PREDICTION OF BREAST CANCER USING MACHINE LEARNING

ABSTRACT

Women are seriously threatened by breast cancer with high morbidity and mortality. The lack of robust prognosis models results in difficulty for doctors to prepare a treatment plan that may prolong patient survival time. Hence, the requirement of time is to develop the technique which gives minimum error to increase accuracy. Four algorithm SVM, Logistic Regression, Random Forest and KNN which predict the breast cancer outcome have been compared in the paper using different datasets. All experiments are executed within a simulation environment and conducted in JUPYTER platform. Aim of research categorises in three domains. First domain is prediction of cancer before diagnosis, second domain is prediction of diagnosis and treatment and third domain focuses on outcome during treatment. The proposed work can be used to predict the outcome of different technique and suitable technique can be used depending upon requirement. This research is carried out to predict the accuracy. The future research can be carried out to predict the other different parameters and breast cancer research can be categorises on basis of other parameters.

INTRODUCTION

GENERAL DESCRIPTION:

The second major cause of women's death is breast cancer (after lung cancer). 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the US during 2016 and 40,450 of women's death is estimated. Breast cancer is a type of cancer that starts in the breast. Cancer starts when cells begin to grow out of control. Breast cancer cells usually form a tumour that can often be seen on an x-ray or felt as a lump. Breast cancer can spread when the cancer cells get into the blood or lymph system and are carried to other parts of the body. The cause of Breast Cancer includes changes and mutations in DNA. There are many different types of breast cancer and common ones include ductal carcinoma in situ (DCIS) and invasive carcinoma. Others, like phyllodes tumours and angiosarcoma are less common.

SYMPTOMS:

New lump in the breast or underarm (armpit), Thickening or swelling of part of the breast, Irritation or dimpling of breast skin, Redness or flaky skin in the nipple area or the breast, Pulling in of the nipple or pain in the nipple area, Nipple discharge other than breast milk, including blood, Any change in the size or the shape of the breast, Pain in any area of the breast.

PROBLEM IDENTIFICATION:

We can notice that SVM takes about 0.07 s to build its model unlike k-NN that takes just 0.01 s. It can be explained by the fact that k-NN is a lazy learner and does not do much during training process unlike others classifiers that build the models. In other hand, the accuracy obtained by SVM (97.13%) is better than the accuracy obtained by C4.5, Naïve Bayes and k-NN that have an accuracy that varies between 95.12 % and 95.28 %. It can also be easily seen that SVM has the

highest value of correctly classified instances and the lower value of incorrectly classified instances than the other classifiers.

PROPOSED METHODOLOGY:

Phase 1 - Pre-Processing Data The first phase we do is to collect the data that we are interested in collecting for pre-processing and to apply classification and Regression methods. Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real world data is often incomplete, inconsistent, and lacking certain to contain many errors. Data pre-processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing. For pre-processing we have used standardization method to pre-process the UCI dataset. This step is very important because the quality and quantity of data that you gather will directly determine how good your predictive model can be. In this case we collect the Breast Cancer samples which are Benign and Malignant. This will be our training data. Phase 2 - DATA PREPRATION Data Preparation, where we load our data into a suitable place and prepare it for use in our machine learning training. We'll first put all our data together, and then randomize the ordering. Phase 3 - FEATURES SELECTION In machine learning and statistics, feature selection, also known as variable selection, attribute selection, is the process of selection a subset of relevant features for use in model construction. Data File and Feature Selection Breast Cancer Wisconsin (Diagnostic):- Data Set from Kaggle repository and out of 31 parameters we have selected about 8-9 parameters. Our target parameter is breast cancer diagnosis – malignant or benign. We have used Wrapper Method for Feature Selection. The important features found by the study are: Concave points worst, Area worst, Area se, Texture worst, Texture mean, Smoothness worst, Smoothness mean, Radius mean, Symmetry mean. FEATURE SELECTION FEATURE PROJECTION DATA PREPROCESSIN G MODEL SELECTION DATA PREPARATIO N FEATURE SCALING

PREDICTION © 2020 JETIR May 2020, Volume 7, Issue 5 www.jetir.org (ISSN-2349-5162) JETIR2005145 Journal of Emerging Technologies and Innovative Research (JETIR) www.jetir.org 19 We have used Wrapper Method for Feature Selection. The important features found by the study are: 1. Concave points worst 2. Area worst 3. Area se 4. Texture worst 5. Texture mean 6. Smoothness worst 7. Smoothness mean 8. Radius mean 9. Symmetry means. Attribute Information: ID number 2) Diagnosis (M = malignant, B = benign) 3–32) Phase 4 - Feature Projection Feature projection is transformation of high-dimensional space data to a lower dimensional space (with few attributes). Both linear and nonlinear reduction techniques can be used in accordance with the type of relationships among the features in the dataset. Phase 5 - Feature Scaling Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidian distance between two data points in their computations. We need to bring all features to the same level of magnitudes. This can be achieved by scaling. Phase 6 - Model Selection Supervised learning is the method in which the machine is trained on the data which the input and output are well labelled. The model can learn on the training data and can process the future data to predict outcome. They are grouped to Regression and Classification techniques. A regression problem is when the result is a real or continuous value, such as “salary” or “weight”. A classification problem is when the result is a category like filtering emails spam” or “not spam”. Unsupervised Learning: Unsupervised learning is giving away information to the machine that is neither classified nor labelled and allowing the algorithm to analyse the given information without providing any directions. In unsupervised learning algorithm the machine is trained from the data which is not labelled or classified making the algorithm to work without proper instructions. In our dataset we have the outcome variable or Dependent variable i.e. Y having only two set of values, either M (Malign) or B (Benign). So Classification algorithm of supervised learning is applied on it. We have chosen three different

types of classification algorithms in Machine Learning. We can use a small linear model, which is a simple. Phase 7 - PREDICTION Machine learning is using data to answer questions. So Prediction, or inference, is the step where we get to answer some questions. This is the point of all this work, where the value of machine learning is real

METHODS USED:

(1) Logistic Regression Logistic regression was introduced by statistician DR Cox in 1958 and so predates the field of machine learning. It is a supervised machine learning technique, employed in classification jobs (for predictions based on training data). Logistic Regression uses an equation like Linear Regression, but the outcome of logistic regression is a categorical variable whereas it is a value for other regression models. Binary outcomes can be predicted from the independent variables. The general workflow is: (1) get a dataset © 2020 JETIR May 2020, Volume 7, Issue 5 www.jetir.org (ISSN-2349-5162) JETIR2005145 Journal of Emerging Technologies and Innovative Research (JETIR) www.jetir.org 20 (2) train a classifier (3) make a prediction using such classifier (2) k-Nearest Neighbour (k-NN) K-Nearest Neighbour is a supervised machine learning algorithm as the data given to it is labelled. It is a nonparametric method as the classification of test data point relies upon the nearest training data points rather than considering the dimensions (parameters) of the dataset. ALGORITHM (1) Input the dataset and split it into a training and testing set. (2) Pick an instance from the testing sets and calculate its distance with the training set. (3) List distances in ascending order. (4) The class of the instance is the most common class of the 3 first trainings instances (k=3). (3) Support Vector machine Support Vector Machine is a supervised machine learning algorithm which is doing well in pattern recognition problems and it is used as a training algorithm for studying classification and regression rules from data. SVM is most precisely used when the number of features and number of

instances are high. A binary classifier is built by the SVM algorithm. In an SVM model, each data item is represented as points in an n-dimensional space where n is the number of features where each feature is represented as the value of a coordinate in the n-dimensional space. Here's how a support vector machine algorithm model works: (1) First, it finds lines or boundaries that correctly classify the training dataset. (2) Then, from those lines or boundaries, it picks the one that has the maximum distance from the closest data points. (3) Random Forest Random forest, like its name implies, consists of many individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction.