

A MACHINE LEARNING APPROACH TO PREDICT BREAST CANCER USING DECISION TREE BASED ALGORITHM

Manasa.K

Department of Computer Science and Engineering, CVR College of Engineering

ABSTRACT: Nowadays one of the major application areas of data mining is medical diagnosis of diseases and treatment. Data mining techniques are and also used to find correlations and associations between different diseases. Present days many people are dying because of sudden heart attack, brain stroke, liver failure, kidney failure, cancer and etc. Breast cancer is one of the leading causes of death in women. In this paper we focused on how data mining techniques can be used to predict breast cancer in advance such that patient is well treated. We used J48 machine learning algorithms supported by WEKA to predict Breast cancer in advance. Experimental analysis is done on breast cancer data set supported by WEKA. Data set contains 286 instances each instance having 10 attributes relevant to breast cancer including class label attribute. We used 65% data for training and 35% data for testing. The algorithm has shown 72% accuracy in prediction.

KEYWORDS: Classification, Breast cancer, machine learning, C4.5, J48 algorithm

1. INTRODUCTION

Healthcare industry today generates large amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices etc [1]. The large amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Breast cancer is the most common malignant tumor for women and it poses serious threat to the lives of people and it is the second leading cause of death in women today [2]. In the past twenty years, the incidence of breast cancer continues to rise. Then, the diagnosis and treatment of the breast cancer have become an extremely urgent work to do. Breast cancer begins with forming tumor in cells of breast forming clusters and spreads in the entire tissue [3]. In this article, we especially focus on ladies bosom growth and procedures for early forecast. Symptoms of Breast Cancer: A lump in a breast, A pain in the armpits or breast that does not seem to be related to the woman's menstrual period, Pitting or redness of the skin of the breast; like the skin of an orange, A rash around (or on) one of the nipples, A swelling (lump) in one of the armpits, An area of thickened tissue in a breast, One of the nipples has a discharge; sometimes it may contain blood, The nipple changes in appearance; it may become sunken or inverted, The size or the shape of the breast changes, The nipple-skin or breast-skin may have started to peel, scale or flake. Types of breast cancer classified basically malignant tumor and benign tumor [3]. The choice of the classification technique is very much important as the accuracy of the classification as malignant or benign varies from algorithm to algorithm. The paper focuses on use of different classification techniques to classify a particular cancer as malignant or benign.

2. MACHINE LEARNING ALGORITHMS

In this paper we our focus is how we can train the Machine to learn from the medical data so it can predict and treat the disease. Learning can be defined in general as a process of gaining knowledge through experience. We humans start the process of learning new things from the day we are born. This learning process continues throughout our life where we try to gather more knowledge from our surroundings and through our experience [4]. Machine Learning (ML) is a sub-field of AI whose concern is the development, understanding and evaluation of algorithms and techniques to allow a computer to learn [4]. ML intertwines with other disciplines such as statistics, human psychology and brain modeling. Human psychology and neural models obtained from brain modeling help in understanding the workings of the human brain, and especially its learning process, which can be used in the formulation of ML algorithms. Since many ML algorithms use analysis of data for building models, statistics plays a major role in this field [5]. ML algorithms need a dataset, which is collection of records are instances where each instance consist of attributes. The input attributes are the information given to the learning algorithm and the output attribute contains the feedback of the activity on that information. The value of the output attribute is assumed to depend on the values of the input attributes.

Machine learning algorithms are broadly classified as Supervised and unsupervised learning algorithms. In supervised learning instance and predefined classes are there. The model predicts the class membership of an instance. In unsupervised learning only instance are there based on the similarities between the instances, they are segmented as groups. In this paper we used supervised learning techniques to predict the class label of test instances. Supervised learning algorithms also called as Classification Algorithms [6]. Here instance can be treated as patient record. Experimental analysis is done using J48 algorithm [6], rest of the chapter explains how j48 classifies the test instances.

2.1 CLASSIFICATION ALGORITHMS

Algorithms that classify a given instance into a set of discrete categories are called classification algorithms [4]. These algorithms work on a training set to come up with a model or a set of rules that classify a given input into discrete output values. Most classification algorithms can take inputs in any form, discrete or continuous. Although some of the classification algorithms

require all of the inputs also to be discrete. The output is always in the form of a discrete value. Decision trees and Baye's nets are examples of classification algorithms

2.1.1 THE J48 DECISION TREE ALGORITHM

J48 is a decision tree learner based on C4.5. Where C4.5 is an update of the ID3 algorithm. A decision tree classifies a given instance by passing it through the tree starting at the top and moving down until a leaf node is reached [5]. The value at that leaf node gives the predicted output for the instance. At each node an attribute is tested and the branches from the node correspond to the values that attribute can take. When the instance reaches a node, the branch taken depends on the value it has for the attribute being tested at the node.

The ID3 algorithm builds a decision tree based on the set of training instances given to it. It takes a greedy top-down approach for the construction of the tree, starting with the creation of the root node. At each node the attribute that best classifies all the training instances that have reached that node is selected as the test attribute. At a node only those attributes are considered which were not used for classification at other nodes above it in the tree. To select the best attribute at a node, the information gain for each attribute is calculated and the attribute with the highest information gain is selected. Information gain for an attribute is defined as the reduction in entropy caused by splitting the instances based on values taken by the attribute.

2.1.2 PERFORMANACE MEASURES USED FOR CLASSIFIER EVALUATION

The classifier's evaluation is most often based on prediction accuracy (the percentage of correct prediction divided by the total number of predictions). If the error rate evaluation is unsatisfactory, we must return to a previous stage of the supervised Machine learning process. A variety of factors must be observed, perhaps relevant features for the problem are not being considered, may need a larger training set is required, the dimensionality of the problem is too high, the selected algorithm may not suitable or parameter tuning is needed [7].

TABLE 1. Measures and Formula

Classifier Accuracy	$\frac{TP + TN}{(P + N)}$
Classifier Error rate	$\frac{FP + FN}{(P + N)}$
Recall	$\frac{TP}{P}$
Precision	$\frac{TP}{(TP + FP)}$
F-Measure	$\frac{(2 \times \text{precision} \times \text{recall})}{(\text{precision} + \text{recall})}$

Where P is total number of positive records, N is total number of negative records, TP refers to the positive records which are correctly labeled by the classifier, TN is the negative records which are correctly labeled by the classifier, FP is the negative records which are improperly labeled as positive, and FN is the positive records which are incorrectly labeled as negative.

3. EXPERIMENTAL SETUP

In this paper we used a decision tree based J48 classification algorithm which is implemented in WEKA (Waikato Environment for Knowledge Analysis). It is a collection of various ML algorithms, implemented in Java, which can be used for data mining problems. Apart from applying ML algorithms on datasets and analyzing the results generated, WEKA also provides options for pre-processing and visualization of the dataset. It can be extended by the user to implement new algorithms. We have taken Breast cancer data set of 286 diagnostic records. Each record consists of attributes like age, menopause, tumor size etc. Data set is available on online at <https://archive.ics.uci.edu/ml/datasets/breast+cancer>. Here 65% records are used to train the model remaining 35% records are used to test the model. Table 3 gives the performance details of J48 algorithm used for classifying BP patient records.

Here Table 2 represents confusion matrix of J48 classifier. It represents out of 67 records whose class label is YES, 66 records are predicted correctly and 6 records predicted in correctly. Out of 33 records whose class label is NO, 22 records are predicted correctly and 11 records are predicted incorrectly. We considered record with class label YES as positive and record with class label NO as Negative.

TABLE 2. Confusion Matrix

Actual Class	Predicted Class	
	Positive	Negative
Positive	61	6
Negative	22	11

TABLE 3. Class wise Accuracy

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
YES	0.910	0.667	0.735	0.910	0.813
NO	0.333	0.090	0.647	0.333	0.440

4. CONCLUSION

In this paper we used J48 classifier to predict whether the patient is suffering with breast cancer or not. The algorithm has shown 72% accuracy in predicting the class label of unknown records. The evaluation criteria proved that, J48 decision trees are more effective and efficient classification techniques for the prediction of breast cancer risks among patients. The considerable point about the algorithm used is it out performs in accuracy while predicting the positive class labels. These machine learning algorithms can be used to predict many disease like heart attack, asthma, diabetes and high blood pressure etc.

REFERENCES

- [1] M. Ghousaini, O. Fletcher, K. Michailidou, C. Turnbull, M. K. Schmidt, E. Dicks, J. Dennis, Q. Wang, M. K. Humphreys, C. Luccarini et al., "Genome-wide association analysis identifies three new breast cancer susceptibility loci, " *Nature genetics*, vol. 44, no. 3, pp. 312-318, 2012.
- [2] A. Jemal, R. Siegel, E. Ward, T. Murray, J. Xu, C. Smigal, and M. J. Thun, "Cancer statistics, 2006, " *CA: a cancer journal for clinicians*, vol. 56, no. 2, pp. 106-130, 2006.
- [3] American Cancer Society. *Breast Cancer Facts & Figures 2005-2006*. Atlanta: American Cancer Society, Inc. (<http://www.cancer.org/>).
- [4] N. Satyanarayana, CH. Ramalingaswamy, and Y. Ramadevi, 2014. *Survey of Classification Techniques in Data Mining*.
- [5] C. Christin, H. C. Hoefsloot, A. K. Smilde, B. Hoekman, F. Suits, R. Bischoff, and P. Horvatovich, "A critical assessment of feature selection methods for biomarker discovery in clinical proteomics, " *Molecular & Cellular Proteomics*, vol. 12, no. 1, pp. 263-276, 2013.
- [6] High blood pressure prediction based on AAA++ using machine-learning algorithms, Satyanarayana Nimmala, Y. Ramadevi, R. Sahith & Ramalingaswamy Cheruku, *Cogent Engineering* (2018), 5: 1497114.
- [7] Bellaachia Abdelghani and Erhan Guven, "Predicting Breast Cancer Survivability using Data Mining Techniques," *Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining*, 2006.