BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 8 2011, pages 1168–1169
doi:10.1093/bioinformatics/btr103

Gene expression Advance Access publication March 2, 2011

SurvJamda: an R package to predict patients' survival and risk assessment using joint analysis of microarray gene expression data

Haleh Yasrebi[1,2]

[1]Swiss Institute for Experimental Cancer Research (ISREC), Swiss Federal Institute of Technology (EPFL), School of Life Sciences (SV) and [2]Swiss Institute of Bioinformatics, EPFL SV ISREC, Station 15, 1015 Lausanne, Switzerland

Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: SurvJamda (Survival prediction by joint analysis of microarray data) is an R package that utilizes joint analysis of microarray gene expression data to predict patients' survival and risk assessment. Joint analysis can be performed by merging datasets or meta-analysis to increase the sample size and to improve survival prognosis. The prognosis performance derived from the combined datasets can be assessed to determine which feature selection approach, joint analysis method and bias estimation provide the most robust prognosis for a given set of datasets.

Availability: The `survJamda` package is available at the Comprehensive R Archive Network, http://cran.r-project.org.

Contact: hyasrebi@yahoo.com

Received on November 18, 2010; revised on January 27, 2011; accepted on February 18, 2011

1 INTRODUCTION

The `survJamda` package was developed for survival prediction and risk assessment based on microarray data. It allows to jointly analyze the datasets through data merging and meta-analysis. Data merging combines the data into one set prior to their analysis, whereas meta-analysis integrates only the results. In addition to different joint analysis methods, `survJamda` contains various feature selection approaches and bias estimation techniques which enable the user to determine the combination of which methods provides the most robust prediction for a given set of datasets.

A few other R packages like `ipdmeta` (Broeze *et al.*, 2009) and `survcomp` (Haibe-Kains *et al.*, 2008) have been created for joint analysis, that are more specifically, related to meta-analysis of censored data with time-to-event outcome.

2 DATA

The functions and algorithms developed in `survJamda` can be assessed on the datasets of the `survJamda.data` package. `SurvJamda.data`, created by the author, is a data package of 18 Mb containing three breast cancer datasets, GSE3143, GSE1992 and GSE4335, which were analyzed in Yasrebi *et al.* (2009). `SurvJamda.data` are also available on Comprehensive R Archive Network.

3 METHODS

3.1 Feature selection

(i) Top-ranking (Yasrebi *et al.*, 2009). The multiple hypothesis testing correction implemented in the `p.adjust` function in the R `stats` package can also be applied to the top-ranking method.

(ii) User-defined method.

3.2 Joint analysis methods

(i) Merging method:

(a) ComBat (Johnson *et al.*, 2007).

(b) *Z*-score normalization. This method is applied in two ways:

(1) *Z*-score1 normalization: in this approach, all datasets are *Z*-score normalized (Larsen *et al.*, 2000) prior to their selection for the training and testing sets and their combination into one set (Yasrebi *et al.*, 2009).

(2) *Z*-score2 normalization: in this method, the datasets are initially selected for the training and testing sets. Then, the datasets composing the training set are merged together and *Z*-score normalized. The testing set is also *Z*-score normalized independently and separately from the training set.

(ii) Meta-analysis. The inverse normal method (Hedges *et al.*, 1985) is used for meta-analysis.

3.3 Validation frameworks

(i) Cross validation (CV) nested in 10 iterations.

(ii) Independent validation.

(a) Pair-wise mode: two datasets are selected at a time, one of which is used as the training set and the other as the testing set. This process is iterated until all datasets are used as the training and testing sets (Yasrebi *et al.*, 2009).

(b) Leave one dataset out: all datasets except one are merged together to form the training set and the left-out set is used as the testing set. Similarly, this process is iterated until all datasets are used as the training and testing sets (Yasrebi *et al.*, 2009).

3.4 Performance measures

(i) Survival prediction is expressed by time-dependent area under the receiver operating characteristic curve (Heagerty *et al.*, 2000) and hazard ratio measures risk assessment (Yasrebi *et al.*, 2009).

(ii) Concordance index (Haibe-Kains *et al*., 2008).

(iii) Brier score (Haibe-Kains *et al*., 2008).

Conflict of Interest: none declared.

REFERENCES

Broeze,K. *et al*. (2009) Individual patient data meta-analysis of diagnostic and prognostic studies in obstetrics, gynaecology and reproductive medicine. *BMC Med. Res. Methodol.*, **9**, 22.

Haibe-Kains,B. *et al*. (2008) A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? *Bioinformatics*, **24**, 2200–2208.

Heagerty,P.J. *et al*. (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. **56**, 337–344.

Hedges,L.V. and Olkin,I. (1985) *Statistical Methods for Meta-Analysis*. Academic Press.

Ishwaran,H. *et al*. (2008) Random survival forests. *Ann. Appl. Statist.*, **2**, 841–860.

Johnson,E.,W. *et al*. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Larsen,R.J. and Marx,M.L. (2000) *An Introduction to Mathematical Statistics and Its Applications*, 3rd edn. Prentice Hall.

Yasrebi,H. *et al*. (2009) Can survival prediction be improved by merging gene expression data sets? *PLoS ONE*, **4**, e7431.