Boolean Decision Rules via Column Generation

—

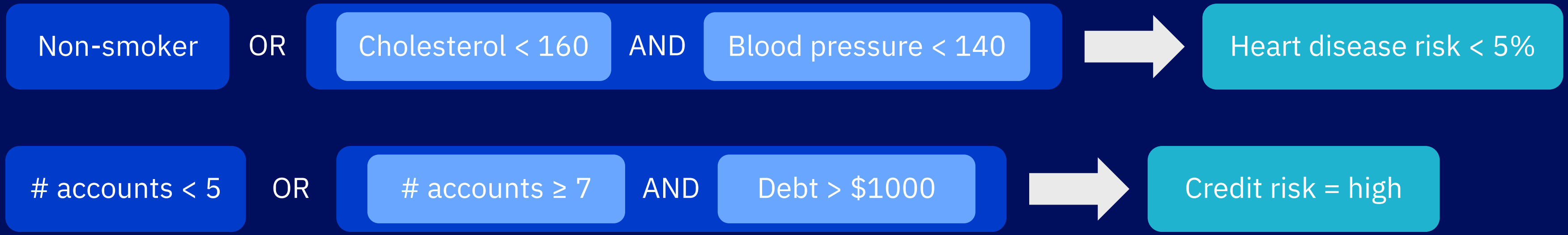
Sanjeeb Dash*
Oktay Günlük*
Dennis Wei*

IBM Research

Problem Statement

Learn Boolean rules for binary classification

- Disjunctive normal form (DNF, OR of ANDs)
- Conjunctive normal form (CNF, AND of ORs)



Non-smoker OR Cholesterol < 160 AND Blood pressure < 140 → Heart disease risk < 5%

# accounts < 5 OR # accounts ≥ 7 AND Debt > $1000 → Credit risk = high

Rules with few clauses and conditions are interpretable

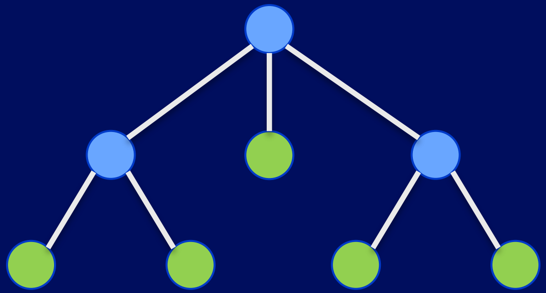
Optimize accuracy vs. simplicity using integer programming (IP)

Related Models

DNF Boolean rule = Decision rule set

IF A THEN Y=1
IF B AND C THEN Y=1
IF D AND E THEN Y=1
ELSE Y=0

Decision tree



Decision list

IF A THEN Y=1
ELSE IF B AND C THEN Y=1
ELSE IF D AND E THEN Y=1
ELSE Y=0

Preliminaries

Assume non-binary features have been binarized

- Categorical: “one-hot” coding (e.g. color=red, color=blue)
- Numerical: comparison with thresholds (e.g. blood pressure ≤ 130, >130)

Main Challenge

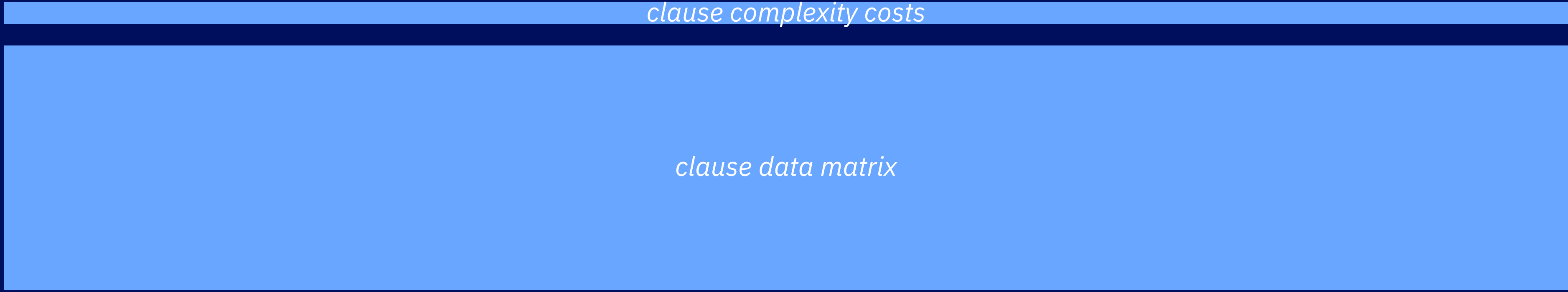
Exponentially many possible clauses

- e.g. # accounts, # accounts AND debt, # accounts AND debt AND months since delinquency, …

Previous works limited search using heuristics

Column Generation

Select clauses from exponentially large set



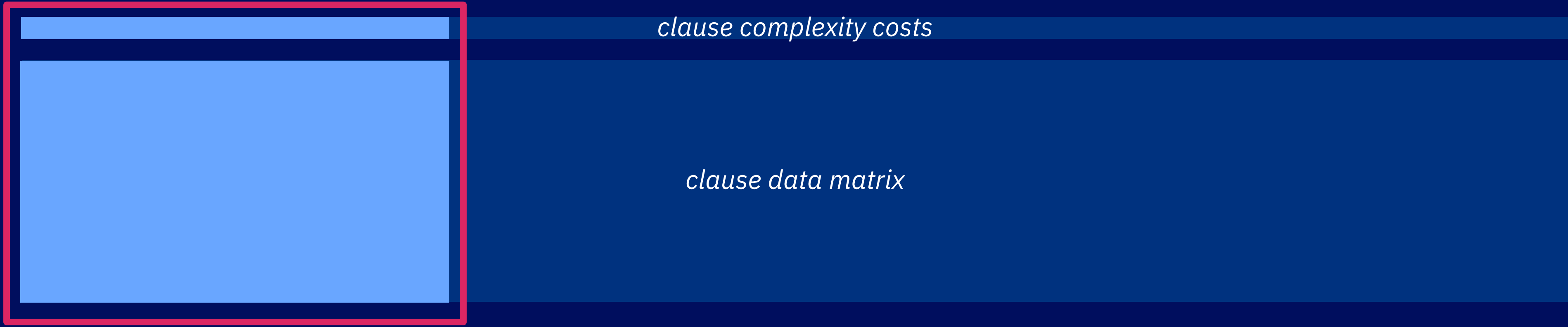
clause complexity costs

clause data matrix

Master IP/LP

Column Generation

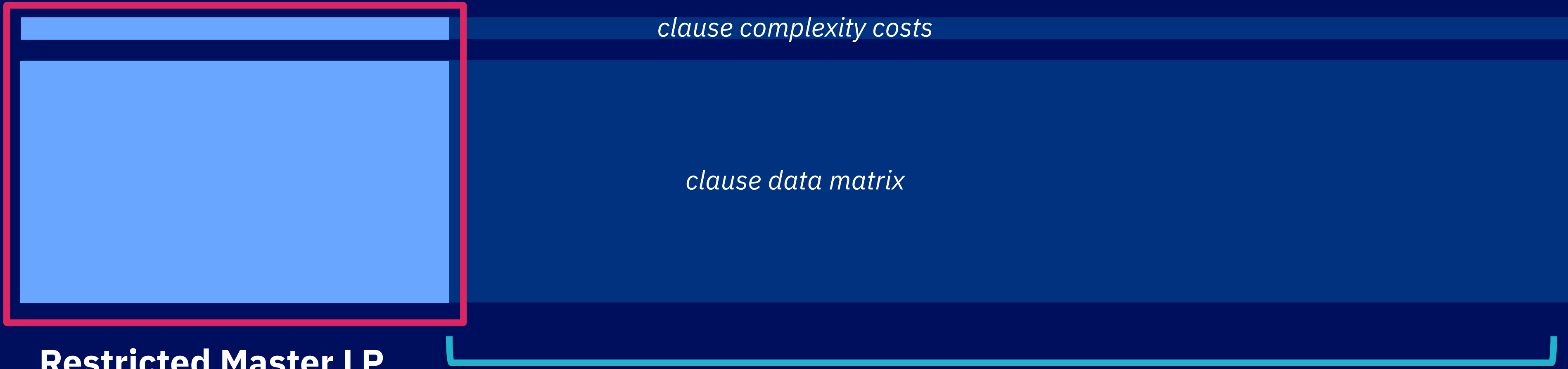
Solve only over small subsets



Restricted Master LP

Column Generation

Solve only over small subsets

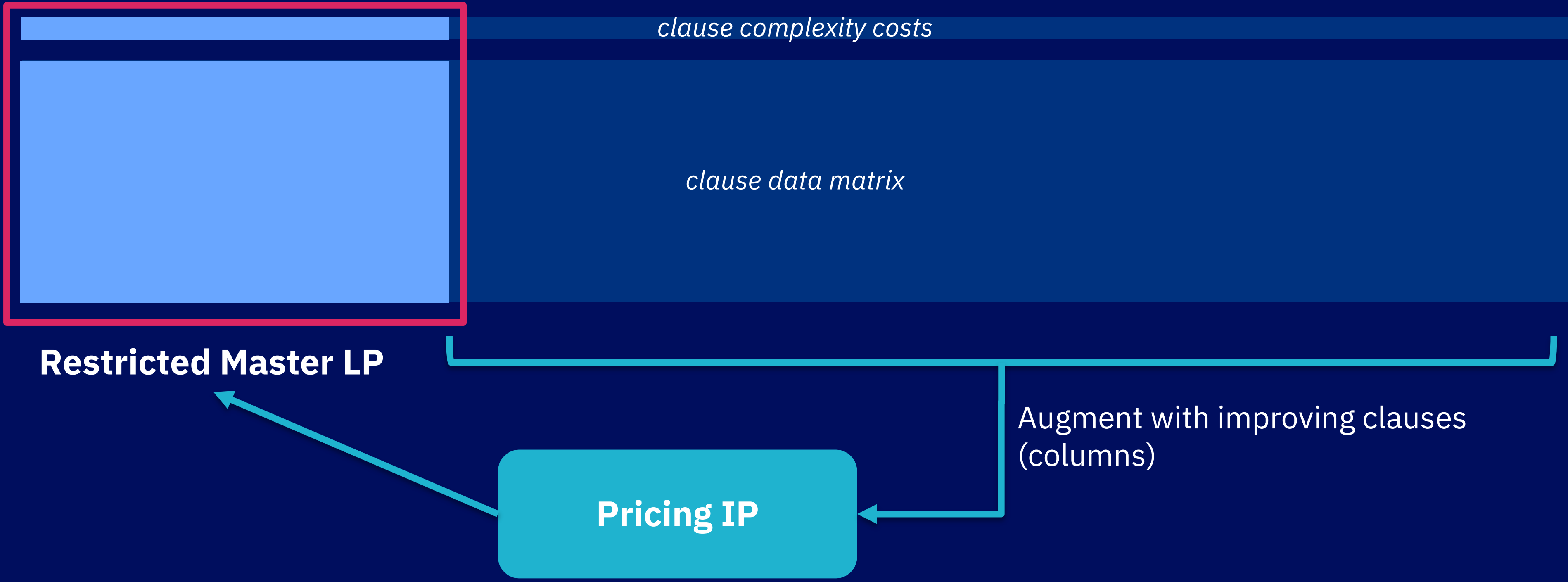


Restricted Master LP

Augment with improving clauses (columns)

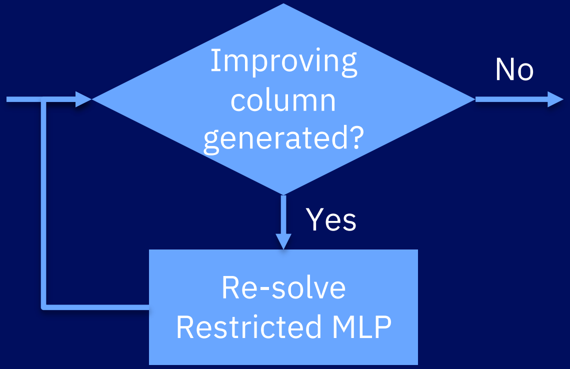
Column Generation

Solve only over small subsets



Also generate columns using heuristic

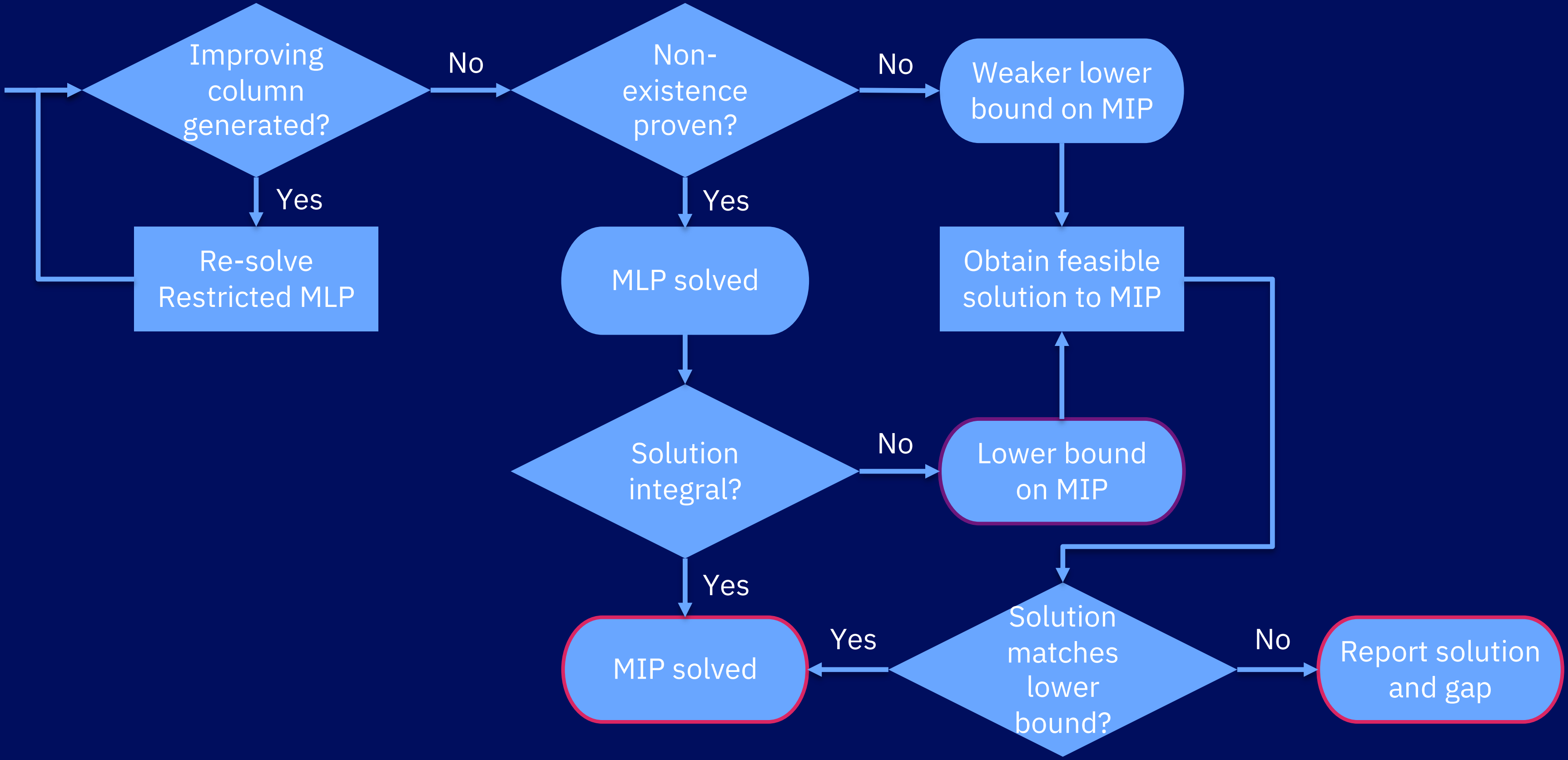
Procedure and Optimality Guarantees



IPs solved using CPLEX

5 min time limit overall

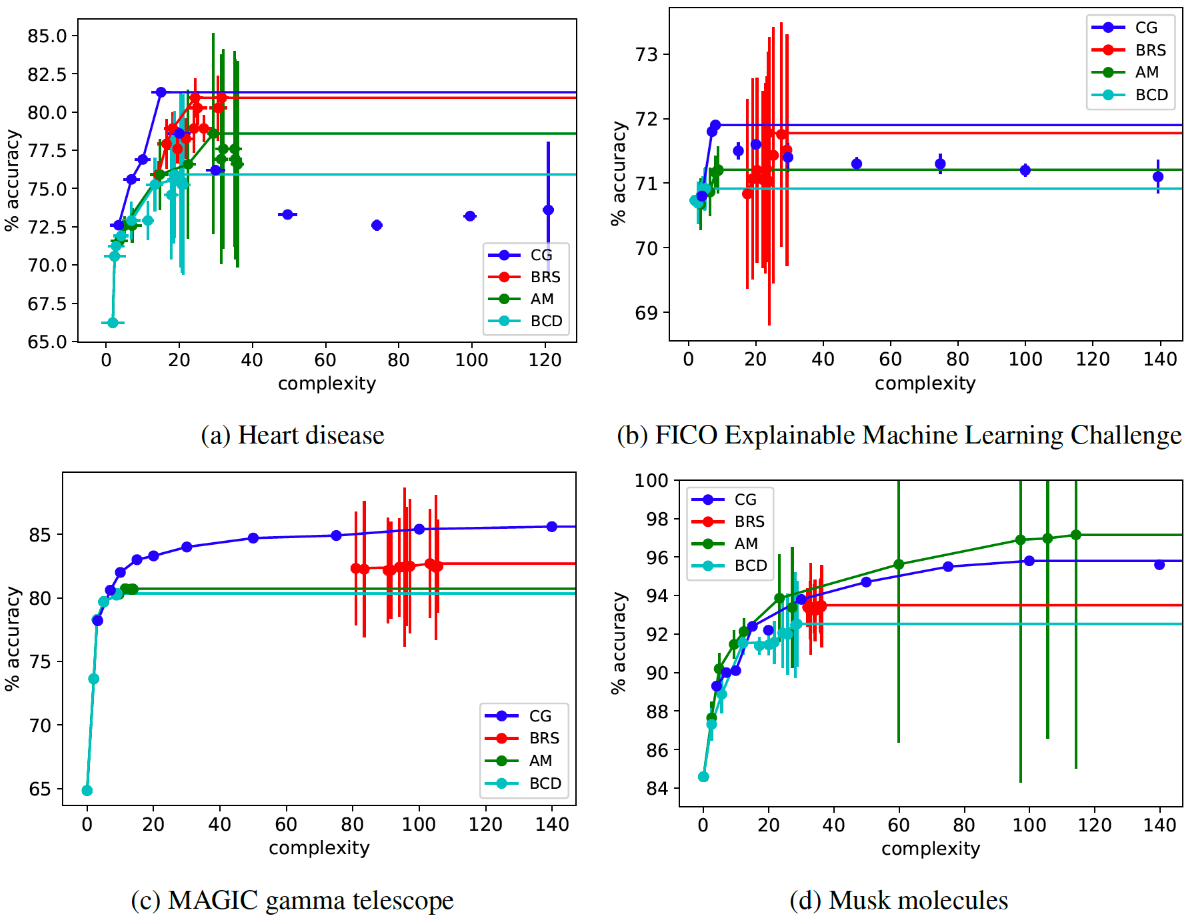
Procedure and Optimality Guarantees



Accuracy-Complexity Trade-Off

Lines connect Pareto-efficient points

Column generation (CG) dominates on 8 of 16 datasets and is close on 2 others



(a) Heart disease

(b) FICO Explainable Machine Learning Challenge

(c) MAGIC gamma telescope

(d) Musk molecules

Accuracy Maximization

CG competitive with RIPPER [Cohen 1995]

CG can find simpler rules that are no less accurate
(adult, bank, magic, FICO)

accuracy

dataset	CG	BRS	AM	BCD	RIPPER	CART	RF
adult	83.5	81.7	83.0	82.4	83.6	83.1	84.7
bank	90.0	87.4	90.0	89.7	89.9	89.1	88.7
gas	98.0	92.2	97.6	97.0	99.0	95.4	99.7
magic	85.3	82.5	80.7	80.3	84.5	82.8	86.6
mushroom	100.0	99.7	99.9	99.9	100.0	96.2	99.9
musk	95.6	93.3	96.9	92.1	95.9	90.1	86.2
FICO	71.7	71.2	71.2	70.9	71.8	70.9	73.1

complexity

adult	88.0	39.1	15.0	13.2	133.3	95.9
bank	9.9	13.2	6.8	2.1	56.4	3.0
gas	123.9	22.4	62.4	27.8	145.3	104.7
magic	93.0	97.2	11.5	9.0	177.3	125.5
mushroom	17.8	17.5	15.4	14.6	17.0	9.3
musk	123.9	33.9	101.3	24.4	143.4	17.0
FICO	13.3	23.2	8.7	4.8	88.1	155.0

Conclusion

Accurate and interpretable Boolean classification rules

Column generation to efficiently search space of rules without restrictions

Optimality guarantees on training set

Superior accuracy-simplicity trade-offs

Poster #79, Room 210, 10:45 – 12:45 today