Biostatistics (2018), **0**, 0, *pp.* 1–9
doi:10.1093/biostatistics/PKB˙supplement˙new

Supplementary materials to A pathway-based kernel boosting method for sample classification using genomic data

LI ZENG

Department of Biostatistics, Yale University, New Haven, CT 06511, USA

ZHAOLONG YU

Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511, USA

HONGYU ZHAO*

Department of Biostatistics, Yale University, New Haven, CT 06511, USA

hongyu.zhao@yale.edu

*To whom correspondence should be addressed.

S1. Solution to the optimization of the regularized loss function

The optimization problem in each boosting iteration is reduced to equation (2.5) in the paper, as follows:

$$\min_{\beta,c} \frac{1}{N}(\eta_t + K_m\beta + 1_N c)^T W_t(\eta_t + K_m\beta + 1_N c) + \Omega(f). \tag{S1.1}$$

We can separate the optimizations of β and c. Since the penalty term $\Omega(f)$ does not involve c, we take the derivative of the objective function with respect to c, and get

$$1_N^T W_t(\eta_t + K_m\beta + c1_N) = 0,$$

from which we can solve

$$c = -\frac{1_N^T W_t\eta_t + 1_N^T W_t K\beta}{tr(W_t)}.$$

Plugging it back to equation (S1.1), the equation becomes only related to β:

$$\begin{aligned}
&\min_{\beta} \frac{1}{N}(\eta_t + K_m\beta)^T \left[I_N - \frac{1_N 1_N^T W_t}{tr(W_t)}\right] W_t \left[I_N - \frac{1_N 1_N^T W_t}{tr(W_t)}\right](\eta_t + K_m\beta) + \Omega(f) \\
=& \min_{\beta} \frac{1}{N}\|W_t^{\frac{1}{2}} \left[I_N - \frac{1_N 1_N^T W_t}{tr(W_t)}\right](\eta_t + K_m\beta)\|_2^2 + \Omega(f) \\
=& \min_{\beta} \frac{1}{N}\|\tilde{\eta} + \tilde{K}_m\beta\|_2^2 + \Omega(f).
\end{aligned}$$

Replacing $\Omega(f)$ with $\lambda\|\beta\|_1$ and $\lambda\|\beta\|_2^2$ yields equations (2.6) and (2.7) in the main article, which can be solved by LASSO and Ridge Regression, respectively.

S2. Automated choice of the regularization parameter λ

• L_1 boosting:

The objective for L_1 boosting is:

$$\min_{\beta} \frac{1}{N}\|\tilde{\eta} + \tilde{K}_m\beta\|_2^2 + \lambda\|\beta\|_1$$

By the KKT condition, optimal β needs to satisfy:

$$\mathbf{0} \in \frac{2}{N}\tilde{K}_m^T\tilde{\eta} + \frac{2}{N}\tilde{K}_m^T\tilde{K}_m\beta + \lambda\partial\|\beta\|_1,$$

where ∂ is the subgradient operator. Therefore, if the ith entry

$$(\frac{2}{N}\tilde{K}_m^T\tilde{\eta})_i \in [-\lambda, \lambda], \tag{S2.2}$$

then $\beta_i = 0$.

In practice, we calculate the left hand of equation (S2.2) for all m's and i's, and obtain a set of

$$\{\frac{2}{N}|\tilde{K}_m^T\tilde{\eta}|_i : m = 1, 2, \ldots, M; i = 1, 2, \ldots, N\}.$$

We use the 85th percentile of this set as the choice of λ, which will shrink 85 % of the β_i's to 0.

- L_2 boosting:

The objective for L_2 boosting is:

$$\min_{\beta} \frac{1}{N}\|\tilde{\eta} + \tilde{K}_m\beta\|_2^2 + \lambda\|\beta\|_2^2,$$

whose solution is

$$\beta = -(\tilde{K}_m^T\tilde{K}_m + N\lambda I_N)^{-1}\tilde{K}_m^T\tilde{\eta}.$$

Assume that $\tilde{K}_m$ has SVD decomposition $\tilde{K}_m = UDV^T$, then

$$\tilde{K}_m\beta = U\Lambda U^T\tilde{\eta},$$

where

$$\Lambda = \text{diag}(\frac{d_1^2}{d_1^2 + N\lambda}, \frac{d_2^2}{d_2^2 + N\lambda}, \ldots, \frac{d_N^2}{d_N^2 + N\lambda}),$$

and d_j is the jth diagonal element of D. Since $\tilde{K}_m\beta$ is the increment of the target function on training data, we want to control it not to be too large:

$$\|\tilde{K}_m\beta\| \leqslant \max_j \frac{d_j^2}{d_j^2 + N\lambda}\|U^T\tilde{\eta}\| = \frac{d^2}{d^2 + N\lambda}\|\tilde{\eta}\|,$$

where $d = \max_j d_j$. To control $\|\tilde{K}_m\beta\| \leqslant C$, it is sufficient to have:

$$\lambda \geqslant \frac{d^2(\|\tilde{\eta}\| - C)}{CN}.$$

In our implementations, we set $C = 2$, calculate this lower-bound for $m = 1, 2, \ldots, M$, and use the 20th quantile as our choice of λ.

S3. Simulation details

S3.1 *Parameter specifications*

• Parameters used for PKB:

- kernels: polynomial kernel with degree 3 (poly3), radial basis function kernel (rbf)

- penalty parameter (λ): calculated penalty $\times$ 0.2, 0.05

- learning rate ν: 0.1, 0.02

• Parameters used for NPR:

- Shrinkage coefficient: 0.01

• Parameters used for EasyMKL:

- kernels: poly3, rbf

We also incorporated random selection of pathways in each iteration to increase computation speed. In each iteration, we randomly chose 1/3 of all pathways, and implemented the boosting algorithm only on the selected pathways.

Data sets	Clinical outcome	Gene number	Source
Metabric	Grade 3 (+1, 259), Grade 2 and 1 (-1, 211)	24368	Link
Glioma	Grade 3 (+1, 265), Grade 2 (-1, 248) Temporal Lobe (+1, 146), Frontal Lobe (-1, 301)	17797	Link
Melanoma	Stage IV, V (+1, 224), Stage I, II, III (-1, 101) Metastatic (+1, 369), Primary (-1, 103)	20437	Link

Table 1. Overview of the real data sets. In the clinical outcome column, +1's and −1's are the codings for the outcome variables, and the number following them are their counts respectively.

S4. Real data applications

S4.1 *Datasets overview*

We list the basic information of the real data sets in Table 1, with the codings and counts of the clinical outcomes, number of input genes, and the links to the data sources.

S4.2 *Pathway database preprocessing*

We used four commonly used pathway databases in real data applications: KEGG, Biocarta, and GO biological process pathways. Considering the large amount of overlapping between pathways, we applied a heuristic algorithm to reduce the number of pathways.

We first sorted the pathways by size, from smallest to largest. We then selected the pathways one by one following the steps below:

- calculate the intersection of the current pathways with the union of selected pathways

- if the intersection is $\geqslant 60\%$ of the current pathway, continue to next, otherwise select it

An overview of the four pathway databases we used in real data applications is provided in Table 2.

Pathway database	# pathways	# genes
KEGG	186	5266
Biocarta	254	1397
GO Biological Process	625	8452

Table 2. The four pathway databases used in real data applications. The second and third columns are the number of pathways and number of genes covered in each database, respectively.

S4.3 *Parameter specifications*

• Parameters used for PKB:

- kernels: poly3, rbf

- learning rate ν: 0.1, 0.02

- penalty parameter (λ): calculated penalty $\times$ 0.2, 0.05

- pathway database: KEGG, Biocarta, GO biological pathway

• Parameters used for NPR:

- pathway database: KEGG, Biocarta, GO biological pathway

- Shrinkage coefficient: 0.01

• Parameters used for EasyMKL:

- kernels: poly3, rbf

- pathway database: KEGG, Biocarta, GO biological pathway

Since some pathway databases were too large, making each boosting iteration slow, we incorporated the same random selection technique in PKB, as employed in simulations, to increase computation efficiency .

S4.4 *Results for Glioma (site) and Melanoma (stage)*

For the glioma and melanoma data sets, we considered two other variables as classification outcomes: tumor site for glioma and Clark level for melanoma. Samples in the glioma data set mainly come from two brain sites, frontal lobe and temporal lobe. Since the two lobes are in charge of different brain functionalities, glioma in them usually leads to different symptoms. The melanoma Clark level measures the stage of tumor. Depending on how many layers of skin the tumor penetrates, samples are classified to five stages from I to V with increasing severity. We binarized the stages into two classes: -1 (stage I, II and III), +1 (stage IV and V).

Using GO Biological Process pathways and rbf kernel function in PKB yielded the best performance on glioma (site), and using the same pathway database with poly3 kernel yielded the

best results on melanoma (stage). The top pathways identified in both models are presented in Table 3.

Among the top pathways selected to be associated with glioma tumor sites, the trophoblast giant cell differentiation pathway has the largest estimated weights. Genes in this pathway include several transcription factors (such as HAND1, E2F7 and E2F8) and other proteins which participate in transcription regulation and cell differentiation. One of the genes in the pathway, nuclear receptor subfamily 2 group F member 2 (NR2F2) has been shown to be related to neuronal differentiation and subtype specification (Stappert *and others*, 2015; Abranches *and others*, 2009).

For the PKB model classifying melanoma stages, pathways including camera type eye photoreceptor cell differentiation and positive regulation of hair cycle came out to be most relevant with the outcome. It has been shown that the genes in camera type eye photoreceptor cell differentiation pathway are implicated in metastasis of melanoma. For example, Thy-1 mediates the metastasis of melanoma cells by adhesion of melanoma cells to endothelial cells (Schubert *and others*, 2013), and PROM1 is a biomarker for cancer stem cells in melanoma cell lines (Zimmerer *and others*, 2013). Genes in the positive regulation of hair cycle pathway have also been reported to be related to melanoma progression. For example, Wnt-10b can repress proliferation and migration of melanoma cells (Misu *and others*, 2015), and contributes to melanoma invasion and metastasis (Murry *and others*, 2006).

	Glioma (site)	Melanoma (stage)
1	Trophoblast giant cell differentiation	Camera type eye photoreceptor cell differentiation
2	Protein carboxylation	Positive regulation of hair cycle
3	Cardiac left ventricle morphogenesis	Ureter development
4	Thyroid hormone generation	Regulation of cell proliferation involved in kidney development
5	Epithelial to mesenchymal transition involved in endocardial cushion formation	Peripheral nervous system neuron differentiation
6	Positive regulation of guanylate cyclase activity	Multicellular organismal macromolecule metabolic process
7	Metanephric renal vesicle morphogenesis	Metanephric renal vesicle morphogenesis
8	Regulation of growth hormone secretion	Cell aggregation
9	Myoblast fusion	Regulation of collateral sprouting
10	Endocardium development	Embryonic skeletal joint development
11	Oxygen transport	Negative regulation of mast cell activation
12	Positive regulation of hormone metabolic process	Negative regulation of jun kinase activity
13	Sterol catabolic process	Positive regulation of tyrosine phosphorylation of stat1 protein
14	Negative regulation of acute inflammatory response	Regulation of water loss via skin
15	Cd4 positive or cd8 positive alpha beta t cell lineage commitment	G protein coupled glutamate receptor signaling pathway

Table 3. Top fifteen pathways in terms of pathway weights fitted by PKB. Pathways in the two columns are from GO Biological Process pathways.

References

Abranches, Elsa, Silva, Margarida, Pradier, Laurent, Schulz, Herbert, Hummel, Oliver, Henrique, Domingos and Bekman, Evguenia. (2009). Neural differentiation of embryonic stem cells in vitro: a road map to neurogenesis in the embryo. *PloS one* **4**(7), e6286.

Misu, Masayasu, Ouji, Yukiteru, Kawai, Norikazu, Nishimura, Fumihiko, Nakamura-Uchiyama, Fukumi and Yoshikawa, Masahide. (2015). Effects of wnt-10b on proliferation and differentiation of murine melanoma cells. *Biochemical and biophysical research communications* **463**(4), 618–623.

Murry, Brian P, Blust, Bryan E, Singh, Amandip, Foster, Timothy P and Marchetti, Dario. (2006). Heparanase mechanisms of melanoma metastasis to the brain: Development and use of a brain slice model. *Journal of cellular biochemistry* **97**(2), 217–225.

Schubert, Kathleen, Gutknecht, Danny, Köberle, Margarethe, Anderegg, Ulf and Saalbach, Anja. (2013). Melanoma cells use thy-1 (cd90) on endothelial cells for metastasis formation. *The American journal of pathology* **182**(1), 266–276.

Stappert, Laura, Roese-Koerner, Beate and Brüstle, Oliver. (2015). The role of micrornas in human neural stem cells, neuronal differentiation and subtype specification. *Cell and tissue research* **359**(1), 47–64.

Zimmerer, Rüdiger M, Korn, Philippe, Demougin, Philippe, Kampmann, Andreas, Kokemüller, Horst, Eckardt, André M, Gellrich, Nils-Claudius and Tavassol, Frank. (2013). Functional features of cancer stem cells in melanoma cell lines. *Cancer cell international* **13**(1), 78.