DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4602

ISSN No. 0976-5697

Volume 8, No. 8, September-October 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

IDENTIFYING PATIENTS AT RISK OF BREAST CANCER THROUGH DECISION TREES

Soumya Kanta Sarkar
Department of Information Technology, University Institute of Technology,
The University of Burdwan
Burdwan, WB, India

Akash Nag
Department of Computer Science,
MUC Women's College, Burdwan
Burdwan, WB, India

Abstract: In this paper, we explore how the C4.5 algorithm can be applied to breast cancer datasets in order to extract and formulate rules for identifying risk factors. For this study, we have used the Wisconsin dataset containing 9 attributes related to various cell features and anomalies. We have then applied the C4.5 algorithm to that dataset to create a decision tree. From the inferred tree, the rules for identifying the patients at risk have been derived. With a training-set size of 200 patient records, our system was found to have an accuracy of 96.7%.

Keywords: breast cancer, decision tree, classification, Wisconsin dataset, C4.5

I. INTRODUCTION

Breast cancer is a type of cancer that develops from breast tissue and is often associated by a lump in the breast, change in breast shape, development of red and patchy skin, or fluid emanating from the nipple. The causes for breast cancer have not been fully understood till date. There are some genetic factors, and some environmental factors associated with its development. Breast cancer is preliminarily detected by a mammogram exam and confirmed by a biopsy.
When a lesion is detected, typically a breast FNA (Fine Needle Aspiration) is performed. It is a simple procedure similar to drawing blood using needles. It is used to remove some fluid or cells from a breast lesion or cyst in order to determine the nature of the lesion. The extracted sample is smeared on a glass slide and sent to a pathological laboratory to be examined under a microscope. During examination of the tissue samples, 9 characteristics are usually considered [1]. Each characteristic is assigned a number in a scale from 1 to 10 by the pathologist; where the larger the number, the greater is the likelihood of malignancy. No single measurement however can be used to determine whether a given sample is benign or malignant.

The 9 characteristics considered by the pathologist are as follows:

1. **Clump Thickness:** This is used to assess if cells are mono-layered or multi-layered. Benign cells tend to be grouped in mono-layers, while cancerous cells are often grouped in multi-layer.
2. **Uniformity of Cell Size:** It is used to evaluate the consistency in the size of cells in the sample. Cancer cells tend to vary in size. That is why this parameter is very valuable in determining whether the cells are cancerous or not.
3. **Uniformity of Cell Shape:** It is used to estimate the equality of cell shapes and identifies marginal variances, because cancer cells tend to vary in shape.
4. **Marginal Adhesion:** Normal cells tend to stick together. Cancer cells tend to loose this ability. So loss of adhesion is a sign of malignancy.
5. **Single Epithelial Cell Size:** It is related to the uniformity. Epithelial cells that are significantly enlarged may be a malignant cell.
6. **Bare Nuclei:** This is a term used for nuclei that is not surrounded by cytoplasm. Those are typically seen in benign tumors.
7. **Bland Chromatin:** Describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells, the chromatin tends to be coarser.
8. **Normal Nucleoli:** Nucleoli are small structures seen in the nucleus. In normal cells the nucleolus is usually very small if visible at all. In cancer cells the nucleoli become much more prominent, and sometimes there are more of them.
9. **Mitoses:** It is an estimate of the number of mitosis that has taken place. Larger the value, greater is the chance of malignancy.

A decision tree is a decision support tool that describes conditions and possible outcomes in the form of a tree-like graph. Each non-terminal node in the tree represents a test or decision on the considered data item. Choice of a certain branch depends upon the outcome of the test. To classify a particular data item, we start at the root node and follow the assertions down until we reach a terminal node (or leaf). A decision is made when a terminal node is approached. Decision trees can also be interpreted as a special form of a rule set, characterized by their hierarchical organization of rules.

There are many popular algorithms that classify a given dataset and construct a decision tree in the process that encodes, in the form of rules, how the classification takes place. ID3 (Iterative Dichotomizer 3) [2] is one such popular algorithm developed by Ross Quinlan. It is typically used in machine learning and natural language processing applications. Quinlan subsequently improved this algorithm to create the

C4.5 algorithm [3], which is one of the most widely used decision-tree algorithms.

The rest of this paper is organized as follows: in Section 2, we discuss some related work that has been done in this field for predicting breast cancer; in Section 3, we present the data used for this study, as well as the methods we have followed; in Section 4, we present our findings and discuss them; and we finally conclude in Section 5.

II. RELATED WORK

A lot of work has been done in the field of classification till date. Abbass [4] has used artificial neural networks for cancer diagnosis. Ratanamahatana and Dimitrios [5] have used decision trees for feature selection and have used the Wisconsin dataset. Mangasarian and Wolberg [6] [7], and Bennett and Mangasarian [8] have used linear programming for cancer diagnosis using the same dataset. Bennett et al. [9] have developed an ensemble method of classification for assembling labelled and unlabelled data. They have also used the breast cancer dataset for testing their methods. Grąbczewski and Włodzisław [10] have used decision tree forests for classification of breast cancer data.

III. DATA AND METHODS

A. *Data*

For our study, we have used the Wisconsin dataset [6] [7] [8] [11] as our sample, which was created by Dr. William H. Wolberg in 1992, from his patient records at the University of Wisconsin Hospitals, Madison, USA. The dataset contains 699 samples with 458 benign (65.5%) and 241 (34.5%) malignant cases. Each record contains 11 attributes as listed in Table 1. Some values are missing from the dataset, and hence preprocessing was required before we could feed the data to the decision-tree algorithm.

Table I. Attributes of the Wisconsin dataset

Attribute Name	***Attribute value range***
Sample code number	61,634 – 13,454,352
Clump Thickness	1 – 10
Uniformity of Cell Size	1 – 10
Uniformity of Cell Shape	1 – 10
Marginal Adhesion	1 – 10
Single Epithelial Cell Size	1 – 10
Bare Nuclei	1 – 10
Bland Chromatin	1 – 10
Normal Nucleoli	1 – 10
Mitoses	1 – 10
Class	2 (benign) or 4 (malignant)

B. *Data Preprocessing*

Since the dataset contains missing values, we have included a preprocessing phase which replaces the missing values by the median of the various values of the corresponding attribute. The median is a holistic measure that is equal to the middle value in a list of values arranged in either ascending or descending order. If the list is of even length, the median is the arithmetic mean of the two middle values.

C. *Methods*

We have used the C4.5 algorithm for classifying our dataset. The splitting test of a node in the C4.5 is defined to be the gain ratio. Here, the classification uses entropy and information gain for tree splitting. It is suitable for handling both categorical as well as continuous data. A threshold value is fixed such that all the values above the threshold are not taken into consideration. The initial step is to calculate information gain for each attribute. The attribute with the maximum gain will be preferred as the root node for the decision tree.

A sample *S* is partitioned as follows:

1. When all records in *S* belong to the same class, it is assigned to be a leaf of the tree.
2. When *S* contains no records, it is assigned to be a leaf of the tree.
3. When *S* contains records belonging to more than one class, *S* must be partitioned or refined into subsamples. A node for *S* is assigned to the tree, and children nodes are created under it which will hold the subsamples.

There are many ways for testing which attribute should be chosen for partitioning the sample, but the most common test is the test of entropy. The entropy of a sample *S* is given by:

$$Info(S) = \sum_{i=1}^{k}\left[\left(\frac{freq(C_i,S)}{|S|}\right).\log_2\left(\frac{freq(C_i,S)}{|S|}\right)\right] \quad (1)$$

Where, *k* is the number of classes; in our case k=2,
$|S|$ represents the number of records in sample *S*,
freq(C_i,*S*) represents the number of records in *S* belonging to class C_i.
After *S* has been partitioned based on the *n* possible outcomes (values) for each attribute X, we compute the following:

$$Info_X(S) = \sum_{i=1}^{n}\left[\frac{|S_i|}{|S|}.Info(S_i)\right] \quad (2)$$

$$Gain(X) = Info(S) - Info_X(S) \quad (3)$$

The attribute *X* having the highest Gain value is selected as the partitioning attribute. The process is repeated for every sub-sample associated with each node, till every sub-sample contains records of the same class.

IV. RESULTS

We have implemented the algorithm in Java 7 and have tested it using a sample size of 200. We have computed the accuracy as shown in Eqn. 4, and was found to be 96.71%. In Figure 1, we present the decision tree obtained using the C4.5 algorithm on the sample.

$$Accuracy = \frac{True\ positives + True\ negatives}{True\ positives + True\ negatives + False\ positives + False\ negatives} \quad (4)$$

The rules generated by the decision tree for identifying patients at risk of breast cancer are as follows (where CLASS=2 refers to BENIGN, and CLASS=4 refers to MALIGNANT):

1) IF Bare-Nuclei<=5.0 AND IF Clump-Thickness<=5.0 AND IF Cell-Size-Uniformity<=2.0 THEN CLASS=2
2) IF Bare-Nuclei<=5.0 AND IF Clump-Thickness<=5.0 AND IF Cell-Size-Uniformity>2.0 AND IF Normal-Nucleoli<=3.0 AND IF Cell-Size-Uniformity<=3.0 THEN CLASS=2
3) IF Bare-Nuclei<=5.0 AND IF Clump-Thickness<=5.0 AND IF Cell-Size-Uniformity>2.0 AND IF Normal-Nucleoli<=3.0 AND IF Cell-Size-Uniformity>3.0 THEN CLASS=4
4) IF Bare-Nuclei<=5.0 AND IF Clump-Thickness<=5.0 AND IF Cell-Size-Uniformity>2.0 AND IF Normal-Nucleoli>3.0 THEN CLASS=4
5) IF Bare-Nuclei<=5.0 AND IF Clump-Thickness>5.0 AND IF Clump-Thickness<=7.0 AND IF Clump-Thickness<=6.0 AND IF Single-Epithelial-Cell-Size<=3.0 THEN CLASS=2
6) IF Bare-Nuclei<=5.0 AND IF Clump-Thickness>5.0 AND IF Clump-Thickness<=7.0 AND IF Clump-Thickness<=6.0 AND IF Single-Epithelial-Cell-Size>3.0 AND IF Bare-Nuclei<=2.0 THEN CLASS=4
7) IF Bare-Nuclei<=5.0 AND IF Clump-Thickness>5.0 AND IF Clump-Thickness<=7.0 AND IF Clump-Thickness<=6.0 AND IF Single-Epithelial-Cell-Size>3.0 AND IF Bare-Nuclei>2.0 THEN CLASS=2
8) IF Bare-Nuclei<=5.0 AND IF Clump-Thickness>5.0 AND IF Clump-Thickness<=7.0 AND IF Clump-Thickness>6.0 THEN CLASS=4
9) IF Bare-Nuclei<=5.0 AND IF Clump-Thickness>5.0 AND IF Clump-Thickness>7.0 AND IF Clump-Thickness<=8.0 AND IF Bland-Chromatin<=3.0 THEN CLASS=2
10) IF Bare-Nuclei<=5.0 AND IF Clump-Thickness>5.0 AND IF Clump-Thickness>7.0 AND IF Clump-Thickness<=8.0 AND IF Bland-Chromatin>3.0 THEN CLASS=4
11) IF Bare-Nuclei<=5.0 AND IF Clump-Thickness>5.0 AND IF Clump-Thickness>7.0 AND IF Clump-Thickness>8.0 THEN CLASS=4
12) IF Bare-Nuclei>5.0 AND IF Cell-Shape-Uniformity<=5.0 AND IF Cell-Size-Uniformity<=4.0 AND IF Bland-Chromatin<=3.0 AND IF Clump-Thickness<=1.0 THEN CLASS=2
13) IF Bare-Nuclei>5.0 AND IF Cell-Shape-Uniformity<=5.0 AND IF Cell-Size-Uniformity<=4.0 AND IF Bland-Chromatin<=3.0 AND IF Clump-Thickness>1.0 AND IF Cell-Shape-Uniformity<=3.0 THEN CLASS=4
14) IF Bare-Nuclei>5.0 AND IF Cell-Shape-Uniformity<=5.0 AND IF Cell-Size-Uniformity<=4.0 AND IF Bland-Chromatin<=3.0 AND IF Clump-Thickness>1.0 AND IF Cell-Shape-Uniformity>3.0 AND IF Clump-Thickness<=5.0 THEN CLASS=2
15) IF Bare-Nuclei>5.0 AND IF Cell-Shape-Uniformity<=5.0 AND IF Cell-Size-Uniformity<=4.0 AND IF Bland-Chromatin<=3.0 AND IF Clump-Thickness>1.0 AND IF Cell-Shape-Uniformity>3.0 AND IF Clump-Thickness>5.0 THEN CLASS=4
16) IF Bare-Nuclei>5.0 AND IF Cell-Shape-Uniformity<=5.0 AND IF Cell-Size-Uniformity<=4.0 AND IF Bland-Chromatin>3.0 AND IF Bare-Nuclei<=7.0 THEN CLASS=2
17) IF Bare-Nuclei>5.0 AND IF Cell-Shape-Uniformity<=5.0 AND IF Cell-Size-Uniformity<=4.0 AND IF Bland-Chromatin>3.0 AND IF Bare-Nuclei>7.0 THEN CLASS=4
18) IF Bare-Nuclei>5.0 AND IF Cell-Shape-Uniformity<=5.0 AND IF Cell-Size-Uniformity>4.0 THEN CLASS=4
19) IF Bare-Nuclei>5.0 AND IF Cell-Shape-Uniformity>5.0 THEN CLASS=4

V. CONCLUSION

In this study we explored how successfully decision trees can be used to diagnose breast cancer from breast FNAC results. We showed that the C4.5 algorithm, when used with cancer datasets like the Wisconsin dataset can produce extremely high accuracy rates. Breast cancer takes thousands of lives with an estimated 533,600 deaths occurring in 2015. Using decision tree based classification systems would result in better diagnosis at an early stage for patients who are potentially at risk of breast cancer; and which in turn would help save thousands of lives each year.

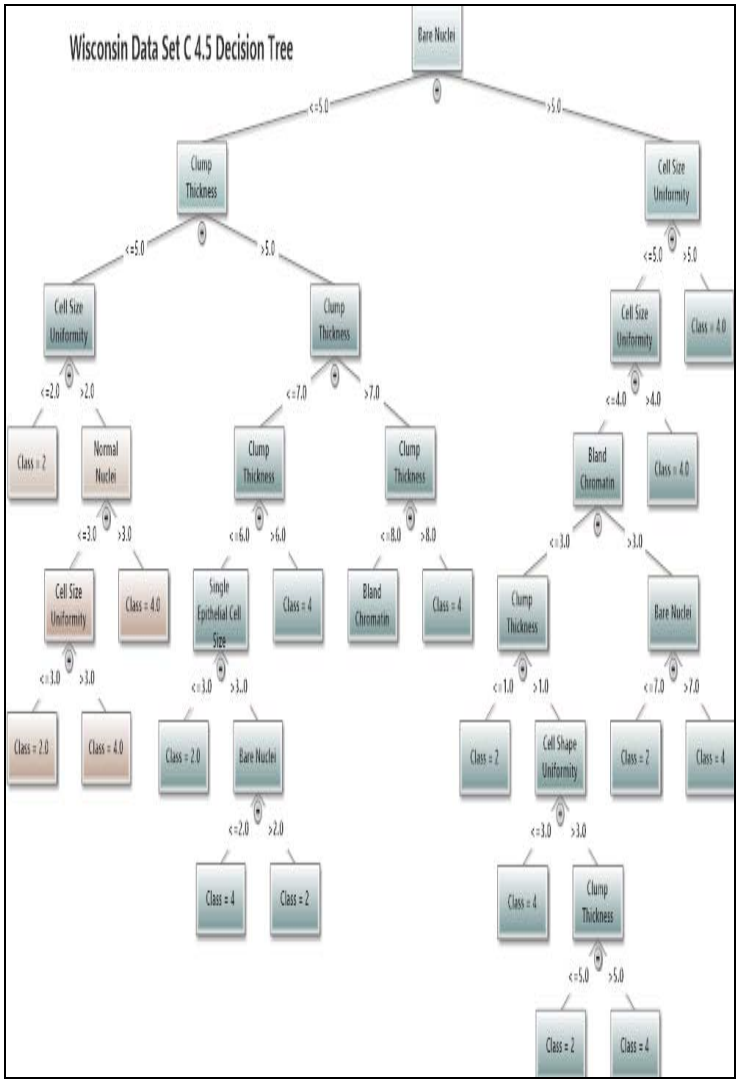

Figure 1. Decision tree obtained using the C4.5 algorithm

VI. REFERENCES

[1] Mouriquand, J., and D. Pasquier. "Fine needle aspiration of breast carcinoma: a preliminary cytoprognostic study." Acta cytologica 24.2 (1980): 153-159.

[2] Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81–106

[3] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[4] Abbass, Hussein A. "An evolutionary artificial neural networks approach for breast cancer diagnosis." Artificial intelligence in Medicine 25.3 (2002): 265-281.

[5] Ratanamahatana, Chotirat Ann, and Dimitrios Gunopulos. "Scaling up the naive Bayesian classifier: Using decision trees for feature selection." (2002).

[6] Mangasarian, Olvi L., W. Nick Street, and William H. Wolberg. "Breast cancer diagnosis and prognosis via linear programming." Operations Research 43.4 (1995): 570-577.

[7] Wolberg, William H., and Olvi L. Mangasarian. "Multisurface method of pattern separation for medical diagnosis applied to breast cytology." Proceedings of the national academy of sciences 87.23 (1990): 9193-9196.

[8] Bennett, Kristin P., and Olvi L. Mangasarian. "Robust linear programming discrimination of two linearly inseparable sets." Optimization methods and software 1.1 (1992): 23-34.

[9] Bennett, Kristin P., Ayhan Demiriz, and Richard Maclin. "Exploiting unlabeled data in ensemble methods." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.

[10] Grąbczewski, Krzysztof, and Włodzisław Duch. "Heterogeneous forests of decision trees." International Conference on Artificial Neural Networks. Springer, Berlin, Heidelberg, 2002.

[11] Mangasarian, Olvi L., R. Setiono, and W. H. Wolberg. "Pattern recognition via linear programming: Theory and application to medical diagnosis." Large-scale numerical optimization (1990): 22-31.