CALIFORNIA SCIENCE & ENGINEERING FAIR
2018 PROJECT SUMMARY

Name(s)

Dhanvee Ivaturi; Philip Kabranov

Project Number

38526

Project Title

Improving Breast Cancer Detection in Fine-Needle Aspirate Biopsies through Machine Learning

Abstract

Objectives/Goals

The intent of this project is to compare the accuracy multiple machine learning (ML) algorithms to improve and/or automate breast cancer detection. This research takes into account various ML algorithms (logistic regression, support vector machines, and neural networks) and PCA dimensionality reduction.

Methods/Materials

Python, along with the Jupyter Notebook development environment are used in this software. The SciKit Learn and Tensorflow libraries are imported to implement the machine learning algorithms. The pandas and NumPy libraries are used for data visualization. The dataset in this project is the Wisconsin Breast Cancer dataset, which contains the records of 569 patients as unique 30-dimensional feature vectors. These features are extracted from measurable visual attributes of tissue samples. The 569 entries are randomly split into test and training subsets, and then used to train and test a machine learning. The algorithms are also trained separately with a dataset processed through PCA in order to eliminate features with low impact on the outcome.

Results

The dataset was found to be linearly separable, meaning that a logistic regressor and support vector machine could be applied to this dataset. The various ML algorithms were all trained using the same data source. For both data processed by PCA (reduced to 5 features) and non-reduced data (30 features), the logistic regression algorithm produced optimal mean accuracy over 20 training cycles: 97.82% for the non-reduced 30-dimensional feature set, and 97.24% for the feature set reduced to 5 dimensions, while the SVM produced a lowest mean accuracy of 96.69% and 96.50%, respectively. These accuracies are comparable to or exceed those of medical professionals.

Conclusions/Discussion

The results demonstrate that the best algorithm for this dataset is the logistic regressor, being the fastest to train and the most accurate. Neural network accuracies are in between those of the logistic regressor and SVM but takes 100 times longer to train. This is likely due to the computational complexity of the backpropagation training algorithm for neural networks. Also, the relatively low performance of the SVM is likely due to the fact that the data is not perfectly separated into two non-overlapping clusters, giving way to classification errors.

Summary Statement

Applied and analyzed the accuracy of multiple machine learning algorithms and a neural network on a dataset consisting of numerical values that relate to image attributes from fine-needle aspirate biopsy samples.

Help Received

The implementation of the machine learning algorithms and results were reviewed by a machine learning professional.