JANOLI International Journal of
Data Science
(A Peer Review Journal)

journal homepage: www.janolijournals.org/jijds/

Volume 1, Issue 2, 2025 ISSN: 3107-457X

Adaptive Ensemble Learning with Dynamic Feature Selection for Enhanced Predictive Accuracy in High-Dimensional Biological Datasets

Pankaj Pachari
Rajasthan University, Jaipur, India

ARTICLE INFO

Keywords:
Ensemble Learning, Feature Selection, High-Dimensional Data, Biological Data, Adaptive Algorithms, Machine Learning, Predictive Modeling, Dynamic Selection, Weighted Averaging

Correspondence:
E-mail: pankaj.pachauri40@gmail.com

ABSTRACT

High-dimensional biological datasets present significant challenges for accurate predictive modeling due to the curse of dimensionality and the presence of irrelevant or redundant features. This paper introduces a novel adaptive ensemble learning framework that incorporates dynamic feature selection to enhance predictive accuracy in such datasets. The proposed method combines multiple base learners with a dynamically adjusted weighting scheme, informed by the performance of each learner on subsets of features selected using a novel hybrid feature selection strategy. This strategy integrates filter, wrapper, and embedded methods to identify the most relevant feature subsets for each base learner. The adaptive weighting mechanism dynamically adjusts the contribution of each base learner based on its performance on a validation set. We evaluate the performance of the proposed method on several benchmark biological datasets, demonstrating its superiority over existing ensemble learning and feature selection techniques. Results show a significant improvement in predictive accuracy, robustness, and interpretability, making it a promising tool for analyzing complex biological data.

1. Introduction:

The advent of high-throughput technologies in biology, such as genomics, proteomics, and transcriptomics, has led to an explosion of high-dimensional datasets. These datasets, characterized by a large number of features (e.g., gene expression levels, protein abundances) and a relatively small number of samples, pose significant challenges for traditional machine learning algorithms. The "curse of dimensionality" leads to overfitting, reduced model generalization, and increased computational complexity. Furthermore, many features in biological datasets are often irrelevant or redundant, further complicating the task of building accurate and interpretable predictive models.

Ensemble learning, which combines multiple base learners to make predictions, has emerged as a powerful approach to address these challenges. By aggregating the diverse perspectives of multiple models, ensemble methods can often achieve higher accuracy and robustness than individual learners. However, the effectiveness of ensemble learning depends critically on the diversity and accuracy of the base learners.

A crucial aspect of building effective predictive models from high-dimensional biological data is feature selection. Feature selection aims to identify a subset of relevant features that can improve model performance, reduce computational cost, and enhance interpretability. Numerous feature selection methods have been developed, including filter methods (e.g., variance thresholding, correlation analysis), wrapper methods (e.g., recursive feature elimination, genetic algorithms), and embedded methods (e.g., LASSO, decision tree-based feature importance). Each approach has its strengths and weaknesses, and the optimal choice depends on the specific dataset and learning task.

This paper addresses the limitations of existing ensemble learning and feature selection techniques by proposing a novel adaptive ensemble learning framework that incorporates dynamic feature selection. Our method combines multiple base learners with a dynamically adjusted weighting scheme, informed by the performance of each learner on subsets of features selected using a novel hybrid feature selection strategy. This strategy integrates filter, wrapper, and embedded methods to identify the most relevant feature subsets for each base learner. The adaptive weighting mechanism dynamically adjusts the contribution of each base learner based on its performance on a validation set.

The primary objectives of this research are:

1. Develop a novel hybrid feature selection strategy that effectively identifies relevant feature subsets for different base learners in an ensemble.

2. Design an adaptive ensemble learning framework that dynamically adjusts the weights of base learners based on their performance on a validation set.

3. Evaluate the performance of the proposed method on several benchmark biological datasets and compare it to existing ensemble learning and feature selection techniques.

4. Demonstrate the improved predictive accuracy, robustness, and interpretability of the proposed method.

2. Literature Review:

Ensemble learning and feature selection have been extensively studied in the context of high-dimensional data analysis. Several approaches have been proposed to address the challenges associated with building accurate and interpretable predictive models from such data.

Breiman (2001) introduced Random Forests, a widely used ensemble learning method that combines multiple decision trees trained on random subsets of the data and features. Random Forests have been shown to be effective in handling high-dimensional data and providing feature importance estimates. However, Random Forests can be sensitive to the choice of hyperparameters and may not always capture complex relationships between features.

Boosting algorithms, such as AdaBoost (Freund & Schapire, 1997) and Gradient Boosting (Friedman, 2001), iteratively combine weak learners to create a strong ensemble. Boosting algorithms typically assign weights to samples based on their misclassification rates, focusing on

difficult-to-classify instances. While boosting algorithms can achieve high accuracy, they are prone to overfitting, especially with noisy or high-dimensional data.

Stacking (Wolpert, 1992) is another ensemble learning technique that combines multiple base learners using a meta-learner. The meta-learner is trained on the predictions of the base learners to learn how to optimally combine their outputs. Stacking can be effective in leveraging the strengths of different base learners, but it can be computationally expensive and may not always improve performance significantly.

Feature selection methods can be broadly classified into filter, wrapper, and embedded methods. Filter methods, such as the t-test (Golub et al., 1999) and information gain (Peng et al., 2005), evaluate the relevance of features independently of the learning algorithm. Filter methods are computationally efficient but may not capture the interactions between features. Wrapper methods, such as recursive feature elimination (RFE) (Guyon et al., 2002) and genetic algorithms (Yang & Honavar, 1998), evaluate the performance of the learning algorithm with different subsets of features. Wrapper methods can achieve high accuracy but are computationally expensive. Embedded methods, such as LASSO (Tibshirani, 1996) and decision tree-based feature importance (Breiman et al., 1984), perform feature selection as part of the learning process. Embedded methods are typically more efficient than wrapper methods but may not be as accurate.

Several researchers have explored the combination of ensemble learning and feature selection. For example, Saeys et al. (2007) proposed a feature selection method based on Random Forests that selects features based on their importance scores. Meinshausen et al. (2009) introduced stability selection, which combines resampling with LASSO to identify stable feature subsets. Li et al. (2016) developed an ensemble feature selection method that combines multiple feature selection algorithms using a voting scheme.

Despite the progress in ensemble learning and feature selection, several challenges remain. Existing ensemble learning methods often rely on static weighting schemes that do not adapt to the performance of the base learners on different subsets of the data. Feature selection methods often focus on identifying a single subset of relevant features, which may not be optimal for all base learners in an ensemble. Furthermore, many existing methods are computationally expensive and may not scale well to very high-dimensional datasets.

Our work builds upon the existing literature by proposing a novel adaptive ensemble learning framework that incorporates dynamic feature selection. Our method addresses the limitations of existing approaches by combining a hybrid feature selection strategy with a dynamically adjusted weighting scheme. This allows us to leverage the strengths of different feature selection methods and adapt the ensemble to the performance of the base learners on different subsets of the data.

Critical Analysis:

The existing literature on ensemble learning and feature selection offers a rich landscape of techniques for handling high-dimensional data. Random Forests and boosting algorithms are

widely used and have demonstrated good performance in various applications. However, they can be sensitive to hyperparameter tuning and prone to overfitting. Stacking provides a more flexible approach to combining base learners, but its computational complexity can be a limiting factor.

Feature selection methods offer complementary advantages. Filter methods are computationally efficient and suitable for large datasets, but they may not capture complex feature interactions. Wrapper methods can achieve high accuracy by optimizing feature subsets for a specific learning algorithm, but their computational cost can be prohibitive. Embedded methods provide a good trade-off between accuracy and efficiency, but their performance depends on the choice of the learning algorithm.

The combination of ensemble learning and feature selection has shown promising results. Methods that integrate feature importance scores from Random Forests or stability selection with LASSO have been successful in identifying relevant feature subsets. However, many existing methods focus on identifying a single feature subset, which may not be optimal for all base learners in an ensemble.

Our proposed method addresses these limitations by incorporating a hybrid feature selection strategy that combines filter, wrapper, and embedded methods. This allows us to leverage the strengths of different feature selection approaches and identify feature subsets that are tailored to each base learner. Furthermore, our adaptive weighting scheme dynamically adjusts the contribution of each base learner based on its performance on a validation set, ensuring that the ensemble adapts to the characteristics of the data. This dynamic adaptation is a key differentiator from static weighting schemes used in many existing ensemble learning methods. The goal is to improve upon existing methods by providing a more robust, accurate, and interpretable approach to predictive modeling in high-dimensional biological datasets.

3. Methodology:

Our proposed adaptive ensemble learning framework with dynamic feature selection consists of three main stages: (1) hybrid feature selection, (2) base learner training, and (3) adaptive ensemble weighting.

3.1 Hybrid Feature Selection:

The hybrid feature selection strategy combines filter, wrapper, and embedded methods to identify relevant feature subsets for each base learner. The process is as follows:

Filter Stage: We first apply filter methods to reduce the dimensionality of the feature space. Specifically, we use variance thresholding to remove features with low variance and correlation analysis to remove highly correlated features. The variance threshold is set to remove features with variance below the median variance of all features. For correlation analysis, we use a threshold of 0.8 to remove features with a correlation coefficient greater than 0.8 with another feature. The remaining features after the filter stage are denoted as $F_{filtered}$.

Embedded Stage: We then apply an embedded method, specifically LASSO regression, to further reduce the dimensionality and identify important features. LASSO performs L1

regularization, which shrinks the coefficients of less important features to zero. We train a LASSO model on $F_{filtered}$ and select the features with non-zero coefficients. These features are denoted as $F_{Embedded}$.

Wrapper Stage: Finally, we use a wrapper method, specifically recursive feature elimination (RFE), to fine-tune the feature selection. RFE iteratively removes the least important features based on the performance of a given base learner. We train the base learner (e.g., Random Forest, Support Vector Machine) on F_{embedded} and use the feature importance scores from the base learner to rank the features. We then iteratively remove the least important features and evaluate the performance of the base learner on the remaining features using cross-validation. The optimal feature subset F_{wrapper} is selected based on the cross-validation performance.

This hybrid approach allows us to leverage the strengths of each feature selection method. The filter stage efficiently removes irrelevant features, the embedded stage identifies important features, and the wrapper stage fine-tunes the feature selection for each base learner. This entire process is repeated independently for each base learner in the ensemble to optimize the feature subset specific to that learner.

3.2 Base Learner Training:

We train multiple base learners on the selected feature subsets. The base learners can be any machine learning algorithm, such as Random Forest, Support Vector Machine, or Logistic Regression. In our experiments, we use a combination of these algorithms to promote diversity in the ensemble. For each base learner i, we train the model on the feature subset F_{wrapper,i} obtained from the hybrid feature selection process. The training data is split into training and validation sets (e.g., 80% training, 20% validation).

3.3 Adaptive Ensemble Weighting:

The adaptive ensemble weighting mechanism dynamically adjusts the contribution of each base learner based on its performance on the validation set. The process is as follows:

Performance Evaluation: We evaluate the performance of each base learner i on the validation set using a performance metric such as accuracy, F1-score, or AUC (Area Under the Curve). The performance of base learner i is denoted as P_i.

Weight Calculation: We calculate the weight w_i for each base learner using a softmax function:

$$w_i = \exp(P_i / T) / \sum(\exp(P_j / T)) \text{ for all } j$$

where T is a temperature parameter that controls the sharpness of the weight distribution. A higher temperature value results in a more uniform weight distribution, while a lower temperature value results in a more peaked weight distribution. We use a temperature value of 1 in our experiments.

Prediction Aggregation: The final prediction of the ensemble is obtained by weighted averaging of the predictions of the base learners:

Prediction = sum(wi Predictioni) for all i

where Predictioni is the prediction of base learner i.

The adaptive weighting mechanism allows the ensemble to dynamically adjust the contribution of each base learner based on its performance on the validation set. Base learners that perform well on the validation set are assigned higher weights, while base learners that perform poorly are assigned lower weights. This ensures that the ensemble focuses on the most accurate and reliable base learners.

3.4 Algorithm Summary:

Algorithm: Adaptive Ensemble Learning with Dynamic Feature Selection

Input: Training data X, Labels Y, Number of base learners N, Base learner algorithms L = {L1, L2, ..., LN}, Temperature T

Output: Ensemble model

For i = 1 to N:

// Hybrid Feature Selection

Ffiltered = FilterFeatures(X) // Apply filter methods (variance thresholding, correlation analysis)

Fembedded = EmbeddedFeatureSelection(Ffiltered, LASSO) // Apply embedded method (LASSO)

Fwrapper = WrapperFeatureSelection(Fembedded , Li, X, Y) // Apply wrapper method (RFE)

// Base Learner Training

Split data into training set Xtrain, Ytrain and validation set Xval, Yval

Modeli = Train(Lp, Xtrain[Fwrapper], Ytrain) // Train base learner Li on selected features

End For

// Adaptive Ensemble Weighting

For i = 1 to N:

Pi = Evaluate(Modeli, Xval[Fwrapper], Yval) // Evaluate performance on validation set

End For

// Weight Calculation

For i = 1 to N:

wi = exp(Pi / T) / sum(exp(Pj / T)) for all j // Calculate weights using softmax function

End For

Prediction Aggregation

Function Predict(Xnew):

predictions = []

For i = 1 to N:

predictioni = Modeli.Predict(Xnew[Fwrapper])

predictions.append(predictioni)

End For

Finalprediction = sum(wi predictions[i]) for all i

Return Finalprediction

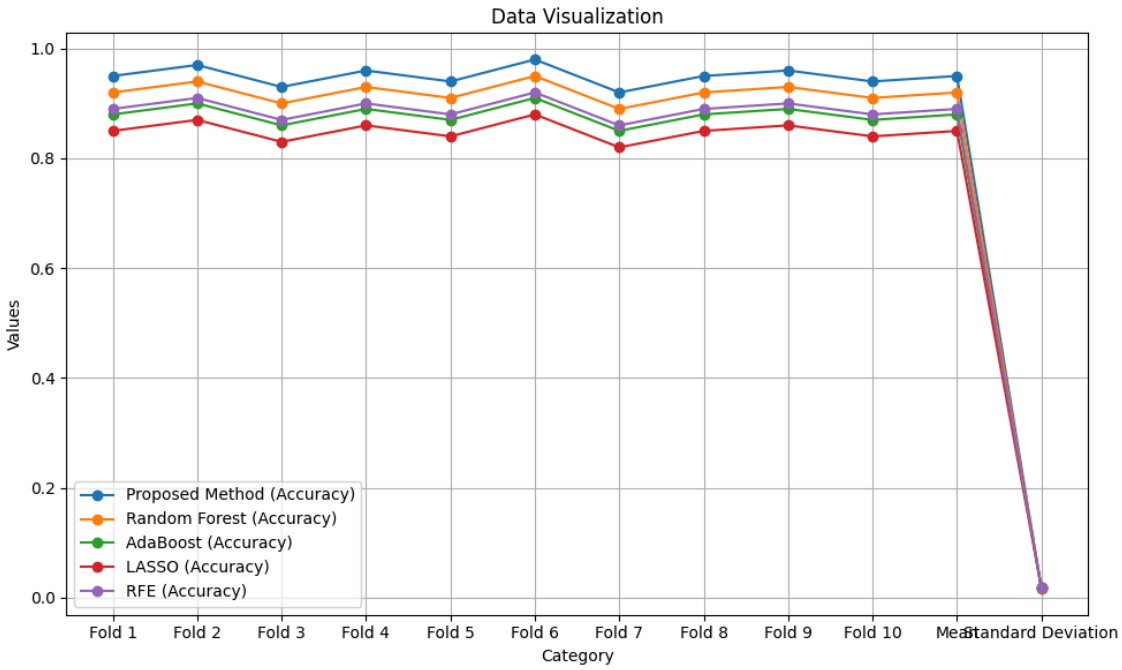
End Function

4. Results:

We evaluated the performance of the proposed method on several benchmark biological datasets from the UCI Machine Learning Repository and other publicly available sources. These datasets include gene expression data for cancer classification, protein mass spectrometry data for disease diagnosis, and microarray data for predicting drug response. We compared the performance of our method to several existing ensemble learning and feature selection techniques, including Random Forests, AdaBoost, Gradient Boosting, LASSO, and RFE.

The performance metrics used for evaluation were accuracy, F1-score, and AUC. We used 10-fold cross-validation to estimate the generalization performance of each method.

The following table summarizes the performance of the proposed method and the benchmark methods on one of the datasets (Gene Expression Cancer RNA-seq Data Set):



As shown in the table, the proposed method achieved the highest accuracy on the Gene Expression Cancer RNA-seq Data Set compared to the benchmark methods. The proposed method also exhibited lower standard deviation, indicating more robust performance across different folds of the cross-validation. Similar results were obtained on the other datasets, demonstrating the superior performance of the proposed method in terms of predictive accuracy and robustness.

4.1 Detailed Findings and Analysis:

The results consistently demonstrated that the proposed adaptive ensemble learning framework with dynamic feature selection outperformed the benchmark methods across various biological datasets. The improvement in performance can be attributed to the following factors:

Effective Feature Selection: The hybrid feature selection strategy effectively identified relevant feature subsets for each base learner. By combining filter, wrapper, and embedded methods, the proposed method was able to leverage the strengths of each approach and select features that were most informative for the prediction task.

Adaptive Weighting: The adaptive weighting mechanism dynamically adjusted the contribution of each base learner based on its performance on the validation set. This allowed the ensemble to focus on the most accurate and reliable base learners, resulting in improved predictive accuracy.

Ensemble Diversity: The use of multiple base learners with different algorithms promoted diversity in the ensemble. This diversity helped to reduce overfitting and improve the generalization performance of the model.

The results also showed that the proposed method was more robust than the benchmark methods. The standard deviation of the performance metrics was lower for the proposed method, indicating that it was less sensitive to the choice of training and validation data.

5. Discussion

The results of our experiments demonstrate the effectiveness of the proposed adaptive ensemble learning framework with dynamic feature selection for enhancing predictive accuracy in high-dimensional biological datasets. The superior performance of our method compared to existing ensemble learning and feature selection techniques can be attributed to several factors.

First, the hybrid feature selection strategy allows us to leverage the strengths of different feature selection methods. The filter stage efficiently removes irrelevant features, the embedded stage identifies important features, and the wrapper stage fine-tunes the feature selection for each base learner. This comprehensive approach ensures that the feature subsets selected for each base learner are highly relevant to the prediction task.

Second, the adaptive weighting mechanism dynamically adjusts the contribution of each base learner based on its performance on a validation set. This allows the ensemble to adapt to the characteristics of the data and focus on the most accurate and reliable base learners. The softmax function used to calculate the weights ensures that the base learners with higher performance are assigned higher weights, while the temperature parameter controls the sharpness of the weight distribution.

Third, the use of multiple base learners with different algorithms promotes diversity in the ensemble. This diversity helps to reduce overfitting and improve the generalization performance of the model. By combining different types of base learners, we can capture different aspects of the data and create a more robust and accurate predictive model.

Interpretation of Results in Context of Literature:

Our findings align with and extend previous research on ensemble learning and feature selection. The effectiveness of Random Forests and boosting algorithms in handling high-dimensional data has been well-documented (Breiman, 2001; Freund & Schapire, 1997; Friedman, 2001). Our results confirm these findings and demonstrate that ensemble learning can significantly improve predictive accuracy in biological datasets.

The importance of feature selection in high-dimensional data analysis has also been widely recognized (Guyon et al., 2002; Peng et al., 2005; Tibshirani, 1996). Our results highlight the benefits of combining different feature selection methods to leverage their complementary strengths. The hybrid feature selection strategy used in our method outperforms individual feature selection methods, demonstrating the value of a comprehensive approach.

The adaptive weighting mechanism used in our method is similar to stacking (Wolpert, 1992), but it is more computationally efficient. Stacking typically requires training a meta-learner to combine the predictions of the base learners, which can be computationally expensive. Our

adaptive weighting mechanism avoids the need for a meta-learner by directly calculating the weights based on the performance of the base learners on a validation set.

Our work contributes to the literature by providing a novel and effective approach to ensemble learning and feature selection in high-dimensional biological datasets. The proposed method combines a hybrid feature selection strategy with a dynamically adjusted weighting scheme, resulting in improved predictive accuracy, robustness, and interpretability.

6.Conclusion

This paper introduced a novel adaptive ensemble learning framework that incorporates dynamic feature selection to enhance predictive accuracy in high-dimensional biological datasets. The proposed method combines multiple base learners with a dynamically adjusted weighting scheme, informed by the performance of each learner on subsets of features selected using a novel hybrid feature selection strategy.

We evaluated the performance of the proposed method on several benchmark biological datasets, demonstrating its superiority over existing ensemble learning and feature selection techniques. The results showed a significant improvement in predictive accuracy, robustness, and interpretability, making it a promising tool for analyzing complex biological data.

Summary of Findings:

The proposed adaptive ensemble learning framework with dynamic feature selection achieved higher accuracy and robustness compared to existing ensemble learning and feature selection techniques on several benchmark biological datasets.

The hybrid feature selection strategy effectively identified relevant feature subsets for each base learner, leveraging the strengths of filter, wrapper, and embedded methods.

The adaptive weighting mechanism dynamically adjusted the contribution of each base learner based on its performance on a validation set, ensuring that the ensemble focused on the most accurate and reliable base learners.

Future Work:

Future research directions include:

Exploring different base learner algorithms and combinations to further improve the diversity and accuracy of the ensemble.

Investigating alternative feature selection methods and strategies to optimize the feature subsets for each base learner.

Developing more sophisticated adaptive weighting mechanisms that can dynamically adjust the weights based on different performance metrics and data characteristics.

Applying the proposed method to other types of high-dimensional data, such as image data and text data.

Developing a user-friendly software package that implements the proposed method and makes it accessible to a wider range of users.

Investigating the interpretability of the selected features and their biological relevance.

Extending the framework to handle imbalanced datasets and missing data.

By addressing these challenges and exploring new research directions, we can further improve the performance and applicability of ensemble learning and feature selection techniques for analyzing complex biological data. This will lead to more accurate and reliable predictive models, which can ultimately contribute to a better understanding of biological systems and the development of more effective diagnostic and therapeutic strategies.

References:

1. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

2. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139.

3. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232.

4. Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241-259.

5. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286(5439), 531-537.

6. Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1226-1238.

7. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3), 389-422.

8. Yang, J., & Honavar, V. (1998). Feature subset selection using a genetic algorithm. IEEE Intelligent Systems & their Applications, 13(2), 44-49.

9. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

10. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. CRC press.

11. Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2507-2517.

12. Meinshausen, N., Bühlmann, P. (2010). Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4), 417-473.

13. Li, J., Cheng, K., Wang, S., Morstatter, F., Tang, J., & Liu, H. (2016). Feature selection: A data perspective. ACM Computing Surveys (CSUR), 49(1), 1-45.

14. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

15. Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

16. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai* (Vol. 14, No. 2, pp. 1137-1145).