
International Journal of Review and Research in Applied Sciences and Engineering (IJRRASE)
Vol.5 No.2. PP 58- 63. 2013
ISSN: 2231 – 0061X

# AN ANALYSIS ON BREAST CANCER USING CLASSIFICATION

T.Balasubramanian
Department of Comp. Sci
Sri Vidya Mandir Arts And Science College
Uthangarai, Krishnagiri (Dt).
Tamilnadu, India
balaeswar123@gmail.com.

## ABSTRACT

Data mining is the process of discovering interesting knowledge from large amount of data stored in databases and these techniques based on advanced analytical methods and tools for handling a large amount of information. Much research is being carried out in applying data mining to a variety of applications in healthcare.This thesis explores data mining techniques in health care. In particularly, it discusses about data mining and its various applications in areas where people are mostly affected rigorously by cancer in Erode District, Tamil Nadu, and India.The people affected by cancer using tobacco, chemical water. This thesis identifies the cancer level using C4.5 classification and Naïve Bayes algorithms and finds meaningful patterns which gives meaningful decision making to this socio – economic real world health venture.

**Keywords: Data mining, Cancer, Classification, Naive bayes,**

### 1. INTRODUCTION

Breast cancer occurs due to an uncontrolled growth of cells in the breast tissues. Tumor is an abnormal cell growth cell growth that can be either benign or malignant. Benign tumors are non – invasive while malignant tumors are cancerous and spread to other parts of the body.

Breast cancer is a malignant tumor, grew from cells of the breast. Hence, cancer of breast tissue is called breast cancer. Worldwide, it is the most common form of cancer in females that is affecting approximately 10% of all women at some stage of their life in the Western world.

Although significant efforts are made to achieve early detection and effective treatment but scientists do not know the exact causes of most breast cancer , they do know some of the risk factors (i.e. ageing, genetic risk factors, family history, menstrual periods, not having children, obesity )that increase the likelihood of developing breast cancer in females.

### 1.1 Objectives

The Main Objective of this research is to asses and classifies the people who all are affected by the cancer and also find the age factor of cancer people. The objective of the research have been :

1. To study and analysis more on cancer types and symptoms.

2. To Design the concepts to carry out data mining on cancer data to classify the people and find the age factor

The Research is undertaken to gauge the seriousness of the impact of cancer due to the utilization of tobacco, smoking containment water available in and around Erode District. It also aims to find out the age group of people who are mostly affected by this cancer. Though it is widely known that cancer is an academic problem.

## 2. DATAMINING TECHNIQUES:

Data mining is the non-trivial process of identifying valid, Novel, Potentially useful, and ultimately understandable pattern in data. With the widespread use of databases and the explosive growth in their sizes, there is problem with information overload. The problem of effectively utilizing these massive volumes of data is becoming a major hurdle for all enterprises.

Data mining**,** also called knowledge discovery in databases, in computer science, the process of discovering interesting and useful patterns and relationships in large volumes of data. The field combines tools from statistics and artificial intelligence (such as neural networks and machine learning) with database management to analyze large digital collections, known as data sets. Data mining is widely used in business (insurance, banking, retail), science research (astronomy, medicine), and government security (detection of criminals and terrorists).

### 2.1. Classification

The Classification of large data sets is an important problem in data mining. The classification problem can be simply stated as follows. For a database with a number of records and for a set of classes such that each record belongs to one of the given classes, the problem of classification is to decide the class to which a given record belongs. But there is much more to this than just simply classifying.

Classification is a form of data analysis that can be used to extract models describing important data classes. Such analysis will provide us a better understanding of the data at large. The classification predicts categorical (discrete, unordered) labels. Classification have numerous applications, including fraud detection, target marketing, performance prediction, manufacturing and medical diagnosis.

### 2.2. C4.5 Classification

C4.5 classification algorithm has been applied to breast cancer dataset to classify patients into either “Carcinoma in situ” (beginning or pre-cancer stage) or“Malignant potential” group.

Classification is one of the most frequently studied problems by DM and machine learning (ML) researchers. It consists of predicting the value of a (categorical) attribute (the class) based on the values of other attributes (the predicting attributes).

## 3. WEKA 3.6.4 Data Miner Tool:

### 3.1. Introduction

In this thesis, WEKA (to find interesting patterns in the selected dataset), a data mining tool has been used for classification and classification and clustering techniques. The selected software is able to provide the required data mining functions and methodologies. The suitable data format for WEKA data mining software are MS Excel and ARFF formats respectively. It provides scalability in minimum number of columns and rows the software can efficiently handle. However, in the selected data set, the number of columns and the number of records were reduced.

WEKA is developed at University of Waikato in New Zealand. “WEKA” stands for the Waikato Environment of Knowledge Analysis. the system is written in java, an object-oriented programming language that is widely available for all major computer platforms, and WEKA has been tested under Linux, Windows, and Macintosh operating systems.

Java provides a uniform interface to many different learning algorithms, salong with methods for pre and post processing and for evaluating the result of learning schemes on any given dataset. WEKA excepts the data to be fed into ARFF format (Attribution Relation File Format).

### 3.2. The WEKA workbench

The WEKA project aims to provide a comprehensive collection of machine learning algorithms and data preprocessing tools to researchers and practitioners a like. It allows users to quickly try out and compare different machine learning methods on new data sets. Its modular, extensible architecture allows sophisticated data mining processes to be built up from the wide collection of base learning algorithms and tools provided. Extending the toolkit is easy; thanks to a simple API, plugin mechanisms and facilities that integration of new learning algorithms with WEKA’s graphical user interfaces.

The workbench includes algorithms for regression, classification, clustering, association rule mining and attribute selection. Preliminary exploration of data is well catered for by data 42 Visualization facilities and many

preprocessing tools. These, when combined with statistical evaluation of learning schemes and visualization facilities schemes and visualization of the results of learning, supports process models of data mining such as CRISP-DM.


Figure 3.1: WEKA Graphical User Interface

**WEKA'S Graphical User Interfaces**

WEKA has several graphical interfaces (figure 3.1) that the enable easy access to the underlying functionality. The main graphical user interface is the "Explorer". In the first panel based interface, where different panels correspond to different data mining tasks. In the first based interface, where different panels correspond to different data mining tasks. In the first panel, called "preprocess" panel, data can be loaded and transformed using WEKA's data preprocessing tools, called "filters".

This panel is shown in figure 3.2 data can be loaded from various sources, including files, URLs and databases. Supported file formats include WEKA's own ARFF format, CSV, Lib SVM's format, and C4.5's format. It is also possible to generate data using an artificial data source and edit data manually using a dataset editor.43

Figure 3.2: the WEKA Explorer user interface

The second panel in the Explorer gives access to WEKA's classification and regression algorithms. The corresponding panel is called "Classify" because regression techniques are viewed as predictors of "continuous classes". By default, the panel runs a cross-validation for a selected learning algorithm on the dataset that has been prepared in the Preprocess panel to estimate predictive performance.

The last panel in the Explorer, called "Visualize", provides a color-coded scatter plot matrix, along with the option of drilling down by selecting individual plots in this matrix and selecting portions of the data to visualize.it is also possible to obtain information regarding individual data points, and to randomly perturb data by a chosen amount to uncover obscured data

The explorer is designed for batch data processing training and completely loaded into memory, then processed. This may not be suitable for problems involving large dataset. However, WEKA does have implementations of some algorithms that allow incremental model building, which can be applied in incremental mode from a command 45 line interface. The incremental nature of these algorithms is ignored in the Explorer, but can be exploited using a more recent addition to WEKA's set of graphical user interfaces, namely the so called "Knowledge Flow", shown in figure 3.4. most tasks that can be tackled with the Explorer can also be handled by the Knowledge Flow. However, in addition to batch-based training, its data flow model enables incremental updates with processing nodes that can load and preprocess individual instances before feeding them into appropriate incremental learning algorithms. It also provides nodes for visualization and evaluation. Once a set-up of interconnected processing nodes has been configured, it can be saved for later re-use. The third main graphical user interface in WEKA is the "Experimenter" (see Figure 3.4). This interface is designed to facilitate experimental comparison of the predictive performance of algorithms based on the many different evaluation criteria that are available in WEKA. Experiments can involve multiple algorithms that are run across multiple dataset; for example, using repeated cross-validation.

Figure .3.3: The WEKA Knowledge Flow user interface

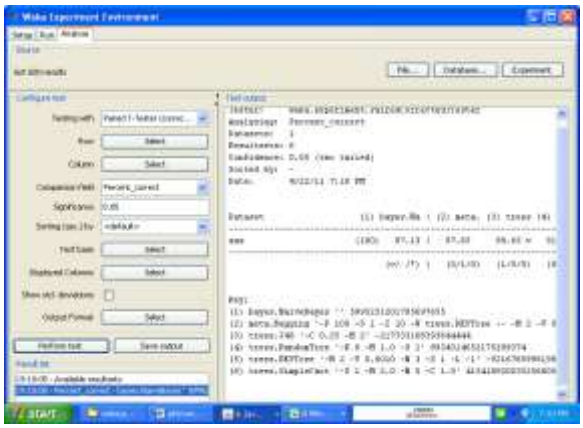

Figure 3.4: The WEKA Experimenter user interface

Compared to WEKA's other user interfaces, the Experimenter is perhaps used less frequently by data mining practitioners. However, once preliminary experimentation has been performed in the Explorer, it is often much easier to identify a suitable algorithm for a particular dataset, or collection of datasets, using this alternative interface.

## 4. Analysis Through C4.5 And Naïve Bayes Algorithm:

### 4.1 Analysis of C4.5 (J48)

The algorithm constructs a decision tree starting from a training set T S, which is asset of cases, or tuples in the database terminology. Each case specifies values for a collection of attributes and for a class. Each attribute may have either discrete or continuous values. Moreover, the special value unknown is allowed, to denote unspecified values. The class may have only discrete values. We denote with C1,……..$C_{NClass}$ the values of the class.

### 4.2 Analysis of Naïve Bayes

Bayesian classification is quite different from the decision tree approach. In Bayesian classification a hypothesis is made that the given data set belongs to a particular class, then the probability for the hypothesis is calculated. This is among the most practical approaches for certain types of problems. The approach requires only one scan of the whole data set.

The expression P (A) refers to the probability that event A will occur. P(A|B) stands for the probability that event A will happen given that event B has already happened. In other words P (A|B) is the conditional probability of A based on the condition that B has already happened. For example, A and B may be probability of passing a course A and passing another 80 course B respectively. P (A|B) then is the probability of passing A when we know that B has been passed.

Now the Bayes theorem

$$P(A|B) = P(B|A)P(A)|P(B)$$

If we consider X to be an object to be classified then Bayes theorem may be read as giving the probability of it belonging to one of the classes C1, C2, C3, etc by calculating P(Ci|X). Once these probabilities have been computed for all the classes, we simply assign X to the class that has the highest conditional probability.

P(Ci/X) may be calculated as
$$P(Ci|X) = [P(X|Ci)P(Ci)]|P(X)$$

- P(Ci|X) is the probability of the object X belonging to class Cr
- P(X|Ci) is the probability of obtaining attribute values X if we know that it belongs to class Cr
- P(Cr) is the probability of any object belonging to class Ci without any other information.
- P(X) is the probability of obtaining attribute values X whatever class the object belongs to.

## 5. Result and Discussion

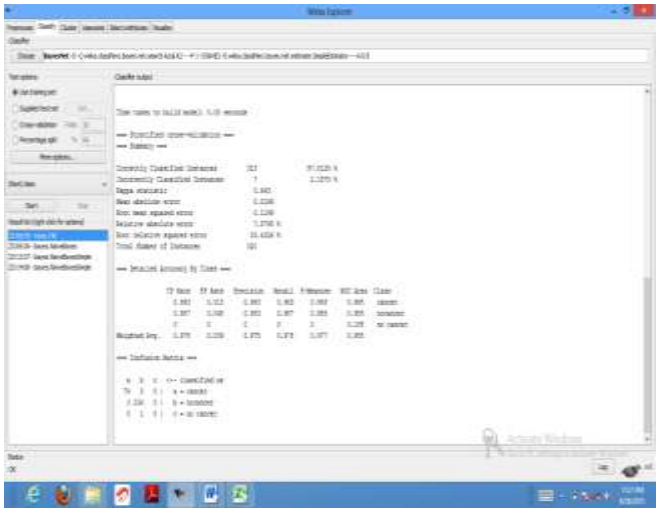

Figure 5.1.Run information in J48 Algorithm

Table 5.1: Comparison between C4.5 and Naïve Bayes algorithms

| Classification Algorithms | % of correctly classified instances | Result Mean squared error | Kappa statistic | Total No. of Instances | Correctly Classified Instances | No.of attributes | Time duration |
|---|---|---|---|---|---|---|---|
| J48 (C4.5) | 97.8125% | 0.11999 | 0.943 | 320 | 113 | 10 | 0.08sec |
| NaïveBaye's | 98.4375% | 0.0937 | 0.9596 | 320 | 315 | 10 | 0.02sec |

Figure5.2.Run Information In Naïve Bayes Algorithm

### 5.1 Performance Analysis

In this Thesis work carried out Classification C4.5 Algorithm and Naïve Bayes Algorithm. The above first and second figure displays the Breast Cancer Databases. The third figure displays J48 (C4.5) Classification Algorithm results. And the fourth figure displays Naïve Bayes Algorithm results. The given two algorithms show same results. J48 classification algorithm takes the processing time 0.08 seconds. Naïve Bayes algorithm takes the processing time 0.02 seconds only. Here the comparison is taken through time based. The same results have two different times. Here the Naive Bayes algorithm has taken less time than J48 (C4.5) Algorithm. So the Naïve Bayes algorithm is best one. That means the Naïve Bayes algorithm is effective and efficient for Healthcare domain.

In this thesis works are compared to predict the best algorithm. Experimental results show the effectiveness of the best algorithm. We used a Breast cancer database of a sample of 320 records and then applied the C4.5 (J48) classification and Naïve Bayes algorithm obtained to the full Breast Cancer Database. We obtained to get accuracy results from these algorithms.J48 classification algorithm takes the processing time 0.08 seconds. Naïve Bayes algorithm takes the processing time 0.02 seconds only. Here the comparison is taken through time based. The same results have two different times. Here the Naive Bayes algorithm has taken less time than J48 (C4.5) Algorithm. So the Naïve Bayes algorithm is best one. That means the Naïve Bayes algorithm is effective and efficient for Healthcare domain.

## 6. Conclusion

In this thesis works are compared to predict the best algorithm. Experimental results show the effectiveness of the best algorithm. We used a Breast cancer database of a sample of 320 records and then applied the C4.5 (J48) classification and Naïve Bayes algorithm obtained to the full Breast Cancer Database. We obtained to get accuracy results from these algorithms.J48 classification algorithm takes the processing time 0.08 seconds. Naïve Bayes algorithm takes the processing time 0.02 seconds only. Here the comparison is taken through time based. The same results have two different times. Here the Naive Bayes algorithm has taken less time than J48 (C4.5) Algorithm. So the Naïve Bayes algorithm is best one. That means the Naïve Bayes algorithm is effective and efficient for Healthcare domain.

## References

1) Breast cancer facts and figures http://www.brest cancer.org/.
2) Breast cancer statics from centers for Disease Control and Prevention, http://www.cdc.gv/cancer/brest/statistics./.
3) K.Gajalaksmi, V.Shanta, R. Swaminathan,R.Sankaranarayanan,andR.J.Black,"A population – based survival study on female breast cancer in Madras, India",Cancer Institute (WIA), Adayar, Madras, India.
4) Quinlian, J. R. C4.5:Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

5) http://en.wikipedia.org /wiki/C4.5 algorithm – C4.5 Algorithm Description.
6) Abdulah H.Wahbeh ,Qasem A. Al – Radaideh,Mohammed N.Al-Kabi, and Emad M. Al-Shawakfa,"A Comparision study between Data Mining Tools over some Classification Methods",Internationals Journal of Advanced Computer Science and Applications.
7) Introduction to Data mining with case studies - G.K.Gupta PHI.
8) Berry Mj Linoff G Data Mining Techniques: for Marketing, Sales and Customer support USA.Wiley,1997.
9) WEKA 3.6.4 data miner manual.
10) Reutemann, Ian H. Witten, Pentaho Corporation, Department of Computer Science.
11) The WEKA Data Mining Software: An Update, White paper, Peter.
12) Reutemann, Ian H. Witten, Pentaho Corporation, Department of Computer Science.
13) Towards the use of C4.5 Algorithm for classifying banking data set – Veronica S.Moertrhni – Integral Vol. 8 No. 2, October 2003.
14) Tan.P.N. Steinbach .M & Kumar.V "Introduction to Data Mining", PearsonEducation, 2006.