Colloquium Biometricum 46
2016, 1−8

LOGIC REGRESSION FOR DIAGNOSTIC CLASSIFICATION BY MEANS OF BIOMARKER PANELS

Jan Bocianowski[1], Kamila Nowosad[2]

[1] Poznań University of Life Sciences, Department of Mathematical and Statistical Methods, Wojska Polskiego 28, 60-637 Poznań, Poland, e-mail: jboc@up.poznan.pl,
[2] Wrocław University of Environmental and Life Sciences, Department of Genetics, Plant Breeding and Seed Production, Grunwaldzki 24A, 53-363 Wrocław, Poland, e-mail: kamila.nowosad@up.wroc.pl

Summary

Biomarkers can indicate a variety of health or disease characteristics, including the level or type of exposure to an environmental factor, genetic susceptibility, genetic responses to exposures, markers of subclinical or clinical disease, or indicators of response to therapy. In this paper we presented using the logic regression for diagnostic classification by means of biomarker panels. The sample collective comprised 389 highly characterized rheumatoid artritis patients and 390 controls, composed of 200 healthy and 190 osteoarthritis samples. The predictive panel consisted of the most promising 11 biomarkers for the early diagnosis of rheumatoid artritis, collected in serum. Obtained results show that the Logic II is the simplest model and performs well, resulting in the lowest misclassification error rate on the test set.

Keywords and phrases: biomarkers, diagnostic, discriminant analysis, logistic model

Classification AMS 2010: 62F10, 62J02, 62F12

1. Introduction

A very promising diagnostic technique, gaining more and more ground in the last years in the pharmaceutical area, is to use molecular biomarkers in the early identification of a disease (Mishra et al., 2003; Bignotti et al., 2006; Vasan 2006; Scherzer et al., 2007; Chen et al., 2008; Lee and Wong 2009; Blennow et al., 2010; Liu et al., 2014; Chinen et al., 2015; Jickling and Sharp 2015). As single markers in the context of complex disease are usually not sensitive and/or specific enough to meet strict diagnostic conditions, the attention of clinicians has shifted to biomarker combinations, which are expected to enhance the diagnostic accuracy. One should select from a list features (initial pool of markers) a feature subset, which complies with the requests of high power to discriminate between cases and control and of parsimony as well.

Logistic regression (Hosmer and Lemeshow, 1989) is the model of choice in many medical data classification tasks. In logistic regression, the model complexity is already low, especially when no or few interaction terms and variable transformations are used. Performing variable selection is a way to reduce a model's complexity and consequently decrease the risk of overfitting (Dreiseitl and Ohno-Machado, 2002).

A variety of computer models have been developed in the area of machine learning and statistics that can be used for predicting clinical outcomes, such as logistic regression, decision trees, artificial neural networks, and Bayesian networks. Logistic regression was developed by the statistics community, whereas the remaining methods were developed by the machine-learning community. Logistic regression, a statistical fitting model, is widely used to model medical problems because the methodology is well established and coefficients can have intuitive clinical interpretations.

The aim of this paper is use the logic regression for diagnostic classification by means of biomarker panels.

2. Methods

Given an initial biomarker pool, three feature selection algorithms were developed in order to choose the biomarker combination with the highest discriminatory power. The selection method used was logic regression. This procedurę assings the patients to one of two given classes according to the corresponding logic value 0/1 of an optimal chosen logic term, which combines

some binary input variables (characteristics of individuals or dummies) from a given set by means of the logic operators and/or.

Since logic models work only with binary predictors, the necessity arises to dichotomize appropriately the quantitative explanatory, e. g. markers. Each algorithm is based on a different dichotozimation idea.

The first algorithm (**Logic I**) divides the sample population into *k* approximately equally sized clusters with respect to all features, hoping to reveal useful disease specific subpopulation structures in the data, which might help in making a reliable diagnosis. The best feature panel in each cluster is chosen by a feature selection algorithm based on **regularized discriminant analysis** (RDA). The optimal RDA-rule in each cluster is used to classify the whole sample, providing a binary input variable for the logic model bulding. The second algorithm (**Logic II**) splits any continuous feature at the thresholds given by a set of **empirical quantiles** (Koenker and Bassett, 1978). The third algorithm (**Logic III**) uses optimal thresholds estimated by **logistic regression**. The receiver operating characteristic (ROC) analysis has been developed to determine an optimal decision threshold for relative costs of false positive and false negative errors (Halpern et al., 1996). This optimal threshold is called the optimal operating point (OOP). This OOP will be of clinical use in guidelines for decision making. There are many different issues to keep in mind when determining this point such as cost issues for false-positive test results or false-negative test results, or pre-assigned thresholds for predicting a true positive or true negative.

All three algorithms are wide applied in this type researches.

3. Application

Algorithms based on logic regression were developed and validated on the data provided by Roche Diagnostics in the framework of the clinical discrimination between rheumatoid artritis (RA) patients and a pooled control collective, including osteoarthritis (OA) patients. The sample collective comprised 389 highly characterized RA-patients and 390 controls, composed of 200 healthy and 190 OA samples (Kellgren and Lawrence, 1957). The predictive panel consisted of the most promising 11 biomarkers (encoded M1-M11: C reactive protein – M1, matrix metalloproteinase 1 – M2, vascular cell adhesion molecule 1 – M3, vascular endothelial growth factor A – M4, intracelular adhesion protein-1 – M5, tumor necrosis factor receptor superfamily member 1A – M6, matrix metalloproteinase 3 – M7, human cartilage glycoprotein 39 – M8, leptin and interleukin 6 – M9, epidermal growth factor – M10 and serum amyloid A – M11) for the early diagnosis of RA, collected in serum.

The original sample was split at random into a training (520) and an independent test data set (259) in a proportion of 2:1, whereas the design was left balanced. The dichotomization and the logic model fitting steps of algorithms Logic I – Logic III, including all optimizations necessary, were carried out on the training dataset.

The test dataset provided more accurate estimates and binomial 95% confidence intervals of the true misclassification error rate, used to compare the performances of the three algorithms with respect to this particular diagnostic setting (Table 1).

Table 1. 5-fold cross-validate (C-V) estimates and test sample estimates with 95% confidence intervals for the misclassification error rates of the logic model chosen by algorithms Logic I – Logic III

Logic model	Size	Training-error rate		Test error rate	
		C-V (5-fold)	AER[#]	TS-estimate[$]	95% CI
Logic I	3	0.1116	0.1039	0.1080	(0.073; 0.153)
Logic II	2	0.1309	0.1174	0.1080	(0.073; 0.153)
Logic III	6	0.1356	0.1144	0.1280	(0.086; 0.170)

[#] AER – apparent error rate
[$] TS-estimate – test sample estimate

Optimizations of the dichotomization rules, concerning the regularization parameters and the best feature panel of each cluster in Logic I as well as the logistic cut-offs in Logic III, rely on minimizations of cross-validate (C-V) misclassification error estimates, computed by inner cross-validation loops (they are part of the algorithm).

For choosing the appropriate logic model size and providing fairer estimates of the misclassification error for the optimal models outer cross-validation loops (the algorithm is run on each C-V-training datasets) were employed. They prevent for over-optimistic results inducted by the training dataset in use. Cross-validation is an effective method for estimating the prediction error of a classifier. The right model size was chosen by comparing the overlaid plots of the 5-fold C-V misclassification error estimated and of the misclassification scores on the whole training data, computed over an adequate range of model sizes (Fig. 1). The small discrepancy between them for the model size 3 indicates stability of the model choises.

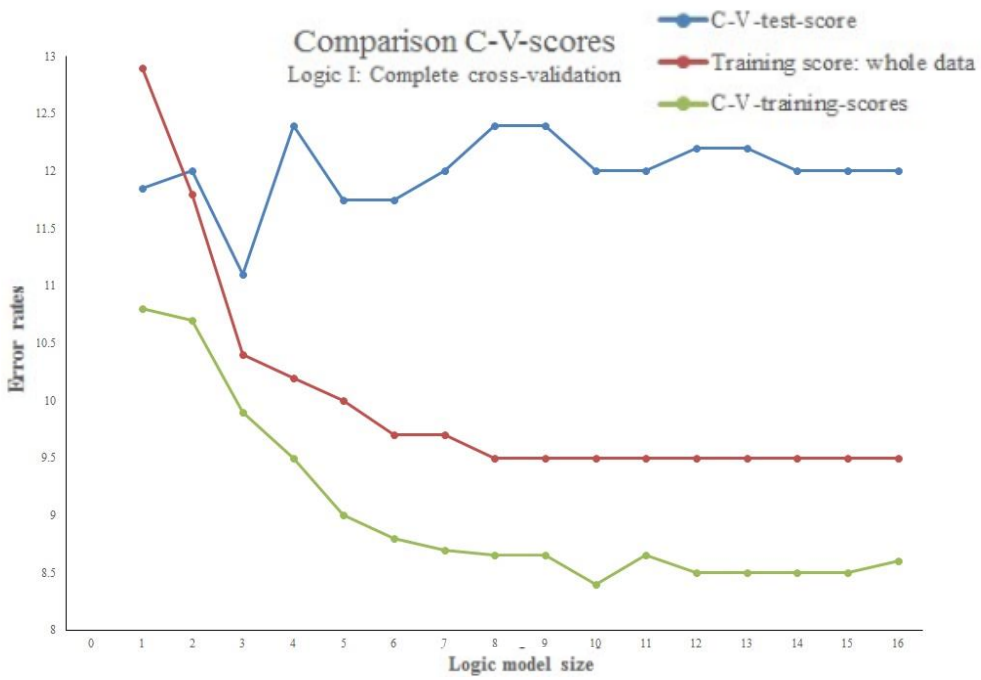

Fig. 1. The cross-validate test score of misclassification, apparent error rate and the C-V-training score of misclassification.

4. Results and Discussion

For Logic I, the training data was split into five equally sized cluster (n=104) and the RDA feature selection algorithm, performer successively on each of them, led to 5 different marker panels, each mirroring the specificity of the associated cluster. Classification by RDA with every marker panel in turn is used to reduce the continuous multivariate predictive data to a binary input variable pro cluster, D1-D5, which are then employed in the logic model building step (Fig. 2).

Logic II was performer with nine empirical quantiles. The quantiles set comprised the 0.1-0.9 percentiles. The Logic I model results in a positive test if D2 and D1 are 1 (=true) or D5 is 1.

The shortcoming of the final logic model by Logic I (Fig. 2) is that it contains seven different biomarkers, but this inconvenience can be surmounted by a sequential diagnosis. Logic I based on regularized discriminant analysis (RDA). The purpose of the regularization is to reduce the variance related to the sample-based estimates at the expense of potentially increased bias (Friedman, 1989).

Obtained result indicate that the method of regularization applied here has the potential to increase the power of discriminant analysis.

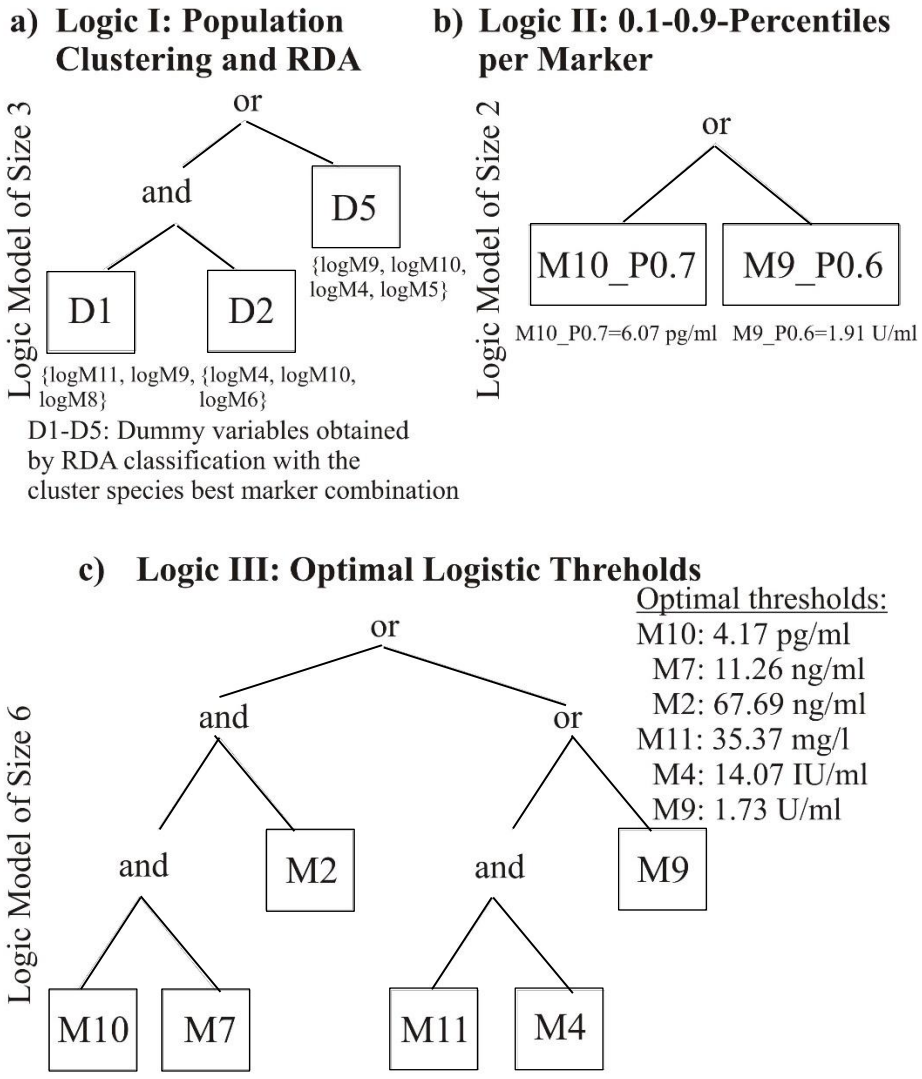

Fig. 2. Logic tree-visualization for the best logic models by: a) Logic I, b) Logic II with nine quantiles and c) Logic III.

The best logic model by Logic II is parsimonious (only markers M9 and M10) and it has an easy evaluation, while the best logic model by Logic III comprises six markers, leading again to the idea of sequential test design.

Logic I and Logic II perform best, achieving 10.81% misclassification rate on the test dataset.

Wolf et al. (2010) used logic forest to analysis of biomarker selection. They present a simulation study to comparing the performance of logic regression and logicFS with logic forest considering data with noise in the predictors, latent predictors and varying true model complexity. Diagnostic classification by means of biomarker panels in the early identification of a rheumatoid artritis and other diseases may be analysed by Student's *t* test, χ^2 test, Fisher's exact test, Pearson product-moment correlation coefficient (McMahon et al., 2014), comprehensive microRNA analysis (Murata et al., 2013), discrimination by computing the area under the receiving operator characteristic curve (Burakoff, et al., 2015) or naive Bayes, linear discriminant analysis, and support vector machines (Ding and Peng, 2005). Logistic regression used in this paper is more informative and unbiased method than should be prefer as a very important tool for diagnostic classification.

5. Conclusions

- Logic regression is potentially useful tool for diagnostic classifications.
- Logic II is the simplest model and performs well, resulting in the lowest misclassification error rate on the test set.

References

Bignotti E., Tassi R. A., Calza S., Ravaggi A., Romani C., Rossi E., Falchetti M., Odicino F.E., Pecorelli S., Santin A.D. (2006). Differential gene expression profiles between tumor biopsies and short-term primary cultures of ovarian serous carcinomas: Identification of novel molecular biomarkers for early diagnosis and therapy. *Gynecologic Oncology* 103(2), 405-416.

Blennow K., Hampel H., Weiner M., Zetterberg H. (2010). Cerebrospinal fluid and plasma biomarkers in Alzheimer disease. *Nature Reviews Neurology* 6, 131-144.

Burakoff R., Pabby V., Onyewadume L., Odze R., Adackapara C., Wang W., Friedman S., Hamilton M., Korzenik J., Levine J., Makrauer F., Cheng C., Smith H.C., Liew C.-C., Chao S. (2015). Blood-based biomarkers used to predict disease activity in Crohn's disease and ulcerative colitis. *Inflammatory Bowel Diseases* 21(5), 1132-1140.

Chen X, Ba Y., Ma L., Cai X., Yin Y., Wang K., Guo J., Zhang Y., Chen J., Guo X., Li Q., Li X., Wang W., Zhang Y., Wang J., Jiang X., Xiang Y., Xu C., Zheng P., Zhang J., Li R., Zhang H., Shang X., Gong T., Ning G., Wang J., Zen K., Zhang J., Zhang C.Y. (2008). Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. *Cell Research* 18, 997–1006.

Chinen A.B., Guan C.M., Ferrer J.R., Barnaby S.N., Merkel T.J., Mirkin C.A. (2015). Nanoparticle probes for the detection of cancer biomarkers, cells, and tissues by fluorescence. *Chemical Reviews* 115 (19), 10530–10574.

Ding C., Peng H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology* 3(2), 185-205.

Dreiseitl S., Ohno-Machado L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics* 35, 352–359.

Friedman J. H. (1989). Regularized discriminant analysis. *J. Amer. Stat. Assoc.* 84, 165-175.

Halpern E.J., Albert M., Krieger A.M., Metz C.E., Maidment A.D. (1996). Comparison of receiver operating characteristic curves on the basis of optimal operating point. *Statistics for Radiologists* 3, 245-253.

Hosmer D.W., Lemeshow S. (1989). *Applied Logistic Regression*. Wiley, New York.

Jickling G.C., Sharp F.R. (2015). Biomarker panels in ischemic stroke. *Stroke* 46, 915-920.

Kellgren J.H., Lawrence J.S. (1957). Radiological assessment of osteo-arthrosis. *Ann. Rheum. Dis.* 16, 494–502.

Koenker R., Bassett G. (1978). Regression Quantiles. *Econometrica* 46, 33-50.

Lee Y.-H., Wong D.T. (2009). Saliva: An emerging biofluid for early detection of diseases. *Am. J. Dent.* 22(4), 241-248.

Liu R., Wang X., Aihara K., Chen L. (2014). Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. *Medicinal Research Reviews* 34(3), 455-478.

McMahon M., Skaggs B.J., Grossman J.M., Sahakian L., FitzGerald J., Wong W.K., Lourenco E.V., Ragavendra N., Charles-Schoeman C., Gorn A., Karpouzas G.A., Taylor M.B., Watson K.E., Weisman M.H., Wallace D.J., Hahn B.H. (2014). A panel of biomarkers is associated with increased risk of the presence and progression of atherosclerosis in women with systemic lupus erythematosus. *Arthritis & Rheumatology* 66(1), 130-139.

Mishra J., Ma Q., Prada A., Mitsnefes M., Zahedi K., Yang J., Barasch J., Devarajan P. (2003). Identification of Neutrophil Gelatinase-Associated Lipocalin as a Novel Early Urinary Biomarker for Ischemic Renal Injury. *J. Am. Soc. Nephrol.* 14, 2534–2543.

Murata K., Furu M., Yoshitomi H., Ishikawa M., Shibuya H., Hashimoto M., Imura Y., Fujii T., Ito H., Mimori T., Matsuda S. (2013). Comprehensive microRNA analysis identifies miR-24 and miR-125a-5p as plasma biomarkers for rheumatoid arthritis. *PLoS ONE* 8(7): e69118. doi:10.1371/journal.pone.0069118

Scherzer C.R., Eklund A.C., Morse L.J., Liao Z., Locascio J.J., Fefer D., Schwarzschild M.A., Schlossmacher M.G., Hauser M.A., Vance J.M., Sudarsky L.R., Standaert D.G., Growdon J.H., Jensen R.V., Gullans S.R. (2007). Molecular markers of early Parkinson's disease based on gene expression in blood. *Proceedings of the National Academy of Sciences of the United States of America* 104(3), 955-960.

Vasan R.S. (2006). Biomarkers of Cardiovascular Disease Molecular Basis and Practical Considerations. *Circulation* 113, 2335-2362.

Wolf B.J., Hill E.G., Slate E.H. (2010). Logic Forest: an ensemble classifier for discovering logical combinations of binary markers. *Bioinformatics* 26(17), 2183-2189.