Safe Pattern Pruning: An Efficient Approach for Predictive Pattern Mining (KDD2016)

Ichiro Takeuchi

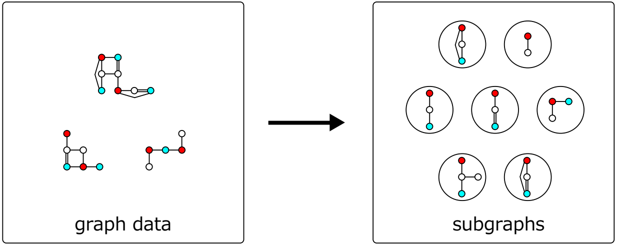
Nagoya Institute of Technology, Japan

Joint work with

Kazuya Nakagawa, Shinya Suzumura, Masayuki Karasuyama, Koji Tsuda

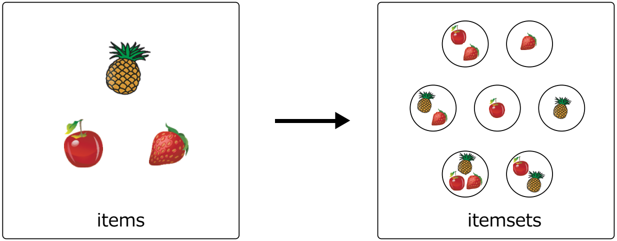
Graph mining

- The goal of frequent subgraph mining is to find a set of subgraphs that frequently appear in a database.
- Some computational tricks for handling exponentially large number of subgraphs are needed for graph mining tasks.
- Many efficient algorithms that exploit anti-monotonicity properties of subgraph frequencies exist in the literature.



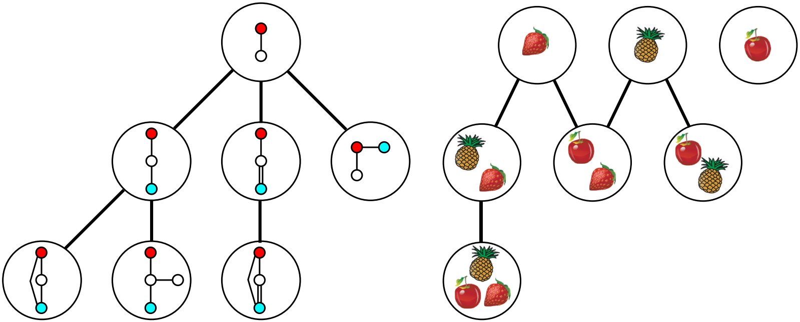
Itemset mining

- The goal of frequent itemset mining is to find a set of items that frequently appear in a database.
- Some computational tricks for handling exponentially large number of itemsets are needed for itemset mining tasks.
- Many efficient algorithms that exploit anti-monotonicity properties of itemset frequencies exist in the literature.



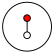
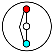
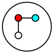
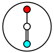
Anti-monotonicity property among patterns

- The frequency of a pattern in an ancestor node is greater than or equal to the frequency of a pattern in its any descendant node.



Predictive pattern mining

- The goal of predictive pattern mining is to learn a classification or regression model f as a function of the existences of patterns.
- A graph classification/regression model looks like

$f = w_1$  $+ w_2$  $+ w_3$  $+ w_4$  $+ \cdots$

- An itemset classification/regression model looks like

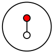
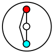
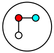
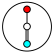
$f = w_1$  $+ w_2$  $+ w_3$  $+ w_4$  $+ \cdots$

- A classification/regression model f is learned over exponentially large number of patterns: some computational tricks are needed.

Can we use existing feature selection algorithms in statistics and ML?

- Feature selection for classification/regression models have been intensively studied in statistics and machine learning.
  - marginal screening
    (e.g.) correlations, χ^2 statistics
  - stepwise selection
    (e.g.) forward selection / backward elimination
  - **sparse modeling**
    (e.g.) L_1-**norm regularization**
- Existing feature selection algorithms in statistics and ML cannot be directly applied to predictive pattern mining problems because there are exponentially large number of features (patterns).

- We study **predictive pattern mining** problems (graph mining and itemset mining) for regression and classification tasks.
- We select a subset of predictive patterns by using **sparse modeling** (L_1 regularization).
- To handle exponentially large number of patterns, we propose a novel algorithm called **safe pruning rule**.
- The safe pruning rule is an extension of **safe screening rules** which have been recently actively studied in ML literature.

$f = w_1$  $+ w_2$  $+ w_3$  $+ w_4$  $+ \cdots$

$f = w_1$  $+ w_2$  $+ w_3$  $+ w_4$  $+ \cdots$

Problem Setup and Basic Idea

Problem setup (dataset)

- Training dataset

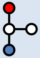
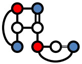
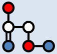
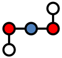
$$\{(G_i, y_i)\}_{i=1}^n$$

- G_i is a graph or a set of items
- $y_i \in \mathbb{R}$ for regression, $y_i \in \{\pm 1\}$ for binary classification
- G_i is represented as a binary vector $\boldsymbol{x}_i \in \mathbb{R}^D$ whose t^{th}-element is defined as

$$x_{it} := I(t \subseteq G_i) \text{ for each pattern } t \text{ in the database,}$$

where D is the number of all possible patterns (too huge to handle naively).

Dataset looks like

y_i	G_i				$\cdots$
-1		1	1	1	$\cdots$
+1		1	0	1	$\cdots$
-1		1	0	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Problem setup (sparse modeling)

- Huge-dimensional linear model for regression or classification:

$$f(\boldsymbol{x}_i) = w_1 x_{i1} + w_2 x_{i2} + \ldots + w_D x_{iD}$$

- Sparse modeling: introduce a mechanism to make coefficients sparse via L_1-regularization (LASSO)

$$\boldsymbol{w}^* := \arg\min_{\boldsymbol{w} \in \mathbb{R}^D} P(\boldsymbol{w}) := \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{w})^2 + \lambda \sum_{j=1}^{D} |w_j|$$

- Active patterns are defined as

$$\mathcal{A}^* := \{j \in \{1, \ldots, D\} \mid w_j^* \neq 0\}$$

Basic idea

- The difficulty lies in the fact that existing LASSO solvers cannot be used because D is too huge.
- Our idea is to find a superset of the active set:

$$\mathcal{A} \supseteq \mathcal{A}^*,$$

and solve the optimization problem over $\mathcal{A}$ (assuming $\mathcal{A}$ is small enough).

- It can be guaranteed that the optimal solution of the problem defined on $\mathcal{A} \supseteq \mathcal{A}^*$ is also optimal for the original problem, i.e.,

$$\boldsymbol{w}^* = P \begin{bmatrix} \boldsymbol{w}^*_{\mathcal{A}} \\ \mathbf{0} \end{bmatrix} \text{ where } P \text{ is the permutation matrix,}$$

and

$$\boldsymbol{w}^*_{\mathcal{A}} := \arg\min_{\boldsymbol{w}_{\mathcal{A}}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_{i\mathcal{A}}^\top \boldsymbol{w}_{\mathcal{A}})^2 + \lambda \sum_{j \in \mathcal{A}} |w_j|$$

Safe screening

is a method to find a superset $\mathcal{A} \supseteq \mathcal{A}^*$ in ordinal LASSO problem (but cannot be applied to predictive pattern mining problems without tricks).

Safe screening for LASSO (1)

- Primal Problem:

$$\boldsymbol{w}^* := \arg\min_{\boldsymbol{w}\in\mathbb{R}^D} P(\boldsymbol{w}) := \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{w})^2 + \lambda \sum_{j=1}^{D} |w_j|$$

- Dual Problem:

$$\boldsymbol{\theta}^* := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n} \frac{1}{2}\left(\theta_i - \frac{1}{\lambda}y_i\right)^2 \texttt{ s.t. } \left|\sum_{i=1}^{n} x_{ij}\theta_i\right| \leq 1, \forall j$$

- Sparsity Condition:

$$\underbrace{\left|\sum_{i=1}^{n} x_{ij}\theta_i^*\right|}_{\text{score}} < 1 \quad \Rightarrow \quad w_j^* = 0,$$

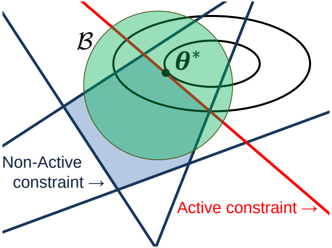
Safe screening for LASSO (2)

- Using convex optimization theory, we can find a ball $\mathcal{B}$ in the dual solution space in which the dual optimal solution $\boldsymbol{\theta}^*$ exists:

$$\mathcal{B} := \{\boldsymbol{\theta} \mid \|\boldsymbol{\theta}^* - \boldsymbol{c}\| \leq r\} \text{ where } \boldsymbol{c} := \hat{\boldsymbol{\theta}}, r := 2\lambda^{-1}\sqrt{P(\hat{\boldsymbol{w}}) - D(\hat{\boldsymbol{\theta}})}$$

- Safe screening exploits the fact that $\boldsymbol{\theta}^* \in \mathcal{B}$ in order to identify sparse features:

$$\underbrace{\max_{\boldsymbol{\theta}\in\mathcal{B}} \left|\sum_{i=1}^{n} x_{ij}\theta_i\right| < 1}_{\text{UB of the score}<1} \;\Rightarrow\; \underbrace{\left|\sum_{i=1}^{n} x_{ij}\theta_i^*\right| < 1}_{\text{the score}<1} \;\Rightarrow\; \underbrace{w_j^* = 0}_{\text{sparse}}$$



Safe pruning

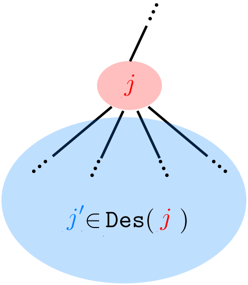
is an extension of safe screening for handling exponentially large number of patterns in predictive pattern mining problems.

Safe pruning rule

- We cannot compute safe feature screening bounds for each of these exponentially large number of patterns.
- We develop a **safe pruning rule** $\texttt{spr}(j)$ for each node j in the tree such that

$$\texttt{spr}(j) \text{ is true } \Rightarrow \left|\sum_{i=1}^{n} x_{ij'}\theta_i^*\right| < 1 \Rightarrow w_{j'}^* = 0 \text{ for all } j' \in \texttt{Des}(j),$$

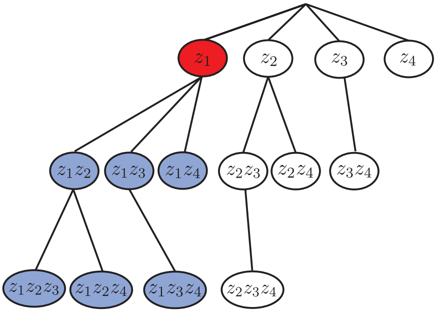
where $\texttt{Des}(j)$ is the set of descendant nodes of the node j.



Safe pruning rule (how it works)

- Finding a superset $\mathcal{A}$ of the active set $\mathcal{A}^* := \{j \mid w_j^* \neq 0\}$:

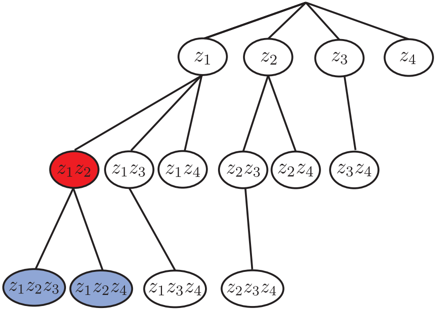
$$\texttt{spr}(z_1) = \text{false}, \quad \mathcal{A} = \{z_1\}$$



Safe pruning rule (how it works)

- Finding a superset $\mathcal{A}$ of the active set $\mathcal{A}^* := \{j \mid w_j^* \neq 0\}$:

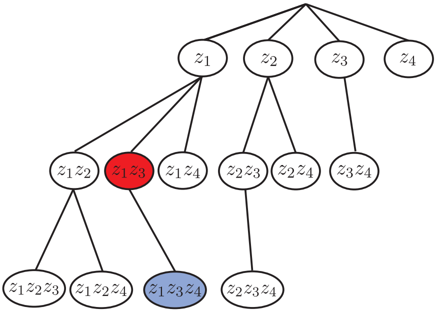
$$\mathtt{spr}(z_1 z_2) = \text{true}, \quad \mathcal{A} = \{z_1\}$$



Safe pruning rule (how it works)

- Finding a superset $\mathcal{A}$ of the active set $\mathcal{A}^* := \{j \mid w_j^* \neq 0\}$:

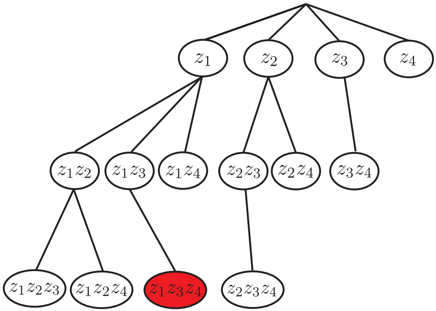
$$\texttt{spr}(z_1 z_3) = \text{false}, \quad \mathcal{A} = \{z_1, z_1 z_3\}$$



Safe pruning rule (how it works)

- Finding a superset $\mathcal{A}$ of the active set $\mathcal{A}^* := \{j \mid w_j^* \neq 0\}$:

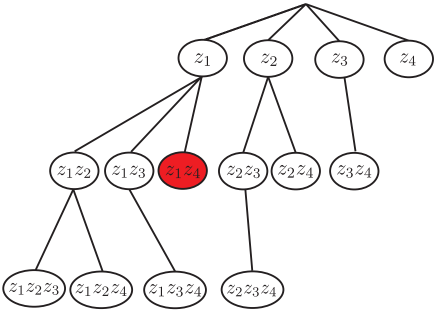
$$\mathtt{spr}(z_1 z_3 z_4) = \text{false}, \quad \mathcal{A} = \{z_1, z_1 z_3, z_1 z_3 z_4\}$$



Safe pruning rule (how it works)

- Finding a superset $\mathcal{A}$ of the active set $\mathcal{A}^* := \{j \mid w_j^* \neq 0\}$:

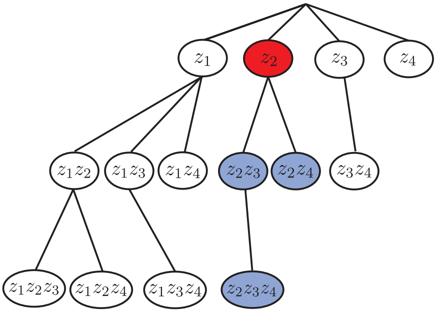
$$\texttt{spr}(z_1 z_4) = \text{true}, \quad \mathcal{A} = \{z_1, z_1 z_3, z_1 z_3 z_4\}$$



Safe pruning rule (how it works)

- Finding a superset $\mathcal{A}$ of the active set $\mathcal{A}^* := \{j \mid w_j^* \neq 0\}$:

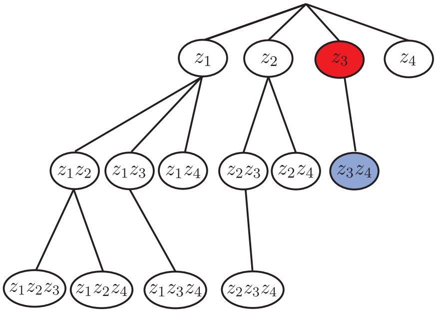
$$\texttt{spr}(z_2) = \text{true}, \quad \mathcal{A} = \{z_1, z_1z_3, z_1z_3z_4\}$$



Safe pruning rule (how it works)

- Finding a superset $\mathcal{A}$ of the active set $\mathcal{A}^* := \{j \mid w_j^* \neq 0\}$:

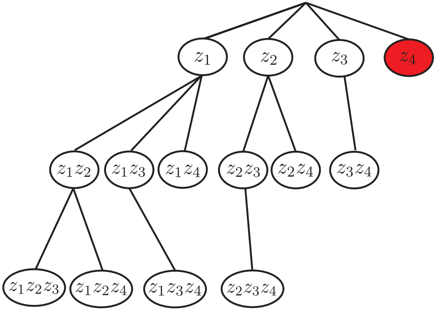
$$\mathtt{spr}(z_3) = \text{true}, \quad \mathcal{A} = \{z_1, z_1 z_3, z_1 z_3 z_4\}$$



Safe pruning rule (how it works)

- Finding a superset $\mathcal{A}$ of the active set $\mathcal{A}^* := \{j \mid w_j^* \neq 0\}$:

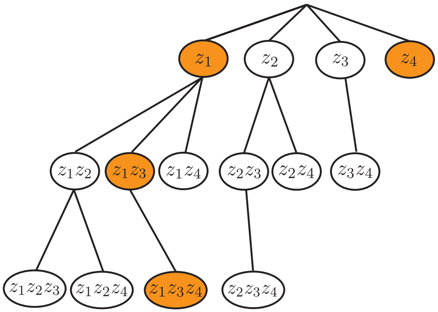
$$\texttt{spr}(z_4) = \text{false}, \quad \mathcal{A} = \{z_1, z_1z_3, z_1z_3z_4, z_4\}$$



Safe pruning rule (how it works)

- Finding a superset $\mathcal{A}$ of the active set $\mathcal{A}^* := \{j \mid w_j^* \neq 0\}$:

$$\mathcal{A} = \{z_1, z_1 z_3, z_1 z_3 z_4, z_4\}$$



Experiments

Problem setup in experiments

Computing a sequence of solutions for various λ values

Input: $\{(G_i, y_i)\}_{i \in [n]}$, $\{\lambda_k\}_{k \in [K]}$

1. Compute λ_0 and $\boldsymbol{w}_0^* = \mathbf{0}$
2. **for** $k = 1, \ldots, K$ **do**
3. Find $\mathcal{A}_{\lambda_k} \supseteq \mathcal{A}_{\lambda_k}^*$ by using the SPP based on $\boldsymbol{w}_{k-1}^*$ and $\boldsymbol{\theta}_{k-1}^*$
4. Solve an optimization problem defined only with the set of patterns in $\mathcal{A}_{\lambda_k}$
5. **end for**

Output: $\{\boldsymbol{w}_k^*\}_{k \in [K]}$, $\{\boldsymbol{\theta}_k^*\}_{k \in [K]}$

Competing method in experiments

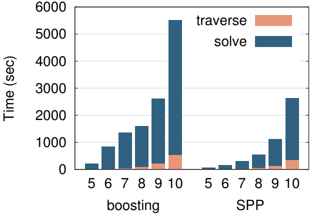
- We compared the computational costs of the proposed safe pattern pruning with g-boost [1] and i-boost [2].
- In these boosting-based approaches, the most active feature (the most violating constraint in the dual) is added one by one.
- These boosting-based approaches also exploit the tree structure among the patterns when it finds the most active feature.
- These boosting-based approaches require multiple searches over the tree (every time a new active feature is added).

---

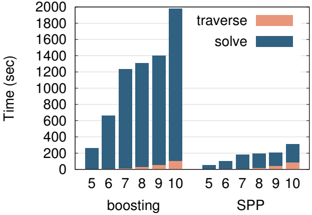
[1] H. Saigo, T. Uno and K. Tsuda. Mining complex genotypic features for predicting hiv-1 drug resistance (Bioinformatics, 2006)

[2] H. Saigo, T. Uno and K. Tsuda. Mining complex genotypic features for predicting hiv-1 drug resistance. (Bioinformatics, 2006)

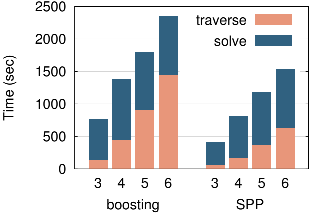
Results in experiments



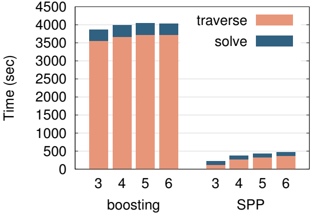
Graph Data1 (classification)



Graph Data2 (regression)



Item Data1 (classification)



Item Data2 (regression)

Summary

- Predictive pattern mining problems are still challenging, i.e., it is still difficult to efficiently identify a subset of patterns that are useful for prediction tasks.
- Safe screening method recently developed in ML literature is a new promising approach for making sparse modeling efficient (although it cannot be directly applied to pattern mining problems).
- The proposed safe pruning method exploits the tree structure and the anti-monotonicity property among the patterns in order to handle exponentially large number of patterns in the database.

Related studies in our group

- Selective inference for discovered patterns Suzumura, Nakagawa, Sugiyama, Tsuda, Takeuchi. Selective Inference Approach for Statistically Sound Predictive Pattern Mining (arXiv, 2016)
- Safe screening method for samples in SVM
  Ogawa, Suzuki and Takeuchi. Safe screening of non-support vectors in pathwise SVM computation (ICML2013)
- Simultaneous safe screening for features and samples
  Shibagaki, Karasuyama, Hatano, Takeuchi. Simultaneous safe screening of features and sample in doubly sparse modeling (ICML2016)
- Quick sensitivity analysis
  Okumura, Suzuki and Takeuchi. Quick sensitivity analysis for incremental data modification and its application to leave-one-out CV in linear classification problems (KDD2015)
- Approximate model selection
  Shibagaki, Suzuki, Karasuyama Takeuchi. Regularization path of cross-Validation error lower bounds (NIPS2015)

Other references

- L. El Ghaoui, V. Viallon and T. Rabbani. Safe feature elimination in sparse supervised learning (Pacific Journal of Optimization, 2012)
- J. Liu, Z. Zhao, J. Wang and J. Ye. Safe screening with variational inequalities and its application to Lasso (ICML2014)
- J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space (Journal of The Royal Statistical Society B, 2008)
- E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap safe screening rules for sparse multi-task and multi-class models (NIPS2015)
- H. Saigo, T. Uno and K. Tsuda. Mining complex genotypic features for predicting hiv-1 drug resistance (Bioinformatics, 2006)
- H. Saigo, S. Nowozin, T. Kadowaki, T. Kudo, K. Tsuda. gboost: a mathematical programming approach to graph classification and regression (Machine Learning, 2009)
- Hanada, Shibagaki, Sakuma, Takeuchi. Efficiently Bounding Optimal Solutions after Small Data Modification in Large-Scale Empirical Risk Minimization (arXiv, 2016)

Thank you