Two Level Diagnosis of Breast Cancer Using Data Mining

Rajkamal kaur Grewal
Department of computer application
Phagwara, Punjab
Lovely Professional University

Babita Pandey
Department of computer application
Phagwara, Punjab
Lovely Professional University

ABSTRACT
Breast Cancer is a dreadful disease. Mostly women affected with breast cancer disease. Mainly problem in medical science is to diagnosis of breast cancer at early stage. So the early detection of breast cancer is important for saving life. In this work, develop method for diagnosis of breast cancer at two levels. At the first level diagnosis is based Wisconsin Breast Cancer dataset (pathological test result) and classified into malignant and benign class. At the second level diagnosis based on pathological and physiological parameters of malignant breast cancer dataset and classified into five breast cancer disease as: Ductal Carcinoma in Situ(DCIS), Lobular Carcinoma in Situ(LCIS), Invasive Ductal Carcinoma(IDC), Invasive Lobular Carcinoma(ILC) and Mucinous Carcinoma(MC). In this paper evaluate the performance based on correct and incorrect element of data classification using J48 classification algorithm. The experiment result shows that classification accuracy, sensitivity and specificity of J48 is good.

Keywords
Breast cancer, J48 decision tree, WEKA, Classification, ROC Curve.

1. INTRODUCTION
Cancer is a harmful type of diseases that causes the cells in a part of body start to grow out of control. The cancer's cell growth is different from normal cell growth and finally the cancer cell becomes a tumor.[7] Breast cancer is a tumor related disease. The initial detection of breast cancer is needed for saving life of people. But the early diagnosis needed accurate and reliable diagnosis technique that allow physician to distinguish malignant tumor from the benign tumor and distinguish the different malignant breast cancer type as: Ductal Carcinoma in Situ (DCIS), Lobular Carcinoma in Situ (LCIS), Invasive Ductal Carcinoma(IDC), Invasive Lobular Carcinoma(ILC) and Mucinous Carcinoma(MC).[15]

Data mining (DM) is known as knowledge discovery that extract and discover hidden features. The data mining have been widely used in prediction, mainly in medical diagnosis [5]. The Data mining function classified as classification, clustering, feature selection and associating mining [14]. In this work data classification function applied and calculate the performance of classifier algorithm with True positive rate, False positive rate that generated after applied the J48 algorithm on the breast cancer dataset.

A lot of work has been done for diagnosis of breast cancer but nobody has classified it at two levels. This paper presents the data mining (J48) classifier that applied on breast cancer dataset at two levels. At the first level diagnosis is based Wisconsin Breast Cancer dataset (pathological test result) and classified into malignant and benign class. At the second level diagnosis based on pathological and physiological parameters of malignant breast cancer dataset (collect from hospital) and classified into five breast cancer disease as: Ductal Carcinoma in Situ(DCIS), Lobular Carcinoma in Situ(LCIS), Invasive Ductal Carcinoma(IDC), Invasive Lobular Carcinoma(ILC) and Mucinous Carcinoma(MC).

The rest of this paper is organized as follow: section 2 describes two level diagnosis of breast cancer. Section 3 deals with experimental work and result. The conclusions are given in section 4.

2. TWO LEVEL DIAGNOSIS OF BREAST CANCER

The diagnosis of breast cancer is done at two levels.

2.1 At the First level diagnosis
The J48 classifier is a Quinlan’s C4.5 decision tree which is develop from a set of labeled training data and permit classification by decision tree or rules produced from them [6].

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Data type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cellularity</td>
<td>Categorical(High, Medium, Low)</td>
</tr>
<tr>
<td>Marginal Adhesion</td>
<td>Categorical(High, Medium, Low)</td>
</tr>
<tr>
<td>Epithelial Cell Size</td>
<td>Categorical(High, Medium, Low)</td>
</tr>
<tr>
<td>Bare Nuclei</td>
<td>Categorical(High, Medium, Low)</td>
</tr>
<tr>
<td>Nucleoli</td>
<td>Categorical(High, Medium, Low)</td>
</tr>
<tr>
<td>Bland Chromatin</td>
<td>Categorical(High, Medium, Low)</td>
</tr>
<tr>
<td>Mitoses</td>
<td>Categorical(High, Medium, Low)</td>
</tr>
<tr>
<td>Swelling</td>
<td>Categorical(High, Medium, Low)</td>
</tr>
<tr>
<td>Lump</td>
<td>Categorical(High, Medium, Low)</td>
</tr>
<tr>
<td>Nipple Discharge</td>
<td>Categorical(High, Medium, Low)</td>
</tr>
<tr>
<td>Pain</td>
<td>Categorical(High, Medium, Low)</td>
</tr>
</tbody>
</table>

Each attribute of the data can be used to make a decision by dividing the data into subsets and leaf nodes defining the final class. While creating tree, it ignores the missing values and handle the continuous and discrete attributes [7].

At the first level, Wisconsin breast cancer(WBC) dataset used that provided by the University of Wisconsin Hospital based on microscopic examination of breast masses with the fine needle aspirate test[8]. The Wisconsin breast cancer(WBC) dataset consists of 699 patient records with nine parameters. The nine pathological parameters are: Clump Thickness (CT), Uniformity of cell size (CS), Uniformity of
cell shape (CShp), Marginal Adhesion (MA), Single Epithelial Cell Size (ECS), Bare Nuclei (BN), Bland Chromatin (BC), Normal Nuclei (NN), Mitosis (M) [3]. The dataset of breast cancer disease divide into 241 malignant case and 458 benign case. At this level Data mining applied using J48 on the breast cancer dataset and classified data in malignant and benign class.

2.2 Second level diagnosis
After first level diagnosis the cases are classified as a malignant and benign. The malignant cases are pass to second level diagnosis for further classification as: Ductal carcinoma in situ (DCIS), Lobular carcinoma in situ (LCIS), Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC) and Mucinous Carcinoma (MC). J48 algorithm was applied at second level.

At the second level data consist of attributes such as:

pathological parameters and physiological parameters. Pathological test result taken from Wisconsin breast cancer dataset from UCI machine and the physiological parameter dataset collect from hospital. The pathological parameters are [2] classified as: Cellularity (C) which further categorized as: Cellularity Scanty (CS) and Cellularity High (CH); Marginal Adhesion (MA) which is further categorized as: Marginal Loose (ML) and Marginal Tight (MT); Epithelial Cell Size (ECS) which is further categorized as: Normal Epithelial cell size (EN), Moderately Epithelial cell size (EM) and Enlarged Epithelial cell size (EE); Bare Nuclei (BN) which is further categorized as: Nuclei Present (NP) and Nuclei Absence (NA); Nucleoli (N) which is further categorized as: Nucleoli Absence (NA), Nucleoli Inconspicuous (NI) and Nucleoli Prominent (NP); Bland Chromatin (BC) further categorized as: Chromatin Stippled (CS), Chromatin Coarse (CC); Mitoses (M) further categorized as: Mitoses Abnormal (MA), Mitoses Present (MP) and Mitoses Absence (MA). The Physiological parameter are classified as: Swelling (SW), Lump (L), Nipple Discharge (ND), Pain (P).[9] The detail of dataset shown in table 1 and hierarchal relationship between the parameter and their sub parameter are shown in figure 1.

The diagnosis processes at first and second level are shown in Figure 2. In this process, firstly enter the dataset of Wisconsin’s breast cancer in the first level and classify it into benign and malignant cases with J48 algorithm and further malignant breast cancer dataset passes again at second level and classified it into different five malignant breast cancer diseases.
3. EXPERIMENT WORK AND RESULT

The classification of breast cancer at first level in malignant or benign and at the second level in Ductal Carcinoma in Situ (DCIS), Lobular Carcinoma in Situ (LCIS), Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC) and Mucinous Carcinoma (MC) done using J48 algorithm of Weka (version 3.6.4) tool.

The pruned tree of first level diagnosis and the second level diagnosis shown in given figure3 and figure4 respectively.

Some rules that generated by J48 pruned tree at second level as shown in figure4 as:

Rule1: If loose cohesive=medium; B nuclei=medium and swelling low then DCIS.

Rule2: If loose cohesive=medium; B nuclei=medium and swelling medium then DCIS

Rule3: If loose cohesive=medium; B nuclei=medium and swelling high then IDC.

Rule4: If loose cohesive=medium and B nuclei=low then DCIS.

Rule5: If loose cohesive=medium and B nuclei=high and lump high then ILC.

Rule6: If loose cohesive=medium; B nuclei=high and lump medium then LCIS.
The decision tree generated at first level and at the second level is shown in figure 5 and figure 6. In figure 6 the High is labeled as H, Low as L and Medium as M.

Confusion matrix generated from Weka tool is shown in Table 2. The results are computed using True positive and false positive with sensitivity and specificity is shown in table 3.
Some Standard terms those have been defined for the matrix like:

**True positive (TP):** Number of positive sample correctly predicted. If the outcome from a prediction is positive and the actual value is also positive, then it is called true positive.

**False positive (FP):** Number of negative sample incorrectly predicted as positive. If the actual value is negative then it is named false positive (FP).

**False negative (FN):** Number of positive sample incorrectly predicted.

**True negative (TN):** Number of negative samples correctly predicted.

### Table 2: Confusion Matrix

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>Classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>a=DCIS</td>
</tr>
<tr>
<td>0</td>
<td>18</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>b=LCIS</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>19</td>
<td>1</td>
<td>0</td>
<td>c=IDC</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>16</td>
<td>1</td>
<td>d=ILC</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>15</td>
<td>e=MC</td>
</tr>
</tbody>
</table>

In the above confusion matrix, the true positive (TP) for the class a = ‘DCIS’ is 19 and the false positive (FP) is 1,0,0,0 . The true positive (TP) for the class b = ‘LCIS’ is 18, while the false positive (FP) for class b is 0,0,2,0. For the class c = ‘IDC’, the true positive (TP) value is 19 and the false positive (FP) value is 0,1,0,0. For the class d = ‘ILC’, the true positive (TP) value is 16 and the false positive (FP) value is 0,1,2,1. For the class e = ‘MC’, true false value 15 and false value is 2,0,2,1.

The true positive rate calculated as:-

**True Positive Rate (TRP)** for class a = \( \frac{19}{19+1} = \frac{19}{20} = 0.95 \)

**False positive Rate (FPR)** for class a = \( \frac{2}{2 + 79} = \frac{2}{81} = 0.0246 \)

**True Positive Rate (TRP)** for class b = \( \frac{18}{18 + 2} = \frac{18}{20} = 0.9 \)

**False positive Rate (FPR)** for class b = \( \frac{2}{2 + 82} = \frac{2}{84} = 0.024 \)

**True Positive Rate (TRP)** for class c = \( \frac{19}{19 + 1} = \frac{19}{20} = 0.95 \)

**False positive Rate (FPR)** for class c = \( \frac{4}{4 + 79} = \frac{4}{83} = 0.05 \)

**True Positive Rate (TRP)** for class d = \( \frac{16}{16 + 4} = \frac{16}{20} = 0.8 \)

**False positive Rate (FPR)** for class d = \( \frac{4}{4 + 84} = \frac{4}{88} = 0.045 \)

**True Positive Rate (TRP)** for class e = \( \frac{15}{15 + 5} = \frac{15}{20} = 0.75 \)

**False positive Rate (FPR)** for class e = \( \frac{1}{1 + 84} = \frac{1}{85} = 0.012 \)

### Table 3: Calculation of TPR and FPR

<table>
<thead>
<tr>
<th>Class</th>
<th>True positive</th>
<th>False positive</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.95</td>
<td>0.025</td>
<td>0.95</td>
<td>0.975</td>
</tr>
<tr>
<td>B</td>
<td>0.9</td>
<td>0.024</td>
<td>0.9</td>
<td>0.976</td>
</tr>
<tr>
<td>C</td>
<td>0.95</td>
<td>0.05</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>D</td>
<td>0.8</td>
<td>0.045</td>
<td>0.8</td>
<td>0.955</td>
</tr>
<tr>
<td>E</td>
<td>0.75</td>
<td>0.012</td>
<td>0.75</td>
<td>0.988</td>
</tr>
</tbody>
</table>

The receiver operating characteristic curve displays the relationship between true positives and false positives [6]. The ROC curve is a graphical plot which demonstrates the performance of a classifier system. It is developed by plotting the fraction of true positive out of total actual positive rate and the fraction of false positive out of the total actual negative, at the threshold settings [7]. The Threshold curve of the five malignant diseases: Ductal Carcinoma in Situ (DCIS), Lobular Carcinoma in Situ (LCIS), Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC) and Mucinous Carcinoma (MC) are defined in given Figure 7,8,9,10 and 11 respectively. In this axis X defines the false positive rate, whereas its Y axis corresponds to the true positive rate.
Figure 8: Threshold Curve of Lobular Carcinoma in Situ

Figure 9: Threshold Curve of Invasive Ductal Carcinoma

Figure 10: Threshold Curve of Invasive Lobular Carcinoma

Figure 11: Threshold Curve of Mucinous Carcinoma

4. CONCLUSION

In this work, breast cancer is diagnosed at two levels. At the first level, diagnosis is based on the Wisconsin Breast Cancer dataset and classified into malignant and benign classes. At the second level, diagnosis is based on pathological and physiological parameters of malignant breast cancer dataset collected from hospitals with the help of experts and classified into five breast cancer diseases: Ductal Carcinoma in Situ (DCIS), Lobular Carcinoma in Situ (LCIS), Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC), and Mucinous Carcinoma (MC).

Data mining (J48 algorithm) generates rules and decision trees for classification of breast cancer diseases. It gives good classification accuracy, sensitivity, and specificity in diagnosis.
of breast cancer dataset at two levels. The sensitivity for all malignant breast cancer diseases are lies between 94-100%. It was observed from the result that data mining can efficiently classify the breast cancer disease. The receiver operating characteristic (ROC) curve calculate the relationship between true positive and false positive with good efficiency. Therefore, the classification of two levels helpful for doctor and patients to diagnosis of breast cancer in early stage and saving the life of effected patient.

5. REFERENCES


[14] Torill Squer, Department of Pathology, Oslo University Hospital.