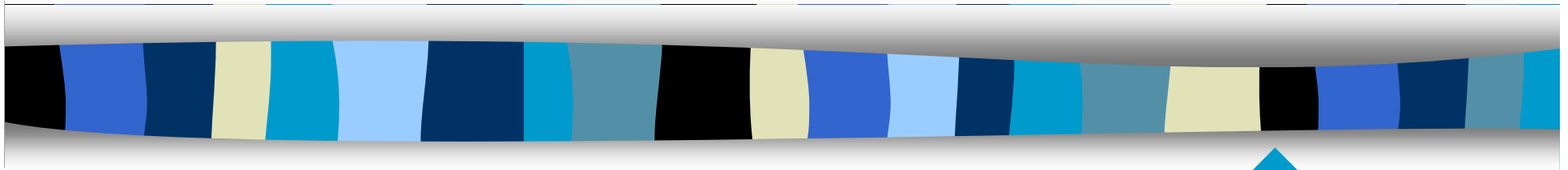


Supervised Learning In Quest (SLIQ)



Qian Wang

Arthurs

Manish Mehta
Rakesh Agrawal
Jorma Rissanen



Introduction

- Classification
- Decision-Tree Classifier
 - Presorting
 - Breadth-First Growth
 - Subsetting Issues
 - Tree Growing & Tree Pruning
- Performance



Presorting Training Data

- To reduce the cost of evaluating numeric attributes.
- Separate list for each attribute
 - Class list
 - Attribute list
- Initialize the lists
- Sort attribute lists independently



Breadth-First Growth

- Processing Node Splits
 - For a value in the current attribute list
 - Get the corresponding entry in class list
 - Update the histograms
- Updating the class list
 - Create child nodes for each of the leaf
 - Update the class list



Tree Growing

- Repeat splitting nodes and updating labels until:
 - Each leaf node becomes a pure node.
 - No further splits are required
- Optimization
 - Condense the attribute list



Subsetting for Categorical Attributes

- Use hybrid approach
 - If the cardinality of possible values is less than a threshold, all of the subsets are evaluated.
 - Otherwise, use a greedy algorithm to obtain the desired subset.
 - Repeat the process until there is no improvement in the splits.



Tree Pruning

- MDL principle
 - $\text{cost}(M, D) = \text{cost}(D|M) + \text{cost}(M)$
- Hybrid method
 - Data Encoding
 - Model Encoding
 - Pruning Algorithms
 - cost on internal node
 - cost on leaf node



Performance

- Easy to implement
- Achieves better classification accuracy
- Produces small decision trees
- Has small execution times
- Achieves good scalability
- Performs well for large datasets

