IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Comparison of Tree Based Supervised Classification Methods with Mammogram Data Set

M. Vasantha

Dept. of Computer Science, Bhaktavatsalm Memorial College for Women, Chennai, India

Corresponding Author: vasanthham@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i4.504506 | Available online at: www.ijcseonline.org

Accepted: 18/Apr/2019, Published: 30/Apr/2019

Abstract: This paper discusses the different classification techniques. It also compares the efficiency of Tree Based Classifiers Random Forest, REP Tree and J48 Classifiers for the detection of masses in mammogram images and compares their robustness through various measures. The mammogram images used in this research have been taken from MIAS database and the classification is performed with the help of open source machine learning tool. Finding the best classifier is a tough task and this paper gives opportunity to researchers to drill down efficient research works for evaluating different classifiers

Keywords*- Mammogram, Classification, Random Forest (RF), REP tree, J48 classifiers*

I. INTRODUCTION

Breast cancer continues to be a public health problem in the world and it remains to be a leading cause of death among women around the world [1].]. In 2015, 40290 woman's death reported due to breast cancer [2]. It is also the largely widespread female cancer in both developing and developed countries [3]. Digital mammography is one of the most promising options to diagnose breast cancer. However, its effectiveness is enfeebled due to the complexity in distinguishing actual cancer lesions from benign abnormalities, which results in unnecessary biopsy referrals. To overcome this issue, computer aided diagnosis using machine learning techniques has been studied globally. The accuracy with which tumors are detected, when large volumes of images are to be read by radiologists tends to decrease, and hence an automated mechanism for reading of digital mammograms is always preferable. However, with the use of proper computer aided systems, we could reduce the number of unnecessary biopsies being conducted. In the present work different algorithms are used for the classification of digital mammograms. They classify the masses into three categories, i.e. normal, benign and malignant, where benign and malignant are considered as abnormal; benign has tumors which are not cancerous and malignant is one in which the tumors are cancerous. The proposed method is to compare different classification technique used for categorizing the mammograms based on some performance measures.

Classification of images is an important area of research and of practical applications in a variety of fields, including pattern recognition, artificial intelligence medicine and vision analysis. Mass classification is a vital stage for the performance of the Computer-aided breast cancer detection.

The decision tree is most widely used classification technique [4] [5]. Random Forest is an ensemble decision tree method used for classification and regression [6]. REP Tree method is an efficient classification algorithm which prunes the tree using reduced-error pruning by applying back fitting process. The rest of this paper is organized as follows: The three methods used for the classification of mammogram images, namely Random Forest Classification, J48 Classification and REP tree classification are discussed in Section 2, Section 3 and Section 4 respectively. The experimental results are discussed in Section 5 and Section 6 gives the conclusion.

II. RANDOM FOREST CLASSIFICATION

Random Forest algorithm is most commonly used supervised classification algorithm. It creates the forest with a number of trees. The number of trees in the forest determines the robustness of the forest. That is, the higher the number of trees gives higher accuracy results. Random Forest constructs a large number of trees and aggregates the results from those trees.

As there is an increased requirement of machine learning techniques in the medical data analysis, Random Forest method which interacts naturally in the learning process is one of the most relevant options in the domain of Biomedicine [7]. RF includes a collection of decision trees and it also incorporates feature selection. With its excellent performance, developing variants of Random Forest became an active research topic in computational biology [8]. The major benefits of RF are,

- Well adapted for both prediction and Variable Importance
- Better prediction
- Readily accessible by novice user

Steps involved in Random Forest Algorithm:

i) It takes a set of samples from the data.
ii) Grows a tree based on the sample taken. At each node, samples a predefined number of predictions randomly and selects the best split among those variables.
iii) Computes the classification error rate using Out of Bag samples [10].

III. J48 CLASSIFIER

J48 classifier is a simple C4.5 decision tree for classification, which generates a binary tree. To classify a new item, first, based on the attribute values of the available training data a decision tree is created. Hence, whenever it comes across a set of items (training set) it recognizes the attribute that categorize the different instances clearly. It is most useful decision tree method for classification problems. This technique constructs a tree to model the classification process. Once the tree is built, the algorithm is applied to each tuple in the database and results in classification for that tuple.

IV. REP TREE

Reduces Error Pruning (REP) Tree Classifier is a high-speed decision tree learning algorithm and is based on the principle of computing the information gain with entropy and minimizing the error arising from variance [9]. This algorithm was first proposed in [10]. REP Tree applies regression tree logic and generates multiple trees in altered iterations. Afterwards it picks the best one from all spawned trees. This algorithm constructs the regression/decision tree using variance and information gain. Also, this algorithm prunes the tree using reduced-error pruning with back fitting method. At the beginning of the model preparation, it sorts the values of numeric attributes once. As in C4.5 Algorithm, this algorithm also deals the missing values by splitting the corresponding instances into subsets. [11].

V. EXPERIMENTAL RESULTS

In this work, to perform the benchmark experiment, WEKA [12] an open source Java based machine learning workbench is used, which can be run on any computer that has a Java run time environment installed. It brings together many machine learning algorithms and tools under a common framework.

To evaluate the performance, 300 digital mammogram images are used and are taken from the Mammogram Image Analysis Society (MIAS) an online database for mammograms available for research from the UK. The MIAS Digital Mammogram Database contains 322 images representing 161 mammogram pairs.

Mammograms are difficult to interpret, and a preprocessing phase of the image is used to improve the quality of the images and make the feature extraction phase more reliable. Background noise elimination is necessary to enhance the visibility and deteectability of tumors such as malignant or benign. In this paper, we performed low pass filter to remove noise. Image enhancement techniques are applied to to improve the interpretability or perception of information in images for human viewers, or to provide better input for other automated image processing techniques[13] and histogram equalization method for contrast enhancement is applied in this work

.Actually MIAS contains only the images and classification cannot be directly applied to the images. So, features are extracted from the image and these features are used for classification.

It can be noted from Table 1, the TP RATE of j48 is better when compared to RF and REP TREE methods. Similarly, the F measure is also greater when compared with other methods. The above values are compared and shown in Figure 1.

Table 1. TP Rate, FP Rate, Precision, Recall and F-Score Values

Measures **Methods**	**TP RATE**	**FP RATE**	**PRECISION**	**RECALL**	**F-Measure**
RF	0.9	0.048	0.907	0.9	0.9
J48	0.95	0.027	0.956	0.95	0.95
REP Tree	0.8	0.102	0.8	0.8	0.8

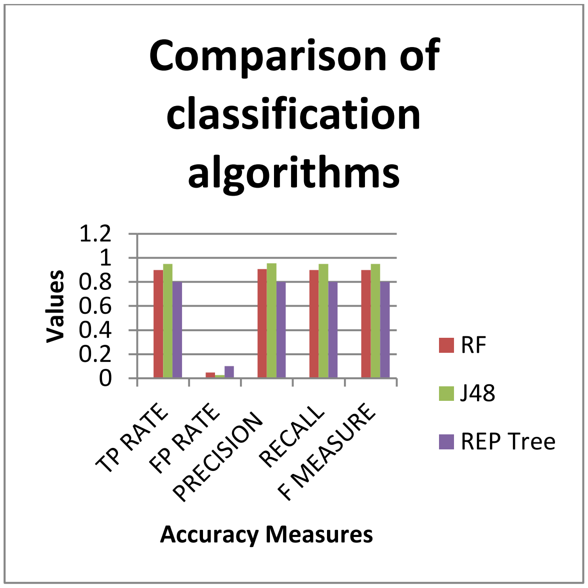

Figure 1. Comparison of RF, J48 and REP TREE (in terms of TPRATE, FPRATE, Precision, Recall, F-Measure)

VI. CONCLUSION

In this paper, we have compared RF, J48 and REP TREE methods that are used for classification with the features extracted from mammogram images. From the experimental results it is concluded that, with respect to highest tprate, recall, precision and F-Measures, J48 classifier seems to be better than the other two algorithms. Hence, we conclude that J48 is an effective method in extracting relationship between entities with the best F-Score value and less fprate.

REFERENCES

[1] DM. Parkin Bray F, Ferlay J, Pisani P. Global cancer statistics, 202. CA Cancer J Clin, 55(2): PP 74- 108 , 2005

[2] American Cancer Society, " *Breast cancer facts & figures 2015-2016*,"Atlanta, American Cancer Society , 2015

[3] F. Fauci, S. Bagnasco, R. Bellotti, D. Cascio, S.C. Cheran, F. De Carlo, G. De Nunzio, M.E. Fantacci, G. Forni, A. Lauria, E.L. Torres, R. Magro, G.L. Masala, P. Oliva, M.Quarta, G. Raso, A. Retico, S. Tangaro: "*Mammogram Segmentation by Contour Searching and Massive Lesion Classification with Neural Network*", *IEEE Nuclear Science Symposium Conference Record, Rome, Italy*, Vol. 5, pp. 2695-2699 2004.

[4] Kella BhanuJyothi, K. Hima Bindu and D. Suryanarayana, " *A Comparative Study of Random Forest & K-Nearest Neighbours on HAR dataset using Caret",* IJIRT, Volume 3, Issue 9 ISSN: 2349-6002., 2017.

[5] Dadye,Harold Buko and Richard Rimiru, "*Effects of Different Pre-processing Strategies :A Comparative Study on Decision Tree Algorithms",* International journal of Digital Content Technology and its Applications 7.7 : pp 935-939,2013

[6] Liaw, Andy and Matthew Wiener, "*Classification and regression by Random Forest*", R News : pp 18-22, 2002

[7] Goldstein, Benjamin .A, Polley Eric. C and Briggs, Farren. B.S, "*Random Forests for Genetic Association Studies*", Statistical Applications in Genetics and Molecular Biology, Vol.10. Iss.1. Article 32, DOI: 10.2202/1544-6115.1691, 2011

[8] H. Hu , "*Mining patterns in disease classification forests"* Journal of Biomedical Informatics Volume 43 pp . 820-827, 2010

[9] I H Witten and E Frank . *"Data mining: practical machine learning tools and techniques* "– 2nd ed. , Morgan Kaufmann series in data management systems, United States of America, 2005

[10] Quinlan, J, "*Simplifying Decision trees*", International Journal of Man Machine Studies, 27(3), pp 221–234, 1987

[11] S.K. Jayanthi and S. Sasikala, " *REP Tree Classifier for identifying Link Spam in Web Search Engines*" , IJSC, Volume 3, Issue 2 , pp 498 – 505, 2013

[12] WEKA: Waikato environment for knowledge analysis .http://www.cs.waikato.ac. nz/ml/weka

[13] Hussam Elbehiery," Optical Fiber Cables Networks Defects Detection using Thermal Images Enhancement Techniques", International Journal of Scientific Research in Computer Science and Engineering Vol.6, Issue.1, pp.22-29 , 2018

Author's Profile

Mrs.M.Vasantha pursed Master of Computer Application from Alagappa University, Master of Engineering from Anna University, India and Doctorate in Computer Science from Mother Teresa University, India in the year 2015. She is currently working as Associate Professor in PG Department of Computer Sciences, Bhaktavatsalm Memorial College For Women, Chennai affiliated with the University of Madras, India since 2016. He has published more than 15 research papers in reputed international journals.. Her main research work focuses on Big Data Analytics, Data Mining, and Machine learning. She has 25 years of teaching experience and 10 years of Research Experience.