

Multi-Task Feature Selection on Multiple Networks via Maximum Flows

Mahito Sugiyama^{1 (,2)}, Chloé-Agathe Azencott³, Dominik Grimm^{2,4}, Yoshinobu Kawahara¹, Karsten Borgwardt^{2,4}

¹Osaka University, ²Max Planck Institutes Tübingen, ³Mines ParisTech, Institut Curie, INSERM, ⁴Eberhard Karls Universität Tübingen

Goal

- Given multiple networks
- Find features (vertices), which are associated with the target response and tend to be connected each other



Main Result

New formulation of multi-task feature selection

$$\begin{aligned} \underset{K \text{ tasks}}{\operatorname{argmax}} & \sum_{i=1}^{K} \left(\underbrace{f_i(S_i)}_{\text{association}} - g_i(S_i) \right) - \sum_{i < j} h(S_i, S_j), \\ \underbrace{f_i(S_i)}_{\text{penalty}} & = \sum_{v \in S_i} q_i(v), \quad g_i(S_i) := \lambda \sum_{\substack{e \in B_i \\ \text{connectivity}}} w_i(e) + \underbrace{\eta |S_i|}_{\text{sparsity}}, \\ h(S_i, S_j) & := \mu |S_i \vartriangle S_j| = \mu |(S \cup S') \setminus (S \cap S')| \end{aligned}$$

- It is efficiently solved by max-flow algorithms
- Its performance is superior to Lasso-based methods

Motivation: Data Mining on Networks

- Networks (graphs) are everywhere
 - Biological pathways (KEGG), chemical compounds (PubChem), social networks, ...
- Which part of the network is responsible for performing a particular function?
 - → Feature selection on networks
 - Features = vertices (nodes)
 - Network topology = *a priori* knowledge about relationships between features
- Multi-task feature selection should be considered for more effectiveness

Existing Approach: Lasso

- Lasso-based regression with:
 - ℓ_1 -regularizer
 - Structured (network) regularizers
- Drawbacks
 - Prediction loss is optimized, different from finding features that are relevant for (associated with) a property of interest
- Drawbacks in multi-task setting
 - Exactly the same features are always selected among different tasks
 - Only one network structure can be employed
 - No method can use multiple networks simultaneously

SCONES (Selecting Connected Explanatory SNPs)

- Single task feature selection on a network [Azencott *et al.* ISMB2013]
- Given a weighted graph G = (V, E)
 - Each $v \in V$ has a relevance score q(v)
 - If you have a design matrix $\mathbf{X} \in \mathbb{R}^{N \times |V|}$ and a response vector $\mathbf{y} \in \mathbb{R}^N$, q(v) is the association of \mathbf{y} and each feature of \mathbf{X}
- Objective: Find a subset $S \subset V$ which maximizes

$$f(S) := \sum_{v \in S} q(v)$$
 (additive score), while

– *S* is small

Vertices in S are connected each other

Formulation of SConES

• $\operatorname{argmax}_{S \subset V} f(S) - g(S)$

$$f(S) := \sum_{v \in S} q(v), \quad g(S) := \underbrace{\lambda \sum_{e \in B} w(e)}_{\text{connectivity}} + \underbrace{\eta |S|}_{\text{sparsity}}$$

- $B = \{ \{v, u\} \in E \mid v \in V \setminus S, u \in S \}$ (boundary)

 $- w : E \rightarrow \mathbb{R}^+$ is a weighting function



Solution of SConES via Maximum Flow

- The *s*/*t*-network $M(G) = (V \cup \{s, t\}, E \cup S \cup T)$ with $S = \{\{s, v\} \mid v \in V, q(v) > \eta\}, T = \{\{t, v\} \mid v \in V, q(v) < \eta\}$ and set the capacity $c : E' \to \mathbb{R}^+$ to $c(\{v, u\}) = \begin{cases} |q(u) - \eta| & \text{if } u \in \{s, t\} \text{ and } v \in V, \\ \lambda w(\{v, u\}) & \text{otherwise} \end{cases}$
- The minimum s/t cut of M(G) = the solution of SConES



Key Contribution: Multi-SConES

- Given K networks $\mathcal{G} = \{G_1, G_2, \dots, G_K\}$
 - They share vertices and have different edges
- Multi-task version of SConES:



Solution of Multi-SConES

 Multi-SConES is solved by the max-flow algorithm on the unified single network



Solution of Multi-SConES

 Multi-SConES is solved by the max-flow algorithm on the unified single network



From Multi-Task to Single-Task

- The unified network $U(\mathcal{G}) = (\tilde{V}, \tilde{E})$ from K networks: $\tilde{V} := \bigcup_{i=1}^{K} V'_i, \ \tilde{E} := \bigcup_{i=1}^{K} E'_i \cup \bigcup_{m=1}^{n} A_m$, where $A_m := \left\{ \{v_i^m, v_j^m\} \mid i, j \in \{1, \dots, K\}, i \neq j \right\}.$
 - The weight \tilde{w} of edges is given as $\tilde{w}(e) = w_i(e)$ if $e \in E'_i$ and $\tilde{w}(e) = \mu/\lambda$ otherwise
 - U(G) has $|\tilde{V}| = Kn$ vertices, $|\tilde{E}| = \sum_{i=1}^{K} |E_i| + nK(K-1)/2$ edges
- Our multi-task problem is exactly equivalent to the single-task over $U(\mathcal{G})$

Empirical Comparison

- Correlation ranking (baseline)
- Single-task:
 - Lasso, Elastic net
 - Group Lasso (with groups formed by edges)
 - Grace, aGrace (Lasso-based state-of-the-art) argmin $\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \underbrace{\lambda_1 \|\beta_1\|}_{\text{sparsity}} + \underbrace{\lambda_2 \beta^T \mathbf{L}\beta}_{\text{connectivity}}$
 - Accelerated by replacing SVD to the incidence matrix
- Multi-task:
 - Multi-task Lasso, Multi-task Grace
- Performance is measured by MCC (Matthews correlation coefficient) and MSE

Synthetic Data

- Gene regulatory networks are simulated
 - 2,200 features (vertices)
 - 200 transcription factors (TFs) and 2,000 genes
 - Each TF is connected to 10 regulatory target genes
 - First 44 features (4 TFs and 40 genes) are causal to the response
 - They are correlated with the response
 - 4 models
 - Models 1, 3 (2, 4) are positively (negatively) correlated
 - Correlation in models 3, 4 is weaker than 1, 2
 - The same data was used in [Li and Li; 2008]

Running Time



Parameter Sensitivity (1/2)



Parameter Sensitivity (2/2)



Performance for Synthetic Data (model 1)



CR: ranking of correlations (baseline), LA: Lasso, EN: the elastic net, GL: group Lasso, GR: Grace, AG: aGrace, and SC: SConES

Performance for Two Tasks (model 1)



Multi-Locus Association Mapping

- <u>Goal</u>: Find SNPs (features) that are associated with phenotype, using a network over SNPs
- Arabidopsis thaliana GWAS data
 - 216,130 SNPs (features)
 - 6 flowering time phenotypes (1 phetnotype = 1 task)
 - Protein-protein interaction network from TAIR
 - SNPs are connected if they belong to the connected genes
- 282 candidate genes are gold standard of causal features



Results of Association Mapping (MCC)

Phenotype		МСС	
	Lasso	Grace	SConES
2W	0.001	-0.001	0.014
2W + 4W	-0.001	-0.003	0.016
2W + FT GH	0.001	0.000	0.024
2W + 4W + FT GH	0.005	0.002	0.027
LDV	0.001	0.000	0.016
LDV + 0W	0.005	0.007	0.020
LDV + FT10	0.001	0.001	0.021
LDV + 0W + FT10	0.003	0.002	0.023

Results of Association Mapping (SNPs)

Phenotype	Hit ratio of SNPs (prec.)		
	Lasso	Grace	SConES
2W	7/126	4/98	42/338
2W + 4W	7/175	6/198	81/802
2W + FT GH	9/173	7/146	106/818
2W + 4W + FT GH	15/183	16/265	101/679
LDV	6/116	7/144	73/667
LDV + 0W	16/196	19/206	86/702
LDV + FT10	12/214	10/191	92/762
LDV + 0W + FT10	18/283	19/323	81/482

Conclusion

- A new formulation, Multi-SConES, for multi-task feature selection with multiple network regularizers
 - Direct optimization of feature relevance scores
 - Exact solution via max-flow algorithms
- It can select different features for different tasks
- It can use different networks for different tasks
- Future work
 - Incorporating more complex task relationships
 - Currently, a single parameter μ

Appendix



Difficulty of Optimization

- Given a positive integer k
- The maximum-weight connected graph (MCG) problem: argmax f(S) s.t. $G|_S$ is connected and |S| = k $S \subset V$
 - is known to be strongly NP-complete

Regularization Path

- η has anti-monotonicity with respect to the number of selected features
 - $S(\eta) \subset S(\eta')$ if and only if $\eta > \eta'$
- The entire regularization path along with the changes in η can be obtained
 - Time complexity does not increase by using parametric maximum flow algorithm [Gallo *et al*. 1989]

Spectral Analysis

- $\mathbf{f} \in \{0, 1\}^{|V|}$: the indicator vector of a subset $S \subset V$
- $\mathbf{c} \in \mathbb{R}^{|V|}$: the vector composed of values q(v)
- Single-task SConES: argmax $\mathbf{c}^{\mathrm{T}}\mathbf{f} - \lambda \mathbf{f}^{\mathrm{T}}\mathbf{L}\mathbf{f} - \eta \|\mathbf{f}\|_{0}$ $\mathbf{f} \in \{0,1\}^{|V|}$
- Multi-task SConES: argmax $\sum_{i=1}^{K} (\mathbf{c}_{i}^{\mathrm{T}} \mathbf{f}_{i} - \lambda \mathbf{f}_{i}^{\mathrm{T}} \mathbf{L}_{i} \mathbf{f}_{i} - \eta \|\mathbf{f}_{i}\|_{0}) - \sum_{i < j} \mu \|\mathbf{f}_{i} - \mathbf{f}_{j}\|_{2}^{2}$, $\mathbf{f}_{1},...,\mathbf{f}_{K}$
- On the unified network, multi-task SConES is written as: argmax $\tilde{\mathbf{c}}^{\mathrm{T}}\tilde{\mathbf{f}} - \lambda \tilde{\mathbf{f}}^{\mathrm{T}}\tilde{\mathbf{L}}\tilde{\mathbf{f}} - \eta \|\tilde{\mathbf{f}}\|_{0}$. $\tilde{\mathbf{f}} \in \{0,1\}^{K|V|}$

Performance for Synthetic Data (model 2)



CR: ranking of correlations (baseline), LA: Lasso, EN: the elastic net, GL: group Lasso, GR: Grace, AG: aGrace, and SC: SConES

A-5/A-11

Performance for Synthetic Data (model 3)



CR: ranking of correlations (baseline), LA: Lasso, EN: the elastic net, GL: group Lasso, GR: Grace, AG: aGrace, and SC: SConES

A-6/A-11

Performance for Synthetic Data (model 4)



CR: ranking of correlations (baseline), LA: Lasso, EN: the elastic net, GL: group Lasso, GR: Grace, AG: aGrace, and SC: SConES

A-7/A-11

Performance for Two Tasks (model 2)



A-8/A-11

Performance for Two Tasks (model 3)



A-9/A-11

Performance for Two Tasks (model 4)



A-10/A-11

Results of Association Mapping (genes)

Phenotype	Hit ratio of genes (prec.) Lasso Grace SConES		
2W	2/112	1/91	7/124
2W + 4W	2/163	2/191	11/240
2W + FT GH	9/162	7/135	13/250
2W + 4W + FT GH	6/174	3/256	13/208
LDV	2/107	2/131	9/202
LDV + 0W	2/183	2/187	10/209
LDV + FT10	1/199	1/181	10/221
LDV + 0W + FT10	2/265	1/307	10/153

A-11/A-11