Sequential Classification for Microarray and Clinical Data

Guenter Tusch
Grand Valley State University, 1 Campus Drive, Allendale, MI 49401 and
Grand Rapids Medical Education & Research Center for Health Professions (MERC),
1000 Monroe NW, Grand Rapids, MI 49503
tuschg@gvsu.edu

Abstract

Sequential classification uses in a stepwise process only part of the data (evidence) for partial classification, i.e., classifying only objects with sufficient evidence and leaving the rest unclassified. In the following steps the procedure is repeated using additional data until all objects are classified. This is especially useful when data become available only at certain points in time, as in surgical decision making, i.e., clinical patient data, lab data, or cDNA microarray expression data from tissue samples become available before, during and after the operation. Surgeons are interested in classifying patients into low or high risk groups, which might need special measures, e.g., prolonged intensive care.

1. Introduction

Sequential classification is especially useful when data become available only at certain time points, as in clinical decision making. In surgery, for instance, tumor markers are available before the operation, tissue samples at the time of surgery, and again tumor markers with hopefully changed values are available in the clinical course after surgery. Data sources can be clinical patient data, lab data, or cDNA microarray expression data from tissue samples. Surgeons are interested in classifying patients according to their risk into groups. High risk patients need special measures, e.g., extended intensive care, which might be detrimental for low risk patients and also for reasons of cost. When applying statistical methods in sequential risk classification to surgery, there is an inherent dilemma: For ethical reasons a risk classification should have the lowest possible error rate. However, only the last decision step in a sequential classification procedure includes all possible available data and has therefore the lowest error rate. On the other hand, the most complete data set represents the smallest possible sample, because data stemming from clinical decisions to stop early are missing in subsequent steps. This means that the sample is always biased if not based on a randomized clinical trial, which itself may be unethical when exposing patients to a procedure that is considered inappropriate in their specific situation. If, for instance, previous evidence would have indicated that the patient was not a suitable candidate for surgery at all, the decision after surgery based on all available data, however, would be worse for the patient than a classification based on only the preoperative data indication a very high risk, so that doctors decide not to operate at all. Therefore, sequential classification actually means to accept a higher error rate than would be possible with a one time classification. The advantage, however, is that the physician can chose an acceptable error rate within a window depending on the data. Here lower error rate means later decision on average, higher error rate earlier decision. Another advantage is that cases with sufficient evidence can be decided early, resulting in a more appropriate and ethical clinical decision.

2. Method

The proposed approach is based on the idea to use the activation function of a single output neuron of an artificial neural network to generate a continuous score at each classification step. This scores used only data values that are available at that point in time. Then two thresholds each are determined to decide if a patient can be classified (above or below thresholds) or not (in between thresholds). If not classified, the classification is attempted at the next step using the very same approach. At the final step a classification will take place in any case (resulting in one threshold only). The (conditional) error rates in each step are summed up and the proposed procedure makes sure that on average the anticipated total conditional error rates will be achieved. The proposed classification procedure is based on, but not restricted to, K different feedforward neural networks (multilayer perceptron and radial basis networks) which form continuous scores for each classification step. For this model non-parametric procedures are developed to guarantee that the whole procedure maintains conditional error rates (based on Brofitt et al.[1]). Our proposal is a search for cut-off points on each step by non-linear optimization constrained to the conditional errors. The resulting rule is based on two cut-off points of the scores at every step k<K and one for the last step. In theory the method gives the same results as backward induction, because both methods are based on an optimal Bayes rule maintaining the error constraints. (See also Tusch[2].) The conditional error rates can be determined by a surgeon independently of each other. The range of possible values, however, depends on the quality of the available data.

3. Results of Evaluation

The advantages of the methods are demonstrated on liver transplantation data from the Hanover Medical School hospital, Hanover, Germany, and HCC cDNA microarray data from the Stanford data base.

The clinical data included 318 patients with liver resection or transplantation presented with hepatocellular carcinoma. Excluded from the study were patients with mixed hepato and cholangiocellular carcinoma. The observations were based on the first operation at the Hanover Medical School hospital. Initially, some 50 variables including demographic, laboratory, and other findings, were available for analysis. It was assumed that risk was related to a survival of less than two years. The sequential classification problem was formulated as follows: There are three classification steps: 1. immediately before the operation, 2. immediately after the operation, 3. in the postoperative course after the first month. Therapeutic consequences could range after step 1 from carefully selecting the donor organ to cancellation of the operation; after step 2, an optimization of the intensive care management and after step 3 the same as step 2, or a final assessment after hospitalization could be the result.

In recent years a number of expression data from microarrays like survivin, clusterin, RhoC, or HLA-DR expression showed promise as potential markers for the prognosis and prediction of outcome of hepatocellular carcinoma. (See, e.g., [3].) We generated an artificial sample combining the clinical HCC data with data from the Stanford microarray database and other public sources.

Artificial neural networks (ANNs) were used to generate the necessary scores. Model selection and regularization are the two main approaches to controlling the complexity of neural networks. For model selection the model with the minimum generalization error estimate was selected, which best can be done by bootstrapping for non-linear models, but Schwarz's Bayesian information criterion (BIC) is reasonable effective for larger samples and much cheaper than bootstrapping. It is also reasonable to check if all connections in the network are necessary. One way of regularization is weight decay that can be viewed as penalized ridging. Regularization reduces to ridge regression in the case of linear models. For medical reasons and reasons of adequate comparison to LDA only *augmented* multilayer perceptrons (AMLP) and radial basis function models (ARBF) were considered.

4. Conclusion

The proposed procedure, if properly implemented combines clinical and cDNA microarray data in a useful way and can lead in most cases to an earlier clinical decision. Therefore, it can help to improve patient management, and also cutting costs in hospitals.

5. References

[1] J.D. Brofitt, R.H. Randles, R.V. Hogg. “Distribution-free partial discriminant analysis”, *J Amer Statist Assoc,* 1976, 71(356), pp. TMS 934-939.

[2] G. Tusch “An optimization model for sequential decision-making - applied to risk prediction after liver resection and transplantation”, *Proc. AMIA Symposium 1999*, pp. 425-429.

[3] K. Matoba et al.“Tumor HLA-DR expression linked to early intrahepatic recurrence of hepatocellular carcinoma”, *Int J Cancer*, 2005, 115(2), pp. 231-40.