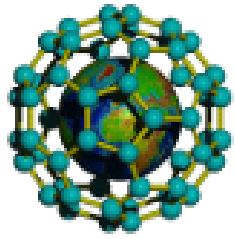


GOL

UNIVERSITÀ
DEGLI STUDI
FIRENZE

DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

An Effective Procedure for Feature Subset Selection in Logistic Regression Based on Information Criteria

AIRO ODS 2021
50th Annual Meeting of AIRO
Rome, 16th September 2021

E. Civitelli, M. Lapucci, F. Schoen, **A. Sortino**

Department of Information Engineering
Università degli Studi di Firenze

Logistic Regression Problem

Given

- $X \in \mathbb{R}^{N \times p}$ a dataset of N examples with p features
- $Y \in \{-1, 1\}^N$ a set of binary labels

The logistic regression model for binary classification defines the probability for an example x of belonging to the class y as

$$\mathbb{P}(y \mid x) = \frac{1}{1 + \exp(y w^\top x)} \tag{1}$$

The model (1) is associated with the following log-likelihood function

$$\ell(w) = -\sum_{i=1}^{N} \log\left(1 + \exp\left(-y_i w^\top x_i\right)\right)$$

Best Subset Selection in Logistic Regression

$$\underset{w \in \mathbb{R}^p}{\arg\min} -2\ell(w) + \lambda \|w\|_0$$

Best Subset Selection in Logistic Regression

$$\underset{w \in \mathbb{R}^p}{\arg\min} -2\ell(w) + \lambda \|w\|_0$$

How to choose the trade-off parameter λ?

Best Subset Selection in Logistic Regression

$$\underset{w \in \mathbb{R}^p}{\arg\min} -2\ell(w) + \lambda \|w\|_0$$

How to choose the trade-off parameter λ?

Goodness-Of-Fit Measures

$$\text{AIC} \Longrightarrow \lambda = 2$$

$$\text{BIC} \Longrightarrow \lambda = \log(N)$$

Cardinality-Penalized Optimization Problem

$$\arg\min_{w \in \mathbb{R}^p} \mathcal{F}(w) = \mathcal{L}(w) + \lambda \|w\|_0$$

where

- $\mathcal{L} : \mathbb{R}^p \to \mathbb{R}$ is convex and continuously differentiable ($\mathcal{L}(w) = -2\ell(w)$)
- $\lambda > 0$ depends on the choice of the GOF measure

Basic Definitions and Properties

- Support Set: $S(w) = \{i \mid w_i \neq 0, i = 1, \ldots, p\}$
- $\bar{S}(w) = \{1, \ldots, n\} \setminus S(w)$
- w^* satisfies Lu-Zhang (LZ) conditions if there exists a support set S for w^* such that $\nabla_S \mathcal{L}(w^*) = 0$
  - w^* local minimizer $\Longleftrightarrow$ w^* satisfies LZ conditions
- w^* is a CW-minimum if $w_i^* \in \arg\min_{w_i} \mathcal{F}(w_i; w_{\neq i}^*)$

Existing Approaches

- Enumerate exhaustively variable subsets (all possible combinations of non-zero variables) and solve each continuous problem
- Employ step-wise heuristics
- Lasso (ℓ_1-norm instead of ℓ_0-seminorm)
- Concave Approximation of ℓ_0-seminorm (Rinaldi et al., 2010)
- Penalty-Decomposition approach (Lu and Zhang, 2013)
- Outer Approximation Method (Bertsimas et al., 2017; Kamiya et. al., 2019)

The MILO Approach (Sato et al., 2016)

- Two key ideas:
  1. Replacement of the ℓ_0 term by the sum of binary indicator variables
  2. Approximation of nonlinearity in $\mathcal{L}$ by a piecewise linear function defined by the pointwise maximum of a family of tangent lines

The MILO Approach (Sato et al., 2016)

- Two key ideas:
  1. Replacement of the ℓ_0 term by the sum of binary indicator variables
  2. Approximation of nonlinearity in $\mathcal{L}$ by a piecewise linear function defined by the pointwise maximum of a family of tangent lines

MILO reformulation

$$
\begin{aligned}
\underset{w,z,t}{\arg\min} \quad & 2\sum_{i=1}^{N} t_i + \lambda \sum_{i=1}^{p} z_i & & \\
\text{s.t.} \quad & -Mz_i \leq w_i \leq Mz_i & & \forall i = 1, \ldots, p \\
& z \in \{0,1\}^p & & \\
& t_i \geq f'(v^k)(y_i w^\top x_i - v^k) + f(v^k) & & \forall k = 1, \ldots, K \\
& & & \forall i = 1, \ldots, N
\end{aligned}
$$

where $f(v) = \log(1 + \exp(-v))$ and $M \gg 0$.

The MILO Approach (Sato et al., 2016)

- Two key ideas:
  1. Replacement of the ℓ_0 term by the sum of binary indicator variables
  2. Approximation of nonlinearity in $\mathcal{L}$ by a piecewise linear function defined by the pointwise maximum of a family of tangent lines

MILO reformulation

$$
\begin{aligned}
\underset{w,z,t}{\arg\min} \quad & 2\sum_{i=1}^{N} t_i + \lambda \sum_{i=1}^{p} z_i \\
\text{s.t.} \quad & -Mz_i \leq w_i \leq Mz_i && \forall i = 1, \ldots, p \\
& z \in \{0,1\}^p \\
& t_i \geq f'(v^k)(y_i w^\top x_i - v^k) + f(v^k) && \forall k = 1, \ldots, K \\
& && \forall i = 1, \ldots, N
\end{aligned}
$$

where $f(v) = \log(1 + \exp(-v))$ and $M \gg 0$.

Drawback

Scales pretty badly as N or p grows.

Our Approach

Main Idea

Apply the *Block Coordinate Descent* (BCD) decomposition strategy at each iteration of MILO approach

- Typical approach when p is large
- The information about all the gradient is not required
- Only a block of b variables, called *working set*, is considered

MILO-BCD Approach

MILO-BCD reformulation

$$
\begin{aligned}
\underset{w_{B_\ell}, z, t}{\arg\min} \quad & 2\sum_{i=1}^{N} t_i + \lambda \sum_{i \in B_\ell} z_i & & \\
\text{s.t.} \quad & -M z_i \leq w_i \leq M z_i & & \forall i \in B_\ell \\
& z_i \in \{0,1\} & & \forall i \in B_\ell \\
& t_i \geq f'(v^k)(y_i w^\top x_i - v^k) + f(v^k) & & \forall k = 1, \ldots, K \\
& & & \forall i = 1, \ldots, N \\
& w_{\bar{B}_\ell} = w^{\ell}_{\bar{B}_\ell} & &
\end{aligned}
$$

where ℓ is the current iteration, $B_\ell \subset \{1, \ldots, p\}$ is the *working set* and $\bar{B}_\ell = \{1, \ldots, p\} \setminus B_\ell$.

The Working Set Selection Rule

- Let w^ℓ be the current iterate; the score function is the following

$$s(w^\ell, i) = \begin{cases} \mathcal{L}(0, w^\ell_{\neq i}) - \lambda + \lambda \|w^\ell\|_0 & \text{if } w^\ell_i \neq 0, \\ \min_{w_i} \mathcal{L}(w_i, w^\ell_{\neq i}) + \lambda + \lambda \|w^\ell\|_0 & \text{if } w^\ell_i = 0. \end{cases}$$

- How the objective function value changes if the variable w_i entering/leaving the support

- Let $b = |B_\ell|$; the working set B_ℓ is chosen by selecting the b lowest scoring variables

The Complete Procedure

Algorithm 1: MILO-BCD

Input: $w^0 = 0$, $b < p$.

for $\ell = 0, 1, \ldots$ **do**

  Choose the working set B^ℓ

  Compute $\nu_{B^\ell}^{\ell+1}$ by solving the MILO-BCD reformulation

  Set $\nu_{\bar{B}^\ell}^{\ell+1} = w_{\bar{B}^\ell}^{\ell}$

  if $\mathcal{F}(\nu^{\ell+1}) \geq \mathcal{F}(w^\ell)$ **then**

    Set

$$\nu^{\ell+1} \in \arg\min_{w} \mathcal{F}(w) \quad \text{s.t. } \|w^\ell - w\|_0 \leq 1, \; w_{\bar{B}^\ell} = w_{\bar{B}^\ell}^{\ell}$$

  Set

$$w^{\ell+1} \in \arg\min_{w} \mathcal{L}(w) \quad \text{s.t. } w_i = 0 \text{ for all } i \in \bar{S}(\nu^{\ell+1})$$

  if $\mathcal{F}(w^{\ell+1}) = \mathcal{F}(w^\ell)$ **then**

    return w^ℓ

Theoretical Analysis

Lemma

Let $\Gamma = \{\mathcal{F}(w) \mid w \text{ is a local minimum}\}$. Then $|\Gamma| \leq 2^p$.

Lemma (Relationship between $\mathcal{F}(w)$ and $s(w, i)$)

Let s be the score function and let $\bar{w} \in \mathbb{R}^p$. Moreover, $\forall h = 1, \ldots, p$ let $\bar{w}^h \in \arg\min_{w_h} \mathcal{F}(w_h; \bar{w}_{\neq h})$. Then:

1. if $\mathcal{F}(\bar{w}^h) = \mathcal{F}(\bar{w})$ then $s(\bar{w}, h) \geq \mathcal{F}(\bar{w})$
2. if $\mathcal{F}(\bar{w}^h) < \mathcal{F}(\bar{w})$ and $\bar{w}$ satisfies LZ conditions, then $s(\bar{w}, h) = \mathcal{F}(\bar{w}^h)$

Proposition (Finite Termination and Optimality Properties)

Let $\{w^\ell\}$ be the sequence generated by MILO-BCD algorithm. Then $\{w^\ell\}$ is a finite sequence and the last element $\bar{w}$ is a CW-minimum.

Escaping from Bad Local Minima

- In order to find good CW-optima we introduce an heuristic, modifying the score function s in this way

$$\hat{s}(w^{\ell}, i) = s(w^{\ell}, i) + 2^{r_i} - 1$$

  where r_i is the number of times the i-th variable was in the working set in the previous attempts

- The variables that were tried more times and could not provide improvements are (exponentially) penalized

- This modification does not alter the theoretical properties

Computational Experiments – Setup

- 11 standard datasets from UCI Repository for binary classification
- AIC and BIC as GOF measures to be minimized
- a total of 22 problems
- time limit of 10 000 seconds
- working set size $b = 20$

Dataset	***N***	***p***
Parkinsons	195	22
Heart (Statlog)	270	25
Breast Cancer Wisconsin (Prognostic)	194	33
QSAR Biodegradation	1055	41
SPECTF Heart	267	44
Spambase	4601	57
Optical Recognition of Handwritten Digits	3823	62
Libras Movement	360	90
a2a	2265	123
w2a	2470	300
Madelon	2000	500

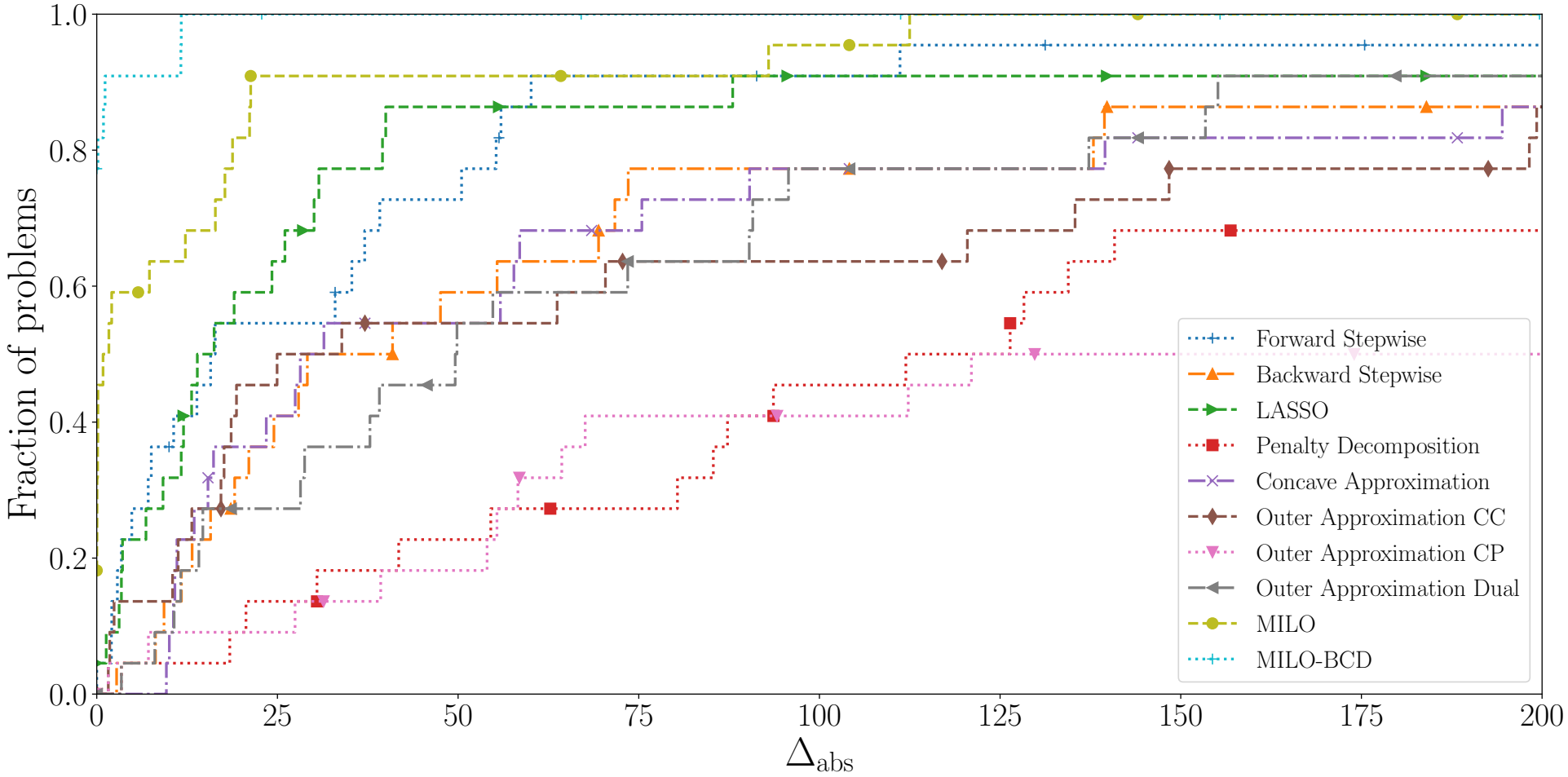
Computational Experiments – Results

Dataset	Method	AIC	ℓ_0	Time (s)
	Forward Stepwise	178.9840	6	0.797
	Backward Stepwise	180.0595	28	0.214
	LASSO	181.4678	13	8.966
	Penalty Decomposition	222.8672	2	55.287
	Concave Approximation	181.8271	17	3.788
`spectf`	Outer Approximation CC	178.8349	12	$\geq$ 10000
	Outer Approximation CP	222.3555	5	$\geq$ 10000
	Outer Approximation Dual	206.1484	38	10.766
	MILO	168.5162	15	1293.650
	MILO-BCD	**168.3443**	15	205.6255

Computational Experiments – Results

Dataset	Method	BIC	ℓ_0	Time (s)
a2a	Forward Stepwise	1741.3958	15	10.727
	Backward Stepwise	2016.2528	64	5.798
	LASSO	1764.5871	15	397.503
	Penalty Decomposition	1860.2444	5	607.869
	Concave Approximation	1873.3706	44	21.709
	Outer Approximation CC	2028.2982	11	$\geq$ 10000
	Outer Approximation CP	2268.7472	4	$\geq$ 10000
	Outer Approximation Dual	1829.5696	14	$\geq$ 10000
	MILO	1754.9999	16	$\geq$ 10000
	MILO-BCD	**1733.8513**	17	2933.3452

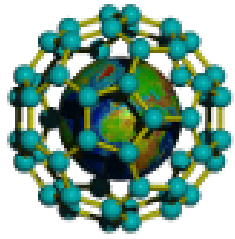
Computational Experiments – Results



Cumulative distribution of absolute distance from the optimum

Conclusions

- Simple BCD decomposition method for GOF-based sparse logistic regression, using MILP solvers
- The features and the behavior of the proposed method have been theoretically characterized
- Numerical experiments shows the effectiveness of our approach, compared to the state-of-the-art methods



GOL

UNIVERSITÀ
DEGLI STUDI
FIRENZE

DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

An Effective Procedure for Feature Subset Selection in Logistic Regression Based on Information Criteria

AIRO ODS 2021
50th Annual Meeting of AIRO
Rome, 16th September 2021

E. Civitelli, M. Lapucci, F. Schoen, **A. Sortino**

Department of Information Engineering
Università degli Studi di Firenze