFICATION METHODS

Naïve Baysein Classifier:

The Naive Bayes is a quick method for creation of statistical predictive models. NB is based on the Bayesian theorem. It is commonly used to solve prediction problems for ease of implementation and usage. This classification technique analyses the relationship between each attribute and the class for each instance to derive a conditional probability for the relationships between the attribute values and the class.

During training, the probability of each class is computed by counting how many times it occurs in the training dataset. This is called the "prior probability" P(C=c). In addition to the prior probability, the algorithm also computes the probability for the instance x given c with the assumption that the attributes are independent. This probability becomes the product of the probabilities of each single attribute. The probabilities can then be estimated from the frequencies of the instances in the training set.

$P (Ci \mid X) > P (Cj \mid X)$ for $1 \leq j \leq m, j \neq i$ (1)

To maximize $P (Ci \mid X)$, Bayes rule is applied as stated in Eq. (2)

$P (Ci \mid X) = P(X \mid Ci) P (Ci)$ (2)

$P (X)$

P (X) is constant for all classes and P (Ci) is calculated as in

Eq. (3),

P(Ci)= Number of training sample in a class(3)

Total number of training samples

To evaluate $P(X \mid Ci)$, the naïve assumption of class

conditional independence is used as in

Eq. (4),

n $P(X|Ci) = \Pi P (xk \mid Ci)$ (4)

k=1

The given sample X is assigned to the class Ci for which

$P(X \mid Ci) P (Ci)$ is the maximum 1.

C5.0 Classifier:

C5.0 Decision Tree algorithm is a software extension of the basic C4.5 algorithm designed by Quinlan recursively visits each decision node selecting the optimal split. The process is continued until no further split is possible 1. The algorithm uses the concept of information gain or entropy reduction to select the optimal split. Information gain is the increase in information produced by partitioning the training data according to the candidate split.

The C5.0 algorithm chooses the split with highest information gain as the optimal split. The information gain measure is used to select the best test attribute at each node in the tree. To avoid over fitting problem, C5.0 uses post pruning method and thus increases the accuracy of the classification. These include avoidance of over fitting the data; reduced error pruning, rule post-pruning, handling continuous attributes and handling data with missing attribute values. In testing phase we used training data with known result and. C5.0 algorithm was applied to obtain the rule set. In the testing phase, the classification rules obtained were applied to the whole pre-processed data. The results obtained are analysed.

Algorithm C5.0

Input: Example, Target Attribute, Attribute

Output: decision tree

Algorithm:

- Check for the base class
- Construct a DT using training data
- Find the attribute with the highest info gain (A_Best)
- For each ti ϵ D, apply the DT to determine its class Since the application of a given tuple to a DT is relatively straightforward.

The performance of a chosen classifier is validated based on error rate and computation time. The classification accuracy is predicted in terms of Sensitivity and Specificity. The computation time is noted for each classifier is taken in to account.

The evaluation parameters are the specificity, sensitivity, and overall accuracy.

The sensitivity or the true positive rate (TPR) is defined by TP / (TP + FN); while the specificity or the true negative rate (TNR) is defined by TN / (TN + FP); and the accuracy is defined by (TP + TN) / (TP + FP + TN + FN)

True positive (TP) = number of positive samples correctly predicted.

False negative (FN) = number of positive samples wrongly predicted.

False positive (FP) = number of negative samples wrongly predicted as positive.

True negative (TN) = number of negative samples correctly predicted.

These values are often displayed in a confusion matrix as be presented in Table 2. Classification Matrix displays the frequency of correct and incorrect predictions. It compares the actual values in the test dataset with the predicted values in the trained model.

Table 3: Confusion Matrix

	Predicted		
Actual	**Positive**	**Negative**	**Total**
Positive	193(TP)	8(FN)	201
Negative	62(FP)	23(TN)	85
Total	255	31	286

Table 4: Performances on Test Data

Method	**Accuracy**	**Sensitivity**	**Specificity**	**Error rate**	**Computation time**
Naïve Baysien	71.32	0.856	0.76	4.5	577 ms
C5.0	75.52	0.96	0.75	4.3	156 ms

Note- ms: mili seconds.

6. CONCLUSION

In this paper the performance of Naïve baysien Classifier and C5.0 analysis on University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia data set. The performance of C5.0 shows the high level compare with other classifiers. Therefore C5.0 decision tree is suggested for predict survivability of Breast Cancer disease based classification to get better results with accuracy, low error rate and performance.

7. FUTURE WORK

The performance of C5.0 shows the high level compare with other data mining algorithm. But in future C5.0 algorithm is suggested for predict survivability of Breast Cancer disease based classification to get better results with accuracy, low error rate and performance as compared to previous results.

REFERENCES

[1] American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).

[2] A.Bellachia and E.Guvan,"Predicting breast cancer survivability using data mining techniques", Scientific Data Mining Workshop, inconjunction with the 2006 SIAM Conference on Data Mining, 2006.

[3] A. Endo, T. Shibata and H. Tanaka (2008), Comparison of seven algorithms cancer survival, Biomedical Soft Computing and Human Sciences, vol.13, pp.11-16.

[4] Breast Cancer Wisconsin Data [online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancerwisconsin/breast-cancer-wisconsin.data.

[5] Brenner, H., Long-term survival rates of cancer patients achieved by the end of the 20th century: a period analysis. Lancet. 360:1131–1135, 2002.

[6] D. Delen, G. Walker and A. Kadam (2005), Predicting breast cancer survivability: a comparison of three data mining methods, Artificial Intelligence in Medicine.

[7] Santi Wulan Purnami, S.P. Rahayu and Abdullah Embong, "Feature selection and classification of breast cancer diagnosis based on support vector machine", IEEE 2008.

[8] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" IJCSE Vol. 3 No. 6 June.