M@CBETH: a microarray classification benchmarking tool

Nathalie L.M.M. Pochet*, Frizo A.L. Janssens, Frank De Smet, Kathleen Marchal,
Johan A.K. Suykens and Bart L.R. De Moor
K.U.Leuven, ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium
Email: {nathalie.pochet,frizo.janssens,frank.desmet,kathleen.marchal,
johan.suykens,bart.demoor}@esat.kuleuven.ac.be

Abstract

Summary: *Microarray classification can be useful to support clinical management decisions for individual patients in for example oncology. However, comparing classifiers and selecting the best for each microarray dataset can be a tedious and nonstraightforward task. The M@CBETH (a MicroArray Classification BEnchmarking Tool on a Host server) web service offers the microarray community a simple tool for making optimal two-class predictions. M@CBETH aims at finding the best prediction among different classification methods by using randomizations of the benchmarking dataset. The M@CBETH web service intends to introduce an optimal use of clinical microarray data classification.*
Availability: *Web service at http://www.esat.kuleuven.ac.be/MACBETH/.*
Contact: Nathalie.Pochet@esat.kuleuven.ac.be

Introduction

Using microarray data allows making predictions on for example therapy response, prognosis and metastatic phenotype of an individual patient. Microarray technology has shown to be useful in supporting clinical management decisions for individual patients (for example breast cancer (van 't Veer et al., 2002), acute myeloid leukemia (Valk et al., 2004), and ovarian cancer (De Smet et al., 2004)) in combination with classification methods (Furey et al., 2000). Finding the best classifier for each dataset can be a tedious and nonstraightforward task for users not familiar with these classification techniques. In this note, a web service is presented that compares, for each microarray dataset introduced to this service, different classifiers and selects the best in terms of randomized independent test set performances.

Systematic benchmarking of microarray data classification revealed that either regularization or dimensionality reduction is required to obtain good independent test set performances (Pochet et al., 2004). Regularization - as is performed in Support Vector Machines (SVM) (Cristianini and Shawe-Taylor, 2000) - already led to the Gist web service, which offers SVM classification on the web (Pavlidis et al., 2004). This note allows comparing different classification methods. By exploring different combinations of nonlinearity and dimensionality reduction, our benchmarking study showed that the optimal classifier can differ for each dataset. Also important, but often underestimated in the model building process, is the fine-tuning of all hyperparameters (e.g. regularization parameter, kernel parameter, number of principal components). Exploring all combinations to find the optimal classifier for each dataset can be complicated.

Website

The M@CBETH website offers two services: benchmarking and prediction. After registration and logging on to the web service, users can request benchmarking or prediction analyses. Users are notified by email about the status of their analyses running on the host server. They can also check this on the analysis results page, which gives an overview of all analyses and contains links to corresponding results pages.

Benchmarking, the main service on the M@CBETH website, involves selection and training of an optimal model based on the submitted benchmarking dataset and corresponding class labels. This model is then stored for immediate or later use on prospective data. Benchmarking results in a table showing summary statistics (leave-one-out cross-validation (LOO-CV), training set accuracy (ACC) and area under the Receiver Operating Characteristic curve performance (AUC), test set ACC and AUC) for all selected classification methods, highlighting the best method. Prospective data can be submitted and evaluated immediately during the same benchmarking analysis. Via the *prediction* service, the M@CBETH website offers a way for later evaluation of prospective data by reusing an existing optimal prediction model (built in a previous benchmarking analysis by the same user). For both services, if the corresponding prospective labels are submitted, the prospective accuracy is calculated. Otherwise, labels are predicted for all prospective samples. This latter application is useful for classifying new unseen patients in clinical practice.

The M@CBETH web service is intended for classifica-

*To whom correspondence should be addressed

tion of patient samples, supposing microarray data is represented by an expression matrix characterized by high dimensionality in the sense of a small number of patients and a large number of gene expression levels for each patient. Two kinds of data formats are accepted: spreadsheet-like tab-delimited text files and matrix-like matlab files. Datasets are not allowed to contain missing values. Class labels are restricted to '+1' or '-1'. Several publicly available microarray datasets are present on the website in correct data format as an example.

Users can select the classification methods that will be compared (default selection set to the best overall and most efficient methods from the benchmarking study), change the number of randomizations (default 20, while keeping in mind that results are more reliable when the number of randomizations is large) and switch off normalization (although performing normalization is better from a statistical viewpoint).

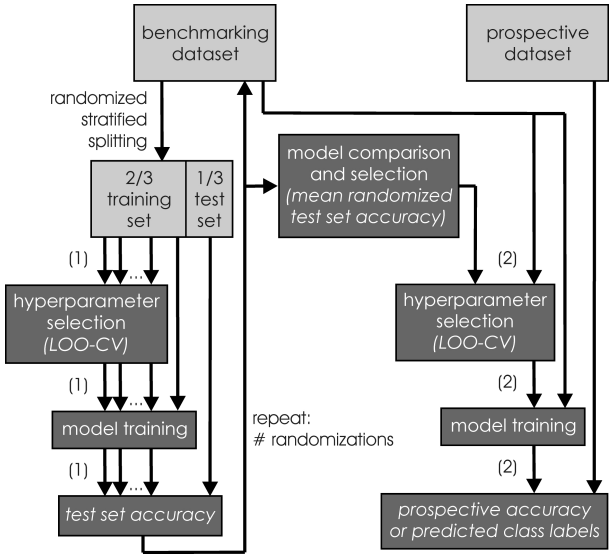

Figure 1: Overview of the algorithm. The benchmarking dataset is reshuffled until the number of requested randomizations is reached. All randomizations are split (2/3 of the samples for training, the rest as test set) in a stratified way (class labels are equally distributed over the training-test split). Iteratively, all selected classification methods (1) are applied to all randomizations. In each iteration, selection of the hyperparameters is first performed by means of LOO-CV, then the model is trained based on the training set and finally this model is then applied onto the test set resulting in a test set ACC. The mean randomized test set ACC is calculated for each classification method. The best generalizing method (2) - with best test set ACC - is then used for building the optimal classifier onto the complete benchmarking dataset, which is stored for application onto prospective datasets.

Algorithm

An overview of the algorithm behind this web sevice is presented in Figure 1. Different classification methods - based on Least Squares SVM (LS-SVM) (Suykens et al., 2002) (based on linear and Radial Basis Function (RBF) kernels), Fisher Discriminant Analysis (FDA), Principal Component Analysis (PCA) and kernel PCA (Schölkopf et al., 1998; Suykens et al., 2002) (based on linear and RBF kernels) - are considered. More detailed information on these methods can be found in (Pochet et al., 2004).

Acknowledgements

Research supported by 1. Research Council KUL: GOA-AMBioRICS, IDO (IOTA Oncology, Genetic networks), several PhD/postdoc & fellow grants; 2. Flemish Government: - FWO: PhD/postdoc grants, projects G.0115.01, G.0407.02, G.0413.03, G.0388.03, G.0229.03; - IWT: PhD Grants, STWW-Genprom, GBOU-McKnow, GBOU-SQUAD, GBOU-ANA; 3. Belgian Federal Government: DWTC (IUAP V-22 (2002-2006)); 4. EU: CAGE; Biopattern.

References

Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines (and other Kernel-Based Learning Methods)*. Cambridge University Press, Cambridge.

De Smet,F., Pochet,N., Engelen,K., Van Gorp,T., Van Hummelen,P., Suykens,J., Marchal,K., Amant,F., Moreau,Y., Timmerman,D., De Moor,B. and Vergote I. (2004) Predicting the clinical behavior of ovarian cancer from gene expression profiles, *Internal Report 04-152*, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2004. Submitted.

Furey,T.S., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Haussler,D. (2000) Support vector machines classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, 16,906-914.

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

Pochet,N., De Smet,F., Suykens,J.A.K. and De Moor,B.L.R. (2004) Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction, *Bioinformatics*, 20,3185-3195.

Schölkopf,B., Smola,A.J. and Müller,K.-R. (1998) Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, 10,1299-1319.

Suykens,J.A.K., Van Gestel,T., De Brabanter,J., De Moor,B. and Vandewalle,J. (2002) *Least Squares Support Vector Machines*. World Scientific, Singapore (ISBN 981-238-151-1).

Valk,P.J.M., Verhaak,R.G.W., Beijen,M.A., Erpelinck,C.A.J., van Doorn-Khosrovani,S., van Waalwijk,B., Boer,J.M., Beverloo,H.B., Moorhouse, M.J., van der Spek,P.J., Lowenberg,B. and Delwel, R. (2004) Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia, *New England Journal of Medicine*, 350,1617-1628.

van 't Veer,L.J., Dai,H., Van De Vijver,M.J., He,Y.D., Hart,A.A.M., Mao,M., Peterse,H.L., Van Der Kooy,K., Marton,M.J., Witteveen,A.T., Schreiber,G.J., Kerkhoven,R.M., Roberts,C., Linsley,P.S., Bernards,R. and Friend,S.H. (2002) Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer, *Nature*, 415,530-536.