Statistical Modeling: The Two Cultures by Leo Brieman

Maricella Velita

December 2, 2024

Abstract

- The goal is to demonstrate how algorithmic models improve prediction accuracy and offer deeper insights into complex data that traditional methods may overlook. The author explains how algorithmic models, such as random forests, are better at dealing with large datasets, identifying key variables, and detecting patterns that simpler models cannot. The paper critiques the over reliance on traditional data models, arguing that the best approach should depend on the nature of the data and the problem itself, not just sticking to a specific model. Though the author is not against data modeling, they advocate for using both data modeling and algorithmic modeling to solve real world problems

Contrasting Statistical Cultures: Data vs. Algorithmic Modeling

Two Cultures in Statistics:

- **Data Modeling:**
  - Uses predefined models (e.g., linear regression, logistic regression).
  - Focuses on understanding variable relationships and data-generating processes.
- **Algorithmic Modeling:**
  - Treats relationships as complex or unknown.
  - Relies on algorithms (e.g., decision trees, neural networks).
  - Measures model performance based on predictive accuracy.

Brieman's Critique

- Dominance of data modeling has led to irrelevant theory and limited practical applications.
- Algorithmic models prioritize predictive accuracy, proving valuable in fields like medicine and genetics.
- The paper explores algorithmic modeling, recent advances, and practical applications in detail.
- Brieman suggests statisticians should embrace algorithmic models and move beyond the constraints of traditional data models.

Data Modeling

Traditional Approach to Data Models:

- Assumes data originates from specific models.
- Often fails to represent real-world mechanisms.
- Overemphasis on hypothetical models.
- Data models are often parametric, assuming a specific structure for relationships in data.
- Rely on assumptions that may not reflect the complexities of nature.

Challenges and Limitations

Limitations:

- **Model fit:** Conclusions depend on how well the model represents reality. Poorly fitting models can lead to erroneous conclusions
- **Goodness-of-fit tests:** These tests, including residual analysis, often lack the power to detect poor fits, especially in high-dimensional data. Can lead to misleading conclusions about model accuracy

Challenges in Statistical Modeling:

- **Multiplicity of models:**
  - Many models can fit the same data but yield different conclusions.
  - Leads to ambiguity in understanding the true relationship between variables.

Predictive Accuracy

- A better measure of model performance is predictive accuracy (how well the model predicts unseen data)
- Techniques like cross-validation are recommended to assess predictive accuracy more reliably, yet they remain underused in traditional statistical practices.
- Cross validation reduces bias in predictive accuracy estimates

Example

Gender Discrimination Study:

- A regression analysis concluded there was gender discrimination in salaries based on a significant coefficient for gender. However, this conclusion might be invalid due to:
    - Lack of validation of the linear model's assumptions.
    - Over-reliance on the 5 percent significance threshold
    - Failure to critically assess the data's capacity to address the research question.
    - Focus on the model rather than the problem itself.

Summary of Challenges in Statistical Modeling

Key issues highlighted:

- Overemphasis on hypothetical models over real-world data
- Inadequate tools for assessing model fit in complex datasets
- Lack of focus on predictive accuracy

Proposed solutions:

- Broaden the statistical toolbox with algorithmic and non-parametric methods
- Prioritize predictive accuracy through cross-validation
- Combine traditional and algorithmic methods

Algorithmic Modeling

- Allows for more accurate and flexible representation of real-world situations compared to traditional modeling techniques.
- Unlike traditional models, it does not assume a specific form or structure for how data is generated.
- Handles complex, non-linear, and high dimensional data more effectively than traditional methods.
- Prioritizes predictive accuracy over interpretability.

Historical context:

- Algorithmic approaches have been applied in fields like industrial statistics and medical data.
- Prominence began in the 1980s with neural networks, decision trees.
- Used where traditional models failed (e.g., speech recognition and financial predictions).

Lessons from Machine Learning

- Rashomon Effect: multiple models can have similar accuracy, creating uncertainty about which model is best.
- Occam's razor: simpler models are interpretable but may sacrifice accuracy.
- Dimensionality: high variable count increases model complexity, posing challenges and opportunities.

Occam and Simplicity vs. Accuracy

- Occam's razor suggests simpler models are better, but in predictive analytics, simplicity often sacrifices accuracy.
  - Example: decision trees are interpretable but less accurate than complex models like random forests or neural networks.
- Occam's Dilemma: accuracy often requires more complex models, but these models may be harder to interpret.
  - RF combines multiple decision trees, making predictions difficult to interpret.

Random Forests

- Random forests are top-performing predictive models.
- Consistently ranked highly across datasets.
- A group of decision trees, significantly improve prediction accuracy by introducing randomness in model construction.
- Reduce test set errors more effectively than single trees or other classifiers.
- Downside: they are nearly impossible to interpret.

Interpretability vs. Predictive Accuracy

- Modern predictive methods, like random forests often sacrifice interpretability for accuracy.
- The goal is not simplicity but obtaining reliable information about the relationship between predictor and response variables.
- Complex models can reveal insights that are not accessible through traditional models like logistic regression.

Dimensionality

Traditional view:

- High dimensional data should be reduced to avoid overfitting.

New perspective:

- Increasing dimensionality by adding features can enhance predictive performance.
- Complex models like RF benefit from large feature sets.

Leveraging Features in High Dimensions

Effective methods:

- Shape Recognition Forest: uses shallow decision trees and geometric features for decisions. Achieved a test set error rate of 0.7 percent using large training and test set.
- Support Vector Machines (SVM): Expands feature space using polynomial monomials of the original predictor variables.
- However, excessive dimensionality may lead to high generalization error.

Examples of High-Dimensional Modeling

Variable importance in Medical Data:

- Random forests provided more reliable information about variable importance compared to logistic regression.
- In a hepatitis dataset, traditional methods showed instability in identifying important variables.
- Applications in medical and microarray data reveal patterns and relationships that simpler models may overlook.

Examples of Random Forest Applications

Medical data (Hepatitis and Liver Disease):

- Highlighted key variables related to liver function and clustered class two patients into two distinct groups based on these tests.
- Demonstrated the stability and utility of random forests in analyzing biomedical datasets.

Microarray data (Lymphoma):

- Analyzed a data set with over 4,600 variables and identified gene expressions with high precision, outperforming data models in accuracy and variable selection.

Algorithmic vs. Data Models

Algorithmic models:

- Focus on predictive accuracy and real-world problem solving.
- Can process massive data sets with high-dimensional features (e.g., genetic data).
- Does not require assumptions about the underlying data distribution.

Data models:

- Often emphasizes simplicity and interpretability but underperforms in complex datasets.
- Tend to delete variables, risking the loss of important interactions.

Flexibility in Model Choice

Algorithmic Modeling vs. Data Modeling.

- Statistical solutions should prioritize the specific problem and data over sticking to rigid models.
- RF and similar models can provide better insights in larger datasets.
- Algorithmic models offer greater accuracy and insights in complex datasets.

Conclusion

- The author urges that statisticians should embrace diverse tools (algorithmic and traditional) to solve real-world problems.
- Real-world problems demand flexible methods that prioritize solving problems over adhering to specific model frameworks.
- Statistical methods should be driven by the problem and the data, rather than the model.
- Breiman does not oppose data modeling but encourages statisticians to explore beyond traditional methods.