BIOINFORMATICS

Vol. 00 no. 00 2015
Pages 1–2

APTER: Aggregated Prognosis Through Exponential Reweighting

Liu Yang [1], Kristiaan Pelckmans [1] *

[1]Department of Information Technology, Polacksbacken, Uppsala, Sweden

ABSTRACT

This paper considers the task of learning how to make a prognosis of a patient based on his/her micro-array expression levels. The method is an application of the aggregation method as recently proposed in the literature on theoretical machine learning, and excels in its computational convenience and capability to deal with high-dimensional data. This paper gives a formal analysis of the method, yielding rates of convergence similar to what traditional techniques obtain, while it is shown to cope well with an exponentially large set of features. Those results are supported by numerical simulations on a range of publicly available survival-micro-array datasets. It is empirically found that the proposed technique combined with a recently proposed preprocessing technique gives excellent performances. All employed software and datasets are available on `http://www.it.uu.se/research/project/survlab`.

APPENDIX B: BENCHMARK DATASETS

This appendix describes the real-world datasets. The datasets range from large-dimensional ($d = O(100)$) to huge-dimensional ($O(10,000)$) and record $n = O(100)$ subjects. We report the performance of different methods on:

- 7 public datasets containing micro-array expression levels and events (occurrence of disease) of the associated subjects as used in (1).
- The micro-array survival dataset as presented in (2) and analysed in the report (3).

Details are given below. The 7 publicly available microarray datasets as used for benchmarking in (1), have the following properties.

(NSBCD): The Norway/Stanford Breast Cancer Data set is given in (4). In this database there are survival data of $n = 115$ women who have breast cancer, and $d = 549$ intrinsic genes introduced in (4) were measured. In the 115 patients, 33% (38) have experienced an event during the study. Missing values were imputed by the 10-nearest neighbour method.

(Veer): The survival data of sporadic lymph-node-negative patients with their gene expression profiles is given in (5). It has $n = 78$ patients with $d = 4751$ gene expressions selected from the 25,000 genes recorded with the microarray. 44 patients remained free of disease after their diagnosis for an interval of at least 5 years. The average follow-up time for these patients was 8.7 years. 34 patients had developed distant metastases within 5 years, and the mean time to metastases was 2.5 years.

(Vijver): The data set of $n = 295$ consecutive patients with primary breast carcinomas is from (5) All patients had stage I or II breast cancer and were younger than 53 years old. They gave the previously determined $d = 70$ marker genes that are associated with the risk of early distant metastases in young patients with lymph-node-negative breast cancer. The median follow-up among all 295 patients was 6.7 years (range, 0.05 to 18.3). There were no missing data. 88 patients have experienced an event during the study.

(DBCD): The Dutch Breast Cancer Data set is described in (6), and is a subset of the data from (5). There are survival data of $n = 295$ women who have breast cancer. The measures of $d = 4919$ gene expression were taken from the fresh-frozen-tissue bank of the Netherlands Cancer Institute. All the ages of the patients are smaller than or equal to 52 years. The diagnosis was made between 1984 and 1995 without previous history of cancer. The median of follow-up time was 6.7 years (range 0.05-18.3). In the 295 patients, 26.78% (79) have experienced an event during the study.

(DLBCL): The diffuse large-B-cell lymphoma data set is described in (7). This contains survival data of $n = 240$ patients who have diffuse large-B-cell lymphoma. $d = 7399$ different gene expression measurements are given. The median of follow-up time was 2.8 years. From the 240 patients, 58% have experienced an event during the study.

(Beer): The survival data of $n = 86$ patients with primary lung adenocarcinomas is from (8) There are $d = 7129$ expressed genes selected from Affymetrix hu6800 microarrays. 76 patients have experienced an event during the study.

(AML): The survival data of acute myeloid leukemia patients is described in (9). It contains $n = 116$ patients with acute myeloid leukemia and the expression levels of $d = 6283$ genes. 71 patients have experienced an event during the study.

The same datasets were used in (1) and (10) to benchmark a state-of-art methods, results that are reproduced here as well. The high-dimensional FL dataset has the following description.

*to whom correspondence should be addressed

(FL): Additionally, we use the micro-array dataset which was used in (2), and analysed in (3). This data set included the survival data of $n = 191$ patients with follicular lymphoma after diagnosis. The median age at diagnosis was 51 years (range, 23 to 81), and the median follow-up time was 6.6 years (range, less than 1.0 to 28.2); the median followup time among patients alive at last follow-up was 8.1 years. It contains $d = 44928$ gene expression levels selected from Affymetrix U133A and U133B microarrays.

REFERENCES

[1]H. M. Bøvelstad, S. Nygård, H. L. Størvold, M. Aldrin, Ø. Borgan, A. Frigessi, and O. C. Lingjærde, "Predicting survival from microarray datała comparative study," *Bioinformatics*, vol. 23, no. 16, pp. 2080–2087, 2007.

[2]S. S. Dave, G. Wright, B. Tan, A. Rosenwald, R. D. Gascoyne, W. C. Chan, R. I. Fisher, R. M. Braziel, L. M. Rimsza, T. M. Grogan *et al.*, "Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells," *New England Journal of Medicine*, vol. 351, no. 21, pp. 2159–2169, 2004.

[3]R. S. Lin, "Re-analysis of molecular features in predicting survival in follicular lymphoma," *report, http://statweb.stanford.edu/~tibs/FL/report/RayLin_Lab_rotation2006s.pdf*, 2006.

[4]T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler *et al.*, "Repeated observation of breast tumor subtypes in independent gene expression data sets," *Proceedings of the National Academy of Sciences*, vol. 100, no. 14, pp. 8418–8423, 2003.

[5]L. J. van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[6]H. C. van Houwelingen, T. Bruinsma, A. A. Hart, L. J. van't Veer, and L. F. Wessels, "Cross-validated cox regression on microarray gene expression data," *Statistics in medicine*, vol. 25, no. 18, pp. 3201–3216, 2006.

[7]A. Rosenwald, G. Wright, W. C. Chan, J. M. Connors, E. Campo, R. I. Fisher, R. D. Gascoyne, H. K. Muller-Hermelink, E. B. Smeland, J. M. Giltnane *et al.*, "The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma," *New England Journal of Medicine*, vol. 346, no. 25, pp. 1937–1947, 2002.

[8]D. G. Beer, S. L. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas *et al.*, "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature medicine*, vol. 8, no. 8, pp. 816–824, 2002.

[9]L. Bullinger, K. Döhner, E. Bair, S. Fröhling, R. F. Schlenk, R. Tibshirani, H. Döhner, and J. R. Pollack, "Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia," *New England Journal of Medicine*, vol. 350, no. 16, pp. 1605–1616, 2004.

[10]V. Van Belle, K. Pelckmans, J. A. Suykens, and S. Van Huffel, "Learning transformation models for ranking and survival analysis." *Journal of machine learning research*, vol. 12, no. 3, 2011.