Boolean Decision Rules via Column Generation  
Sanjeeb Dash, Oktay Günlük, Dennis Wei  
IBM Research, Thomas J. Watson Research Center, Yorktown Heights, NY, USA

**Problem Statement**  
Learn Boolean rules for binary classification:  
- Disjunctive normal form (DNF, OR of ANDs)  
- Conjunctive normal form (CNF, AND of ORs)

# accounts < 5 OR # accounts ≥ 7 AND Debt > $1000

Rules w/ few clauses and conditions are **interpretable**. Optimize accuracy vs. simplicity using integer programming (IP).

**Related models:**  
- DNF Boolean rule = Decision rule set  
- Decision lists, decision trees

**Master IP/LP**  
Minimize Hamming loss subject to complexity bound:  
- loss on positive instances  
- loss on negative instances  
- cover positives  
- complexity bound  
- whether to select clause

**Procedure and Optimality Guarantees**  
IPs/LPs solved using CPLEX. Time limit for training ≤ 5 min. Row/column sampling if Pricing IP is large.

**Main Idea: Column Generation**  
Binarize categorical and numerical features.  
Exponentially many possible clauses (conjunctions). Previous works limited search using heuristics (e.g., association rule mining).  
We implicitly consider all clauses without enumeration using **column generation (CG)**.

**Pricing IP**  
Reduced cost of clause incl. complexity penalty:

\[
\min \sum_{i,j:a} a_i - \sum_{j:y=1} \mu a_i + \lambda \left(1 + \sum_{j=1} z_j \right) \\
\] 

Clause acts as conjunction of features:

\[
\begin{align*}
& a_i + \sum_{j:x=1} z_j \geq 1, \quad a_i \geq 0, \quad i: y_i = 0 \\
& a_i + \sum_{j:x=1} z_j \leq 1, \quad i: y_i = 1, \quad j: x_j = 0 \\
& z_j \in \{0,1\}, \quad j = 1, \ldots, d \\
& \text{whether to select feature } j
\end{align*}
\]

Also use heuristic to find solutions.

**Accuracy-Complexity Trade-Off**  
CG dominates on 8 of 16 datasets, close on 2 others.

**Accuracy Maximization**  

---

**References**