Recursive Binary Partitioning
Old Dogs with New Tricks

KDD Conference 2009

David J. Slate and Peter W. Frey

Background

- D. J. Slate and L. R. Atkin, “Chess 4.5 – the Northwestern University chess program”. In P. W. Frey (Ed.), Chess Skill in Man and Machine, Springer Verlag, 1977, 1978, 1983.

- P.W. Frey,”Algorithmic Strategies for Improving the Performance of Game-Playing Programs”. In D. Farmer, A. Lapedes, N. Packard and B. Wendroff (Eds.), Evolution, Games and Learning, North-Holland Physics Publishing, Amsterdam, 1986.

- P. W. Frey and D. J. Slate, “Letter Recognition Using Holland-Style Adaptive Classifiers”, Machine Learning, 6, 1991, 161-182.

Database Characteristics

- Hundreds of Thousands of Records
- Missing Data
- Erroneous Data Entries

Forecasting Challenges

- Categorical Attributes and/or Outcomes
- Non-Monotonic Relationships between Attributes and the Outcome
- Skewed or Bimodal Numerical Distributions
- Non-Additive Attribute Influence on Outcomes
- Multiple Attribute Combinations that Produce Desirable Outcomes

Recursive Binary Partitioning

J.A. Sonquist and J.N. Morgan, “The Detection of Interaction Effects”, Institute of Social Research Monograph no. 35, Chicago: University of Michigan, 1964

G. V. Kass, An Exploratory Technique for Investigating Large Quantities of Categorical Data. Journal of Applied Statistics, 29:2, 1980, 119-127.

L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, Classification and Regression Trees, Pacific Grove, CA: Wadsworth, 1984.

Advantages of RBP

- Rational Treatment of Missing Data
- Numerical Distribution Is Not Relevant
- Monotonic Relationship Not Required
- Okay with Multiple “Flavors” of a Good Outcome
- Non-Additive Relationships Are Not a Problem
- Large Data Sets Are an Advantage
- Computational Time Is Reasonable
- Methodological Transparency

Problems With RBP

- A Greedy, Myopic Algorithm
- Overfits the Training Sample
- Overshadowing of Useful Attributes

Attacking the Problems

- Look-Ahead Search
- Minimum Record Count for Leaf Node
- Minimum Split Score for Leaf Node
- Random Perturbation of Attribute Availability at Each Node
- Random Perturbation of Record Availability at Each Node

Ensemble RBP

- Split Rule
- Terminal Nodes
- Leaf Node Values
- Missing Values
- Ensemble of Decision Trees
- Parameter Tuning

KDD Cup: Preprocessing

- Removed Attributes with a Constant Value
- No Normalization
- Retained Missing Values
- No Limit on Range of Numerical Attributes
- Retained Duplicate Attributes
- No Generation of Additional Features
- No Modification of Categoric Attributes

KDD Cup: Attribute Selection

- Preliminary Ensemble Construction for Selection of Attributes
- Preliminary Traditional RBP for Selection of Attributes

KDD Cup: Model Building

- Ensemble RBP methodology using Random Attribute Omission at Each Node
- 40,000 Record Construction Set
- 10,000 Record Test Set
- 5-Fold Cross Validation to Select Parameters
- Final Models Built on 50,000 records

Observations

- 15,000 Attributes and 50,000 records
- Binary rather than Numeric Outcomes
- Categoric Attributes without Identifying Information