International Journal of Recent Engineering Research and Development (IJRERD)
Volume No. 02 – Issue No. 02, ISSN: 2455-8761
www.ijrerd.com, PP. 13-19

CLASSIFICATION OF BREAST CANCER DATA USING C4.5 CLASSIFIER ALGORITHM

A.Kathija[1], S. Shajun Nisha[2], Dr .M .Mohamed Sathik[3]

[1]M.Phil. (PG Scholar) Dept of Computer Science, Sadakathullah Appa College, Tirunelveli, Tamil Nadu, India
[2]Prof& Head, P.G Dept of Computer Science, Sadakathullah Appa College, Tirunelveli, Tamil Nadu, India
[3]Principal, Sadakathullah Appa College, Tirunelveli, TamilNadu, India

Abstract: The classification of breast cancer patients is of great importance in cancer diagnosis. Breast cancer is the most common neoplasm in women worldwide and one of the leading causes of cancer-related death in women, with approximately 1.38 million new cases and 458,000 deaths each year around the world. To handle this type of situations we have to examine the breast tissue. Machine learning is fast growing field in computer science which provides better prediction methodologies for diseases in health care management, hence it was applied in the area of the breast cancer and lot of results produced by several researchers. Early detection of breast cancer is far easier to cure. This paper presents a decision tree based data mining technique for early detection of breast cancer. To find the performance of classification algorithms, we used the Wisconsin Diagnostic Breast Cancer (WDBC) datasets with C4.5 classifiers. This work concludes the best algorithm for the chosen input data on decision tree supervised learning algorithms to predict the best classifier.

Keywords: Breast Cancer, Wisconsin Diagnostic of Breast Cancer, Decision Tree, C4.5 Algorithm

1. INTRODUCTION

Breast cancer is the main leading cause of death for the woman in world. It is observed that early detection of malignancy can help in the diagnosis of the disease for woman and it can help strongly to enhance the expectancy of survival. For the detection of breast cancer, various techniques are used in which mammography is the most promising technique and used by radiologist frequently. The classification of breast cancer data can be useful to predict the outcome of some diseases or discover the genetic behaviour of tumors.

This paper looks at the breast cancer diagnosis problem using the Wisconsin Diagnostic Breast Cancer (WDBC) data set which is available publicly on the web [18]. The data set involves recordings from a Fine Needle Aspirate (FNA) test. The aim of the classification is to provide a distinction between the malignant and the benign masses. Some of the common classification methods used in data mining is: decision tree classifiers, Bayesian classifiers, k-nearest-neighbor classifiers, case-based reasoning, genetic algorithms, rough sets, and fuzzy logic techniques. Among these classification algorithms decision tree algorithms is the most commonly used because of it is easy to understand and cheap to implement. In decision tree algorithm C4.5 has additional features such as tree pruning, improved use of continuous attributes, missing values handling and inducing ruleset. In this paper it analyze the performance of supervised learning algorithm such as Decision trees algorithm is used for classifying the breast cancer dataset WDBC, Breast tissue from UCI Machine learning depository . Finally, the Decision Tree created for the UCI Breast Cancer (Wisconsin) dataset using the C4.5 algorithm. 10-fold cross validation [4] is used to prepare training and test data. After data pre-processing, the C4.5 algorithm is employed on the dataset after which data are divided into "benign" or "malignant" depending on the final result of the decision tree that is constructed.

2. RELATED WORK

Classification is a data mining technique based on machine learning which is used to classify each item in a set of data into a set of predefined classes or groups [3]. In the paper [12], by Sujatha and Usha Rani have proposed evaluation of Decision Tree Classifiers it is observed that C4.5 performs well for tumor datasets. In the paper [10] by Liu Ya-Qin et.al have experimented on breast cancer data using C5 algorithm with bagging to predict breast cancer survivability. In the paper [11] by Lavanya and Usha Rani have proposed classification of medical data they employed decision tree algorithm because it produce human readable classification rules which are easy to interpret. In the paper [17] by Badr HSSINA et.al, have proposed a comparative study of decision tree ID3 and C4.5 algorithm which led us to confirm that the most powerful and preferred method in machine learning is certainly C4.5

In the paper [8] by Bellachia et al have proposed the SEER data to compare three prediction models for detecting breast cancer. They have reported that C4.5 algorithm gave the best performance.In the paper [15] by Kavitha and DoraiRangasamy have proposed the performance of Naïve baysein Classifier and C4.5 analysis on SEER data set in survivability of breast cancer is done. The performance of C4.5 shows the high level

compare with other classifiers. In the paper [13] by Syed Shajahaan, et al. have proposed the application of the decision tress in order to predict the presence of the breast cancer and also, the performance measurement of conservative supervised learning algorithms via, CART, C4.5, ID3 and Naive Byes. In the paper [16] by Venkatesan and Velmurugan have proposed the performances in terms of classification accuracy of J48, AD Tree, BF Tree and regression trees (CART) algorithms. In the paper [14] by Ahmad LG et.al have proposed machine learning techniques, i.e., Decision Tree (C4.5), Support Vector Machine (SVM), and Artificial Neural Network (ANN) to develop the predictive models.

In the paper [6] by Kemal Polat et al, proposed a new classification algorithm feature selection-Artificial Immune Recognition System (FS-AIRS) on breast cancer data set. To reduce the data set C4.5 decision tree algorithm is used as a feature selection method. In the paper [9] by Deisy.C et al experimented breast cancer data using three feature selection methods Fast correlation based feature selection, Multi thread based FCBF feature selection and Decision dependent-decision independent correlation further the data is classified using C4.5 decision tree algorithm. In the paper [2] by Mark A. Hall et al have done experiments on various data sets using Correlation based filter feature selection approach further the reduced data is classified using C4.5 decision tree algorithm.

3. MOTIVATION AND JUSTIFICATION

Breast cancer is a leading cause of cancer deaths among women. For women in US and other developed countries, it is the most frequently diagnosed cancer. Efficient detection is the most effective way to reduce mortality, and currently a screening programme based on mammography is considered one the best and popular method for detection of breast cancer. Classification is one of the most studied problems in machine learning and data mining [7] [5]**.** Building accurate and efficient classifiers for Medical databases is one of the essential tasks of data mining and machine learning research. Building effective classification systems is one of the central tasks of data mining**.** Decision tree induction is a very popular and practical approach for pattern classification. There are several algorithms to classify the data using decision trees. The frequently used decision tree algorithms are ID3, C4.5 and CART [19]. These classifiers provide support for many health care areas in decision making. Out of these C4.5 has been proved to the best classifier for medical data. . It develops the classification model as a decision tree. Speed of C4.5 is significantly faster than ID3 (it is faster in several orders of magnitude)**.** C4.5 is more memory efficient than ID3**.** Size of decision Trees in C4.5 gets smaller decision trees. Ruleset of C4.5 can give ruleset as an output for complex decision tree. Missing values of C4.5 algorithm can respond on missing values by **_**?'**.** C4.5 solves overfitting problem through reduce error pruning technique.

Motivated by all these facts, it's recommended to classify the WDBC dataset by using classification algorithm. Hence it justify that the classification of breast cancer dataset with C4.5 algorithm is suitable for this application.

4. ORGANIZATION OF THE PAPER

The remaining paper is organized as follows: - Section 5 define proposed algorithm which includes outline of the framework, Section 6 includes performance Evaluation, Section 7 includes Experimental results and Section 8 includes conclusion of the paper.

5. PROPOSED ALGORITHM

5.1. Outline of the Proposed Work

The processing steps applied to WDBC data are given in Figure I.

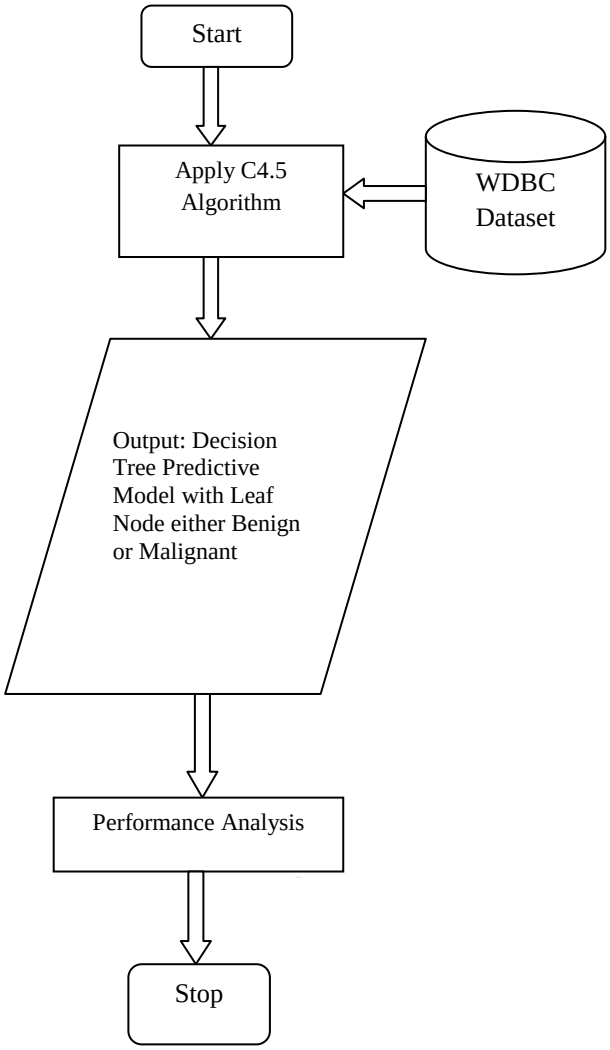

Fig. I Processing Steps

5.2. C4.5 Algorithm:

C4.5 algorithm is an improvement of IDE3 algorithm, Developed by Quinlan Ross in 1986 [1]. It is based on Hunt's algorithm and also like IDE3, it is serially implemented. Pruning takes place in C4.5 by replacing the internal node with a leaf node thereby reducing the error rate. Unlike IDE3, C4.5 accepts both continuous and categorical attributes in building the decision tree. It has an enhanced method of tree pruning that reduces misclassification errors due noise or too much detail in the training data set. Like IDE3 the data is sorted at every node of the tree in order to determine the best splitting attribute. C4.5 uses gain ratio as an attribute selection measure to build a decision tree. The root node will be the attribute whose gain ratio is very high. C4.5 uses pessimistic pruning for deleting of unnecessary branches in the decision tree due to that accuracy was increased.

ALGORITHM C4.5
Input: Example, Target Attribute, Attribute
Output: Classified Instances
Pseudocode: C4.5 (described by Quinlan) general algorithm for building decision trees is:
1. Check for any base cases
2. For each attribute a
3. Find the normalized information gain from splitting on a
4. Let a_best be the attribute with the highest normalized information gain
5. Create a decision node that splits on a_best
6. Recur on the sublists obtained by splitting on a_best, and add those nodes as children of node.

Base cases are the following:
1. All the examples from the training set belong to the same class (a tree leaf labeled with that class is returned).
2. The training set is empty (returns a tree leaf called failure).
3. The attribute list is empty (returns a leaf labelled with the most frequent class or the disjunction of all the classes).

OUTPUT: decision tree which classifies the data correctly.

6. PERFORMANCE EVALUATION

6.1. Measures for performance evaluation

1. Accuracy=$\frac{TP + TN}{TP + FP + FN + TN}$
2. Sensitivity = $\frac{TP}{TP + FN}$
3. Specificity = $\frac{TN}{TN + FP}$
4. Positive Predictive Value: PPV =$\frac{TP}{TP + FP}$
5. Negative Predictive Value: NPV = $\frac{TN}{TN + FN}$
6. Receiver Operating Characteristic:

$$ROC = \frac{Sensitivity + Specificity}{2}$$

Where,
1. The recall or true positive rate (TP) is the proportion of positive cases that were correctly identified
2. The false positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive
3. The true negative rate (TN) is defined as the proportion of negatives cases that were classified correctly
4. The false negative rate (FN) is the proportion of positives cases that were incorrectly classified as negative
5. The accuracy (AC) is the proportion of the total number of predictions that were correct.
6. The Sensitivity or Recall the proportion of actual positive cases which are correctly identified.
7. The Specificity the proportion of actual negative cases which are correctly identified.
8. The Positive Predictive Value or Precision the proportion of positive cases that were correctly identified.
9. The Negative Predictive Value the proportion of negative cases that were correctly identified.

7. EXPERIMENTAL RESULTS

7.1. Data Description and Pre-Processing

The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used to differentiate benign (non-cancerous) from malignant (cancerous) samples. To evaluate the effectiveness of our method, experiments on WDBC is conducted. This database was obtained from the university of Wisconsin hospital, Madison from Dr. William H. Wolberg. This is publicly available dataset in the Internet.
Table 1 shows a brief description of the dataset that is being considered.

Table 1 Description of Breast Cancer Dataset

Dataset	NO. Of Attributes	No. Of Instances	No. Of Classes
Wisconsin Diagnosis Breast Cancer (WDBC)	11	699	2

Details of attributes present in the dataset are shown in Table 2.

Table 2 Attribute Information

No	Attribute	Domain
1.	Sample code number	id number
2.	Clump Thickness	1 -10
3.	Uniformity of Cell Size	1 -10

4.	Uniformity of Cell Shape	1 -10
5.	Marginal Adhesion	1 -10
6.	Single Epithelial Cell Size	1 -10
7.	Bare Nuclei	1 -10
8.	Bland Chromatin	1-10
9.	Normal Nucleoli	1-10
10.	Mitoses	1-10
11.	Class	(2 for benign, 4 for malignant)

- **Clump Thickness:** Monolayer grouping in benign and multi layer grouping for cancerous cells.
- **Marginal Adhesion:** Normal cells stick together while cancer cells lose their ability. This is also relating factor to a single epithelial cell size, which is enlarged for a malignant cell.
- **Bare Nuclei:** Benign tumors have nuclei, which are not surrounded by cytoplasm.
- **Bland Chromatin:** Cancer cells have coarse chromatin.
- **Mitoses:** Uncontrollable levels of mitoses (cell-division) are seen in cancer cells.

The dataset comprises of 699 instances of breast cancer patients with each, either having malignant or benign type of tumor.

7.2. Performance Evaluation

1. Performance Evaluation of C4.5 Algorithm are shown in Table 3.

Table 3 Performance Evaluation

C4.5 Algorithm Performance Evaluation		
No	**Performance Matrices**	**Values (%)**
1.	Accuracy	97.30
2.	Sensitivity	1.720
3.	Specificity	0.500
4.	PPV	0.583
5.	NPV	0.012
6.	ROC	1.114

2. Confusion Matrix

The Confusion Matrix of C4.5 are shown in Figure II

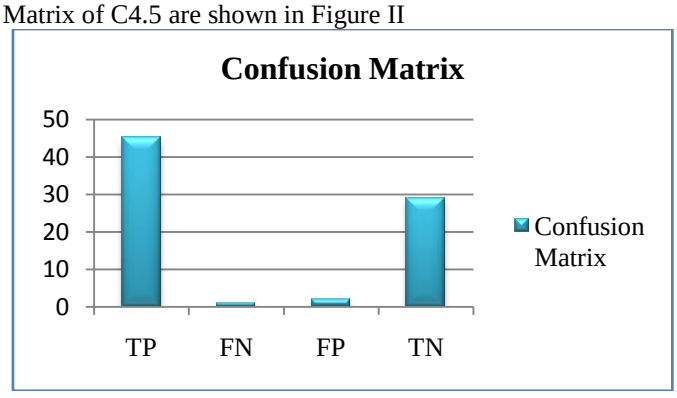

Fig II Confusion Matrix of C4.5 Algorithm

8. CONCLUSION

In this paper the performance of C4.5 analysis on WDBC data set in survivability of breast cancer is done. The performance of C4.5 shows the high level accuracy with other classifiers. Therefore C4.5 decision tree is suggested for predict survivability of Breast Cancer disease based classification to get better results with accuracy and performance.

REFERENCES

[1]. J.R.Quinlan, "Induction of decision tree". Journal of Machine Learning 1, 1986, Pg.no:81-106.
[2]. Mark A. Hall, Lloyd A. Smith, Feature Subset Selection: A Correlation Based Filter Approach, In 1997 International Conference on Neural Information Processing and Intelligent Information Systems (1997), pp. 855-858.
[3]. Han and Kamber, - "Data Mining; Concepts and Techniques", Morgan Kaufmann Publishers, 2000.
[4]. H. Blockeel and J. Struyf.,"Efficient algorithms for decision tree cross-validation", Proceedings of the Eighteenth *International Conference on Machine Learning* (C. Brodley and A. Danyluk, eds.), Morgan Kaufmann, 2001, pp. 11-18.
[5]. Witten H.I., Frank E., ―Data Mining: Practical Machine Learning Tools and Techniques‖ , Second edition, Morgan Kaufmann Publishers, 2005.
[6]. Kemal Polat, Seral Sahan, Halife Kodaz and Salih Günes, A New Classification Method for Breast Cancer Diagnosis: Feature Selection Artificial Immune Recognition System (FS-AIRS), In Proceedings of ICNC (2)'2005. pp.830~838.
[7]. J. Han and M. Kamber, ―Data Mining—Concepts and Technique‖ (The Morgan Kaufmann Series in Data Management Systems), 2nd ed. San Mateo, CA: Morgan Kaufmann, 2006.
[8]. 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.
[9]. Deisy.C, Subbulakshmi.B, Baskar S, Ramaraj.N, Efficient Dimensionality Reduction Approaches for Feature Selection, Conference on Computational Intelligence and Multimedia Applications, 2007.
[10]. Liu Ya-Qin, Wang Cheng, Zhang Lu," Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data", 3rd International Conference on Bioinformatics and Biomedical Engineering, 2009.
[11]. Lavanya and Usha Rani, "Ensemble Decision Tree Classifier for Breast Cancer Data" International Journal of Information Technology Convergence and Services, vol. 2, no. 1, pp. 17-24,2012.
[12]. Sujatha ,Usha Rani, "Evaluation Of Decision Tree Classifiers On Tumor Datasets" , International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 2, Issue 4, July – August 2013.
[13]. Syed SS, Shanthi S, Chitra VM. Application of Data Mining techniques to model breast cancer data. International Journal of Emerging Technology and Advanced Engineering. 2013 Nov; 3(11):362–9.
[14]. Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR (2013) Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. J Health Med Inform 4: 124. doi:10.4172/2157-7420.1000124.
[15]. R.K.Kavitha, Dr. D.DoraiRangasamy, "Predicting Breast Cancer Survivability Using Naïve Baysein Classifier And C4.5 Algorithm", Elysium Journal**,** Volume-1, Special Issue-1 September 2014.
[16]. E. Venkatesan and T. Velmurugan, "Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification", Indian Journal of Science and Technology, Vol 8(29), DOI: 10.17485/ijst/2015/v8i29/84646, November 2015.
[17]. Badr HSSINA, Abdelkarim MERBOUHA,Hanane EZZIKOURI,Mohammed ERRITALI, " A comparative study of decision tree ID3 and C4.5", International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications (IJACSA).
[18]. Wisconsin Diagnostic Breast Cancer (WDBC) Dataset and Wisconsin Prognostic Breast Cancer (WPBC) Dataset. http://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer-wisconsin/
[19]. Matthew N Anyanwu et al Matthew N.Anyanwu, Sajjan G.Shiva, "Comparative Analysis of Serial Decision Tree Classification Algorithms", International Journal of Computer Science and Security, volume 3.

BIOGRAPHY

A. Kathija is currently pursuing M.Phil degree in computer science in Sadakathullah Appa College, Tirunelveli. She has done her M.Sc degree in Computer Science from V.O.Chidambaram College, Thoothukudi and the B.Sc in Computer Science from St.Mary'sCollege(Autonomous),Thoothukudi under Manonmaniam Sundaranar University,Tirunelveli.

S. Shajun Nisha Professor and Head of the Department of P.G Computer Science, Sadakathullah Appa College, Tirunelveli. She has completed M.Phil. (Computer Science) and M.Tech (Computer and Information Technology) in Manonmaniam Sundaranar University Tirunelveli. She has involved in various academic activities. She has attended so many national and international seminars, conferences and presented numerous research papers. She is a member of ISTE and IEANG and her specialization is Image Mining.

Dr.M.Mohamed Sathik M.Tech., M.Phil., M.Sc., M.B.A., M.S., Ph.D (CS)., Ph.D(Engineering) has so far guided more than 35 research scholars. He has published more than 100 papers in International Journals and also two books. He is a member of curriculum development committee of various universities and autonomous colleges of Tamil Nadu. His specializations are VRML, Image Processing and Sensor Networks.