A Regularisation Path-Following Approach for Discovering Interactions in High-Dimensional Survival Data

David A. duVerle[1], Ichiro Takeuchi[2] and Koji Tsuda[1]

[1] Computational Biology Research Center, AIST, Tokyo,
dave.duverle@aist.go.jp
[2] Department of Computer Science, Nagoya Institute of Technology, Nagoya

Abstract

Motivation

Several efficient methods exist to relate high-dimensional gene expression data to various clinical phenotypes, however finding *combinations* of features in such input remains a challenge, particularly when fitting complex statistical models such as those used for survival studies. While such models based on linear combinations of gene expression values can be used to approximate more complex interactions, they usually falls short, when trying to identify certain combinatorial effects (e.g. synthetic lethal genes [4]). Identifying such gene combinations could lead to promising therapeutic applications, yet remains a practical impossibility with current fitting methods due to serious computational complexity issues, as they rely on the ability to enumerate potential input variables.

Method

Gene data, with its typically large dimension and small sample size, invite the use of a penalisation component in its statistical models, with ℓ_1-norm often preferred for its ability to drive sparsity of the model and select a concise set of variables (gene expression values, mutation types, etc.) [1]. Park & Hastie [2], in particular, proposed a method to compute the *regularisation path* of an ℓ_1-penalised Cox model [3] for survival data: efficiently estimating a series of Cox models with different levels of complexity and sparsity. Their approach, however, can only consider the impact of individual input variables, introduced one at a time, on the model.

In our work [5], we extended this single-variable regularisation path-following approach to handle combinatorial covariates (arbitrarily complex patterns of input variables). Using this technique, we can efficiently compute regression parameters for complex statistical models, at the optimal amount of penalisation, with the only requirement of a convex loss function. We solved the combinatorial explosion issue by taking advantage of itemset mining techniques [6] that brought the average computational complexity of the algorithm within practical reach.

With our fitting method, virtually limitless combinations of genes and phenotypes, grouped in itemsets of boolean variables, are used as single predictor variables in the model. Running multiple iterations of the algorithm on subsampled dataset, we can produce ordered lists of candidate interactions with strong predicting power.

Results

The interactions found by applying our method to cancer study survival data (breast cancer and neuroblastoma) include many genes that could not be found through linear models, yet show up in literature as strongly tied to these conditions, confirming the crucial importance of taking interaction effects into account in order to detect some of the weaker signal in gene expression data.

Beyond proportional hazards models, our itemset-based method can be applied to any regression model with convex loss, each time making use of the input's structure and sparsity to sidestep complexity issues, while at the same time guaranteeing that events along the regularisation path (values of the regularisation parameter for which a change occurs in the model structure) are exhaustively explored.

Acknowledgment

The authors would like to thank Yuko Murakami-Tonami and Kenji Kadomatsu, of the Department of Molecular Biology at Nagoya University's Graduate School of Medicine, for their suggestions and guidance regarding the medical aspects of our research into synthetic lethals for cancer cells, as well as Hiroshi Mamitsuka and Timothy Hancock, of Kyoto University, for their helpful feedback and suggestions in summarising our work.

References

1. R. Tibshirani et al. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.
2. M.Y. Park and T. Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007.
3. D. Ghosh. Penalized discriminant methods for the classification of tumors from gene expression data. *Biometrics*, 59(4):992–1000, 2003.
4. W.G. Kaelin. The concept of synthetic lethality in the context of anticancer therapy. *Nature reviews cancer*, 5(9):689–698, 2005.
5. David duVerle, Ichiro Takeuchi, Yuko Murakami-Tonami, Kenji Kadomatsu, Koji Tsuda, et al. Discovering combinatorial interactions in survival data. *Bioinformatics*, 29(23):3053–3059, 2013.
6. T. Uno, T. Asai, Y. Uchida, and H. Arimura. An efficient algorithm for enumerating closed patterns in transaction databases. In *Discovery Science*, pages 57–59. Springer, 2004.