Comparative Analysis of Support Vector Machines and Decision Tree-Based Classifiers using the Breast Cancer Wisconsin Dataset

Author: HARISH GADDAM (Student ID: 24023883) | GitHub

1. Introduction

In the field of data mining, classification algorithms are pivotal for predictive analytics. This report presents a comparative study of two prominent classification techniques: Support Vector Machines (SVM) and Decision Tree-based classifiers, including Random Forests. The Breast Cancer Wisconsin dataset serves as the foundation for evaluating the performance, interpretability, and applicability of these algorithms in a medical diagnosis context.

2. Dataset Description

The Breast Cancer Wisconsin (Original) dataset, sourced from the UCI Machine Learning Repository, comprises 699 instances, each with 10 attributes, such as clump thickness and mitosis. The target variable classifies tumors as benign or malignant. The dataset includes some missing values, which were addressed through appropriate preprocessing techniques.

3. Data Preprocessing

Effective data preprocessing is crucial for model performance. The following steps were undertaken:

- **Handling Missing Values**: Instances with missing values were removed to maintain data integrity.
- **Encoding Categorical Variables**: The target variable was encoded into a binary format assigning 'benign' to 0, 'malignant' to 1. This encoding ensures compatibility with machine learning algorithms, which require numerical input for classification tasks.
- **Feature Scaling**: Features were normalised to ensure uniformity across attributes. Normalisation eliminates scale-related biases, particularly for algorithms like SVM that are sensitive to the magnitude of input values.
- **Train-Test Split**: The dataset was divided into training (80%) and testing (20%) subsets to evaluate model performance. This division allows for robust evaluation of model performance on unseen data, thereby reducing the risk of overfitting and ensuring generalisability.

4. Methodology

In this study, three classification algorithms-Support Vector Machine (SVM), Decision Tree Classifier, and Random Forest Classifier - were employed to evaluate their efficacy in diagnosing breast cancer using the Wisconsin dataset. The selection of these algorithms is substantiated by their documented performance in medical diagnostics:

4.1 Support Vector Machine (SVM)

SVM is a supervised learning model that identifies the optimal hyperplane separating different classes in the feature space. It is particularly effective in high-dimensional spaces and is robust against overfitting, especially in cases where the number of dimensions exceeds the number of samples. Studies have demonstrated that SVM achieves high accuracy rates in breast cancer classification tasks.

4.2 Decision Tree Classifier

Decision Tress classifies instances by sorting them down the tree from the root to some leaf node, which provides the classification. They are intuitive and easy to interpret but can be prone to overfitting. However, when appropriately tuned, Decision Trees can offer valuable insights into feature importance and decision-making processes.

4.3 Random Forest Classifier

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification tasks. It mitigates the overfitting problem associated with individual decision trees and improves predictive accuracy. Research has shown that Random Forest Classifiers can achieve accuracies exceeding 95% in breast cancer diagnosis.

5. Results and Discussion

The comparative analysis of Support Vector Machines (SVM), Decision Tree Classifier, and Random Forest Classifier on the Breast Cancer Wisconsin dataset yielded insightful findings:

- **Support Vector Machines (SVM)**: The SVM demonstrated robust performance, achieving an accuracy of 96.58%. This high accuracy is attributed to SVM's capability to find optimal hyperplane that maximizes the margin between classes, effectively handling the dataset's high-dimensional feature space.
- **Decision Tree Classifier**: The Decision Tree Classifier achieved an accuracy of 95%. While offering interpretability through its decision rules, its performance was slightly lower than that of SVM. This reduction in accuracy may be due to the model's susceptibility to overfitting, especially without proper pruning or tuning.
- **Random Forest Classifier**: As an ensemble method, the Random Forest Classifier constructed multiple decision trees, resulting in enhanced accuracy and robustness. It achieved an accuracy of 96%, demonstrating its effectiveness in reducing overfitting and improving generalisation compared to individual decision tress.

These results align with existing literature, which indicates that ensemble methods like Random Forest often outperform single classifiers by mitigating overfitting and enhancing predictive performance. However, while Random Forests provides improved accuracy, they may sacrifice the interpretability offered by single decision trees. Therefore, the choice of classifier should balance the need for predictive accuracy and model transparency, depending on the specific requirements of the application.

5. Conclusion

The comparative analysis revealed that SVM achieved the highest classification performance on the Breast Cancer Wisconsin dataset, making it a suitable choice for medical diagnosis tasks where accuracy is paramount. However, Random Forest also provided competitive results with the added benefit of reduced overfitting compared to a single decision tree. Decision Trees, while less accurate, offer simplicity and interpretability, which can be valuable in scenarios where transparency is essential.

6. References

1. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297. https://doi.org/10.1007/BF00994018

2. Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106. https://doi.org/10.1007/BF00116251

3. Breiman, L., (2001). Random Forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

4. Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast Cancer Diagnosis and Prognosis via Linear Programming. Operations Research, 43(4), 570-577. https://doi.org/10.1287/opre.43.4.570