Title : Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer

Abstract:

Breast cancer disease is the most common cancer in US women and the second cause of cancer death among women. The prediction of cancer category during its early stage has become an essential area in cancer research, as it can simplify the subsequent clinical requirements of patients and determines the effective treatments. A benign tumor is a non-invasive type of tumor and it rarely causes life-threatening issues. On the contrast, a malignant tumors is an invasive kind that can affect the surrounding tissues and metastasize to distant tissues in the body

Objectives: To compare and evaluate the performance and accuracy of the key supervised and semi-supervised machine learning algorithms for breast cancer prediction.

Methods Used:

We have used nine machine learning classification algorithms for supervised (SL) and semi-supervised learning (SSL),

1. Logistic regression
2. Gaussian Naive Bayes
3. Linear Support vector machine
4. RBF Support vector machine
5. Decision Tree
6. Random Forest
7. Xgboost
8. Gradient Boosting
9. KNN

To ensure the robustness of the model, we have applied K-fold cross-validation and optimized hyperparameters. We have evaluated and compared the models using accuracy, precision, recall, F1-score, and ROC curves.

Supervised learning:

SL is the most widely used machine learning technique. Machine learning requires learning of a function that fits the input pairs of values to output. The function extracts knowledge from labeled training data and each input pair corresponds to a labeled value. SL algorithms detect the pattern in the training data and produce a function that can predict new input pairs or never seen observations.

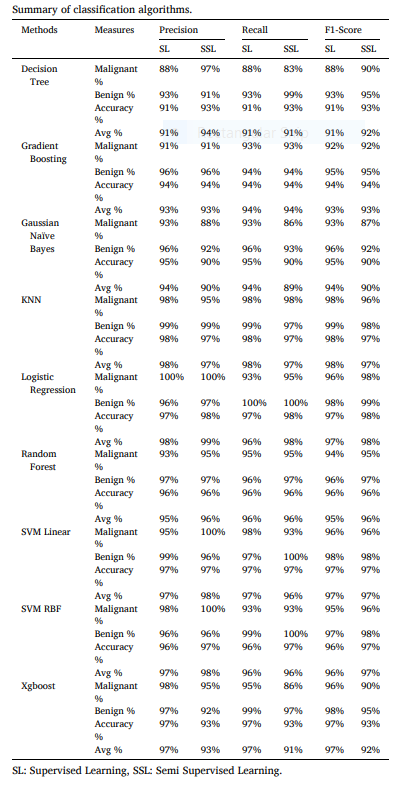
Unsupervised learning:

SSL is considered as a hybrid approach of SL and unsupervised learning. The algorithm is provided with unlabeled data along with the supervision information in a small quantity. The output of SSL contains target variables that are used to train and predict the targets for the unlabeled data.

t-SNE (t-distributed stochastic neighbor embedding):

t-SNE is an unsupervised machine learning algorithm that finds the pattern in the data, and a non-linear dimensionality reduction technique unlike PCA for reducing and visualizing high dimensional space into two or three dimensions. t-SNE selects two similarity measures between pairs of points - one measure for the high dimensional data and another for the two-dimensional embedding.

Dataset: The Wisconsin Diagnosis Cancer dataset was used to apply the machine learning algorithms. The dataset consists of patient ID, cell nuclei features, and diagnosis.



Results:

The results of all models are inspiring using both SL and SSL. The SSL has high accuracy (90%–98%) with just half of the training data. The KNN model for the SL and logistic regression for the SSL achieved the highest accuracy of 98%

Conclusion:

The accuracies of SSL algorithms are very close to the SL algorithms. The accuracies of all models are in the range of 91–98%. SSL is a promising and competitive approach to solve the problem. Using a small sample of labeled and low computational power, the SSL is fully capable of replacing SL algorithms in diagnosing tumor type.