COMPARISON OF DECISION TREE METHODS FOR BREAST CANCER DIAGNOSIS

Emina Alickovic, Abdulhamit Subasi

International Burch University, Faculty of Engineering and Information Technologies
Sarajevo, Bosnia and Herzegovina
ealickovic@ibu.edu.ba, asubasi@ibu.edu.ba

Abstract

In almost all parts of the world, breast cancer is one of the major causes of death among women. But at the same time, it is one of the most curable cancers if it is diagnosed at early stage. This paper tries to find a model that diagnoses and classifies breast cancer with high accuracy and that will help to both patients and doctors in the future. Here we present several different decision tree methods in order to classify breast cancer with high accuracy. The results achieved in this research are very promising (accuracy is 96.49 %). It is very promising result compared to previous researches where decision tree techniques were used. As benchmark test, Breast Cancer Wisconsin (Original) was used.

Keywords - Breast cancer, Decision trees, Random Forest, Random Tree, C 4.5, Simple CART.

1 INTRODUCTION

This paper talks about breast cancer diagnosis. Cancer is a set of illnesses where body cells grow, alter, and multiply without control. As a rule, name of the cancer comes from the part of the body where it originated. Due to this, breast cancer refers to the unpredictable and often fast enlargement of cells that begin in the breast tissue. A cluster of rapidly separating cells may form a mass of extra tissues, called tumors. Tumors can either be cancerous (malignant) or non-cancerous (benign). Malignant tumors travel through healthy body tissues and destroy them.

The term, breast cancer, refers to a malignant tumor that has developed from cells in the breast. It is the most common cancer among women in almost all parts of the world. But if it is discovered in the earlier stages, chance to cure it is very high. According to statistics, early stage detection and treatment results in 98% survival rate but if it is detected in metastases, this plummets to 27% [1].

In reality, one in eight women in the USA might expect to develop breast cancer during the life time [2]. Although in Bosnia and Herzegovina we do not have single register at government level, according to the reports from cantonal health care and hospital registers, breast cancer is the most common malignant illness among woman in our country [3]. Therefore there is great need to develop a technique that will diagnose and classify breast cancer with high accuracy.

Till now, several different techniques have been used for breast cancer diagnosis. One of the most widely used techniques is mammography, but in literature, radiologists show significant differences in interpreting it [4]. Another widely used technique is Fine Needle Aspiration Cytology (FNAC) but bad side is its modest accuracy rate (around 90%). Therefore, there is a need to develop another technique that will provide better performances for classification of breast cancer. The response to this need is usage of statistical and artificial intelligence techniques. Here we divide all data into two groups, either benign (that does not have cancer) or malignant group (strong evidence of having breast cancer). Due to this reason, breast cancer diagnosis can be discussed as classification problem [5-7]. Many researchers used different statistical and artificial intelligence techniques to predict and classify breast cancer techniques.

Decision tree is style of representation uses a “divide and conquer” approach to solve the difficulty of learning from autonomous variables. It is a diagram that helps to select the best action possible out of several actions we have. The main advantage of using this approach is its assigning of direct values to all outcomes, minimizing ambiguity.

In this research, we apply several decision tree methods to classify types of breast cancer. These are Simple CART, C4.5, Random Forest and Random Tree. We use Wisconsin breast cancer (Original) dataset as benchmark test.

This paper is organized as follows. In section 2, we give description of dataset used. In section 3, theoretical background about decision trees and methods applied in this research is given. In section 4, we present out experimental results. In section 5, we give final conclusion and possible future improvements.

2 WISCONSIN BREAST CANCER DATABASE OVERVIEW

Breast cancer is one of the most spread cancers among women. Based on rates from 2005-2007, one in eight women is affected by this cancer during their lifetimes [1]. Breast cancer can also occur in man, although it is not that common. Although some of the risks such as gene, genetic risk factors, family history, menstrual periods, not having children, obesity, etc, that increase chances for development of breast cancer are known, it is not known how these risk factors causes cells to become cancerous. Many researches are trying currently to answer this question and understand how certain alterations in DNA can cause normal breast cells to develop into cancerous with a great progress [11].

Performance is evaluated based on the model using the Wisconsin breast cancer dataset (Original) to classify the types of breast cancer as either benign or malignant. This dataset contains nine features and each of these features is represented by some number between 1 and 10. These data is collected by Dr. William H. Wolberg (1989–1991) at the University of Wisconsin–Madison and this dataset can be found on UCI Machine Learning Repository. It contains 699 records taken from 699 different persons and 241 (65.5%) records are malignant and 458 (34.5%) records are benign. Out of these 699 records, it contains 16 instances with missing attribute values. We tested out proposed method on set *containing 683 data to prove efficiency of our methods.* This database consists of nine attributes obtained from fine needle aspirates; every feature is represented as an integer numbers between 1 and 10 and each of the values in this database is numerical. The measured variables are as follows:

1. clump thickness (c1);
2. uniformity of cell size (c2);
3. uniformity of cell shape (c3);
4. marginal adhesion (*c*4);
5. single epithelial cell size (c5);
6. bare nucleoi (c6);
7. bland chromatin (c7);
8. normal nuclei (c8);
9. mitoses (c9).

3 THEORETICAL BACKGROUND

A “divide-and-conquer” technique to the problem of learning from a group of autonomous parts leads to a representation technique named as *decision tree.* Decision tree nodes include testing a particular attribute. Typically, this node test compares value of attribute with some constant. On the other hand, there are trees that compare two attributes with each other, or employ some function with one or more attributes. Classification is provided by leaf nodes and this classification applies to all instances attaining the leaf. To classify an unknown instance, it is propagated down the tree based on the values of the attributes tested in succeeding nodes, and when a leaf is reached, based on the class assigned to the leaf, we classify the instance [12].

If we have a nominal value of attribute being tested at some node, the number of offspring is equal to the number of attainable attribute’s values. If we have numeric attribute, node test usually decides if this value is greater or bigger than some preset constant giving two-way splitting. Obvious problem here is missing values. It is not easily seen which branch should be taken in consideration when an attribute with missing value is being tested. An easy answer to this problem is to evidence the amount of instances in the training set that propagate down each branch and to employ the most well-liked branch if we have missing value for a test instance (Witten & Frank, 2005). Decision tree learning has

been applied to many practical problems such as medical classification, equipment malfunctioning, and loan applications [13].

3.1 Simple CART

CART is recognized decision tree algorithm. It is a type of binary recursive partitioning. The term "binary" refers to all sets of patients, illustrated as a "node" in a decision tree, can only be divided into two sets. So, every node can be separated into two child nodes, and original node is named as a parent node. The term "recursive" denotes the binary partitioning procedure which can be used many times. As a result, each parent node can gives two child nodes and, in turn, any of these child nodes may possibly be separated, forming further generations. The term "partitioning" refers to the fact that the dataset is divided into sections [14].

This method has advantage over many other classification methods such as: handling numerical data that are highly skewed or categorical data with either ordinal or non-ordinal structure; easiness to deal with missing variables, its relatively automatic "machine learning" methods and it is simple for interpretations [14].

It consists of four steps. First step is tree building using recursive separation of nodes where predicted class is assigned to each node. The second step is stopping tree building process. After this step, "maximal" tree is formed, that probably over fits data contained in learning data set. Third step is "pruning" of tree, resulting in series of simpler trees. The last step is optimal tree selection, through which the tree which suits data in the learning dataset, but does not over fit data, is chosen from the sequence of pruned trees [14].

3.2 C 4.5

C 4.5 algorithm learns using top approach. It starts with the finding attribute to be selected for testing at the tree's root. To find such attribute, each attribute is evaluated by means of a statistical test to find out how well it by itself classifies the training examples. The most excellent attribute is selected and used as the test at the tree's root node. A root node offspring is created after that for each potential value of this attribute, and the training examples are sorted to the suitable offspring node(i.e., down the branch matching the example's value for this attribute). This process is repeated again by means of the training examples connected with every offspring node to pick the most excellent attribute to test at that point in the tree. Essential choice in the C 4.5 algorithm is selecting proper attribute to test at every node. It is selected according to information gain, which measures how well a given attribute separates the training examples according to their target classification [13, 15].

C 4.5 involves followings steps: (1) reducing decision tree from the training set, increasing the tree until the training data is fit and allowing over fitting to happen; (2) alter the learned tree into an equivalent set of rules by making single rule for all paths from the root node to a leaf node; (3) generalize all rules by deleting whichever preconditions that outcome in improving its predicted accuracy; (4) sort the pruned rules by their predicted accuracy, and consider them in this chain while classifying successive instances [13, 15].

3.3 Random Forest

It is a bagging classification algorithm developed by Leo Breiman [9] and it uses an ensemble of T classification or regression trees, in our case classification trees. All of classification trees t are created using a different bootstrap instance I_t having N_I randomly taken cases. In addition to bagging Random Forests uses random feature selection as well. At every split of decision tree, m variables are chosen randomly from a set of all input variables and the finest split is chosen from m variables. Each tree is grown fully to obtain low-bias trees using CART methodology. In order to classify an instance, we need to place the input variables down the T trees in the forest. Every tree chooses the predicted class. Lastly the bagged predictor is received by majority election that is the instance is classified into the class with the highest number of votes over whole T trees in the forest [9, 10].

In this paper we used Random Forest for classification of breast cancer diseases and we achieved satisfying accuracy.

3.4 Random Tree

Random tree is rooted tree. It follows Markov process [16]. Random tree nodes are connected with the n-dimensional vector arrays. A random tree might be acknowledged with a set of vector arrays [17]:

$$\{S_0, S_1[b],\ldots,S_N\overbrace{[b]\ldots[b]}^{N}\}$$

The array indices symbolize tree creation process, and we will now explain this [17].

Costs of n properties at time kΔt are enclosed inside n-dimensional vector $\{S_k[i_1]\ldots[i_k]|\ i_1,\ldots,\ i_k = 1, \ldots, b\}$. All nodes have same number of branches, b. All branches have direction, showing time transition. Node at branch head is originated from tail node. Initial vector S_0, connected to the root, has initial cost of n properties. For given *kth* vector, $S_k[i_1]..[i_k] = (S_k[i_1]..[i_k](j))_{j=1}^{n}(i_{K+1} = 1,\ldots,b)$ is created in accordance with [17]:

$$R(t+\Delta t) = \exp\begin{bmatrix}\Delta t(rI - diag(\delta_1 + \sigma_1^2/2,\ldots,\ \delta_n + \sigma_n^2/2) + \\ \sqrt{\Delta t}diag(\sigma_1 W_1(t),\ldots,\ \sigma_1 W_n(t)\end{bmatrix} R(t) \quad (1)$$

Condition is $S(k\Delta t) = S_k[i_1]\ldots[i_k]$ at time kΔt [17]:

$$\begin{aligned} &S_{k+1}[i_1]\ldots[i_k][i](j) = \\ &= S_k[i_1\}\ldots[i_k](j)\exp[(r - \delta_j - \sigma_j^2/2)\Delta t + \sigma_j\sqrt{\Delta t}W_i(j)] \end{aligned} \quad (2)$$

where i=1,...,b; j=1,...,n, r is the risk-free interest rate, δ_j is share rate for jth property, σ_j is the unpredictability for jth property, and W(t) = (W_1(t), …, W_n(t)) is vector with normal random variables whose average is zero and correlation matrix R = (ρ_{ij}) (ρ_{ii} = 1, i=1, …,n) (Morohosi & Fushimi, 2002).

Single random tree generates a pair of high and low predictions. To be able to acquire a consistent prediction of the cost, the sample average of each prediction must be calculated by autonomous replications of random trees [17]. .

4 EXPERIMENTAL RESULTS

Different decision tree methods have been applied in variety of areas and they showed very good performances. In this research, Wisconsin breast cancer dataset was used in order to test the efficiency of proposed methods: Random Forest, Random Tree, Simple CART and C 4.5. As it was mentioned earlier, out dataset contain 683 records from 683 different persons.

We evaluated performances of four different decision tree methods in terms of precision and overall accuracy. Overall accuracy can be calculated as the number of correctly classified instances divided by the number of all instances. Precision states how close measured values are to each other and t can be calculated as the number of true positives divided by the sum of true positives and false positives. Results we achieved are shown in Table 1 and 2. Table 1 gives precision results and Table 2 gives accuracy results and these results are achieved for 10 – fold cross validations. 10 – fold cross validations means that entire dataset is partitioned into 10 mutually exclusive subsets (folds) of roughly same size. 9 folds are employed for training and testing is done on the remaining one fold. The classification model is trained and tested 10 times.

We first tested Random Forest. This method showed very high classification and precision rate. Average classification rate was 96.49%. Especially high classification rate was achieved for benign type of cancer (97.07%). Time taken to build this model (0.04 seconds) was very low what is confirming the efficiency of Random Forest in classification of breast cancer cases. Precision rate is very high and equal to 0.965, stating that possible deviations are very rare. Achieved performance results are very good for this small classifier. These results show us that Random forest is very convenient for breast cancer classification since it is both very accurate and also very precise (accuracy and precision rates are very high

Second tested method is Random Tree. This method also showed lower results compared to Random Forest. Classification accuracy rate was 95.9% and precision rate was 0.962 in average. Accuracy for benign cancer classification was higher than classification accuracy for malignant type of cancer.

Third method we applied in this research is Simple CART method. Average accuracy achieved using this method is 95.32 % and average precision is 0.953. This method gave the lowest accuracy, but although this accuracy is the lowest out of four proposed methods, it is also very high, over 95%.

The last method we tested in this research is C 4.5 and gave an accuracy of 96.05 % and precision of 0.961. This method gave somewhat lower results then Random Forest.

Besides giving very good accuracies and precision rates, these methods have very low computational complexity which is also an additional advantage of using these four decision tree methods.

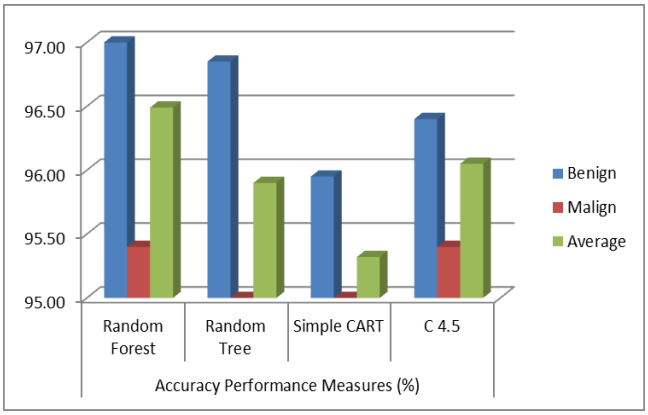

Figure 1 Comparison of Classification Results

Table 1 Precision Performance Results

DECISION TREE METHODS	Precision Performance Measures			
	Random Forest	Random Tree	Simple CART	C 4.5
Benign	0.975	0.973	0.968	0.975
Malign	0.946	0.942	0.926	0.934
Average	0.965	0.962	0.953	0.961

Table 2 Overall Accuracy Performance Results

DECISION TREE METHODS	Accuracy Performance Measures (%)			
	Random Forest	Random Tree	Simple CART	C 4.5
Benign	97.07	96.85	95.95	96.40
Malign	95.40	94.14	94.14	95.40
Average	96.49	95.90	95.32	96.05

5 CONCLUSION

Breast cancer is one of the major killers in our country just like in the rest of the world. According to some statistics, breast cancer is the most spread type of malignant cancer in Bosnia and Herzegovina. Breast cancer diagnosis is the focus of this paper and our research. In this paper, four different decision tree methods very selected for breast cancer classification. These four methods are: Random Tree, Random Forest, C4.5 and Simple CART. Accuracies achieved by these four methods are very

high. Many powerful and complex methods very used for breast cancer classification without giving high classification rate or having very high training and testing time what are two major drawbacks. This research showed that simple decision tree methods can give very high accuracy rates and very low training and testing time if proper parameters are selected. These four methods can be applied for other medical diagnosis classification what will be the main focus of our future work and researches.

References

[1] SEER Web Sit, Last accessed on: April, 13, 2012. Available:

http://www.seer.cancer.gov/statfacts/html/breast.html..

[2] «American Cancer Society | Information and Resources for Cancer: Breast, Colon, Lung, Prostate, Skin,» last accessed on: 31 October 2012. Available:

http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-key-statistics. [Último acceso: 03 December 2012].

[3] M. Mušanović, M. Đapo, "RANA detekcija raka dojke, Sarajevo: Ministarstvo Zdravstva Kantona Sarajevo", Zavod zdravstvenog Osiguranja Kantona Sarajevo, 2009.

[4] J. Elmore, M. Wells, M. Carol, H. Lee, D. Howard, A. Feinstein, "Variability in radiologists interpretation of mammograms", New England Journal of Medicine, vol. 22, p. 1493–1499, 1994.

[5] T. W. Anderson, "An introduction to multivariate statistical analysis", New York: Wiley, 1984.

[6] W. R. Dillon, M. Goldstein, "Multivariate analysis methods and applications", New York: Wiley, 1984.

[7] D. J. Hand, "Discrimination and classification", New York: Wiley, 1981.

[8] R. A. Johnson, D. W. Wichern, "Applied multivariate statistical analysis", 5 ed., Upper Saddle River, NJ : Prentice-Hall, 2002.

[9] L. Breiman, "Random forests", Machine Learning, vol. 45, nº 1, p. 5–32, 2001.

[10] A. Prinzie, D. Van den Poel, "Random Forests for multiclass classification: Random MultiNomial Logit", Expert Systems with Applications, vol. 34, p. 1721–1732, 2008.

[11] M. Karabatak, M. C. Ince, "An expert system for detection of breat cancer based on association rules and neural network", Expert Systems with Applications, vol. 36, pp. 3465-3469, 2009.

[12] I. H. Witten, E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques", 2nd ed., Elsevier Inc., 2005.

[13] T. M. Mitchell, Machine Learning, McGraw-Hill Science/Engineering/Math, 1997.

[14] R. J. Lewis, "An Introduction to Classification and Regression Tree (CART) Analysis", de The 2000 Annual Meeting of the Society for Academic Emergency Medicine, San Francisco, California, 2000.

[15] J. R. Quinlan, «Improved Use of Continuous Attributes in C4.5, "Journal of Artificial Intelligence Research", vol. 4, pp. 77-90, 1996.

[16] M. Broadie, P. Glasserman, “Pricing American-style securities using simulation”, Journal of Economic Dynamics and Control, vol. 21, pp. 1323-1352, 1997.

[17] H. Morohosi, M. Fushimi, “Quasirandom tree method for pricing American Style Derivatives”, Journal of the Operations Research, vol. 45, nº 4, pp. 426-434, 2002.