Optimizing Breast Cancer Classification: A Comparative Analysis of Supervised and Unsupervised Machine Learning Techniques

Prithwish Ghosh, Debolina Banerjee

December 2023

Abstract

This study focuses on the comprehensive analysis of machine learning algorithms for the classification of breast cancer into benign and malignant categories using the Wisconsin breast cancer dataset. Two distinct approaches, supervised and unsupervised, were employed to evaluate the effectiveness of various algorithms in discerning the nature of cancerous growth based on diverse physical properties.

In the supervised learning realm, the study employed three powerful algorithms: Support Vector Machine (SVM), Random Forest, and XGBoost. Additionally, unsupervised learning techniques, specifically Linear Discriminant Analysis and Gaussian Finite Mixture Model for classification, were investigated. Notably, the XGBoost algorithm emerged as the most promising candidate in the supervised category, exhibiting superior classification performance when applied to the testing dataset (20% of the total data). Results indicated that XGBoost achieved an impressive precision of 98.23%, outperforming both Gaussian Finite Mixture Model for classification (97.5%) and Linear Discriminant Analysis (96.48%). The XGBoost algorithm, implemented on a subset of the data, demonstrated its efficacy in accurately identifying the nature of breast cancer, highlighting its potential as a robust tool for predicting malignancy based on distinct physical properties.

This study underscores the significance of supervised and unsupervised learning, particularly the XGBoost algorithm and also Gaussian Finite Mixture Model for classification, in optimizing breast cancer classification. The findings contribute valuable insights into the selection of appropriate machine learning techniques for the accurate and efficient identification of benign and malignant breast cancer, thereby facilitating improved diagnostic practices.