Generalized Linear Rule Models

Dennis Wei
Sanjeeb Dash
Tian Gao
Oktay Günlük

IBM Research

Generalized Linear Models + Rules

Generalized Linear Models + Rules

- Linear regression
- Logistic regression
- Poisson regression
- …

- Conjunctions of conditions on individual features

Age > 50 AND Blood pressure > 140

- Can capture nonlinearities and interactions

Generalized Linear Models + Rules

- Linear regression
- Logistic regression
- Poisson regression
- …

- Conjunctions of conditions on individual features

Age > 50 AND Blood pressure > 140

- Can capture nonlinearities and interactions

Model interpretation

Problem Formulation

$$\min_{\beta} \quad \overbrace{\frac{1}{n}\sum_{i=1}^{n} \text{Loss}\left(\sum_{k\in\mathcal{K}} \underset{\substack{\uparrow \\ \text{conjunction values}}}{a_{ik}}\,\underset{\substack{\uparrow \\ \text{coefficients}}}{\beta_k}\,, y_i\right)}^{\text{performance}} + \overbrace{\sum_{k\in\mathcal{K}} \underset{\substack{\uparrow \\ \text{conjunction complexity}}}{\lambda_k}\,|\beta_k|}^{\text{complexity}}$$

Problem Formulation

$$\min_{\beta} \quad \overbrace{\frac{1}{n}\sum_{i=1}^{n} \mathrm{Loss}\left(\sum_{k\in\mathcal{K}} \underset{\substack{\uparrow\\ \text{conjunction values}}}{a_{ik}}\,\underset{\substack{\uparrow\\ \text{coefficients}}}{\beta_k}\,, y_i\right)}^{\text{performance}} + \overbrace{\sum_{k\in\mathcal{K}} \underset{\substack{\uparrow\\ \text{conjunction complexity}}}{\lambda_k}|\beta_k|}^{\text{complexity}}$$

Challenge: Set of conjunctions $\mathcal{K}$ is exponentially large

- e.g. age alone, age AND blood pressure, age AND blood pressure AND body mass index, ...

Problem Formulation

$$\min_{\beta} \quad \overbrace{\frac{1}{n}\sum_{i=1}^{n} \mathrm{Loss}\left(\sum_{k\in\mathcal{K}} a_{ik}\beta_k \,, y_i\right)}^{\text{performance}} + \overbrace{\sum_{k\in\mathcal{K}} \lambda_k |\beta_k|}^{\text{complexity}}$$

conjunction values coefficients conjunction complexity

Challenge: Set of conjunctions $\mathcal{K}$ is exponentially large

- e.g. age alone, age AND blood pressure, age AND blood pressure AND body mass index, …

We avoid limitations of existing methods

- Pre-select (large) subset e.g. [Friedman & Popescu, 2008]
- Boost but do not revise rules e.g. [Cohen & Singer, 1999; Dembczynski et al., 2010]

Column Generation

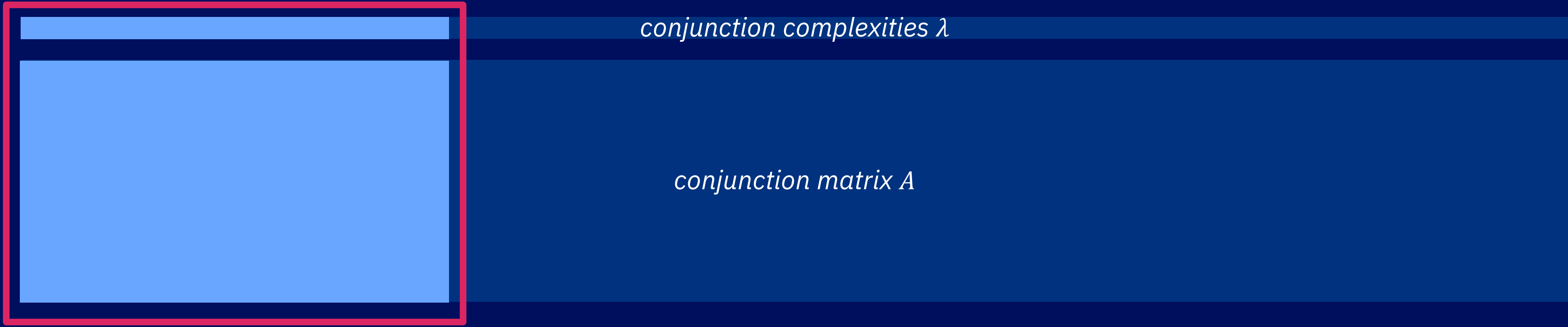
conjunction complexities λ

conjunction matrix A

Column Generation

$$\min_{\beta} \quad \frac{1}{n}\sum_{i=1}^{n} \mathrm{Loss}\left(\sum_{k\in\mathcal{S}} a_{ik}\beta_k\,, y_i\right) + \sum_{k\in\mathcal{S}} \lambda_k|\beta_k|$$

Solve over small subsets $\mathcal{S}$



conjunction complexities λ

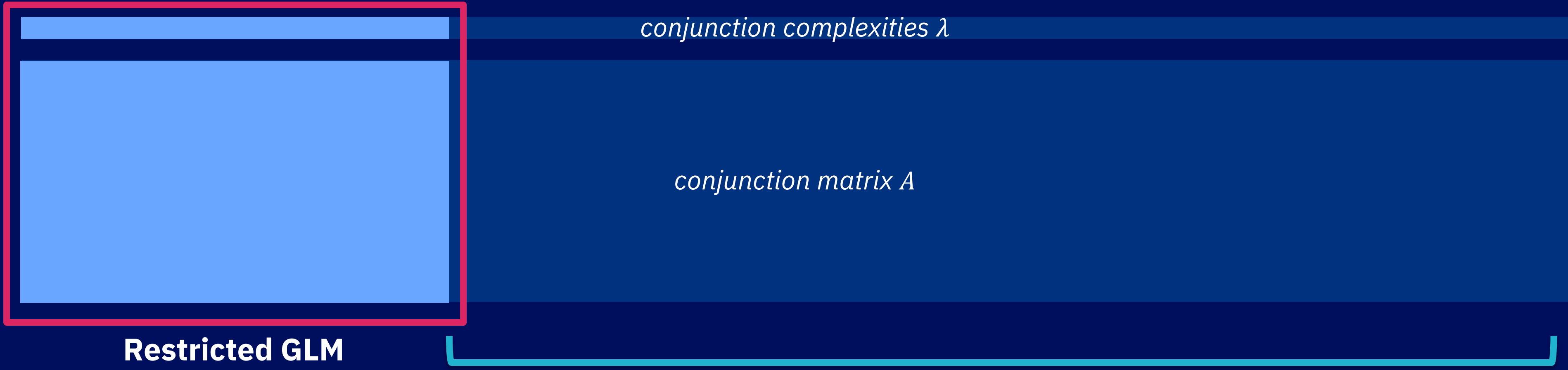
conjunction matrix A

Restricted GLM

Column Generation

$$\min_{\beta} \quad \frac{1}{n}\sum_{i=1}^{n} \text{Loss}\left(\sum_{k\in\mathcal{S}} a_{ik}\beta_k\,, y_i\right) \;+\; \sum_{k\in\mathcal{S}} \lambda_k|\beta_k|$$

Solve over small subsets $\mathcal{S}$



conjunction complexities λ

conjunction matrix A

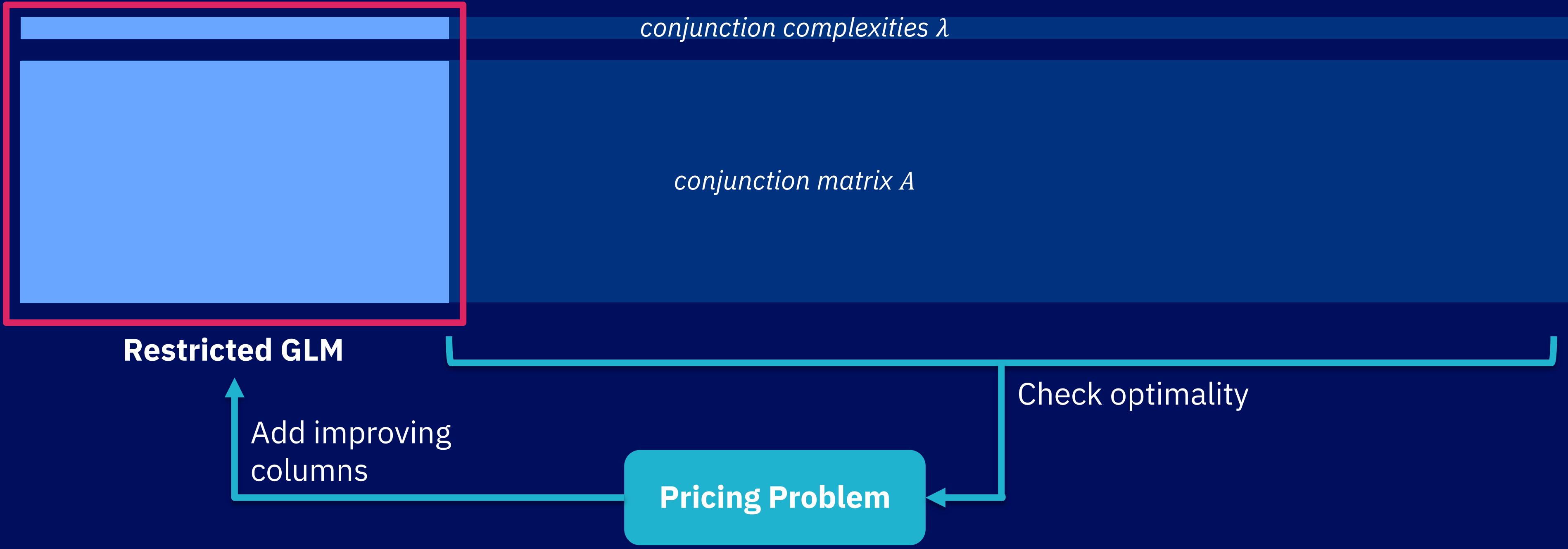
Restricted GLM

Check optimality

Column Generation

$$\min_{\beta} \quad \frac{1}{n}\sum_{i=1}^{n} \text{Loss}\left(\sum_{k\in\mathcal{S}} a_{ik}\beta_k\,, y_i\right) + \sum_{k\in\mathcal{S}} \lambda_k|\beta_k|$$

Solve over small subsets $\mathcal{S}$

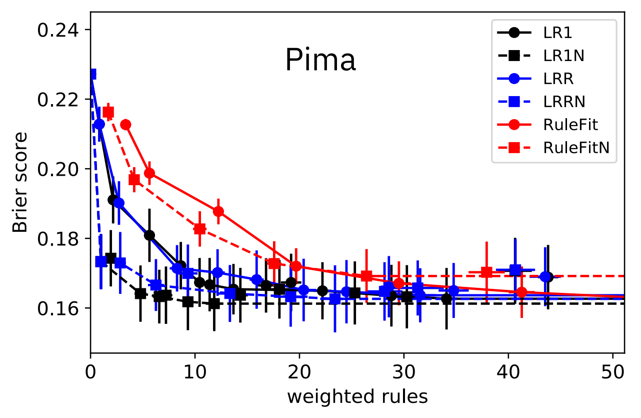


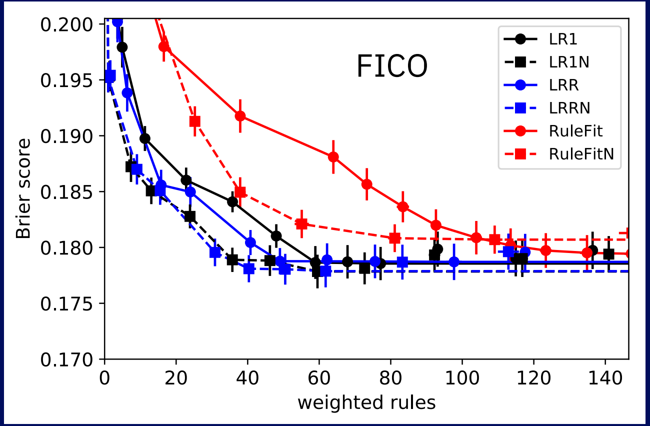
Integer program (solved using CPLEX) gives optimality guarantee
Heuristic also effective and more practical

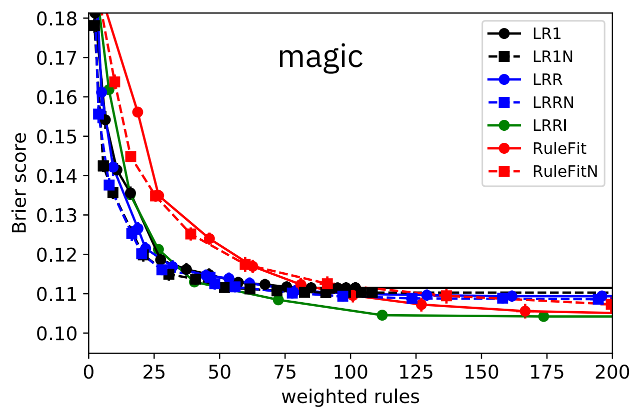
Performance-Complexity Trade-Offs

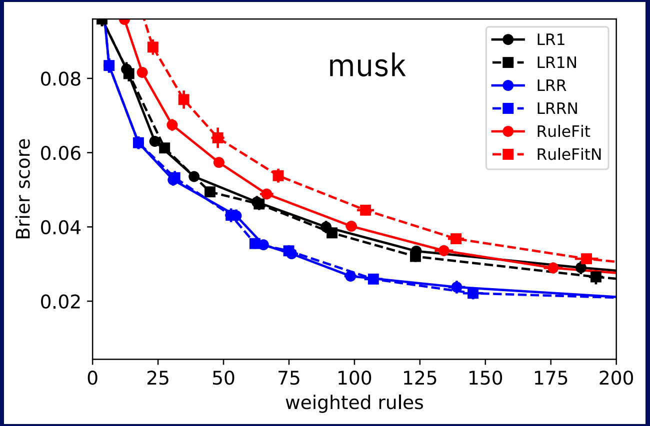
Logistic and linear regression experiments

Logistic Rule Regression with CG (LRR, LRRN) obtains better trade-offs than (RuleFit, RuleFitN) on most of 16 classification datasets









Performance Maximization

Logistic/Linear Rule Regression with CG (LRR, LRRN) is highly competitive when tuned to maximize performance

method	LRR	LRRN	RuleFit	RuleFitN	GBM	SVM
logistic regression mean rank	4.1	**3.6**	4.8	**3.6**	5.3	4.0
linear regression mean rank	4.9	**3.0**	4.5	3.5	3.4	5.0

and uses 2-4 times fewer rules than RuleFit [Friedman & Popescu, 2008]

GLRM: Generalized Linear Models + Rules

Flexible and interpretable models

Probabilistic classification and regression

Column generation to efficiently search space of rules without restrictions

Poster #264, today 6:30-9:00 PM