BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 17 2004, pages 3284–3285
doi:10.1093/bioinformatics/bth376

A web portal for classification of expression data using maximal margin linear programming

Alexey V. Antonov[1], Igor V. Tetko[1,*], Volodymyr V. Prokopenko[2], Denis Kosykh[1] and Hans W. Mewes[1,3]

[1]GSF National Research Center for Environment and Health, Institute for Bioinformatics, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany, [2]Biomedical Department, Institute of Bioorganic & Petroleum Chemistry, Ukrainian Academy of Sciences, Murmanskaya 1, 02094, Kyiv, Ukraine and [3]Department of Genome-Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, 85350 Freising, Germany

Received on May 19, 2004; revised on June 15, 2004; accepted on June 17, 2004
Advance Access publication June 23, 2004

ABSTRACT

Summary: The Maximal Margin (MAMA) linear programming classification algorithm has recently been proposed and tested for cancer classification based on expression data. It demonstrated sound performance on publicly available expression datasets. We developed a web interface to allow potential users easy access to the MAMA classification tool. Basic and advanced options provide flexibility in exploitation. The input data format is the same as that used in most publicly available datasets. This makes the web resource particularly convenient for non-expert machine learning users working in the field of expression data analysis.

Availability: The web software developed is available at http://mips.gsf.de/proj/mdcs/. A stand-alone version of the server can be downloaded from this site too.

Contact: i.tetko@gsf.de

The Maximal Margin (MAMA) linear programming classification algorithm (Antonov *et al.*, 2004) has been developed for classification of microarray expression data. The algorithm was tested with two and multiclass supervised learning problems. The distinctive quality of the algorithm is a clear biological interpretation of the results. The number of classification features corresponds to the number of analyzed classes. Each feature represents a linear combination of a relatively small set of genes (10–25), which provides a simple biological interpretation of the feature as a model of gene interaction. Disruptions in gene interactions between different classes (e.g. in normal and pathological states of patients) makes it possible to perform the classification task. The algorithm demonstrates superior classification performance on publicly available cancer expression datasets (Antonov *et al.*, 2004).

*To whom correspondence should be addressed.

MAMA exploits a linear programming package as its computational engine. The current version employs LINDO software, which is commercially available (www.lindo.com). It was chosen because of its high speed of data analysis, its ability to analyze very large datasets and the absence of freely distributed packages of similar quality. Having to obtain the commercial LINDO software may prevent some users (particularly first-time users) from validating the MAMA approach on their data. To overcome this problem, we received permission from LINDO Inc. to develop a web-based interface, available free-of-charge to Web users. Users who are interested in using MAMA as a stand-alone application can download it from our website. They also need to acquire a license or download a free trial version from LINDO. The restrictions of the trial version make it possible to perform a classification task with ∼100 genes.

This note describes a publicly accessible Web interface that allows users to perform different classification tasks. Users might also be interested in comparing the performance of MAMA with several traditional classification procedures at the caGEDA site http://bioinformatics.upmc.edu/GE2/GEDA.html or with SVM machines (Pavlidis *et al.*, 2004) at http://svm.sdsc.edu. Hyperlinks to these sites are available on our website. Links to new sites for on-line microarray data classification appearing in the future will also be included.

The web server software, available at http://mips.gsf.de/proj/mdcs/ or http://mips.gsf.de/projects/expression, was developed using Java Servlet technology (Hunter and Crawford, 2001) and tested using Tomcat version 5. The server stores input parameters and the complete path to input data files in an xml file inside a local directory ('arhiv'). The Perl script parses this file and then starts the analyses of data using MAMA software. Both processes work asynchronously. If several tasks are submitted, they form a queue, and are executed using the First-in-First-out (FIFO) principle. The

number of tasks in the queue and current status of analysis are displayed to the user. If several tasks are submitted, the user may have to wait some time before the task starts and finishes. Alternatively, he/she can provide an e-mail address and after the analysis is finished a message with a link to the calculated results will be sent to the specified e-mail address.

The separation of data submission and processing tasks also makes it possible to use our application without the Web interface. In order to do this, the user can create an xml file with data and parameters and execute a Perl script to run the analysis. An xml file with results will be stored inside the 'arhiv' directory on the disk. This protocol also makes possible easy debugging of the program. If a task crashes, the data stored in the xml file can be used to repeat the calculations. The source code of the servlet is included in the distribution and can be modified to develop similar servers for other data analysis tools.

The Web version of MAMA operates in three basic modes: test set classification (normal mode), leave one out (LOO) and *n*-fold cross-validation modes. The parameters required to run the analysis can be specified on the 'Basic Options' or 'Advanced Options' HTML pages.

The 'Basic Options' allow the user to access the minimum amount of information to get started. The user has to specify at a minimum the data files containing the expression and classification data. Training data information only is required. Supplying the test data information is optional. The input data format used in MAMA analysis was designed particularly for gene expression data. It is an extension of the format used in most publicly available datasets for cancer classification (Golub *et al*., 1999; Ramaswamy *et al*., 2001; http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi). The user need only additionally specify first row and column indexes corresponding to the first expression value. An example of a dataset is available on the website as well as an example classification analysis that can be run to familiarize the user with the procedure.

The 'Advanced Options' provide a number of parameters that can be optionally set by the user. They include different data preprocessing and gene prefiltering utilities which could affect classification performance. For example, expression data frequently has negative values and it is natural to make a data shift (to add some positive value to each expression value), thus removing most of them. The user can also perform log transformation of the input expression data. For more details on using the software and explanations of possible options we refer readers to the website and references listed there.

Upon transmitting the data the user is directed to a page that reports the current progress of the analysis. When the classification analysis is completed a report is provided. The report consists of three general parts. This includes values of all parameters and details of processed data as well as classification results. The classification results provide details of analysis of each sample, overall classification statistics and feature gene content information.

ACKNOWLEDGEMENTS

We thank Dr Louise Riley for her help with English and LINDO Inc. for their permission to use LINDO in this web interface. This work was supported by grant 031U118A from NGFN to HWM, BFAM and INTAS 00-0363 grant.

REFERENCES

Antonov,A.V., Tetko,I.V., Mader,M.T., Budczies,J. and Mewes,H.W. (2004) Optimization models for cancer classification: extracting gene interaction information from microarray expression data. *Bioinformatics*, **20**, 644–652.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hunter,J. and Crawford,W. (2001) *Java Servlet Programming*. O'Reilly & Associates Inc., Sabastopol, USA.

Pavlidis,P., Wapinski,I. and Noble,W.S. (2004) Support vector machine classification on the web. *Bioinformatics*, **20**, 586–587.

Ramaswamy,S., Tamayo,P., Rifkin,R., Mukherjee,S., Yeang,C.H., Angelo,M., Ladd,C., Reich,M., Latulippe,E., Mesirov,J.P. *et al*. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.