Advances in Computational Research
ISSN: 0975-3273 & E-ISSN: 0975-9085, Volume 7, Issue 1, 2015, pp.-185-186
Available online at http://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BPJ0000187

MINING OF BREAST CANCER DATABASE FOR CLASSIFICATION USING DECISION TREES

PUNDE P.A.[1] AND JADHAV M.E.[2]*

[1]Department of Computer Science, Vivekanand College, Aurangabad - 431 001, MS, India.
[2]Department of Computer Science, Marathwada Institute of Technology, Aurangabad- 431 028, MS, India.
*Corresponding Author: Email- muktijadhav@gmail.com

Received: December 18, 2014; Revised: January 05, 2015; Accepted: January 15, 2015

Abstract- Breast cancer is one of the most common reasons for death of women in developed countries including India. The high incidence of breast cancer in women has increased significantly in last few years. The breast cancer diagnosis is classified from Benign to Malignant breast lumps. In this paper, classification of breast-cancer database is done to construct diagnostic rules for breast cancer. Firstly, we constructed classifiers on the database using data mining tool See5. These classifiers can be expressed as decision trees or set of rules. For each instance of the database, there are two possible classes namely, benign or malignant. This paper summarizes breast cancer diagnosis using decision trees in two predefined classes.

Keywords- Breast cancer, Diagnosis, Data Mining, Classification, See5, Decision Tree

Citation: Punde P.A. and Jadhav M.E. (2015) Mining of Breast Cancer Database for Classification using Decision Trees. Advances in Computational Research, ISSN: 0975-3273 & E-ISSN: 0975-9085, Volume 7, Issue 1, pp.-185-186.

Introduction

Breast cancer has become the primary reason of death in women in developed countries. The most effective way to reduce breast cancer deaths is to detect it earlier. Early diagnosis needs an accurate and reliable diagnosis procedure that can be used by physicians to distinguish '*benign*' breast tumors from '*malignant*' ones without going for surgical biopsy[2]. The objectives of these predictions is to assign patients to one of the two groups either a '*benign*' that is non cancerous or a '*malignant*' that is cancerous.

Data mining is an essential step in the process of knowledge discovery in databases in which intelligent methods are applied in order to extract patterns. Predicting the outcome of a disease is one of the most interesting and challenging tasks to develop data mining applications. The use of computers with automated tools, large volumes of medical data are being collected and made available to the medical research groups[2].

Overview

Data mining models can be one of the two types: predictive or descriptive. The classification task uses predictive model. It makes a prediction about values of data using known results found from different data. Other predictive model tasks are regression, time series analysis, prediction.

Motivation and Aim

Breast cancer is the most common cancer among women. The malignant tumor develops when cells in the breast tissue divide and grow without the normal controls on cell death and cell division. With early diagnosis, 97% of women survive for more years. The two main types are chemotherapy and hormone therapy are systematic therapies [4].

In this paper, we considered breast cancer data base[1]. Our aim is to construct classifiers that make the prediction. The cases of the database can belong to one of the two classes; Benign and Malignant.

Definition: Given a database D = { $t_1, t_2, t_3, \ldots\ldots\ldots\ldots, t_n$ } of tuples (items, records) and a set of classes C = { $C_1, \ldots\ldots, C_m$}, the classification problem is to define a mapping f : D à C where each t_i is assigned to one class. A class, C_j, contains precisely those tuples mapped to it. It means Cj = {t_i | f (t_i) = C_j, 1<=i<=n, and t_i ϵ D }

Methodology

Our definition of classification views it as a mapping from the database to the set of classes. Here, the classes are predefined (Benign or Malignant) as well as non-overlapping and partition the entire database. The problem is implemented in two phases:

Firstly, we create a specific model by evaluating training data. This step has training data as a input and a definition of the model as a output. The model created classifies the training data most accurately.

Apply the model developed in step 1 by classifying tuples from the target database.

According to the above mentioned definition, the second step actually does the classification, most research has been applied to step 1. Step 2 is often straight forward. Several algorithms can be used

in such type of classification. One can perform classification using one of the following techniques Statistical based algorithms, Distance based algorithms, Decision tree based algorithms, Neural network based algorithm, Rule based algorithms, Combining techniques.

For classification of breast cancer data base, we used decision tree based algorithm. The decision tree based approach is most useful in classification problems. With this technique, a tree is constructed to model the classification process. When the tree is built , it is applied to each tuple in the database and results in a classification for that tuple. There are two basic steps in this technique, first is building the tree and then applying the tree to the database.

Decision Tree

The decision tree approach divides the search space into rectangular regions. A tuple is classified based on the region into which it falls. Decision tree learning is one of the most widely used and practical methods for classification. In this method, learned trees can be represented as asset of if-then rules that improve human readability. Decision trees are very simple to understand and interpret by domain experts. A decision tree consists of nodes

A decision tree or classification tree is a tree associated with database D has the following properties: Each internal node is labeled with an attribute, Ai, Each arc is labeled with a predicate that can be applied to the attribute associated with the parent, Each leaf node is labeled with a class, Cj, For classification of breast cancer database, following two steps are used:

Decision tree induction: Construct a DT using training data.

For each $t_i \in D$, apply the DT to determine its class.

In the proposed work, in order to classify the breast cancer database, we used a data mining tool See5 by Rulequest Research Private Ltd. The database was obtained from the University of Wisconsin Hospital, Madison. Using this See5 tool, we applied classifier on breast cancer database. Here, the output is a decision tree.

Evaluation of the Result

Interpretation of Decision Tree

The DT employs a case's attribute values to map it to a leaf designating one of the two classes: benign and malignant. Every leaf of the tree is followed by a cryptic (n) or (n/m). The value of n is the number of cases in the database that are mapped to this leaf, and m (if appears) is the number of them that are classified incorrectly by the leaf, as well as it indicates class.

Evaluation on Training data (400 cases):

Decision Tree	
Size	Errors
8	10(2.5%)
(a)	(b)<- classified as
253	3 (a): class 2(Benign)
7	137 (b): class 4 (Malignant)

These values are computed for the particular classifier and training cases. If we change any of this, it would give different values.

Conclusion

To help physicians in the diagnostic of breast cancer, recent research has looked into the development of computer aided diagnostic tools. Various data mining techniques have been widely used for breast cancer diagnostics. In this paper, our study has shown that decision tree is the most useful approach in classification of breast cancer database because only two classes are there. With this technique using see5, each case belongs to one of a small number of mutually exclusive classes Benign and Malignant. Properties of every case that may be relevant to its class are provided. There are 10 attributes in this case, but see5 can deal with any number of attributes.

Conflicts of Interest: None declared.

References

[1] Sarvestani A.S., Safavi A.A., Parandeh N.M. & Salehi M. (2010) *IEEE 2nd International Conference on Software Technology and Engineering*, 2, V2-227.

[2] Delen D., Walker G. & Kadam A. (2005) *Artificial Intelligence in Medicine*, 34(2), 113-127.

[3] Lim T.S., Loh W.Y. & Shih Y.S. (2000) *Machine Learning*, 40(3), 203-228.