

Evaluating Microarray-based Classifiers: An Overview

A.-L. Boulesteix¹, C. Strobl², T. Augustin² and M. Daumer¹

¹Sylvia Lawry Centre for MS Research (SLC), Hohenlindenerstr. 1, 81677 Munich, Germany.

²Department of Statistics, University of Munich (LMU), Ludwigstr. 33, 80539 Munich, Germany.

Abstract: For the last eight years, microarray-based class prediction has been the subject of numerous publications in medicine, bioinformatics and statistics journals. However, in many articles, the assessment of classification accuracy is carried out using suboptimal procedures and is not paid much attention. In this paper, we carefully review various statistical aspects of classifier evaluation and validation from a practical point of view. The main topics addressed are accuracy measures, error rate estimation procedures, variable selection, choice of classifiers and validation strategy.

Keywords: accuracy measures, classification, conditional and unconditional error rate, error rate estimation, validation data, variable selection, gene expression, high-dimensional data

1 Introduction

In the last few years, microarray-based class prediction has become a major topic in many medical fields. Cancer research is one of the most important fields of application of microarray-based prediction, although classifiers have also been proposed for other diseases such as multiple sclerosis (Bomprezzi et al. 2003). Important applications are molecular diagnosis of disease subtype and prediction of future events such as, e.g. response to treatment, relapses (in multiple sclerosis) or cancer recidive. Note that both problems are usually treated identically from a statistical point of view and related from a medical point of view, since patients with different disease subtypes also often have different outcomes.

Let us consider a standard class prediction problem where expression data of p transcripts and the class information are available for a group of n patients. From a statistical point of view, patients are *observations* and *transcripts* are *variables*. Note that a particular gene might be represented several times. To avoid misunderstandings, we prefer the statistical term “variable” to the ambiguous term “gene”. In microarray studies, the number of variables p is huge compared to n (typically, $5000 \leq p \leq 50000$ and $20 \leq n \leq 300$), which makes standard statistical prediction methods inapplicable. This dimensionality problem is also encountered in other fields such as proteomics or chemometrics. Hence, the issues discussed in the present article are not specific to microarray data. The term *response class* refers to the categorical variable that has to be predicted based on gene expression data. It can be, e.g. the presence or absence of disease, a tumor subtype such as ALL/AML (Golub et al. 1999) or the response to a therapy (Ghadimi et al. 2005; Rimkus et al. 2008). The number of classes may be higher than two, though binary class prediction is by far the most frequent case in practice.

Note that gene expression data may also be used to predict survival times, ordinal scores or continuous parameters. However, class prediction is the most relevant prediction problem in practice. The interpretation of results is much more intuitive for class prediction than for other prediction problems. From a medical point of view, it is often sensible to summarize more complex prediction problems such as, e.g. survival prediction or ordinal regression as binary class prediction. Moreover, we think that the model assumptions required by most survival analysis methods and methods for the prediction of continuous outcomes are certainly as questionable as the simplification into a classification problem. However, one has to be aware that transforming a general prediction problem into class prediction may lead to a loss of information, depending on the addressed medical question.

Beside some comparative studies briefly recalled in Section 2, several review articles on particular aspects of classification have been published in the last five years. For example, an extensive review on machine learning in bioinformatics including class prediction can be found in Larranaga et al. (2006), whereas Chen (2007) reviews both class comparison and class prediction with emphasis on univariate

Correspondence:



Copyright in this article, its metadata, and any supplementary data is held by its author or authors. It is published under the Creative Commons Attribution By licence. For further information go to: <http://creativecommons.org/licenses/by/3.0/>.

validation policy developed by the Sylvia Lawry Centre for Multiple Sclerosis Research (Daumer et al. 2007).

Choice of the validation data set

The impact of the reported classifier accuracy in the medical community increases with the differences between validation data set and open data set. For example, it is much more difficult to find similar results (and thus much more impressive when such results are found) on a validation data set collected in a different lab at a different time and for patients with different ethnic, social or geographical background than in a validation set drawn at random from a homogeneous data set at the beginning of the analysis. An important special case is when the learning and validation sets are defined chronologically. In this scheme, the first recruited patients are considered as learning data and used for classifier construction and selection *before* the validation data set is collected, hence warranting that the validation data remain unopened until the end of the learning phase. Obviously, evaluating a classifier on a validation data set does not provide an estimate of the error rate which would be obtained if both learning and validation data set were used for learning the classifier. However, having an untouched validation data set is the only way to simulate prediction of new data. See Simon (2006) for considerations on the validation problem in concrete cancer studies and Buyse et al. (2006) for details on the validation experiments conducted to validate the well-known 70-gene signature for breast cancer outcome prediction by (van'tVeer et al. 2002).

Furthermore, if the learning and test sets are essentially different (e.g. from a geographical or technical point of view), bad performance may be obtained even with a classifier that is optimal with respect to the learning data. The error rate on the validation set increases with i) the level of independence between Y and \mathbf{X} in both learning and validation sets, ii) the difference between the joint distribution \mathbf{F} of Y and \mathbf{X} in the learning and validation sets, iii) the discrepancy between the optimal Bayes classifier and the constructed classifier. Whereas the components i) and iii) are common to all methods of accuracy estimation, component ii) is specific to validation schemes in which “validation patients” are different from “learning patients”.

Thus, it does make a difference whether the learning and test sets are (random) samples from the same original data set, or if the test set is sampled, e.g. in a different center in a multi-center clinical trial or at a different point in time in a long-term study. The first case—ideally with random sampling of the learning and test sets—corresponds to the most general assumption for all kinds of statistical models, namely the “i.i.d.” assumption that all data in the learning and test sets are randomly drawn independent samples from the same distribution and that the samples only vary randomly from this distribution due to their limited sample size. This common distribution is often called the data generating process (DGP). A classifier that was trained on a learning sample is supposed to perform well on a test sample from the same DGP, as long as it does not overfit.

A different story is the performance of a classifier learned on one data set and tested on another one from a different place or time. If the classifier performs badly on this kind of test sample this can have different reasons: either important confounder variables were not accounted for in the original classifier, e.g. an effect of climate when the classifier is supposed to be generalized over different continents (cf Altman and Royston, 2000, who state that models may not be “transportable”), or—even more severe for the scientist—the DGP has actually changed, e.g. over time, which is an issue discussed as “data drift”, “concept drift” or “structural change” in the literature. In this latter case, rather than discarding the classifier, the change in the data stream should be detected (Kifer et al. 2004) and modelled accordingly—or in restricted situations it is even possible to formalize conditions under which some performance guarantees can be proved for the test set (Ben-David et al. 2007).

When on the other hand the ultimate goal is to find a classifier that is generalizable to all kinds of test sets, including those from different places or points in time, as a consequence we would have to follow the reasoning of “Occam’s razor” for our statistical models: the sparsest model is always the best choice other things being equal. Such arguments can be found in Altman and Royston (2000) and, more drastically, in Hand (2006), who uses this argument not only with respect to avoiding overfitting and the inclusion of too many covariates, but also, e.g. in favor of linear models as opposed to recursive partitioning, where it is, however, at least questionable from our point of

view, if the strictly linear, parametric and additive approach of linear models is really more “sparse” than, e.g. simple binary partitioning.

Recommendations

With respect to the first question posed at the beginning of this subsection we therefore have to conclude that there are at least one clinical and one—if not a dozen—statistical answers, while for the second question we have a clear recommendation. Question 2 should be addressed based on the open learning data set only via cross-validation, repeated cross-validation, Monte-Carlo cross-validation or bootstrap approaches. The procedure is as follows:

1. Define N_{iter} pairs of learning and test sets ($\mathbf{I}^{(j)}$, $\mathbf{t}^{(j)}$), $j = 1, \dots, N_{iter}$ following one of the evaluation strategies described in Section 4 (LOOCV, CV, repeated CV, MCCV, bootstrap, etc). For example, in LOOCV, we have $N_{iter} = n$.
2. For each iteration ($j = 1, \dots, N_{iter}$), repeat the following steps:
 - Construct classifiers based on $\mathbf{I}^{(j)}$ using different methods M_1, M_2, \dots, M_q successively, where M_r ($r = 1, \dots, q$) is defined as the combination of the variable selection method (e.g. univariate Wilcoxon-based variable selection), the number of selected variables (e.g. $\tilde{p} = 50, 100, 500$) and the classification method itself (e.g. linear discriminant analysis).
 - Predict the observations from the test set $\mathbf{t}^{(j)}$ using the constructed classifiers $C_{\mathbf{D}_1^{(j)}}^{M_1}, \dots, C_{\mathbf{D}_1^{(j)}}^{M_q}$ successively.
3. Estimate the error rate based on the chosen procedure for all methods M_1, \dots, M_q successively.
4. Select the method yielding the smallest error rate. It should then be validated using the observations from the independent validation set.

A critical aspect of this procedure is the choice of the “candidate” methods M_1, \dots, M_q . On the one hand, trying many methods increases the probability to find a method performing better than the other methods “by chance”. On the other hand, obviously, increasing the number of methods also increases the chance of finding the right method, i.e. the method that best reflects the true data structure and is thus expected to show good performance on independent new data as well.

CV, MCCV or bootstrap procedures might also be useful in medical studies for accuracy estimation, but their results should not be over-interpreted. They give a valuable preview of classifier accuracy when the collected data set is still not large enough for putting aside a large enough validation set. In this case, one should adopt one of the following approaches for choosing the method parameters:

- Using the default parameter values.
- Selecting parameter values by internal cross-validation (or a related approach) within each iteration (Varma and Simon, 2006). The computational complexity is then n^2 , which makes it prohibitive if the chosen classification method is not very fast, especially when it involves variable selection.
- Selecting parameter values based on solid previous publications analyzing other data sets.

Trying several values and reporting only the error rate obtained with the optimal value is an incorrect approach. Studies and discussions on the bias induced by this approach can be found in Varma and Simon (2006); Wood et al. (2007). In all cases, it should be mentioned that such an analysis does not replace an independent validation data set.

6 Summary and Outlook

For fair evaluation of classifiers, the following rules should be taken into account.

- The constructed classifier should ideally be tested on an independent validation data set. If impossible (e.g. because the sample is too small), the error rate should be estimated with a procedure which tests the classifier based on data that were not used for its construction, such as cross-validation, Monte-Carlo cross-validation or bootstrap sampling.
- Variable selection should be considered as a step of classifier construction. As such, it should be carried out using the learning data only.
- Whenever appropriate, sensitivity and specificity of classifiers should be estimated. If the goal of the study is, e.g. to reach high sensitivity, it is important to design the classifier correspondingly.

Note that both the construction and the evaluation of prediction rules have to be modified if the outcome is not, as assumed in this paper, nominal, but ordinal, continuous or censored. While ordinal variables are

very difficult to handle in the small sample setting and thus often dichotomized, censored survival variables can be handled using specific methods coping with the $n \ll p$ setting. Since censoring makes the use of usual criteria like the mean square error impossible, sophisticated evaluation procedures have to be used, such as the Brier score (see Van Wieringen et al. (2007) for a review of several criteria).

Another aspect that has not been treated in the present paper because it would have gone beyond its scope is the stability of classifiers and classifier assessment. For instance, would the same classifier be obtained if an observation were removed from the data set? How does an incorrect response specification affect the classification rule and the estimation of its error rate? Further research is needed to answer these most relevant questions, which affect all microarray studies.

Further research should also consider the fact that due to the many steps involved in the experimental process, from hybridization to image analysis, even in high quality experimental data severe measurement error may be present (see, e.g. Rocke and Durbin, 2001; Tadesse et al. 2005; Purdom and Holmes, 2005). As a consequence, prediction and diagnosis no longer coincide, since prediction is usually still based on the mismeasured variables, while diagnosis tries to understand the material relations between the true variables. While several powerful procedures to correct measurement error are available for regression models (see, e.g. Wansbeek and Meijer, 2000; Cheng and Ness, 1999; Carroll et al. 2006; Schneeweiß and Augustin, 2006, for surveys considering linear and nonlinear models, respectively), in the classification context well-founded treatment of measurement error is still in its infancy.

A further problem which is largely ignored by many statistical articles is the incorporation of clinical parameters into the classifier and the underlying question of the additional predictive value of gene expression data compared to clinical parameters alone. Although “adjustment for other classic predictors of the disease outcome [is] essential” (Ntzani and Ioannidis, 2003), this problem is largely ignored by most methodological articles. Specific evaluation and comparison strategies have to be developed to answer this question.

Acknowledgement

We are grateful to the two anonymous referees for their helpful comments. ALB was supported by the

Porticus Foundation in the context of the International School for Clinical Bioinformatics and Technical Medicine.

References

- Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96:6745–50.
- Altman, D.G. and Royston, P. 2000. What do we mean by validating a prognostic model? *Statistics in Medicine*, 19:453–73.
- Ambrose, C. and McLachlan, G.J. 2002. Selection bias in gene extraction in tumour classification on basis of microarray gene expression data. *Proceedings of the National Academy of Science*, 99:6562–6.
- Asyali, M.H., Colak, D. Demirkaya, O. and Inan, M.S. 2006. Gene expression profile classification: A review. *Current Bioinformatics* 1: 55–73.
- Ben-David, S., Blitzer, J., Crammer, K. and Pereira, F. 2007. Analysis of Representations for Domain Adaptation. *Advances in Neural Information Processing Systems 20*. Cambridge, MA: MIT Press.
- Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, Z. 2000. Tissue classification with gene expression profiles. *Journal of Computational Biology*, 7:559–84.
- Berger, J.O. 1980. *Statistical Decision Theory and Bayesian Analysis* (2nd edition). New York: Springer.
- Berrai, D., Bradbury, I. and Dubitzky, W. 2006. Avoiding model selection bias in small-sample genomic datasets. *Bioinformatics*, 22:1245–50.
- Binder, H. and Schumacher, M. 2007. Adapting prediction error estimates for biased complexity selection in high-dimension bootstrap samples. *Technical Report, University of Freiburg*, Nr. 100.
- Bo, T.H. and Jonassen, I. 2002. New feature subset selection procedures for classification of expression profiles. *Genome Biology*, 3:R.17.
- Bomprezzi, R., Ringnér, M., Kim, S., Bittner, M.L., Khan, J., Chen, Y., Elkahoul, A., Yu, A., Bielekova, B., Meltzer, P.S., Martin, R., McFarland, H.F. and Trent, J.M. 2003. Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease. *Human Molecular Genetics*, 12:2191–9.
- Boulesteix, A.L. 2004. PLS dimension reduction for classification with microarray data. *Statistical Applications in Genetics and Molecular Biology*, 3:33.
- Boulesteix, A.L. 2006. Reader’s reaction to ‘Dimension reduction for classification with gene expression microarray data’ by Dai et al. (2006). *Statistical Applications in Genetics and Molecular Biology*, 5:16.
- Boulesteix, A.-L. 2007. WilcoxCV: An R package for fast variable selection in cross-validation. *Bioinformatics*, 23:1702–4.
- Boulesteix, A.L. and Strimmer, K. 2007. Partial Least Squares: A versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8:32–44.
- Bradley, R.A. and Terry, M.E. 1952. Rank analysis of incomplete block designs, I. The method of paired comparisons. *Biometrika*, 39:324–45.
- Braga-Neto, U. and Dougherty, E.R. 2004. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20:374–80.
- Breiman, L. 1996. Bagging predictors. *Machine Learning*, 24:123–40.
- Breiman, L. 2001. Random Forests. *Machine Learning*, 45:5–32.
- Buysse, M., Loi, S., van’t Veer, L., Viale, G., Delorenzi, M., Glas, A.M., Saghatchian d’Assignies, M., Bergh, J., Lidereau, R., Ellis, P., Harris, A., Bogaerts, J., Therasse, P., Floore, A., Amakrane, M., Piette, F., Rutgers, E., Sotiriou, C., Cardoso, F. and Piccart, M.J. 2006. Validation and Clinical Utility of a 70-Gene Prognostic Signature for Women With Node-Negative Breast Cancer. *Journal of the National Cancer Institute*, 98:1183–92.
- Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd edition). New York: Chapman and Hall/CRC.

- Chen, J.J. 2007. Key aspects of analyzing microarray gene-expression data. *Pharmacogenomics*, 8:473–82.
- Cheng, C.L. and Van Ness, J.W. 1999. *Statistical Regression with Measurement Error*. London: Arnold.
- Chianga, I.-J. and Hsub, J.Y.-J. 2002. Fuzzy classification trees for data analysis. *Fuzzy Sets and Systems*, 130:87–99.
- Culhane, A., Thioulouse, J., Perriere, G. and Higgins, D.G. 2005. MADE4: an R package for multivariate analysis of gene expression data. *Bioinformatics*, 21:2789–90.
- Daumer, M., Held, U., Ickstadt, K., Heinz, M., Schach, S. and Ebers, G. 2007. Reducing the probability of false positive research findings by prepublication validation. *Nature Precedings*.
- DeLong, E.R., DeLong, D. and Clarke-Pearson, D.L. 1988. Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44:837–45.
- Detting, M. 2004. BagBoosting for tumor classification with gene expression data. *Bioinformatics*, 20:3583–93.
- Detting, M. and Bühlmann, P. 2003. Boosting for tumor classification with gene expression data. *Bioinformatics*, 19:1061–9.
- Diaz-Uriarte, R. and Alvarez de Andrés, S. 2006. Gene selection and classification of microarray data using random forests. *BMC Bioinformatics*, 7:3.
- Ding, B. and Gentleman, R. 2005. Classification using generalized partial least squares. *Journal of Computational and Graphical Statistics*, 14:280–98.
- Dudoit, S., Fridlyand, J. Speed, T.P. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97:77–87.
- Dupuy, A. and Simon, R. 2007. Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting. *Journal of the National Cancer Institute*, 99:147–57.
- Efron, B. and Gong, G. 1983. A leisurely look at the bootstrap, the Jackknife and cross-validation. *The American Statistician*, 37:36–48.
- Efron, B. and Tibshirani, R. 1997. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92:548–60.
- Fort, G. and Lambert-Lacroix, S. 2005. Classification using partial least squares with penalized logistic regression. *Bioinformatics*, 21:1104–11.
- Freund, Y. and Schapire, R.E. 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55:119–39.
- Fu, W.J., Carroll, R.J. and Wang, S. 2005. Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics*, 21:1979–86.
- Gerds, T.A. and Schumacher, M. 2007. Efron-type measures of prediction error for survival analysis. *Biometrics*, 63:1283–7.
- Ghadimi, B.M., Grade, M., Difilippantonio, M.J., Varma, S., Simon, R., Montagna, C. and Fuzesi, L. 2005. Effectiveness of gene expression profiling for response prediction of rectal adenocarcinomas to pre-operative chemoradiotherapy. *Journal of Clinical Oncology*, 23:1826–38.
- Ghosh, D. 2003. Penalized discriminant methods for the classification of tumors from gene expression data. *Biometrics*, 59:992–1000.
- Goeman, J. 2007. An efficient algorithm for L1 penalized estimation. www.msbi.nl/goeman preprint.
- Goldberg, D.E. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. New York: Addison-Wesley.
- 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*, 286:531–7.
- Guo, Y., Hastie, T. and Tibshirani, R. 2007. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8:86–100.
- Hand, D.J. 2006. Classifier technology and the illusion of progress. *Statistical Science*, 21:1–14.
- Hanley, J.A. and McNeil, B.J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36.
- Hastie, T., Tibshirani, R. and Friedman, J.H. 2001. *The Elements of Statistical Learning*. New York: Springer-Verlag.
- Hold, D. and Smith, T.M.F. 1979. Post stratification. *Journal of the Royal Statistical Society: Series A*, 142:33–46.
- Hornik, K. and Meyer, D. 2007. Deriving consensus rankings from benchmarking experiments. *Advances in Data Analysis (Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V., Freie Universität Berlin, March 8–10, 2006.) Studies in Classification, Data Analysis, and Knowledge Organization*. Ed. R. Decker and H.-J. Lenz Springer, 163–70.
- Hothorn, T., Leisch, F., Zeileis, A. and Hornik, K. 2005. The Design and Analysis of Benchmark Experiments. *Journal of Computational and Graphical Statistics*, 14:675–99.
- Huang, X., Pan, W., Grindle, S., Han, X., Chen, Y., Park, S.J., Miller, L.W. and Hall, J. 2005. A comparative study of discriminating human heart failure etiology using gene expression profiles. *BMC Bioinformatics*, 6:205.
- Ioannidis, J.P. 2005. Microarrays and molecular research: noise discovery? *The Lancet*, 365:454–55.
- Jäger, J., Sengupta, R. and Ruzzo, W.L. 2003. Improved gene selection for classification of microarray. *Proceedings of the 2003 Pacific Symposium on Biocomputing*, 53–64.
- Jeffery, I.B., Higgins, D.G. and Culhane, A.C. 2006. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, 7:359.
- Kifer, D., Ben-David, S., Gehrke, J., Nascimento, M.A., Özsu, M.T., Kossman, D., Miller, R.J., Blakeley, J.A. and Schiefer, K.B. 2004. Detecting Change in Data Streams. Morgan Kaufmann. *Proceedings of the Thirtieth International Conference on Very Large Data Bases, Toronto, Canada, August 31–September 3, 2004*. Ed. 180–91.
- Kohavi, R. and Frasca, B. 1994. Useful feature subsets and rough set reducts. *Proceedings of The Third International Workshop on Rough Sets and Soft Computing, San Jose, California, U.S.A.*, 310–17.
- Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armananzas, R., Santafe, G., Perez, A. and Robles, V. 2006. Machine Learning in bioinformatics. *Briefings in Bioinformatics*, 7:86–112.
- Lee, J., Lee, J., Park, M. and Song, S. 2005. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics and Data Analysis*, 48:869–85.
- Li, D. and Zhang, W. 2006. Gene selection using rough set theory. *Lecture Notes in Computer Science: Rough Sets and Knowledge Technology*, 778–85.
- Man, M.Z., Dyson, G., Johnson, K. and Liao, B. 2004. Evaluating methods for classifying expression data. *Journal of Biopharmaceutical Statistics*, 14:1065–84.
- Molinaro, A., Simon, R. and Pfeiffer, R.M. 2005. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21:3301–7.
- Natsoulis, G., El Ghaoui, L., Lanckriet, G.R.G., Tolley, A.M., Leroy, F., Dunleo, S., Eynon, B.P., Pearson, C.I., Tugendreich, S. and Jarnagin, K. 2005. Classification of a large microarray data set: Algorithm comparison and analysis of drug signatures. *Genome Research*, 15:724–36.
- Nguyen, D.V. and Rocke, D. 2002. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18:39–50.
- Ntzani, E.E. and Ioannidis, J.P.A. 2003. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *The Lancet*, 362:1439–44.
- Ooi, C.H. and Tan, P. 2003. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, 19:37–44.

- Opgen-Rhein, R. and Strimmer, K. 2007. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statistical Applications in Genetics and Molecular Biology*, 6:9.
- Park, M.Y. and Hastie, T. 2007. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B.*, 69:659–77.
- Pawlak, Z. 1991. Rough Sets: Theoretical Aspects of Reasoning About Data. Dordrecht: Kluwer Academic Publishers.
- Purdom, E. and Holmes, S.P. 2005. Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 4: Article 16.
- R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rimkus, C., Friederichs, J., Boulesteix, A.L., Theisen, J., Mages, J., Becker, K., Nekarda, H., Rosenberg, R., Janssen, K.P. and Siewert, J.R. 2008. Microarray-based prediction of tumor response to neoadjuvant radiochemotherapy of patients with locally advanced rectal cancer. *Clinical Gastroenterology Hepatology*, 6:53–61.
- Ripley, B.D. 1996. Pattern Recognition and Neural Networks. Cambridge, U.K.: Cambridge University Press.
- Rocke, D. and Durbin, B. 2001. A Model for Measurement Error for Gene Expression Arrays. *Journal of Computational Biology*, 8(6):557–69.
- Romualdi, C., Campanaro, S., Campagna, D., Celegato, B., Cannata, N., Toppo, S., Valle, G. and Lanfranchi, G. 2003. Pattern recognition in gene expression profiling using DNA array: a comparison study of different statistical methods applied to cancer classification. *Human Molecular Genetics*, 12:823–36.
- Ruschhaupt, M., Huber, W., Poustka, A. and Mansmann, U. 2004. A compendium to ensure computational reproducibility in high-dimensional classification tasks. *Statistical Applications in Genetics and Molecular Biology*, 3:37.
- Schäfer, H. Efficient confidence bounds for ROC curves. *Statistics in Medicine* 13 (1994): 1551–61.
- Schneeweiß, H. and Augustin, T. Some recent advances in measurement error models and methods. *Allgemeines Statistisches Archiv—Journal of the German Statistical Association* 90 (2006): 183–197; also printed in: Hubler, O. and Frohn, J. eds. (2006): Modern Econometric Analysis—Survey on Recent Developments, 183–98.
- Simon, R. 2006. Development and validation of therapeutically relevant multi-gene biomarker classifiers. *Journal of the National Cancer Institute*, 97:866–7.
- Simon, R., Radmacher, M.D., Dobbin, K. and McShane, L.M. 2003. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, 95:14–8.
- Smolinski, T.G., Buchanan, R., Boratyn, G.M., Milanova, M. and Prinz, A.A. 2006. Independent Component Analysis-motivated approach to classificatory decomposition of cortical evoked potentials. *BMC Bioinformatics*, 7:S8.
- Smyth, G. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:3.
- Soukup, M., Cho, H. and Lee, J.K. 2005. Robust classification modeling on microarray data using misclassification penalized posterior. *Bioinformatics*, 21:i423–i30.
- Soukup, M. and Lee, J.K. 2004. Developing optimal prediction models for cancer classification using gene expression data. *Journal of Bioinformatics and Computational Biology*, 1:681–94.
- Spiegelhalter, D.J. 1986. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, 5:421–33.
- Statnikov, A., Aliferis, C.F., Tsamardinos, I., Hardin, D. and Levy, S. 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21:631–43.
- Stolovitzky, G. 2003. Gene selection in microarray data: the elephant, the blind man and our algorithms. *Current Opinion in Structural Biology*, 13:370–6.
- Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:25.
- Swets, J.A. 1988. Measuring the accuracy of diagnostic systems. *Science*, 240:1285–93.
- Swiniarski, R.W. and Skowron, A. 2003. Rough set methods in feature selection and recognition. *Pattern Recognition Letters*, 24:833–49.
- Tadesse, M.G., Ibrahim, J.G., Gentleman, R., Chiaretti, S., Ritz, J. and Foa, R. 2005. Bayesian error-in-variable survival model for the analysis of genechip arrays. *Biometrics*, 61:488–97.
- Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99:6567–72.
- Trevino, V. and Falciani, F. 2006. GALGO: an R package for multivariate variable selection using genetic algorithms. *Bioinformatics*, 22:1154–6.
- Troyanskaya, O.G., Garber, M.E., Brown, P.O., Botstein, D. and Altman, R.B. 2002. Non-parametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18:1454–61.
- 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*, 415:530–6.
- VanWieringen, W., Kun, D., Hampel, R. and Boulesteix, A.-L. 2007. Survival prediction using gene expression data: a review and comparison. *Submitted*.
- Vapnik, V.N. 1995. The Nature of Statistical Learning Theory. New York: Springer.
- Varma, S. and Simon, R. 2006. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91.
- Wansbeek, T. and Meijer, E. 2000. Measurement Error and Latent Variables in Econometrics. *Amsterdam: Elsevier*.
- Wickenberg-Bolin, U., Göransson, H., Fryknäs, M., Gustafsson, M.G. and Isaksson, A. 2006. Improved variance estimation of classification performance via reduction of bias caused by small sample size. *BMC Bioinformatics*, 7:127.
- Wood, I.A., Visscher, P.M. and Mengersen, K.L. 2007. Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics*, 23:1363–70.
- Zaffalon, M. 2002. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105:5–21.
- Zaffalon, M., Wesnes, K. and Petrini, O. 2003. Reliable diagnoses of dementia by the naive credal classifier inferred from incomplete cognitive data. *Artificial Intelligence in Medicine*, 29(1–2):61–79.
- Zhang, M. and Yao, J.T. 2004. A rough sets based approach to feature selection. *Proceedings of The 23rd International Conference of NAFIPS, Banff, Canada*, 434–9.
- Zhu, J. 2004. Classification of gene expression microarrays by penalized logistic regression. *Biostatistics*, 5:427–43.

Appendix A

Overview of software implementing classification methods in R

Most methods for microarray-based classification are implemented in R (www.R-project.org) which has become the standard statistical tool for handling high-dimensional genomic data. Simple univariate variable selection might be performed, e.g. based on the t-test (`t.test`) or the Mann-Whitney test (`wilcox.test`). Usual classifiers like logistic regression (R function `glm`), linear discriminant analysis (R function `lda`), quadratic discriminant analysis (R function `qda`) are also accessible in R without loading any particular package. The same holds for PCA dimension reduction (R function `prcomp`). Here is a list of specific R packages that are of particular interest for microarray-based classification and freely available without registration.

- `pamr` package for PAM (Tibshirani et al. 2002)
- `penalized` package for penalized regression approaches: LASSO, L_2 (Goeman, 2007)
- `glmpath` package for LASSO regression (Park and Hastie, 2007)
- `rda` package for regularized discriminant analysis (Guo et al. 2007)
- `plsgenomics` package for PLS-based classification (Boulesteix, 2004; Fort and Lambert-Lacroix, 2005)
- `gpls` package for generalized partial least squares classification (Ding and Gentleman, 2005)
- `e1071` package for SVM
- `randomForest` for random forests classification (Diaz-Uriarte and de Andrés, 2006)
- `logitBoost` package for logitBoost (Dettling and Bühlmann, 2003)
- `BagBoosting` package for bagboosting (Dettling, 2004)
- `MADE4` package for classification by the “between-group analysis” (BGA) dimension reduction method (Culhane et al. 2005)
- `pdmclass` package for classification using penalized discriminant methods (Ghosh, 2003)
- `MLInterfaces` package including unifying functions for cross-validation and validation on test data in combination with various classifiers
- `MCREstimate` package for fair comparison and evaluation of classification methods (Ruschhaupt et al. 2004)

Packages including functions for gene selection are

- `genefilter` package including a function that computes t-tests quickly
- `WilcoxCV` package for fast Wilcoxon based variable selection in cross-validation (Boulesteix, 2007)
- `varSelRF` R package for variable selections with random forests (Diaz-Uriarte and de Andrés, 2006)
- `GALGO` R package for variable selection with genetic algorithms (Trevino and Falciani, 2006) (<http://www.bip.bham.ac.uk/vivo/galgo/AppNotesPaper.htm>).
- `MiPP` package to find optimal sets of variables that separate samples into two or more classes (Soukup and Lee, 2004; Soukup et al. 2005)

Appendix B

Summary of six comparison studies of classification methods

Table 2. Summary of six comparison studies of classification methods. This summary should be considered with caution, since not detailing the used variants of the considered methods.

Dudoit et al. (2002) 3 data sets MCCV 2:1	<ul style="list-style-type: none"> • Included: LDA, DLDA, DQDA, Fisher, <i>k</i>NN, trees, tree-based ensembles • Variable selection: F-statistic <p><i>Conclusion:</i> DLDA and <i>k</i>NN perform best</p>
Romualdi et al. (2003) 2 data sets CV	<ul style="list-style-type: none"> • Included: DLDA, trees, neural networks SVM, <i>k</i>NN, PAM combined with: • Variable selection/dimension reduction: PLS, PCA, soft thresholding, GA/<i>k</i>NN <p><i>Conclusion:</i> PLS transformation is recommendable, No classifier uniformly better than the other</p>
Man et al. (2004) 6 data sets LOOCV, bootstrap	<ul style="list-style-type: none"> • Included: <i>k</i>NN, PCA+LDA, PLS-DA, neural networks, random forests, SVM • Variable selection: F-statistic <p><i>Conclusion:</i> PLS-DA and SVM perform best</p>
Lee et al. (2005) 7 data sets LOOCV, MCCV 2:1	<ul style="list-style-type: none"> • Included: 21 methods (e.g. tree ensembles, SVM, LDA, DLDA, QDA, Fisher, PAM) • Variable selection: F-statistic, rank-based score, soft thresholding <p><i>Conclusion:</i> No classifier uniformly better than the other, rank-based variable selection performs best</p>
Statnikov et al. (2005) 11 data sets LOOCV, 10-fold CV	<ul style="list-style-type: none"> • Included: SVM, <i>k</i>NN, probabilistic neural networks, backpropagation neural networks • Variable selection: BSS/WSS, Golub et al. (1999), Kruskal-Wallis test <p><i>Conclusion:</i> SVM performs best</p>
Huang et al. (2005) 2 data sets LOOCV	<ul style="list-style-type: none"> • Included: PLS, penalized PLS, LASSO, PAM, random forests • Variable selection: F-statistic <p>Random forests perform slightly better</p> <p><i>Conclusion:</i> No classifier uniformly better than the other</p>