Technical report

Visualization and Exploration of Breast Cancer Data Using Tree Decision, Random Forest, and Self-Training

Aulia Samudra Jimananda - 1103201264

Breast cancer is one of the most common cancers in women worldwide. Early detection of breast cancer is essential for successful treatment and survival. This report describes the use of tree determination, random forests, and self-training algorithms for visualization and exploration of breast cancer data. The aim is to identify important features and patterns in the data that will aid in the early detection of breast cancer.

1. Data

The data used in this report is the Breast Cancer Wisconsin (Diagnostic) Dataset, which contains information about breast cancer diagnosis. The dataset consists of 569 instances and 30 features, including patient ID, diagnosis, and 28 different measurements related to the breast tissue. The diagnosis feature, which is the target variable, indicates whether the diagnosis is malignant or benign.

2. Methods

Three algorithms used in this report are Tree decision, random forest, self-training. A tree decision algorithm was used to build a decision tree based on the properties of the data. Decision trees were visualized using a graphics library to aid in interpreting the tree structure. We used the random forest algorithm to build an ensemble of decision trees to improve classification accuracy. Finally, we improved the performance of our random forest by adding unlabeled data to the training set using a self-learning algorithm.

3. Visualization

To visualize the data, seaborn's countplot was used to differentiate between the different sums of patients' target variable. The pairplot was used to compare different symptoms to understand the correlation between them. But before visualizing the breast cancer dataset, it was first converted to a Pandas DataFrame for ease of use with the visualization libraries.

4. Results

The decision tree generated by the tree decision algorithm showed that the most important feature for breast cancer diagnosis was mean concavity, followed by worst perimeter, mean radius, and worst concavity. After using the random forest and the self-training algorithm on the breast cancer dataset, the obtained accuracy is 97.08%, which is the same for both algorithms. On the other hand, the decision tree algorithm gave an accuracy of 94.15%. Count plots showed that there were more patients diagnosed as benign than malignant. Pair plots showed strong correlations between various symptoms such as mean radius and mean girth.

5. Conclusion

The use of tree decisions, random forests, and self-training algorithms are useful for visualizing and studying breast cancer data. Decision trees help identify features that are most important for diagnosis, and random forests can improve classification accuracy. Self-training algorithms can further improve random forest accuracy by adding unlabeled data to the training set. The visualization techniques used in this report are: Logging techniques such as Seaborn's countplot and pairplot can help you interpret the data and identify important patterns and correlations.