iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** III **Month of publication:** March 2017

DOI: http://doi.org/10.22214/ijraset.2017.3074

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Utilization of Data Mining Techniques for Analysis of Breast Cancer Dataset Using R

Keerti Yeulkar[1], Dr. Rahila Sheikh[2]
[1]PG Student, [2]Head of Computer Science and Studies
Rajiv Gandhi College of Engineering, Chandrapur

Abstract: ***for the diagnosis of cancer medical professionals need an accurate & reliable prediction techniques. There are various techniques that is being used for the diagnosis purpose. Classification is a data mining function that assigns items in a collection to target groups or classes. Two algorithms c4.5 and naive bayes has been applied to the breast cancer dataset to analyse the accuracy of algorithm. Pre-processing techniques have been applied to prepare the formatted dataset from the raw dataset and identify the relevant attribute for classification. Test samples has been randomly selected from the dataset. The results are presented and discussed.***
Keywords: c4.5 algorithm, naive bayes algorithm , chi.squared selection process, seer dataset, breast cancer diagnosis.

I. INTRODUCTION

Data mining (DM) comprises the core algorithms that enable to gain fundamental insights and knowledge from massive data. Data mining is all about the larger knowledge discovery process which involves analyzing Breast cancer. One of the new researches in data mining application, deadliest disease and most common of all cancers in the leading cause of cancer deaths in women worldwide. Classification plays a vital role in data mining research. 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. Data mining includes the fields like machine learning, statistics, pattern recognition, artificial intelligence. There are many different types of breast cancer, with different stages or spread, aggressiveness, and genetic makeup. Survival rates for breast cancer may be increased when the disease is detected in its earlier stage[1].

To detect whether the patient had breast cancer or not the patient's history is used. This information is about their menopause condition, age, age of menopause, history of breast cancer or family member having cancer other than breast cancer. These may be the reason of breast cancer. High probability of breast cancer Patient who has first degree relative with family member having cancer .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 happens as a result of the aging process**.**

II. RELATED WORK

Shiv Shakti Shrivastava[2] et al. in his work he made a conclusion that neural network and decision approaches are mostly used by various researchers to create a predictive model and decision rules from the breast cancer data. Dataset was taken from UCI machine learning data repository. Dataset consist 10 attributes and 699 instances.

J. Padmavati(2011) [3] performed a comparative study on WBC dataset for breast cancer prediction using RBF and MLP along with logistic regression. Logistic regression was performed using logistic regression in SPSS package and MLP and RBF were constructed using MATLAB. Observed results were neural networks took slightly higher time than logistic regression.

Santi Wulan Purnami et al. Research study used Support Vector Machines (SVM)[4].Their conclusion is SVM is a new algorithm of data mining technique that has received popularity in machine learning community. Their paper emphasizes how 1-norm SVM can be used in feature selection and smooth SVM (SSVM) for classification.Implementation was a breast cancer diagnosis like First, feature selection for support vector machines was utilized to determine the important features.SSVM was used to classify the state as(benign or malignant) of breast cancer. As a result, SVM can achieve the state of the art performance on feature selection and classification.

A research paper by Abdelghani Bellaachia and Erhan Guven, presents an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques [5]. In this paper, they used the SEER Public-Use Data and the preprocessed data set

consists of 151,886 records, available with 16 fields from the SEER database. They have analyzed the SEER data set using three data mining techniques namely Naïve Bayes, back-propagated neural network, and the C4.5 decision tree algorithms. Several experiments were conducted using these algorithms. Finally, they conclude that C4.5 algorithm has a much better performance than the other two techniques

A research work by Sujatha G. and K. Usha Rani, survived on effectiveness of data mining techniques on cancer data sets [6]. They state that the tumor is an abnormal cell growth that can be either Benign or Malignant. The use of machine learning and data mining techniques has revolutionized the whole process of cancer diagnosis. Many researchers contributed their effective and accurate diagnosis of the breast cancer diseases in various data mining as basic techniques and various review and technical articles on Tumor and Breast cancer data sets.

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. Pre-processing techniques have been applied to prepare the raw dataset and identify the relevant attributes for classification. Random test samples have been selected from the pre-processed data to obtain classification rules. The rule set obtained was tested with the remaining data.

Application of data mining techniques to model breast cancer data is explored by Syed Shajahaan et al. in [8]. In this work, they explore the applicability of decision trees to predict the presence of breast cancer. Also it analyzes the performance of conventional supervised learning algorithms viz. Random tree, ID3, CART, C4.5 and Naive Bayes. Experimental results prove that Random Tree serves to be the best one with highest accuracy. It is found that among various classification techniques random tree outperforms of all other algorithms with highest accuracy rate.

Muhammad Umer Khan et al(2008) [9] they investigated a hybrid scheme based on fuzzy decision trees on SEER data, they performed experiments using different combinations of number of decision tree rules, types of fuzzy membership functions and inference techniques. They compared the performance of each for cancer prognosis and found hybrid fuzzy decision tree classification is more robust and balanced than the independently applied crisp classification.

Jong Pill Choi et al (2009) [10] they compared the performance of an Artifical Neural Network, a Bayesian Network and a Hybrid Network used to predict breast cancer prognosis. The hybrid Network combined both ANN and Bayesian Network. The Nine variables of SEER data which were clinically accepted were used as inputs for the networks. The accuracy of ANN (88.8%) and Hybrid Network (87.2%) were very similar and they both outperformed the Bayesian Network. They found the proposed Hybrid model can also be useful to take decisions.

Chih-Lin Chi et al(2007)[11] they used the Street's ANN model for Breast Cancer Prognosis on WPBC data and Love data. In their research they used recurrence at five years as a cut point to define the level of risk. The applied models successfully predicted recurrence probability and separated patients with good (>5 yrs) and bad(<5yrs) prognoses.

Sudhir D. Sawarkar et al(2006) [12] in their study they applied SVM and ANN on the WBC data .The results of SVM and ANN prediction models were found comparatively more accurate than the human being. The 97% high accuracy of these prediction models can be used to take decision to avoid biopsy.

III. PROPOSED METHOD

A. Seer Dataset

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute provides information on cancer statistics[13]. SEER is used for data mining and classification .SEER database has been used for cancer statistics in the United States. The dataset contains the cancer cases records for the period 1973-2013.The process involved in manipulating SEER dataset contains data pre-processing ana classification using algorithms.

Volume 5 Issue III, March 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

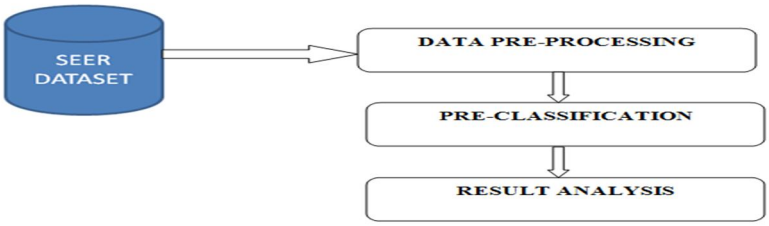


Fig: Processing steps

B. Data Pre-Processing

One of the Data mining technique involves Data preprocessing which involves transforming raw data into a desired understandable format. Some Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a effective method of resolving such issues.[14]

Noisy dataset needs several levels of pre-processing such as:

Data Cleaning

Data Integration

Data Transformation

Data Reduction

In this step the data pre-processing is applied to the SEER dataset to prepare data for the further processing. The data is then transformed in to the particular format. In this step the data which is not that much relevant for the further processing is eliminated from the record and formatted in a particular format. Data cleansing is the very important step in any data mining related tasks. Here, the first step to eliminate non-cancerous data.For example the data related to the Social factor, color, racial, geographical conditions are removed. Near about 5-6 irrelevant attributes has been removed from the dataset.

C. Pre-Classification

Next step is carried out for the primary classification of the 699 data records, the algorithm is designed that checks the condition of malignancy and being stage tumor. The attribute selection process has been carried out using chi.squared selection process.

```
weights <- chi.squared(CLASS~., data)
> print(weights)
                        attr_importance
SNO                     0.2478952
ID                      0.0000000
CLUMP_THICKNESS         0.7370529
CELL_SIZE               0.8739176
CELL_SHAPE              0.8583576
MARGINAL_ADHESION       0.7404031
SINGLE_EPI              0.7957127
BARE_NUCLEI             0.8434223
BLAND_CHROMATIN         0.8071249
NORMAL_NUCLEOI          0.7706075
MITOSIS                 0.5238247
AGE                     0.6576498
subset <- cutoff.k(weights, 9)
> f <- as.simple.formula(subset, "CLASS")
> print(f)
CLASS ~ CELL_SIZE + CELL_SHAPE + BARE_NUCLEI + BLAND_CHROMATIN +
 SINGLE_EPI + NORMAL_NUCLEOI + MARGINAL_ADHESION + CLUMP_THICKNESS +     MITOSIS
```

Table i: Attribute Classification using chi.squared selection process.

STEP 1: Start the pointer with the very first cell and apply the below conditions
condition1: If the clump thickness>=5 && cell Size>=5 ,then the cancer is “malignant”;
consider AGE between18-45,menopause condition for AGE>=45
condition2: if else Bare Nuclei>=5 && Mitosis>=5,then the cancer is “malignant”.
consider AGE between18-45,menopause condition for AGE>=45
Condition 3: if all the cases are below 5 then check for clump thickness and cell shape,
STEP 2: If clump thickness >=5 && cell shape>=5 then the cancer is malignant. consider AGE between18-45,menopause condition for AGE>=45
STEP 3:Else consider the case as “Benign cancer”(NON-CANCEROUS CELL)
End If
End If
End If

Table Ii Classification Rule For 699 Records

D. Result Analysis

The result is analysed using C4.5 and Naive Bayes algorithm using R. C4.5 algorithm builds decision trees from a set of training data , using the concept of information entropy. The training data is a set of classified sample datasets. An open source Java implementation of the C4.5 algorithm in the Weka data mining tool is J48 [15].

Prior knowledge and current evidence can be used to make prediction for Bayes theorem. With evidence, the prediction is changed.The prediction is the probability that investigators are interested in Bayes theorem is formally expressed by the equation $P(A|B)=P(B|A)\times P(A)P(B)$ where P(A) and P(B) are probability of events A and B without regarding each other. P(A|B) is the probability of A conditional on B and P(B|A) is the probability of B conditional on A[16].

S.NO	C4.5	NAIVE BAYES
1	98.0966%	95.8523%

Table Iii Comparative Result Analysis Of C4.5 & Naive Bayes Algorithms

IV. CONCLUSION

In this research work, we have tried to classify the SEER dataset in to “malignant tumor” and “benign tumor” using C4.5 and Naive Bayes algorithm. We used the random sample of 699 records by applying decision tree classification rule. We obtained the accuracy of 98.0966% accuracy from C4.5 and 95.8523% accuracy from Naive Bayes algorithm. We concluded that the performance of C4.5 algorithm is better than Naive Bayes algorithm.

V. ACKNOWLEDGEMENT

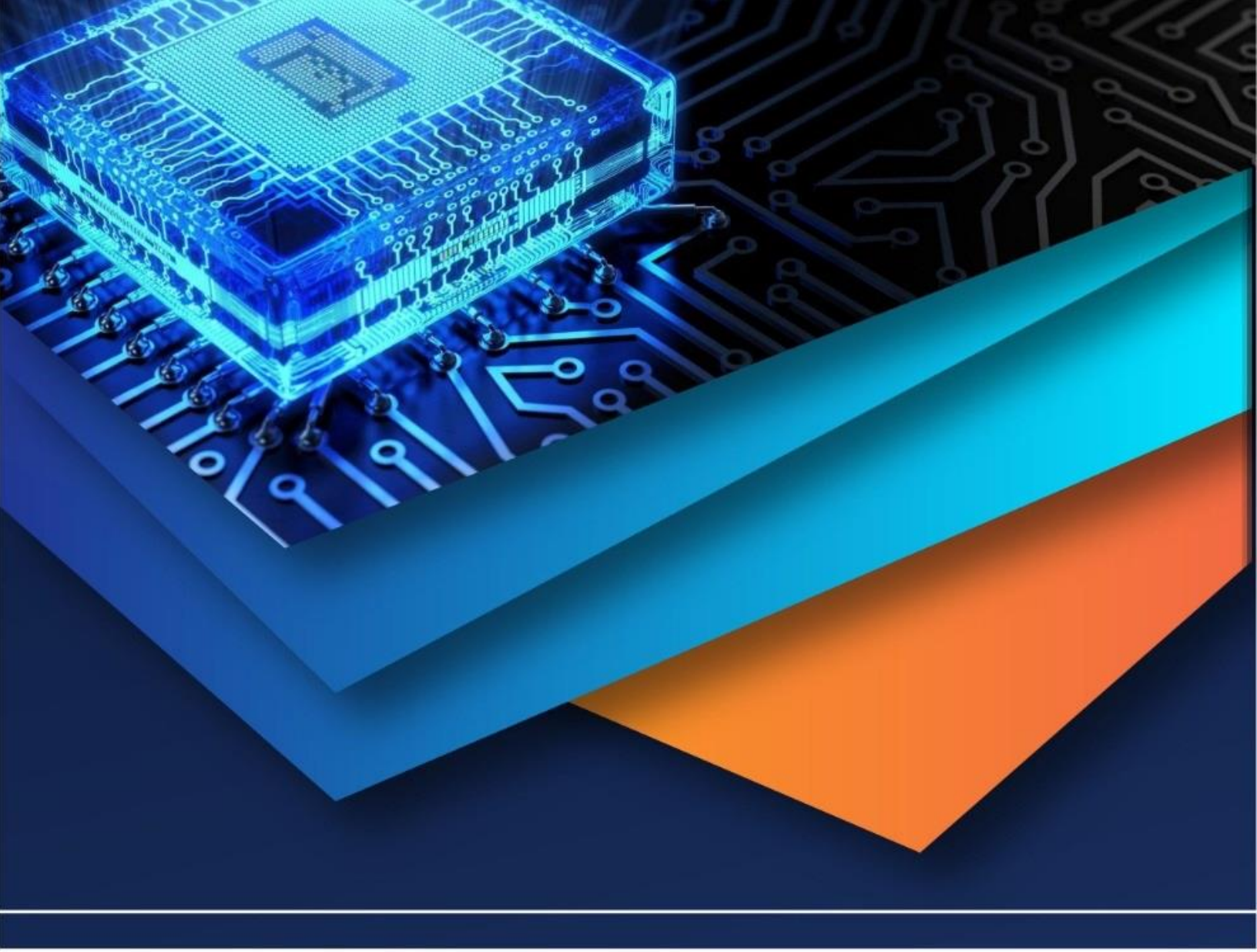
We are sincerely thankful to the National Cancer Institute, USA for permitting us to use SEER cancer database.

REFERENCES

[1] www.arpapress.com/Volumes/Vol10Issue1/IJRRAS_10_1_02.pdf
[2] http://accentsjournals.org/PaperDirectory/Journal/IJACR/2013/12/38.pdf
[3] Padmavati J. (2011), ―A Comparative study on Breast Cancer Prediction Using RBF and MLP,‖ International Journal of Scientific & Engineering Research, vol. 2, Jan. 2011.
[4] http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Santi%20Wulan%20Purnami.QT.&newsearch=true
[5] Bellaachia, Abdelghani, and Erhan Guven, "Predicting breast cancer survivability using data mining techniques", Age, Vol. 58, Issue 13, 2006, pp. 10-110.
[6] Sujatha, G., And K. Usha Rani. “A Survey On Effectiveness Of Data Mining Techniques On Cancer Data Sets”, Int. Journal of Engineering Sciences Research, 2013, Vol. 04, Issue 1, pp. 1298-1304
[7] Rajesh K., and 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, 2012, pp. 72-77.
[8] Syed Shajahaan. S, S. Shanthi, V. ManoChitra, “Application of data mining techniques to model breast cancer data”, International Journal of Emerging Technology and Advanced Engineering, Vol. 3, Issue 11, 2013, pp. 362-369

Volume 5 Issue III, March 2017
ISSN: 2321-9653

[9] Khan M.U., Choi J.P., Shin H. and Kim M (2008), ―Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare‖, Conf Proc IEEE Eng Med Biol Soc., 2008, pp. 48-51.
[10] Choi J.P., Han T.H. and Park R.W.(2009), ― A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis‖, J Korean Soc Med Inform, 2009, pp. 49-57.
[11] Chi C.L., Street W.H. and Wolberg W.H.(2007), ―Application of Artifical Neural Network- based Survival Analysis on Two Breast Cancer Datasets‖, Annual Symposium Proceedings / AMIA Symposium, 2007.
[12] Sudhir D., Ghatol Ashok A., Pande Amol P(2006)., ―Neural Network aided Breast Cancer Detection and Diagnosis‖,7 th WSEAS International Conference on Neural Networks, 2006
[13] https://seer.cancer.gov/
[14] 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
[15] http://iasri.res.in/ebook/win_school_aa/notes/Data_Preprocessing.pdf
[16] http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html
[17] https://en.wikipedia.org/wiki/Naive_Bayes_classifier



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY