Breast cancer prediction system using Data mining methods

Dr. C Nalini1 D.Meera2
1Professor, 2Mtech Student, Computer Science and Engineering,
BIST, BIHER,BharathUniversity,Chennai – 73.
nalini.cse@bharathuniv.ac.in

ABSTRACT

Cancer is the second most common cause of death worldwide with an estimated 7.9 million of deaths in 2007. This number is projected to rise further and reach 12 million deaths in 2030 which makes cancer a major public health issue. Data mining is defined as a process used to extract usable data from a large data set. There are many different areas under data mining and one of them is classification or the supervised learning. In the health care industry, the data mining is predominantly used for disease prediction. Classification also can be implemented through a number of different approaches or algorithms. There are many different areas under Data Mining and one of them is Classification or the supervised learning. We have conducted the comparison between two algorithms with help of WEKA (The Waikato Environment for Knowledge Analysis), which is an open source software. It contains different type’s data mining algorithms. This paper explains discussion of Bayesian Network and J48 algorithms. Here, for comparing the result, we have used the analysis of classification accuracy and execution time and various parameters.

KEYWORDS

Classification, Decision tree [J48], Naïve Bayes, Breast cancer

1. INTRODUCTION

The second leading cause of death among women is breast cancer and it comes directly after lung cancer [1-9]. The health and medical sector is more in need of data mining today. When certain data mining methods used, valuable information can be extracted from large data base and that can help to medical practitioner to take decision, and improve health services. There are a few arguments that can support the use of data mining in health sector for breast cancer like early detection, early avoidance, and indication based medication, rectifying hospital data errors [10-15]. WEKA is a powerful tool as it contains supervised learning as well unsupervised learning methods. It contains Classification, Clustering, Association Mining, Feature Selection, Data Visualization, etc. The main reason behind using WEKA is helps researchers like us to implement and compare data mining techniques very easily on real or synthetic data. It is also well suited for developing new machine learning techniques [5]. WEKA comes under the open source software issued under GNU General Public License [16-21].
Breast cancer is an uncontrolled growth of breast cells. It refers to a malignant tumor that has developed from cells in the breast. Breast Cancer constitutes a major public health issue globally with over 1 million new cases diagnosed annually, resulting in over 400,000 annual deaths and about 4.4 million women living with the disease. It also affects one in eight women during their lives. It is the commonest site specific malignancy affecting women and the most common cause of cancer mortality in women worldwide. It is also found in men but not very common. Statistics available in Nigeria are largely unreliable because of many factors that have not allowed adequate data collection and documentation; but according to Dr Chiinyere Akpanika, a Gynaecologist at the University of Calabar Teaching Hospital (UCTH) said Nigeria recorded over 100,000 new cases of cancer annually, added that early detection of the disease increased the chance of survival of an affected person. According to her, 85 per cent of women who have breast cancer do not have a family history of breast cancer. The Medical Director, Optimal Cancer Care Foundation, Dr Femi Olaleye, said breast cancer killed one in every 25 Nigerian women.

Breast cancer is a malignant or benign tumor, inside breast, wherein cells divide and grow without control [30-34]. Scientists have tried to know the exact reason behind breast cancer, as there are a few risk factors which increase the like hood of a woman developing breast cancer. Age, genetic risk and family history are some such factors being considered for breast cancer[3]. Treatments of breast cancer are divided into two types, local and systematic. Surgery and radiation are local type of treatments whereas chemotherapy and hormone therapy are examples of systematic therapies. For getting best results, both treatments are used together in different variations as per the patient and disease intensity [35-39]. These records are cleaned and filtered with the intention that the irrelevant data from the warehouse would be removed before mining process occurs.

The aim of this paper is to predict the Breast Cancer effectively by applying the attribute subset evaluator namely Consistency Subset Evaluator. Using these significant attributes the classification of dataset is performed using naïve bayesand J48. The remaining portion of the paper is discussed as follows. The proposed methodology is given in Section 3. Section [4] analyses the experimental results. Section [40-45] gives conclusion.

2. LITERATURE SURVEY:

Dr. Rajesh et al (2012) who used SEER dataset for the diagnosis of breast cancer using the C4.5 classification algorithm. The algorithm was used to classify patients into either pre-cancer stage or potential breast cancer cases. During the testing phase, the C4.5 classification rules were applied to a test sample and the algorithm showed had an accuracy of 92.2%, sensitivity of 46.66% and a specificity of 97.4%. Future enhancement of the work will require the improvisation of the C4.5 algorithm to improve classification rate to achieve greater accuracy.

Shajahan et al (2013) worked on the application of data mining techniques to model breast cancer data using decision trees to predict the presence of cancer. Data collected contained 699 instances (patient records) with 10 attributes and the output class as either benign or malignant. Input used contained sample code number, clump thickness, cell size and shape uniformity, cell growth and other results physical examination[6]. The results of the supervised learning
algorithm applied showed that the random tree algorithm had the highest accuracy of 100% and error rate of 0 while CART had the lowest accuracy with a value of 92.99% but naïve bayes’ had the an accuracy of 97.42% with an error rate of 0.0258.

Chaurasia and Pal6 offered three popular data mining algorithms: CART, ID3 (iterative dichotomized 3), and DT for diagnosing heart diseases, and the results presented demonstrated that CART obtained higher accuracy within less time.

Chaurasia and Pal5 compare the performance criterion of supervised learning classifiers, such as Naïve Bayes, SVM-RBF kernel, RBF neural networks, Decision Tree (Dt) (J48), and simple classification and regression tree (CART), to find the best classifier in breast cancer datasets. The experimental result shows that SVM-RBF kernel is more accurate than other classifiers; it scores at the accuracy level of 96.84% in the Wisconsin Breast Cancer (original) datasets.

Qi Fanet.al [6] the authors of this paper focus on SEER public use-data to predict Brest Cancer. They use pre-classification method and find a possible solution to discover their formation of Brest Cancer.H.S.Hota [2] built a classification model using various intelligent techniques such as ANN(Artificial Neural Network),Unsupervised Artificial Neural Network, Statistical technique and decision tree. Experimental results show a testing accuracy of 97.73% from which the efficiency of the ensemble model was highlighted.

Ryan Potter has performed the comparison of classification algorithms on breast cancer dataset to perform the diagnosis of the patients.

3. METHODOLOGY:

3.1 Dataset:
The breast cancer dataset have been created for analysis of diabetic disease. This dataset contains seven hundred and sixty eight instances and eight attributes are used in this comparative analysis.

Menstrual and reproductive history:
Earlymenstruation (before age 12), late menopause(after age 55), having your first child at an olderage, or never having given birth can also increase the risk of breast cancer.

Certain genome:
This is cause by mutation in certain genes and can be determined through genetic test as individual with certain gene mutation can pass it onto their children.

Dense breast tissue:
This can increase the risk of having breast cancer and makes lumps harder to detect.

Population
This study was performed on TMA images from 164 fe- male patients with an invasive ductal carcinoma of the breast, treated at the Oncology Institute of Vilnius University and investigated at the National Center of Pathology, during the period of 2007 to 2009. The study was approvedby the Lithuanian Bioethics Committee. The patients’ con- sent to participate in the study was obtained.

Tissue preparation
The TMAs were constructed, stained and scanned as described previously [3]. Briefly, onemillimetre-diameter cores were punched from tumour areas randomly selected by the pathologist and paraffin sections were cut at 3μm-thickness[3].
Immunohistochemistry (IHC)

IHC for Ki67 was performed with a multimer-technology based detection system, ultraView Universal DAB (Ventana, Tucson, AZ, USA). The Ki67 antibody (clone MIB-1; DAKO, Glostrup, DK) was applied at a 1:200 dilution for 32 minutes, followed by the VentanaBenchMark XT automated immunostainer (Ventana) standard Cell Conditioner1 (CC1, a proprietary buffer) at 95°C for 64 minutes[5]. Finally, the sections were developed in DAB at 37°C for eight minutes, counterstained with Mayershematoxylin and mounted.

Image acquisition

Digital images were captured using the Aperio Scan Scope XT Slide Scanner(Aperio Technologies, Vista, CA, USA) under 20x objective magnification (0.5μm resolution). One TMA spot image per patient was used for the study.

Quantification with stereology test grid

RV were obtained by marking Ki67-positive and negativetumour cell profiles, using a stereological method for 2Dobject enumeration [4,5] implemented by the Stereology module (ADCIS, Caen, France) with a test grid of systematically sampled frames (frame size - 125 pixels, spacing of frames - 250 pixels) overlaid on a spot image in ImageScope (Aperio Technologies, USA).

Visual evaluation (VE)

A global subjective impression for the Ki67 LI on the same images was performed by five pathologists independently and provided semi-quantitative values (Ki67-VE-1, 2, 3, 4 and 5) expressed as the percentage of Ki67-positivetumour cell profiles. Counting was not included in the procedure.

3.2 Classification

Classification is a process that is used for partitioning the data into different classes according to some constrains or it classify each item in a dataset into one of predefined set of classes or groups. It is a supervised learning approach having known class categories. Several major kinds of classification algorithms including C4.5, J48, k-nearest neighbor classifier, Naive Bayes, SVM, Apriori, and AdaBoost. In this research work Naïve Bayes, decision tree[J48] algorithms are used to predict the breast cancer disease.

3.2.1 Naïve Bayes

Naive Bayes is a machine learning algorithm for classification problems. It is based on Thomas Bayes’s probability theorem. It is primarily used for text classification which involves high-dimensional training datasets. A few examples are spam filtration, sentimental analysis, and classifying news articles. It is not only known for its simplicity but also for its effectiveness. It is fast to build models and make predictions with Naive Bayes algorithm. Naive Bayes algorithm is the algorithm that learns the probability of an object with certain features belonging to a particular group/class. In short, it is a probabilistic classifier. The Naïve Bayes algorithm is called “naive” because it makes the assumption that the occurrence of a certain feature is independent of the occurrence of other features. The “Bayes” part refers to the statistician and philosopher Thomas Bayes and the theorem was named after him, Baye’s theorem, which is the base for Naïve Bayes algorithm. Naïve Bayes algorithm is Baye’s theorem or alternatively known as Baye’s rule or Baye’s law. It gives us a method to calculate the conditional probability, i.e. the probability of an event based on previous knowledge available on the events. More formally, Baye’s theorem is stated as the following equation 

\[
P(A|B) = \frac{P(B|A)P(A)}{P(B)}
\]
\[ P(A/B) = \frac{P(B/A) P(A)}{P(B)} \]

- \( P(A/B) \): Probability (conditional probability) of occurrence of event A given the event B is true.
- \( P(A) \) and \( P(B) \): Probabilities of the occurrence of event A and B, respectively.
- \( P(B/A) \): Probability of the occurrence of event B given the event A is true.

### 3.2.2 Decision Trees (J48)

J48 decision trees classifier is a simple decision learning algorithm, it accepts only categorical data for building a model. The basic idea of ID3 is to construct a decision tree by employing a top down greedy search through the given sets of training data to test each attribute at every node. It uses statistical property known as information gain to select which attribute to test at each node in the tree. Information gain measures how well a given attribute separates the training samples according to their classification.

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.

Given a set \( S \) of breast cancer cases, J48 first grows an initial tree using the divide-and-conquer algorithm as follows:

- If all the cases in \( S \) belong to the same class or \( S \) is small, the tree is a leaf labeled with the most frequent class in \( S \);
- Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition \( S \) into corresponding subsets \( S_1, S_2, \ldots, S_n \) for a dataset containing \( n \) cases according to the outcome for each case, and apply the same procedure recursively to each subset. It uses a statistical property known as information gain to select which attribute to test at each node in the tree. It measures how well a given attribute separates the training samples according to their classification.

### 4 EXPERIMENTAL RESULTS

This work is implemented in Weka tool. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. The experimental comparison of classification algorithms are done based on the performance measures of classification accuracy and execution time.

#### 4.1 Classification Accuracy

**Accuracy:**

Accuracy is defined in the terms of correctly classified instances divided by the total number of instances present in the dataset.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

Where TP- True Positive, FP- False Positive, TN- True Negative, FN- False Negative

**TP Rate:**
It is the ability which is used to find the high true-positive rate. The true-positive rate is also called as sensitivity.

$$TPR = \frac{TP}{TP + FN}$$

**Precision:**

Precision is given the correlation of number of modules correctly classified to the number of entire modules classified fault-prone. It is quantity of units correctly predicted as faulty.

$$\text{Precision} = \frac{TP}{TP + FP}$$

**F-Measure:**

F-Measure is the one has the combination of both precision and recall which is used to compute the score.

$$\text{F-Measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

Table: Accuracy Measure for Classifier Algorithms

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Correctly classified instances (%)</th>
<th>Incorrectly classified instances (%)</th>
<th>TP rate</th>
<th>F Measure</th>
<th>IR precision</th>
<th>IR Recall</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve bayes</td>
<td>215</td>
<td>71</td>
<td>0.85</td>
<td>0.81</td>
<td>0.78</td>
<td>0.85</td>
<td>64%</td>
</tr>
<tr>
<td>J48</td>
<td>217</td>
<td>69</td>
<td>0.95</td>
<td>0.84</td>
<td>0.75</td>
<td>0.95</td>
<td>60%</td>
</tr>
</tbody>
</table>

From the table, in performance measures Decision tree(J48) performs best in classifying process than Naïve Bayes algorithm and In accuracy naïve bayes algorithm is better than J 48.

In explorer:

[Figure: decision tree]
In knowledge flow environment:

4.2: Execution time

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Execution time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve bayes</td>
<td>0</td>
</tr>
<tr>
<td>J48</td>
<td>0.1</td>
</tr>
</tbody>
</table>

From the table naïve Bayes takes less execution time than the J48 algorithm.

5. RESULT AND DISCUSSION

The simulation results are partitioned into several sub items for easier analysis and evaluation. The algorithm will be compared by analyzing their performance; execution time. The algorithm which has the higher accuracy with the minimum execution time has chosen as the best algorithm. In this classification, naïve bayes has the maximum classification accuracy, performance factors and minimum execution time. So it is considered as the best classification algorithm. From the results, it is very clear that data mining techniques can be used in predicting breast cancer risks and that the J48 decision trees has a better accuracy than the naïve bayes’ model which is a statistical tool.

6. CONCLUSION

In this paper, we applied two prediction models for breast cancer. Here, we used two popular data mining methods: Naive Bayes and J48. In this research work classification process is used to classify four types of kidney diseases. Comparison of J48 and Naive Bayes classification algorithms is done based on the performance factors classification accuracy and execution time. From the results, the Naive bayes that are accurate, hence it is considered as best classifier when compared with J48 algorithm. Perhaps, Naïve Bayes classifier classifies the data with minimum execution time.
REFERENCES
