Statistical Applications in Genetics and Molecular Biology

Volume 5, *Issue* 1 2006 *Article* 16

Reader's reaction to "Dimension Reduction for Classification with Gene Expression Microarray Data" by Dai et al (2006)

Anne-Laure Boulesteix*

*Department of Medical Statistics and Epidemiology, Technical University of Munich, anne-laure.boulesteix@tum.de

Reader's reaction to "Dimension Reduction for Classification with Gene Expression Microarray Data" by Dai et al (2006)

Anne-Laure Boulesteix

Abstract

This note is a comment on the article "Dimension Reduction for Classification with Gene Expression Microarray Data" that appeared in Statistical Applications in Genetics and Molecular Biology (Dai et al., 2006).

KEYWORDS: dimension reduction, classification, microarray data

I wish to take the opportunity provided by *Statistical Applications in Genetics and Molecular Biology* to make some comments on the interesting article by Jian Dai, Linh Lieu and David Rocke entitled 'Dimension Reduction for Classification with Gene Expression Microarray Data'.

There have been hundreds of articles on microarray-based (tumor) classification from which a few tens are based on dimension reduction. Hence, the need for comparison studies is probably as least as stringent as the need for novel methods. Unfortunately, many journals require methodological innovation as a crucial publication criterion. Of course, comparison studies included in articles presenting novel methods should be considered with caution, since possibly biased in favor of the authors' approach. Thus, I believe that studies like that of Dai et al. (2006) deserve much attention and can help readers to find their way within the plethora of methods available for microarray data analysis.

Important ingredients of all credible comparison articles are the study design, the level of detail of its description and the representation of the results. In this respect, the authors provided a very good work. Summarizing the study design as five distinct steps in a separate section is a very good idea that might inspire authors of future studies. Systematic comparison tools such as the compendium by Ruschhaupt et al. (2004) might also help to improve the transparency and objectivity of comparison studies. Another nice aspect of the article by Dai et al. (2006) is the comparison of the different methods with respect to the computation time. The importance of computation times is sometimes undervalued in comparison studies, although computational aspects may contribute to make a method popular (or not), especially when cross-validation is needed to choose the meta parameters.

I agree with the authors that PLS has the highest 'performance/computation-time' ratio, probably even if some of the competing methods mentioned below are included in the study. I would like to add a remark on the use of PLS dimension reduction for classification. It has sometimes been said that PLS should not be used in this case, since it is designed for a continuous response. This criticism is addressed in a very interesting paper by Barker and Rayens (2003): they show that PLS dimension reduction with a categorical response is actually closely related to PCA performed on the between-group covariance matrix. This result can be seen as a theoretical argument in favor of the use of PLS for classification and might explain the good performance of PLS outlined by Dai et al. (2006).

Let me add some comments on a three critical issues arising from the article of Dai et al. (2006). The first point is the problem of *separation* when logistic regression is used on 'too good' predictors. If the classes are perfectly separated, likelihood converges but the odds ratios estimated by logistic regression are infinite, as first noted by Albert and Anderson (1984). In the context of microarray data, this problem is also mentioned by Boulesteix (2004) and Fort and Lambert-Lacroix

(2005). In own experiments, I remarked that classical linear discriminant analysis often performs better than logistic regression when used on the PLS components, especially in 'easy' data sets like the leukemia data. As noted by Dai et al. (2006), the PLS or SIR components discriminate between the classes better than the principal components. Thus, using, e.g., linear discriminant analysis instead of logistic regression might possibly advantage supervised methods and thus influence the results of the comparison study. The conclusions of the authors about the superiority of supervised methods would be reinforced.

The second point is related to gene selection. As the authors note in Section 2.3.2, the considered dimension reduction methods can handle a large number of genes. Variable selection introduces two types of meta parameters: the selection criterion (e.g., the t-statistic, Wilcoxon's rank sum statistic) and the number of variables to be selected. As can be seen from the results by Dai et al. (2006), the classification accuracy depends on the number of included genes when the genes are selected randomly. Dai et al. (2006) did not vary the number of selected genes based on the t-test, but it is also likely to influence the classification accuracy, though maybe not as much as when the genes are selected randomly. Since the number of included genes may be seen as a meta parameter, it should ideally be chosen by cross-validation, which makes the procedure very complicated. It turns out that the PLS dimension reduction method does *not* require any preliminary variable selection, both from a theoretical and computational point of view. Moreover, variable selection is most often unnecessary to achieve an excellent classification accuracy, at least for the data sets considered here which contain many relevant variables and 'only' a few thousands of variables. See for example the results by Boulesteix (2004) obtained without gene selection. As an unsupervised method, PCA probably benefits more from preliminary variable selection. For the SIR dimension reduction approach, variable selection is advantageous from a computational point of view, since solving the eigenvalue problem of Equation (9) might be hazardous in huge dimension. In my opinion, the variable selection aspect can be seen as a plus of the PLS approach.

My third comment is on the estimation of the covariance matrix $\mathbf{S_X}$ in the SIR approach. The empirical covariance matrix $\mathbf{S_X}$ is still used as an estimator for the true covariance matrix in most articles. However, it is not well-conditioned if the number of variables is very large compared to the sample size. This might pose some problems when solving a problem like Equation (9). Recently, Ledoit and Wolf (2003) suggested an alternative well-conditioned estimator for large-dimensional covariance matrices, which was also used by Schäfer and Strimmer (2005) for the reconstruction of genetic networks from high-dimensional microarray data. Such estimators could potentially be used for classification tasks, for instance in SIR.

At last, I would like to discuss briefly some other dimension reduction methods

that are related to PCA, PLS and SIR and have been used in the context of tumor classification with microarray data analysis. Some of them are cited by Dai et al. (2006). Culhane et al. (2002) suggest a supervised dimension reduction method which may be seen as a 'supervised PCA' procedure performed on the between-group covariance matrix. It is well-known that supervised methods perform better than PCA in the context of high-dimensional microarray data, see for instance Nguyen and Rocke (2002a) and Nguyen and Rocke (2002b) for comparative studies on PCA and PLS. Obviously, other unsupervised methods like independent component analysis (ICA) used by, e.g., Saidi et al. (2004) are likely to have the same pitfall as PCA.

From a chronological point of view, SIR is the first of the family of 'sufficient dimension reduction' methods. These methods, which include 'sliced average variance estimation' (SAVE), 'principal Hessian directions' (PHD) and 'difference of variance' (DOC) are reviewed by Cook and Lee (1999) in the context of binary classification. The SIR and SAVE approaches to sufficient dimension reduction are employed by Bura and Pfeiffer (2003) for graphical tasks and for classification. An implementation can be found in the `R` package `dr`. Chiaromonte and Martinelli (2002) suggest a two-stage dimension reduction approach combining principal component analysis and SIR. The approach taken by Antoniadis et al. (2003) is based on MAVE, a more recent related method proposed by Xia et al. (2002), which can be seen as an extension of SIR.

At last, two recent papers suggest sophisticated PLS-inspired classification procedures. These papers are based on the idea that performance may be improved by modifying PLS dimension reduction in order to adapt it to the specific case of categorical responses. The first paper (Fort and Lambert-Lacroix, 2005) proposes a two-stage method combining PLS and ridge penalized logistic regression, which is implemented in the `R` package `plsgenomics`. The second paper (Ding and Gentleman, 2005) is based on on the approach by Marx (1996) embedding partial least squares into generalized linear models. The problem of (quasi)separation is avoided by applying bias correction to the likelihood. This method is implemented in the `R` package `gpls`. All these approaches have been successfully applied to microarray-based tumor diagnosis. Their relative performance, which remains largely unknown, might be investigated in future research.

References

Albert, A., Anderson, J., 1984. On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71, 1–10.

Antoniadis, A., Lambert-Lacroix, S., Leblance, F., 2003. Efficient dimension reduction methods for tumor classification using gene expression data. Bioinformatics 19, 563–570.

Barker, M., Rayens, W., 2003. Partial least squares for discrimination. Journal of Chemometrics 17, 166–173.

Boulesteix, A. L., 2004. PLS dimension reduction for classification with high-dimensional microarray data. Statistical Applications in Genetics and Molecular Biology 3, Issue 3, Article 33.

Bura, E., Pfeiffer, R. M., 2003. Graphical methods for class prediction using dimension reduction techniques on dna microarray data. Bioinformatics 19, 1252–1258.

Chiaromonte, F., Martinelli, J., 2002. Dimension reduction strategies for analyzing global gene expression data with a response. Mathematical Biosciences 176, 123–144.

Cook, R. D., Lee, H., 1999. Dimension reduction in binary response regression. Journal of the American Statistical Association 94, 1187–1200.

Culhane, A. C., Perriere, G., Considine, E., Gotter, T., Higgins, D., 2002. Between-group analysis of microarray data. Bioinformatics 18, 1600–1608.

Dai, J. J., Lieu, L., Rocke, D., 2006. Dimension reduction for classification with gene expression data. Statistical Applications in Genetics and Molecular Biology 5, Issue 1, Article 6.

Ding, B., Gentleman, R., 2005. Classification using generalized partial least squares. Journal of Computational and Graphical Statistics 14, 280–298.

Fort, G., Lambert-Lacroix, S., 2005. Classification using partial least squares with penalized logistic regression. Bioinformatics 21, 1104–1111.

Ledoit, O., Wolf, M., 2003. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88, 365–411.

Marx, B. D., 1996. Iteratively reweighted partial least squares. Technometrics 38, 374–381.

Nguyen, D., Rocke, D. M., 2002a. Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics 18, 1216–1226.

Nguyen, D., Rocke, D. M., 2002b. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 18, 39–50.

Ruschhaupt, M., Huber, W., Poustka, A., Mansmann, U., 2004. A compendium to ensure computational reproducibility in high-dimensional classification tasks. Statistical Applications in Genetics and Molecular Biology 3, Issue 1, Article 37.

Saidi, S. A., Holland, C. M., Kreil, D. P., MacKay, D. J., Charnock-Jones, D. S., Print, C. G., Smith, S. K., 2004. Independent component analysis of microarray data in the study of endometrial cancer. Oncogene 23, 6677–6683.

Schäfer, J., Strimmer, K., 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology 4, Issue 1, Article 32.

Xia, Y. C., Tong, H., Li, W. K., Zhu, L. X., 2002. An adaptive estimation of dimension reduction space. Journal of the Royal Statistical Society B 64, 363–388.