8-10-2020

Big Data Analytics using Small Datasets: Machine Learning for Early Breast Cancer Detection

Dulani Jayasuriya
University of Auckland, d.jayasuriya@auckland.ac.nz

Johnny Chan
The University of Auckland, jh.chan@auckland.ac.nz

David Sundaram
University of Auckland, d.sundaram@auckland.ac.nz

Big Data Analytics using Small Datasets: Machine Learning for Early Breast Cancer Detection

TREO Talk Paper

Dulani Jayasuriya Daluwathumullagamage
The University of Auckland
d.jayasuriya@auckland.ac.nz

Johnny Chan
The University of Auckland
jh.chan@auckland.ac.nz

David Sundaram
The University of Auckland
d.sundaram@auckland.ac.nz

Abstract

In US breast cancer happens to possess the highest death rate apart from lung cancer. As of 2019, on average, 1 in 8 US women (approx. 12%) would develop invasive breast cancer at some point during her life. These statistics highlight the importance of early detection for increasing the mortality of patients. In recent years, machine learning (ML) techniques begin to play a key role in healthcare, especially as a diagnostic aid. In the case of breast cancer, ML techniques can be used to distinguish between malignant and benign tumours for enabling early detection. Moreover, accurate classification can assist physicians to guide patients and prescribe relevant treatment. Given this background, the objective of this paper is to apply ML algorithms to classify breast cancer outcomes. In this study, we build a platform using Ridge, AdaBoost, Gradient Boost, Random Forest, Principle Component Analysis (PCA) plus Ridge, and Neural Network ML algorithms for early breast cancer outcome detection. As a traditional benchmark technique, we use logistic regression model to compare against our chosen ML algorithms. We utilise the Wisconsin Breast Cancer Database (WBCD) dataset (Dua and Graff 2019). Although ML is generally deployed with large datasets, we highlight their usefulness and feasibility for small datasets in this study of only 30 features. We contribute to literature by providing a platform that will enable (a) big data analytics using small datasets and (b) high accuracy breast cancer outcome classifications. Specifically, we identify most important features in breast cancer outcome classification from a wide range of ML algorithms with a small dataset. This would enable health practitioners and patients to focus on these key features in their decision making for future breast cancer tests and subsequent early detection thus reducing analysis and decision latencies.

In our ML based breast cancer classification platform, the user is required to make three function calls: data pre-processor, model generator and a single test. The pre-processor cleans the raw dataset from the user by removing 'NaN' and empty values, and it follows further instructions from a configuration file. After the pre-processing, the platform can train ML models from model generator based on two inputs, a cleaned dataset and a configuration file. Model generator creates different models from different ML algorithms specified in the study and generates corresponding evaluations. As such, the user can call single test to use the generated models in making predictions.

In Table 1, we observe that the Random Forest model and the Neural Network model (also supported by the receiver operating characteristic curves and confusion matrix results) have reached a 98.2% accuracy rate, and the least number of fault predictions in classifying breast cancer. Feature importance results show concave_points_worst and perimeter_worst to be the most important features for classifying breast cancer outcomes.

Model Type	***Accuracy (%)***
Ridge	94.7
AdaBoost	93.0
Gradient Boost	96.5
Random Forest	**98.2**
PCA+Ridge	94.7
Neural Network	**98.2**

Table 1. Model Accuracy

Reference

Dua D, Graff C (2019) UCI machine learning repository. http://archive.ics.uci.edu/ml