SUPERVISED LEARNING APPROACH FOR BREAST CANCER CLASSIFICATION

P.Sivagami

1Assistant Professor, PSGR Krishnammal college for Women, Coimbatore, India

Abstract: Breast cancer is a cancer that starts in the breast, usually in the inner lining of the milk ducts or lobules. Breast cancer is always caused by a genetic abnormality (a “mistake” in the genetic material). The term “breast cancer” refers to a malignant tumor that has developed from cells in the breast. It is the second most common type of cancer after lung cancer and the fifth most common cause of cancer death. As breast cancer recurrence is high, good diagnosis is important.

The classification of breast cancer is based on a large number of parameters that characterize the tumor's appearance, age and menopause. This helps the physician for diagnosis of breast cancer more easily. This paper presents the implementation of supervised learning algorithms for Classification such as Multilayer Perceptron, OneR, Decision Tree induction and Support vector machine. The prediction accuracy of the classifiers was evaluated using 10-fold cross validation and the results were compared. Finally, it was found out that Support Vector Machines has better performance than the other algorithms.

Keywords: Machine Learning, Breast Cancer Classification, WEKA, Support Vector Machine

1. INTRODUCTION

Machine learning is a technique that can discover previously unknown regularities and trends from diverse datasets, in the hope that machines can help in the often tedious and error-prone process of acquiring knowledge from empirical data, and help people to explain and codify their knowledge. It encompasses a wide variety of techniques used for the discovery of rules, patterns and relationships in sets of data and produces a generalization of these relationships that can be used to interpret new unseen data. The output of a learning scheme is some form of structural description of a dataset, acquired from examples of given data. These descriptions encapsulate the knowledge learned by the system and can be represented in different ways. Today machine learning provides several indispensable tools for intelligent data analysis.

Presently, the digital revolution provided relatively inexpensive and available means to collect and store the data. Modern hospitals are well equipped with monitoring and other data collection devices, and data is gathered and shared in large information systems [1]. Machine learning technology is currently well suited for analyzing medical data, and in particular there is a lot of work done in medical diagnosis in small-specialized diagnostic problems. The derived classifier can then be used either to assist the physician when diagnosing new patients in order to improve the diagnostic speed, accuracy and/or reliability, or to train students or physicians non-specialists to diagnose patients in a special diagnostic problem.

2. MOTIVATION

The motivation behind the research reported in this paper is the results obtained from extensions of an ongoing research effort. The work reported here builds on the initial work by, first, using machine learning techniques to study and understand the accurate prediction of breast Cancer diseases and it helps physician to easily identify suggestive remedies based on the classification schemes or models.

2.1 RELATED WORK

Ryan Potter carried out the related work in preoperative patient classification. They have used weka for to classify the dataset [2]. So, far, a literature survey showed that there has been several studies on the survivability predict on problem using statistical approaches and artificial neural networks. However, we could only find a few studies related to medical diagnosis and survivability using data mining approaches like decision trees [3]. Here in this study, the latest tool Weka and svm light is used to classify the breast cancer dataset with 10-fold cross validation method. In biomedicine, researchers try to calculate various outcomes. The aim of the study is to apply and analyze different machine-learning techniques for classification of Breast Cancer.

3. BREAST CANCER CLASSIFICATION

Breast cancer happens when cells in the breast begin to grow out of control and can then invade nearby tissues or spread throughout the body [4]. Large collection of this out of control tissue called tumors. However, some tumors are not really cancer because they cannot spread or threaten someone’s life. These are called benign tumors. The tumors that can spread throughout the body or invade nearby tissues are considered cancer and are called malignant tumors. The average incidence rate varies from
22-28 per 100,000 women per year in urban settings to 6 per 100,000 women per year in rural areas.

Presently, 75,000 new cases occur in Indian women every year over the course of a lifetime, 1 in 22 women will be diagnosed with breast cancer. Early detection is your best protection. Close to 90% of breast cancer can be detected early, when they are most treatable. About 12 - 13% of women develop breast cancer in their lifetime. Experts estimate that about 178,480 women will be newly diagnosed with invasive breast cancer in the United States in 2007. Another 2,030 men will be diagnosed with breast cancer during the year. Although breast cancer in men is rare, the incidence has been increasing, and men are diagnosed at a later stage than women [5]. An estimated 40,460 women and 450 men will die from breast cancer in 2007. The earlier breast cancer is diagnosed, the earlier the opportunity for treatment.

According to the American Cancer Society, over 2 million women who have been treated for breast cancer are alive today. Age is a major identifiable risk factor. More than 80% of breast cancer cases occur in women over age 50 and especially in women over age 65. Breast Examination by a Health Professional. Early detection of breast cancer significantly reduces the risk of death. Women ages 20 - 49 should have a physical examination by a health professional every 1 - 2 years. Those over age 50 should be examined annually. A breast exam by a health professional can find 10 - 25% of breast cancers that are missed by mammograms [6]. Between 6 - 46% of the lumps detected by examination are malignant. (The yield is lowest in younger women and highest in older women.). Based on attributes of the Breast the Classification is done as recurrence events or non-recurrence events.

4. SUPERVISED LEARNING ALGORITHMS

The algorithms used to classify breast cancer data are SVM, MultilayerPerceptron, One R and Decision tree Induction.

4.1 Support Vector Machine

The machine is presented with a set of training examples, (xi, yi) where the xi is the real world data instances and the yi are the labels indicating which class the instance belongs to. For the two class pattern recognition problem, yi = +1 or yi = -1. A training example (xi, yi) is called positive if yi = +1 and negative otherwise. SVMs construct a hyperplane that separates two classes and tries to achieve maximum separation between the classes. Separating the classes with a large margin minimizes a bound on the expected generalization error.

The simplest model of SVM called Maximal Margin classifier, constructs a linear separator (an optimal hyperplane) given by w † x - y = 0 between two classes of examples. The free parameters are a vector of weights w which is orthogonal to the hyper plane and a threshold value. These parameters are obtained by solving the following optimization problem using Lagrangian duality.

\[
\text{Minimize } J(\alpha) = \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \text{ subject to } D_i(w^T x_i - y_i) \geq 1, i = 1, \ldots, l.
\]

where Dii corresponds to class labels +1 and -1. The instances with non null weights are called support vectors. In the presence of outliers and wrongly classified training examples it may be useful to allow some training errors in order to avoid over fitting. A vector of slack variables \(\xi\) that measure the amount of violation of the constraints is introduced and the optimization problem referred to as soft margin is given below. In this formulation the contribution to the objective function of margin maximization and training errors can be balanced through the use of regularization parameter C. The following decision rule is used to correctly predict the class of new instance with a minimum error.

\[
f(x) = \text{sgn}[w^T x - \gamma]
\]

The advantage of the dual formulation is that it permits an efficient learning of non–linear SVM separators, by introducing kernel functions. Technically, a kernel function calculates a dot product between two vectors that have been (non-linearly) mapped into a high dimensional feature space. Since there is no need to perform this mapping explicitly, the training is still feasible although the dimension of the real feature space can be very high or even infinite. The parameters are obtained by solving the following non linear SVM formulation (in Matrix form),

\[
\text{Minimize } L(\alpha) = \frac{1}{2} \|u\|^T Qu - c^T u
\]

\[
D^T u = 0 \quad 0 \leq u \leq Ce
\]
where \( u \) - the Lagrangian multipliers. In general the larger the margin the lower the generalization error of the classifier.

### 4.2 Multilayer Perceptron

Multilayer Perceptron (MLP) network is the most widely used neural network classifier. MLPs are universal approximators. MLPs are valuable tools in problems when one has little or no knowledge about the form of the relationship between input vectors and their corresponding outputs.

### 4.3 One R

The OneR algorithm creates one rule for each attribute in the training data, and then selects the rule with the smallest error rate as its ‘one rule’. To create a rule for an attribute, the most frequent class for each attribute value must be determined. The most frequent class is simply the class that appears most often for that attribute value. A rule is simply a set of attribute values bound to the \( r \) majority class. OneR selects the rule with the lowest error rate. In the event that two or more rules have the same error rate, the rule is chosen at random. This algorithm is chosen to be a base algorithm for comparing the strength of prediction with other algorithms, and due to its simplicity and single attribute requirement.

### 4.4 J48 Decision Tree Induction

J48 algorithm is an implementation of the C4.5 decision tree learner. This implementation produces decision tree models. The algorithm uses the greedy technique to induce decision trees for classification. A decision-tree model is built by analyzing training data and the model is used to classify unseen data. J48 generates decision trees, the nodes of which evaluate the existence or significance of individual features.

### 5. EXPERIMENTAL SETUP

The breast cancer data analysis and classification study was done using WEKA and Svm light tool. WEKA is a collection of machine learning algorithms for data mining tasks[8]. Svm light provides the extensive support for the whole process of experiment including preparing the input data, evaluating learning schemes statistically and visualizing the input data, and the result of learning.

The dataset is trained using SVM with linear, polynomial and RBF kernel and with different parameter settings for \( d \), gamma and \( C \) –regularization parameter[9]. The parameters \( d \) and gamma are associated with polynomial kernel and RBF kernel respectively.

The datasets are grouped into two broad classes to facilitate their use in experimentally determining the presence or absence of Breast Cancer [7]. The Breast Cancer dataset has 10 attributes, there are 286 instances, and as indicated above, 2 classes. The 10-fold cross validation was performed to test the performance of the three models. The prediction accuracy of the models was compared.

### 6. RESULTS AND IMPACTS

The results of the experiment are summarized in Table 1, and comparison of the accuracy (or number of correctly classified instances) and learning time (or time taken to build the model) on the dataset using Supervised Learning Algorithms are given below.

#### 6.1 Classification Using SVM

The performances of the three kinds of SVMs with linear, polynomial and RBF kernels were evaluated based on the two criteria, the prediction accuracy and time taken to build the model.

The Table 1 shows the results of the classification model based on SVM with linear kernel.

<table>
<thead>
<tr>
<th>Linear SVM</th>
<th>( C=0.1 )</th>
<th>( C=0.2 )</th>
<th>( C=0.4 )</th>
<th>( C=0.5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>80</td>
<td>86</td>
<td>90</td>
<td>83</td>
</tr>
<tr>
<td>Time</td>
<td>0.01</td>
<td>0.23</td>
<td>0.01</td>
<td>0.02</td>
</tr>
</tbody>
</table>

The Table 2 shows the results of the classification model based on SVM with polynomial kernel and with parameters \( d \) and \( C \).

<table>
<thead>
<tr>
<th>( d )</th>
<th>( C=0.1 )</th>
<th>( C=0.2 )</th>
<th>( C=0.4 )</th>
<th>( C=0.5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>83</td>
<td>84.2</td>
<td>82</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>91</td>
<td>87</td>
<td>85</td>
</tr>
<tr>
<td>Time</td>
<td>0.03</td>
<td>0.25</td>
<td>1.05</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.05</td>
<td>0.17</td>
<td>0.03</td>
</tr>
</tbody>
</table>

The Table 3 shows the results of the classification model based on SVM with RBF kernel and with parameters \( g \) and \( C \).
The Table 4 shows the average performance of the SVM based classification model in terms of predictive accuracy and training and shown in FIGURE.1a and FIGURE.1b.

### Table 4 Average Performance of Three Models

<table>
<thead>
<tr>
<th>Kernel Type</th>
<th>Accuracy</th>
<th>Time taken to build model (secs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>90</td>
<td>0.01</td>
</tr>
<tr>
<td>Polynomial</td>
<td>91</td>
<td>0.05</td>
</tr>
<tr>
<td>RBF</td>
<td>95</td>
<td>2.45</td>
</tr>
</tbody>
</table>

The predictive accuracy shown by SVM with RBF kernel with parameter C=0.4 and g=1 is higher than the linear and polynomial kernel.

### 6.2 Classification Using WEKA

The results of the three classifiers Multilayer Perceptron, OneR and Decision tree induction are shown in Table 5.

### Table 5 Predictive Performance

<table>
<thead>
<tr>
<th>Evaluation Criteria</th>
<th>Classifiers</th>
<th>MLP</th>
<th>OneR</th>
<th>J48</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time taken to build model(secs)</td>
<td></td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>Correctly classified instances</td>
<td></td>
<td>212</td>
<td>233</td>
<td>230</td>
</tr>
<tr>
<td>Incorrectly classified instances</td>
<td></td>
<td>74</td>
<td>53</td>
<td>56</td>
</tr>
<tr>
<td>Prediction accuracy</td>
<td></td>
<td>74.1%</td>
<td>83%</td>
<td>80%</td>
</tr>
</tbody>
</table>
The time taken to build the model and the prediction accuracy is high OneR when compared to other two algorithms in WEKA environment.

7. CONCLUSION

In this paper, supervised learning methods were applied on the task of classifying Breast cancer and the most accurate learning methods was evaluated. The key is to getting the best outcome is being able to analyze and compare the results of the different classification algorithms. The study shows that Support Vector Machine has high accuracy rate than other supervised learning algorithms.

REFERENCES

[1]."Machine Learning Approach for pre-operative anaesthetic risk Prediction” Karpagavalli S, Jamuna KS, and Vijaya MS.
[2]."Comparison of Classification Algorithms Applied to Breast Cancer Diagnosis and Prognosis”Ryan Potter.
[3]."Predicting Breast Cancer Survivability Using Data MiningTechniques” Abdelghani Bellaachia, Erhan Guven.
[7].Breast Cancer Dataset http://www.uci.edu/
[8].WEKA http://www.WEKA.com/

AUTHOR

P.Sivagami, Assistant Professor, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore. She received her M.Sc.,in the year 2009 and M.Phil.,in the year 2011 from Bharathiar University. She has an academic experience of 2 years. Her area of research is Data mining.