

On Novel Approaches for Classification

A Proposal for an Interdisciplinary Debate

A. Ziegler

Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Lübeck, Germany

A group of four statistical papers has been published in this issue of *Methods of Information in Medicine*, all dealing with the problem of classification in different areas of application and at different levels.

Classification, also termed class prediction, is one of the major challenges in many areas of application. These do not only include clinical medicine [1–6] but also econometrics, where credit scoring is one of the standard tasks, and genetics [7–8]. Some recent examples published in this journal include approaches for differentiating between malign, benign, and normal tissue in women undergoing digital mammography [9], the prognosis of patients after stroke [10], the prediction of prolonged hospital stay in an intensive care unit [11], or the detection of glaucomas [3].

Although this task is very important for clinical routine, some of the classification rules perform so badly that they cannot be recommended to be used in any real life application [12]. As this task is important and since some classifiers show poor performance, scientists from many different disciplines have developed and still develop novel classifiers, new methods for variable selection and approaches for investigating the validity of the classification scheme. However, Hand [13] has disenchanted many scientists by showing that great

improvement of classifiers cannot be expected if some reasonable classifier is already available. Nevertheless, small improvements are possible, and simpler classification schemes may also be derived. The latter could consist in fewer variables that are required for classification while the accuracy is maintained.

An important aspect though is that often a fair comparison of methods is not intended. In fact, researchers have to demonstrate the superiority of their novel classification rule over a standard classifier to get a paper published. To avoid unfair comparison, the recommendations on the design of benchmark experiments are extremely helpful [14]. An alternative is that different classification schemes are derived by different researchers who are experts for the specific classifier [10]. This idea has led to interesting workshops or competitions in many different areas of applications, such as the Genetic Analysis Workshops [15] or the international competition on mass spectrometry proteomic diagnosis [16].

Furthermore, standard statistics can be used to judge whether a novel classification scheme performs significantly better than the standard classifier [17]. In detail, if two different classification schemes are applied to the same data set, each subject can be judged to be correctly classified by each of the two classifiers. Subsequently, a 2×2 table of correct classification can be created with classifier 1 representing rows and classifier 2 representing columns. The equality of the proportions of correct classifications of the two classifiers can be tested using standard tests for dependent samples, such as McNemar's test, and valid confidence intervals can be derived [18]. However, standard benchmark experiments and statistical comparisons of two different

Methods Inf Med 2010; 49: 205–206

Correspondence to:

Andreas Ziegler
Institut für Medizinische Biometrie und Statistik
Universität zu Lübeck
Universitätsklinikum Schleswig-Holstein
Campus Lübeck
Maria-Goeppert-Str. 1
23562 Lübeck
Germany
E-mail: ziegler@imbs.uni-luebeck.de

classification schemes are hard to find in scientific papers.

This critique also applies to some of the papers on classification schemes and variable selection procedures published in this issue of *Methods of Information in Medicine*. Furthermore, while some of the papers report some standard measures of diagnostic accuracy, such as sensitivity, specificity, accuracy, or predictive values, they typically do not provide confidence intervals which could be used to assess the variability of the classification performance. In fact, most of these measures are proportions, and it would be very simple to provide valid confidence intervals using standard recommendations [19].

Finally, it would be very helpful if researchers all speak the same language of science. If we all use the same terminology, dictionaries such as “artificial neural networks – statistics” [20] would be superfluous. This can be achieved in interdisciplinary work which, of course, requires a substantial amount of willingness to discuss scientific problems with colleagues who speak a different language because they have a different background. Two such fora of great success are the German Society of Medical Informatics, Biometry and Epidemiology (gmde) and the German Umbrella Organization of Statistics (DAGStat) who both meet on a regular basis. I would be most grateful if my opinion which I have expressed in this edi-

torial is the starting point for a discussion in *Methods of Information in Medicine*, and Letters to the Editor are very welcome.

References

1. Stollhoff R, Sauerbrei W, Schumacher M. An experimental evaluation of boosting methods for classification. *Methods Inf Med* 2010; 49 (3): 219–229 (this issue).
2. Fraiwan L, Lweesy K, Khasawneh N, Fraiwan M, Wenz H, Dickhaus H. Classification of sleep stages using multi-wavelet time frequency entropy and LDA. *Methods Inf Med* 2010; 49 (3): 230–237 (this issue).
3. Adler W, Peters A, Lausen B. Comparison of classifiers applied to confocal scanning laser ophthalmoscopy data. *Methods Inf Med* 2008; 47 (1): 38–46.
4. Tortajada S, Garcia-Gomez JM, Vicente J, Sanjuan J, de Frutos R, Martin-Santos R, et al. Prediction of postpartum depression using multilayer perceptrons and pruning. *Methods Inf Med* 2009; 48 (3): 291–298.
5. Castellani U, Cristiani M, Daducci A, Farace P, Marzola P, Murino V, et al. DCE-MRI data analysis for cancer area classification. *Methods Inf Med* 2009; 48 (3): 248–253.
6. Karvounis EC, Tsipouras MG, Papaloukas C, Tsilikakis DG, Naka KK, Fotiadis DI. A non-invasive methodology for fetal monitoring during pregnancy. *Methods Inf Med* 2010; 49 (3): 238–253 (this issue).
7. Chuang L-Y, Yang C-S, Wu K-C, Yang C-H. Correlation-based gene selection and classification using Taguchi-BPSO. *Methods Inf Med* 2010; 49 (3): 254–268 (this issue).
8. Bielza C, Robles V, Larranaga P. Estimation of distribution algorithms as logistic regression regularizers of microarray classifiers. *Methods Inf Med* 2009; 48 (3): 236–241.
9. Campos LF, Silva AC, Barros AK. Independent component analysis and neural networks applied for classification of malignant, benign and normal tissue in digital mammography. *Methods Inf Med* 2007; 46 (2): 212–215.
10. Linder R, König IR, Weimar C, Diener HC, Pöppel SJ, Ziegler A. Two models for outcome prediction – a comparison of logistic regression and neural networks. *Methods Inf Med* 2006; 45 (5): 536–540.
11. Verduijn M, Peek N, Voorbraak V, de Jonge E, de Mol BA. Modeling length of stay as an optimized two-class prediction problem. *Methods Inf Med* 2007; 46 (3): 352–359.
12. Janssens AC, Gwinn M, Bradley LA, Oostra BA, van Duijn CM, Khoury MJ. A critical appraisal of the scientific basis of commercial genomic profiles used to assess health risks and personalize health interventions. *Am J Hum Genet* 2008; 82 (3): 593–599.
13. Hand D. Classifier technology and the illusion of progress. *Stat Sci* 2006; 21 (1): 1–14.
14. Hothorn T, Leisch F, Zeileis A, Hornik K. The design and analysis of benchmark experiments. *J Comput Graph Statist* 2005; 14 (3): 675–699.
15. Cupples LA, Beyene J, Bickeböller H, Daw EW, Fallin MD, Gauderman WJ, et al. Genetic Analysis Workshop 16: Strategies for genome-wide association study analyses. *BMC Proc* 2009; 3 (Suppl 7): S1.
16. Mertens B. International competition on mass spectrometry proteomic diagnosis. *Stat Appl Genet Mol Biol* 2008; 7 (2): article1.
17. König IR, Malley JD, Pajevic S, Weimar C, Diener H-C, Ziegler A, et al. Patient-centered yes/no prognosis using learning machines. *Int J Data Min Bioinform* 2008; 2 (4): 289–341.
18. Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. *Stat Med* 1998; 17 (22): 2635–2650.
19. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med* 1998; 17 (8): 857–872.
20. Arminger G, Enache D. Statistical models and artificial neural networks. In: Bock HH, Polasek W, editors. *Data Analysis and Information Systems*. Heidelberg: Springer; 1996. pp 243–260.