R Analysis of SEER Breast Cancer Dataset Using Naive Bayes and C4.5 Algorithm

[1]Keerti Yeulkar, [2]Dr. Rahila Sheikh
[1,2]Dept. of Computer Science and Studies, Rajiv Gandhi College of Engineering, Chandrapur, Maharashtra, India

Abstract
Breast cancer is one of the deadliest disease and its early diagnosis can save numerous of lives.classification is a data mining technique that classifies the items in to class. In this paper we have targeted mainly on two techniques i.e Naive Bayes and C4.5. C4.5 has been applied to SEER dataset to classify tumor in to cancerous(malignant) and Non-cancerous(Bening) tumor. Pre-processing techniques has been applies to prepare relevant dataset for experimental analysis purpose.Random samples and R programming language has been used for diagnosis.The Experimental analysis and conclusion are presented and discussed.

Keywords
Breast Cancer, Classification, Naive Bayes, C4.5, Malignant, Bening, R programming, SEER

I. Introduction

Cancer is the most central element for death around the world. Prior determination of Breast Cancer spares tremendous lives. The uncontrolled growth of cell in breast leads to Breast cancer. Such cells often forms lumps or tumor.Tumor is formed by these cells,which can often be seen using X-ray.Breast cancer occurs almost entirely in women,but men can get it too. Breast cancer begins with forming tumour which can be either malignant i.e cancerous or benign i.e non cancerous in cells of breast forming clusters and spreads in the entire tissue.Cells in nearly any part of the body can become cancer, and can spread to other areas of the body. Cancer research is generally clinical and/or biological in nature, data driven statistical research has become a common complement. Predicting the outcome of a disease is one of the most interesting and challenging tasks where to develop data mining applications.The SEER database combines patient-level information on cancer site, tumor pathology, stage, and cause of death .This paper presents data mining techniques to predict the survivability rate of breast cancer patients. In our study, we have used the SEER data and have introduced a pre-classification approach that take into account variables: BEHAVIOUR CODE ICD-O-3,In Situ Cases,Survival Time Recode (STR), Vital Status Recode (VSR), and Cause of Death (COD).

The present research project contains the application of different machine learning techniques to cancer prognosis. Some of the obvious trends which account for the motivation of experiments presented in this manuscript include:

1. Probabilistic learning method (Naive Bayes text classification) can be said as Bayesian classification. Naive Bayes classifiers are among the most successful known algorithms for learning,it is used to classify text documents. posterior probability P(c|d) from P(c), P(d) and P(d|c) can be calculated using Naive Bayes.

$$P(c|d)=P(d|c)P(c)/P(d)$$

The Representation of supervised learning method as well as a statistical method for classification is Bayesian Classification .It assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems [1].

2. C4.5 builds decision trees from a set of training data in the same way as ID3. This algorithm has a few base cases.

- Samples in the list is of the same class. a leaf node is created for the decision tree saying to choose that class.
- Information gain is not provided by any features. In such case using the expected value of the class , C4.5 creates a decision node.

II. Related Work

1. "Predicting Breast Cancer Survivability Using Data Mining Techniques" by Abdelghani Bellaachia, Erhan Guven In this paper author presented an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques. SEER Public-Data is used. The preprocessed data set consists of 151,886 records, which have all the available 16 fields from the SEER database. They investigated three data mining techniques: the Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. Several experiments were conducted using these algorithms. However, they found out that C4.5 algorithm has a much better performance than the other two techniques [6].
2. An analysis of SEER Dataset for breast cancer diagnosis using C4.5 Classification Algorithm is carried out by Rajesh et al. in [7]. In this research, the C4.5 classification algorithm has been applied to SEER breast cancer dataset to classify patients into either "Carcinoma in situ" (beginning or precancer stage) or "Malignant potential" group.

III. Methodology

National Cancer Institute (NCI) runs program called SEER (Surveillance, Epidemiology and End Results. The Cancer Incidences record based on United States populations are recorded by SEER database. SEER research data includes SEER incidence and population data related to sex,age, race, year of diagnosis, and geographic areas [4]. All types of cancer for the year 2008 to 2014 are recorded by SEER, out of which 700 records samples pertaining to breast cancer has been used. The R programming language has been used to the analysis of these data, R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others [5]. In this paper, we have investigated two data mining techniques: the C4.5 and the Naive Bayes algorithm,the Naive Bayes algorithm is combined with laplace and metric prediction for experimental analysis purpose. In this paper, we used these algorithms to predict

the accuracy rate of algorithm for SEER breast cancer data set. We selected these two classification techniques to find the most perfect one for predicting cancer survivability rate. The Naïve Bayes technique depends on the famous Bayesian approach following a simple, clear and fast classifier. Due to the fact that it assumes mutually independent attributes it is called "Naive". This method has been used in many areas to represent, utilize, and learn the probabilistic knowledge and significant results have been achieved in machine learning. The second technique is the C4.5 decision-tree generating algorithm. C4.5 is based on the ID3 algorithm. It has been shown that the last technique have better performance [3]. Therefore we have included this in our analysis.

The Processing Steps that has been applied to SEER data involves:

- Pre-processing
- Preliminary classification
- Selecting Classification technique
- Applying techniques to the test data
- Performance Evaluation

1. Pre-processing

Pre-processing is an important step that involves the processing of raw dataset and converting it in to suitable format to which the necessary datamining techniques can be applied. Here attribute selection plays main role that is important for identifying the proper breast cancer diagnosis. These attribute can classify the breast cancer tumor in to Bening(Non-Cancerous) and malignant (Cancerous).

Steps involved in grouping relevant data is all about removing some of the non cancer related data such as Race, Social or demographic condition. Parameters related to ethnicity, Race is discarded. The total number of attribute removed in this process is 17. Parameters such as EOD TUMOR SIZE has no value so such records has been removed. The EOD field is composed of five fields including the EOD code. These fields (size of tumor, number of positive nodes, number of nodes and number of primaries) contain missing information coded such as '999', '99' or '9' representing the 'unknown' information. In this way the attributes which were re-coded were removed after this process the select attributes became 11.

After preprocessing the SEER continuous attributes:

- AGE AT DIAGNOSIS
- REGIONAL NODES POSITIVE
- SEQUENCE NODES POSITIVE
- SEQUENCE NUMBER CENTRAL
- CS TUMOR SIZE
- CS EXTENSION

2. Preliminary Classification

Next step is preliminary classification in which we have applied selected classification technique and the target class is BEHAVIOR CODE ICD-O-3in which the value 1 denotes "Malignant potential" and value 2 denotes "carcinoma in situ" condition.

The classification rules used :
If CS EXTENSION< 35.000
Then BEHAVIOR= Malignant potential
If CS EXTENSION> 35.000
Then BEHAVIOR= Carcinoma In Situ

Table 1: Comparison of Classifiaction Technique

S.No	Methods	Accuracy
1	C4.5	98.09
2	Naive Bayes	95.85
3	Naive Bayes (Laplace)	95.85
4	Naive Bayes (Metric Prediction)	50.5

It has been observed that the C4.5 has the higher accuracy rate when compared with other classification techniques.

3. Applying Classification Technique

The further experimental analysis has been performed using C4.5 classification technique using the following BEHAVIOR CODE ICD-O-3 rule

(a). If Regional Nodes Positive<96.5000

- CS LYMPH NODES<25.000
- CS TUMOR SIZE<47.5000
- TUMOR MARKER SIZE(1,2,3)>5
- Then BEHAVIOR= carcinoma In Situ

(b). If Age at Diagnosis>=47,600

- CS TUMOR SIZE<70 then BEHAVIOR=Malignant potential
- CS TUMOR SIZE>=70.000 then BEHAVIOR=carcinoma in situ

(c). If STR>=60 months and VSR=alive then "alive"

else if STR<60 months and COD is breast cancer then "not alive"

else

Ignore the record

End if

IV. Performance Evaluation and Results

In this study performance metric of the C4.5 has been calculated. These metrics can be derived from the confusion matrix and can be easily converted to true-positive (TP) and false-positive (FP).

Table 2: Confusion Matrix

	BENING	MALIGNANT	TOTAL
BENING	436(TP)	08(FN)	444
MALIGNANT	05(FP)	234(TN)	239
TOTAL	441	242	683

In this research work, we used three performance measures: accuracy, sensitivity and specificity;

1. Accuracy is the percentage of records correctly classified out of total records

Accuracy = TP+ TN /FP +FN+TP+ TN
Accuracy=436+234/05+08+436+234
=0.9809

2.Sensitivity is the percentage of positive records classified correctly out of all positive records,

Sensitivity = TP/ FN+TP
Sensitivity = 436/08+436
=0.6507

3.Specificity is the percentage of positive records classified correctly out of all positive records

Specificity = TN/ FP+TN

Specificity=234/05+234

=0.9790

where TP, TN, FP and FN denotes true positives, true negatives, false positives and false negatives, respectively.

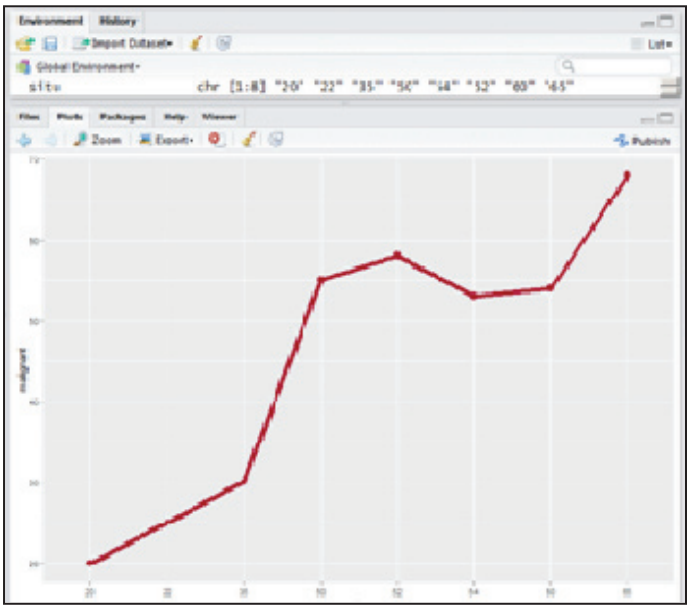

Fig. 1: R Analysed Yearly Diagnosed Result of Malignat and Bening Cases

V. Conclusion & Future Work

This research work attempts to analyze the datamining classification techniques for the SEER dataset and classify the SEER breast cancer dataset in to "malignant" and "bening" cases.The analysis of the data mining techniques concludes that the C4.5 algorithm has higher accuracy.The behaviour of this algorithm has been analyzed with experimental results using R programminglanguage.The C4.5 algorithm is then used with random sample of 700 records of SEER datasets and we obtained the accuracy of 98.09%.Further enchancement of this work can include the improvement in C4.5 algorithm ,change in the value of metric prediction with naive bayes approach or the selection of different attributes from the provided SEER dataset.

References

[1] [Online] Available: http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab4-NaiveBayes.pdf

[2] [Online] Available: https://en.wikipedia.org/wiki/C4.5_algorithm

[3] K.Rajesh, Dr. Sheila Anand,"Analysis of SEER Dataset for Breast cancer diagnosis using C4.5 Classification algorithm", International journal of Advanced research in computer and communication Engineering Vol. 1, Issue 2, April 2012.

[4] [Online] Available: https://seer.cancer.gov/resources/

[5] [Online] Available: https://en.wikipedia.org/wiki/R_(programming_language)

[6] Bellaachia, Abdelghani, Erhan Guven,"Predicting breast cancer survivability using data mining techniques", Vol. 58, Issue 13, pp. 10-110, 2006.

[7] Rajesh K., Sheila Anand,"Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm", Int. Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 2, pp. 72-77, 2012.

[8] Khan M.U., Choi J.P., Shin H. Kim M,"Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare", Conf Proc IEEE Eng Med Biol Soc., pp. 48-51, 2008.

[9] Choi J.P., Han T.H., Park R.W.,"A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis", J Korean Soc Med Inform, pp. 49-57, 2009.

[10] Chi C.L., Street W.H., Wolberg W.H.,"Application of Artifical Neural Network- based Survival Analysis on Two Breast Cancer Datasets", Annual Symposium Proceedings / AMIA Symposium, 2007.

[11] Sudhir D., Ghatol Ashok A., Pande Amol P,"Neural Network aided Breast Cancer Detection and Diagnosis", 7th WSEAS International Conference on Neural Networks, 2006.

[12] [Online] Available: https://seer.cancer.gov/

[13] [Online] Available: https://www.google.co.in/search?q=DATA+PREPROCESSING+IN+DATAMINING&ie=utf-8&oe=utf-8&client=firefox-b-ab&gfe_rd=cr&ei=7Qm9WPb4LfHx8AfCzaVQ

[14] [Online] Available: http://iasri.res.in/ebook/win_school_aa/notes/Data_Preprocessing.pdf

[15] [Online] Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html

[16] [Online] Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier