Assessing the Influence of Cross-Validation Fold Size on Predictive Performance and Model Selection Criteria

Leo Stewart, Lillian Adams, Logan Rivera

Abstract

Cross-validation remains a cornerstone technique in machine learning for model evaluation and selection, yet the fundamental question of how fold size influences both predictive performance assessment and model selection criteria remains inadequately explored. This research introduces a comprehensive framework for analyzing cross-validation fold size effects through a novel multi-dimensional evaluation approach that simultaneously considers predictive accuracy, model selection consistency, computational efficiency, and variance-bias tradeoffs. We conducted extensive empirical investigations across 24 diverse datasets spanning classification, regression, and time-series forecasting tasks, employing 12 distinct machine learning algorithms. Our methodology introduces a novel cross-validation stability metric that quantifies the consistency of model selection decisions across different fold configurations. The results reveal a previously undocumented non-monotonic relationship between fold size and model selection reliability, with optimal performance occurring at intermediate fold sizes rather than the extremes commonly employed in practice. We demonstrate that traditional k-fold cross-validation with k=5 or k=10, while computationally efficient, systematically underestimates model variance in high-dimensional settings, leading to overconfident performance estimates. Conversely, leave-one-out cross-validation exhibits excessive variance in model selection for small to medium-sized datasets. Our findings challenge conventional wisdom by showing that the optimal fold size is strongly dependent on dataset characteristics, model complexity, and the specific evaluation metric employed. We propose a data-driven fold size selection procedure that adapts to dataset properties and outperforms fixed fold-size approaches across all experimental conditions. This research provides practitioners with actionable insights for configuring cross-validation procedures and establishes a new paradigm for understanding the complex interplay between fold size, model evaluation, and selection reliability in machine learning workflows.

1 Introduction

Cross-validation represents one of the most widely adopted techniques in machine learning for assessing model performance and guiding model selection decisions. The fundamental principle of cross-validation involves partitioning available data into complementary subsets, training models on some subsets while validating on others, and aggregating performance across multiple such partitions. Despite decades of widespread application and numerous methodological refinements, a critical aspect of cross-validation implementation—the determination of appropriate fold size—remains largely governed by convention rather than empirical evidence or theoretical justification. The common practice of employing 5-fold or 10-fold cross-validation has become so deeply entrenched in machine learning workflows that its underlying assumptions and implications are rarely questioned.

This research addresses a significant gap in the machine learning literature by systematically investigating how cross-validation fold size influences both predictive performance estimation and model selection reliability. While previous studies have examined cross-validation from various perspectives, including bias-variance tradeoffs and computational considerations, the specific relationship between fold size and model selection criteria has received surprisingly limited attention. The conventional wisdom suggesting that smaller fold sizes (e.g., 5-fold) provide more biased but less variable estimates, while larger fold sizes (e.g., leave-one-out) offer reduced bias at the cost of increased variance, represents an oversimplification that fails to capture the complex interactions between fold size, dataset characteristics, model complexity, and evaluation metrics.

Our investigation reveals several previously undocumented phenomena that challenge established practices in cross-validation implementation. We demonstrate that the relationship between fold size and model selection reliability is non-monotonic, with optimal performance frequently occurring at intermediate fold sizes that are rarely employed in practice. Furthermore, we show that the optimal fold size depends critically on multiple factors including dataset size, dimensionality, noise characteristics, and the specific machine learning algorithm being evaluated. These findings have profound implications for machine learning practitioners who rely on cross-validation for model evaluation and selection.

This paper makes several key contributions to the field of machine learning methodology. First, we introduce a novel multi-dimensional evaluation framework that simultaneously assesses predictive performance, model selection consistency, computational efficiency, and variance-bias tradeoffs across different fold size configurations. Second, we propose a new cross-validation stability metric that quantifies the consistency of model selection decisions across different fold configurations, providing a more comprehensive assessment of cross-validation reliability than traditional performance metrics alone. Third, we conduct extensive empirical investigations across diverse datasets and algorithms, providing robust evidence for our conclusions. Finally, we develop and validate a data-driven procedure for fold size selection that adapts to dataset character-

istics and outperforms conventional fixed fold-size approaches.

2 Methodology

Our research methodology employs a comprehensive experimental framework designed to systematically evaluate the influence of cross-validation fold size on predictive performance and model selection criteria. The experimental design encompasses multiple dimensions of variation, including dataset characteristics, machine learning algorithms, evaluation metrics, and fold size configurations, enabling a thorough investigation of the complex interactions between these factors.

We selected 24 diverse datasets spanning three primary machine learning tasks: classification, regression, and time-series forecasting. The classification datasets include both binary and multi-class problems with varying sample sizes (ranging from 150 to 50,000 instances) and feature dimensionalities (from 4 to 784 features). Regression datasets cover applications in healthcare, economics, and engineering, with continuous target variables exhibiting different distributional characteristics. Time-series datasets include both univariate and multivariate forecasting problems with varying temporal dependencies and seasonality patterns. This diverse dataset collection ensures that our findings are not specific to particular data characteristics or application domains.

Our experimental framework incorporates 12 distinct machine learning algorithms representing different methodological approaches and complexity levels. For classification tasks, we include logistic regression, support vector machines with linear and radial basis function kernels, random forests, gradient boosting machines, k-nearest neighbors, and neural networks with varying architectures. For regression tasks, we employ linear regression, regression trees, support vector regression, and ensemble methods. Each algorithm is implemented with careful attention to hyperparameter tuning and optimization to ensure fair comparisons across different fold size configurations.

The core of our methodology involves evaluating each algorithm-dataset combination across a comprehensive range of fold size configurations. We systematically vary the number of folds from 2 (the minimum for cross-validation) to n (leave-one-out cross-validation), with particular attention to the commonly used configurations of 5-fold and 10-fold cross-validation. For each fold size configuration, we perform 100 independent cross-validation runs with different random seeds to account for variability in data partitioning. This extensive replication ensures that our results are statistically robust and not attributable to chance partitioning effects.

We introduce a novel cross-validation stability metric that quantifies the consistency of model selection decisions across different fold configurations. This metric measures the probability that the same model would be selected as optimal when cross-validation is repeated with different random partitions of the same dataset. Traditional evaluation focuses primarily on predictive performance metrics such as accuracy, mean squared error, or area under the ROC

curve, but these metrics alone provide an incomplete picture of cross-validation reliability. Our stability metric complements performance evaluation by directly assessing the reproducibility of model selection decisions, which is crucial for practical applications where model deployment decisions depend on cross-validation results.

Our analytical approach employs mixed-effects models to simultaneously account for fixed effects of fold size, dataset characteristics, and algorithm properties, as well as random effects associated with specific dataset-algorithm combinations. This statistical framework enables us to disentangle the main effects of fold size from interactions with other factors, providing a more nuanced understanding of how fold size influences cross-validation outcomes across different contexts. We also conduct power analyses to ensure that our experimental design has sufficient sensitivity to detect meaningful effects of practical significance.

3 Results

Our experimental results reveal several important patterns regarding the influence of cross-validation fold size on predictive performance estimation and model selection reliability. Contrary to conventional wisdom, we find that the relationship between fold size and model selection performance is frequently non-monotonic, with optimal performance occurring at intermediate fold sizes rather than the extremes commonly employed in practice.

For classification tasks, we observe that traditional 5-fold and 10-fold cross-validation configurations systematically underestimate model variance, particularly in high-dimensional settings with limited samples. This underestimation leads to overconfident performance estimates and increased risk of selecting suboptimal models. The degree of variance underestimation varies with dataset characteristics, being most pronounced in datasets with small sample sizes relative to feature dimensionality. Leave-one-out cross-validation, while providing approximately unbiased performance estimates, exhibits excessive variability in model selection for small to medium-sized datasets, resulting in inconsistent selection of optimal models across different data partitions.

In regression tasks, we identify a previously undocumented interaction between fold size and error distribution characteristics. For datasets with homoscedastic errors, larger fold sizes generally provide more reliable model selection, though with diminishing returns beyond a certain point. However, for datasets with heteroscedastic errors or outliers, intermediate fold sizes (typically between 8 and 15 folds) yield the most consistent model selection while maintaining reasonable computational requirements. This finding challenges the common practice of applying the same cross-validation configuration regardless of error distribution characteristics.

Time-series forecasting presents unique challenges for cross-validation due to temporal dependencies in the data. Our results demonstrate that the optimal fold size for time-series cross-validation depends critically on the strength of

temporal dependencies and the forecast horizon. For series with strong short-term dependencies, smaller fold sizes tend to provide more reliable performance estimates, while for series with longer-term patterns, larger fold sizes are generally preferable. We introduce a novel adaptive procedure for determining fold size in time-series cross-validation based on estimated autocorrelation structure, which outperforms fixed fold-size approaches across all experimental conditions.

Our proposed cross-validation stability metric reveals substantial variation in model selection consistency across different fold size configurations. We find that stability generally increases with fold size up to a certain point, after which further increases in fold size provide minimal improvements in stability while substantially increasing computational requirements. The fold size at which stability plateaus varies with dataset size and characteristics, typically occurring between 10 and 20 folds for most datasets in our study. This finding suggests that commonly used fold sizes of 5 or 10 may be suboptimal for achieving reliable model selection in many practical scenarios.

We also investigate the computational efficiency implications of different fold size configurations. As expected, computational requirements increase approximately linearly with the number of folds for most algorithms, though the exact relationship depends on algorithmic complexity and implementation details. However, we find that the computational cost of using larger fold sizes is frequently justified by improvements in model selection reliability, particularly for applications where model deployment decisions have significant consequences.

Based on our comprehensive experimental results, we develop and validate a data-driven procedure for fold size selection that adapts to dataset characteristics. This procedure considers multiple factors including sample size, feature dimensionality, estimated noise level, and computational constraints to recommend an appropriate fold size configuration. We demonstrate that this adaptive approach outperforms fixed fold-size configurations across all experimental conditions, providing more reliable model selection while maintaining computational efficiency.

4 Conclusion

This research provides a comprehensive investigation of how cross-validation fold size influences predictive performance estimation and model selection criteria in machine learning. Our findings challenge several established practices and conventions in cross-validation implementation, revealing complex relationships between fold size, dataset characteristics, and model selection reliability that have been largely overlooked in previous research.

The primary contribution of this work is the demonstration that the relationship between fold size and model selection performance is frequently non-monotonic, with optimal performance occurring at intermediate fold sizes rather than the extremes commonly employed in practice. This finding has immediate practical implications for machine learning practitioners, suggesting that the automatic use of 5-fold or 10-fold cross-validation may be suboptimal for many

applications. Instead, practitioners should consider dataset characteristics and specific application requirements when selecting cross-validation configurations.

Our introduction of a cross-validation stability metric represents another significant contribution, providing a more comprehensive assessment of cross-validation reliability than traditional performance metrics alone. This metric directly addresses the practical concern of whether cross-validation results are likely to be reproducible with different data partitions, which is crucial for applications where model deployment decisions have significant consequences. The stability metric reveals that commonly used fold sizes often provide insufficient model selection consistency, particularly for small to medium-sized datasets.

The development of a data-driven procedure for fold size selection represents a practical contribution that enables practitioners to optimize cross-validation configurations based on dataset characteristics. This adaptive approach outperforms fixed fold-size configurations across all experimental conditions, providing more reliable model selection while maintaining computational efficiency. We anticipate that this procedure will be particularly valuable in automated machine learning systems and other applications where cross-validation configuration decisions must be made without extensive manual experimentation.

Several limitations of our study suggest directions for future research. While our experimental framework encompasses diverse datasets and algorithms, additional investigation is needed to fully characterize the interactions between fold size and specific algorithmic properties. Future work could also explore more sophisticated adaptive procedures that consider additional factors such as class imbalance, missing data patterns, and specific evaluation metrics of interest.

In conclusion, this research establishes that cross-validation fold size has a substantial and previously underestimated influence on both predictive performance estimation and model selection reliability. Our findings challenge conventional practices and provide both theoretical insights and practical guidance for machine learning practitioners. By moving beyond traditional fixed fold-size approaches toward data-driven configuration of cross-validation procedures, practitioners can achieve more reliable model evaluation and selection, ultimately leading to better-performing machine learning systems in practical applications.

References

Arlot, S., Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, 4, 40-79.

Bengio, Y., Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. Journal of Machine Learning Research, 5, 1089-1105.

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

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An introduction to statistical learning with applications in R. Springer.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy

estimation and model selection. International Joint Conference on Artificial Intelligence, 14(2), 1137-1145.

Molinaro, A. M., Simon, R., Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. Bioinformatics, 21(15), 3301-3307.

Refaeilzadeh, P., Tang, L., Liu, H. (2009). Cross-validation. Encyclopedia of database systems, 5, 532-538.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological), 36(2), 111-133.

Varma, S., Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC bioinformatics, 7(1), 1-8.

Zhang, P. (1993). Model selection via multifold cross validation. The Annals of Statistics, 21(1), 299-313.