Improving Feature Selection in the Data Science Pipeline: A Hybrid Approach Combining Tree-Based Importance and Feature Clustering

Itay derazon Or Hazav

February 17, 2025

Abstract

This paper presents a novel approach to automated feature selection that combines tree-based importance measures with correlation-based feature clustering. Our method addresses the challenge of selecting optimal features for machine learning models while maintaining interpretability and reducing redundancy. The approach integrates Random Forest importance scores with feature clustering to identify groups of related features and select the most representative feature from each group. We evaluate our method on four diverse datasets and demonstrate comparable or superior performance to traditional feature selection approaches while providing better interpretability. Experimental results show our method successfully reduces feature dimensionality while maintaining model performance and providing clear justification for feature selection decisions.

1 Problem Description

Feature selection is a critical step in the data science pipeline that significantly impacts model performance, computational efficiency, and interpretability. The current challenges in feature selection include:

- Manual feature selection is time-consuming and requires domain expertise
- Automated methods often select redundant features
- Many approaches lack interpretability

- There's often a trade-off between feature reduction and model performance

These challenges can lead to suboptimal model performance, increased computational costs, and difficulty in explaining model decisions to stakeholders. Our work aims to address these issues by developing an automated feature selection method that balances performance with interpretability.

2 Solution Overview

Our solution combines tree-based feature importance with clustering to identify and group related features while selecting the most representative feature from each group. The method consists of three main components:

2.1 Tree-based Feature Importance

We utilize Random Forest models to calculate initial feature importance scores. Random Forests are chosen for their ability to:

- Capture both linear and non-linear relationships with the target variable
- Handle different types of features (numerical and categorical)
- Provide robust importance scores based on multiple decision trees

2.2 Feature Clustering

Features are grouped based on their correlations using K-means clustering. This step:

- Identifies groups of related features
- Reduces redundancy in the selected feature set
- Provides interpretable feature groups for better understanding

2.3 Representative Feature Selection

From each cluster, we select the feature with the highest importance score as the representative feature. This ensures:

- Maintained information value through highest importance features

- Reduced redundancy by selecting one feature per cluster
- Clear justification for selection decisions

3 Experimental Evaluation

We evaluated our method on four diverse datasets:

- Wine Quality Dataset (Classification)
- California Housing Dataset (Regression)
- Breast Cancer Dataset (Classification)
- Diabetes Dataset (Regression)

3.1 Evaluation Metrics

We used the following metrics to evaluate our method:

- Model Performance (Accuracy/R^2 Score)
- Feature Reduction Ratio
- Feature Group Coherence
- Statistical Significance of Improvements

[Include your comparison tables and graphs here]

3.2 Results Analysis

Our experimental results show:

- Comparable or better performance than baseline methods
- Significant reduction in feature dimensionality
- Clear interpretability through feature grouping
- Consistent performance across different types of datasets

4 Related Work

Several approaches to feature selection have been proposed in the literature:

- **Filter Methods**: Such as correlation-based feature selection (CFS) and information gain [1]
- **Wrapper Methods**: Including recursive feature elimination (RFE) [2]
- **Embedded Methods**: Like Lasso and Ridge regression [3]

Our approach differs by combining the strengths of multiple methods while addressing their individual limitations. We build upon the work of [4] for tree-based importance and extend it with clustering-based feature grouping.

5 Conclusion

This project demonstrates that combining tree-based importance measures with feature clustering can effectively improve the feature selection process. Key findings include:

- Successful automation of feature selection while maintaining interpretability
- Effective reduction of feature redundancy
- Consistent performance across different types of datasets
- Clear explanation of feature selection decisions

Future work could explore:

- Dynamic cluster size determination
- Integration with other importance measures
- Application to time-series data

References

[1] Hall, M.A. (1999). Correlation-based Feature Selection for Machine Learning.

[2] Guyon, I., et al. (2002). Gene Selection for Cancer Classification using Support Vector Machines.

[3] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso.

[4] Breiman, L. (2001). Random Forests.