Discussion

Variable selection in large margin classifier-based probability estimation with high-dimensional predictors

Seung Jun Shin[1,2] and **Yichao Wu***[,3]

[1] Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[2] Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[3] Department of Statistics, North Carolina State University, NC 27695, USA

Received 30 October 2013; revised 6 January 2014; accepted 8 January 2014

This is a discussion of the papers: "Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory" by Jochen Kruppa, Yufeng Liu, Gérard Biau, Michael Kohler, Inke R. König, James D. Malley, and Andreas Ziegler; and "Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications" by Jochen Kruppa, Yufeng Liu, Hans-Christian Diener, Theresa Holste, Christian Weimar, Inke R. König, and Andreas Ziegler.

Keywords: Max-type penalty; Regularization; Variable selection.

The authors are to be congratulated for providing a comprehensive and thorough review for probability estimation in classification problems, one of the most widely used statistical tools in a variety of biomedical applications. The authors have nicely summarized several well-established machine learning methods as means of the probability estimation from both theoretical and practical perspectives.

High dimensional data are now becoming more and more common in biomedical sciences due to the rapid advances of related technologies for date generation and storage. The needs of statistical methods for analyzing such high-dimensional data have attracted lots of attention. For example, it is not uncommon to predict a patient's risk of having a certain cancer based on microarray or sequencing data with possibly tens of thousands of covariates. However, in general most of statistical methods including the probability estimation methods some of which are thoroughly summarized by Kruppa et al. (2014a); Kruppa et al. (2014b) are not directly applicable when the number of predictor is large. Here we provide some discussion on the use of the regularization framework in the probability estimation methods discussed in Kruppa et al. (2014a); Kruppa et al. (2014b) while mainly focussing on the SVM-based probability estimation readily extendable to any large margin classifiers (Wang et al., 2007).

1 Regularization

Regularization is a general technique used in statistics and machine learning to prevent overfitting by using a penalty on model complexity. Since the introduction of LASSO (Tibshirani, 1996), regularization has often been regarded as an effective way of variable selection due to its sparse solution with an

*Corresponding author: e-mail: wu@stat.ncsu.edu, Phone: +1-919-513-7677, Fax: +1-919-515-7591

appropriate model complexity penalty and gained great popularity. The main critique of LASSO is that it produces biased estimates for the large coefficients and there have been a vast of sparsity-inducing penalties to tackle this problem. Examples include adaptive LASSO, the smoothly clipped absolute deviation (SCAD), and minimax concave penalty (MCP) among many others. These three are known to possess *oracle properties* (Fan and Li, 2001). See Fan and Lv (2010) and references therein for both selective overviews and further details of variable selection methods including the aforementioned penalties.

The regularization which is originally introduced under the conventional linear regression model can be straightforwardly applied to the logistic regression by adding appropriate penalties to the logistic likelihood. As a simplest choice, L_1 (LASSO) penalized logistic regression can be considered while different penalties such as SCAD and MCP can also be employed as reasonable alternatives to L_1 penalty, although the corresponding estimation requires more attentions in computation since they are nonconvex penalties (Breheny and Huang, 2011).

2 Variable selection in SVM-based probability estimation via max penalty

The support vector machine (SVM) itself is a regularization method and performs quite competitively even when the dimensionality is large, but its solution is not sparse due to the use of an L_2 penalty. Wang et al. (2007) proposed a large margin classifier-based probability estimation method by training weighted large-margin classifiers, such as weighted SVMs, and aggregating information from different weighted classification boundaries. In order to achieve variable selection and probability estimation simultaneously, we may couple Wang et al.'s method with L_1 penalized weighted SVM or SCAD penalized weighted SVM instead of the standard weighted SVM. In this case, it is important to have an identical sparsity pattern across all the weighted SVMs when using different weights in order to identify important variables. To achieve an identical sparsity structure across the different WSVM problems, one can solve them together by using a max-type penalty as described in the following. Denote $\beta_m = \{\beta_{mj} : j = 0, \ldots, q\}^T$ and $f(\mathbf{x}; \pi_m) = \beta_{0m} + \sum_{j=1}^{q} \beta_{jm} h_j(\mathbf{x})$, $m = 1, \ldots, M$, where $\{h_j(\mathbf{x}), j = 1, \ldots, q\}$ denotes possible candidates of the basis functions often referred to as to *dictionary*. The joint estimation minimizes the following objective function with respect to β_m, $m = 1, \ldots, M$:

$$\min_{\beta_1,\ldots,\beta_M} \sum_{m=1}^{M} \left\{ (1-\pi_m) \sum_{i:y_i=1} H_1(f(\mathbf{x}_i; \pi_m)) + \pi_m \sum_{i:y_i=-1} H_1(f(\mathbf{x}_i; \pi_m)) \right\} + \sum_{j=1}^{q} \rho_\lambda \left(\max_{1\le m\le M} |\beta_{jm}| \right), \tag{1}$$

where $\rho_\lambda(\cdot)$ denotes a penalty function such as the L_1 and SCAD penalties and $\lambda > 0$ is the regularization parameter which controls sparsity of the solution. We remark that the set of solutions of (1), denoted by $\hat{\beta}_m$, $m = 1, \ldots, M$, share an identical sparsity structure due to the use of a max-type penalty. Note that a similar approach can be employed in order to achieve variable selection for the penalized multiple logistic regression which requires to solve $J - 1$ optimization problems simultaneously.

For multicategory classification, many different sparse large-margin classifiers have been proposed (Zhang et al., 2008 and references therein). A large margin classifier-based probability estimation scheme was proposed for multicategory classification in Wu et al. (2010). If variable selection is desired, an extension of the aforementioned max-type penalty can be applied to this multicategory large margin classifier-based probability estimation scheme. Yet things become much more complicated since

the number of weighted multicategory large margin classifiers escalates as the number of categories increases. This can potentially be an interesting further research topic.

3 Concluding remarks

Penalized logistic regression is a straightforward extension from the conventional linear regression and very easy to handle while it requires predominant logistic assumption which may not be valid in some application. On the other hand, the aforementioned penalized versions of the SVM-based probability estimates using a max-type penalty may be computationally intensive especially when response is not binary, but they are model-free approaches and hence desirable when there is no prior knowledge on $p(\mathbf{x})$.

Acknowledgements Wu's research is partially supported by NSF grant DMS-1055210 and NIH/NCI Grant R01-CA149569.

Conflict of interest
The author has declared no conflict of interest.

References

Breheny, P., Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics* **5**, 232–253.

Fan, J. and Li, R. (2001). Variable section via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20**, 101–148.

Kruppa, J., Liu, Y., Biau, G., Kohler, M., König, I. R., Malley, J. D. and Ziegler, A. (2014a). Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biometrical Journal* **56**, 534–563.

Kruppa, J., Liu, Y., Diener, H.-C., Holste, T., Weimar, C., König, I. R. and Ziegler, A. (2014b). Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications. *Biometrical Journal* **56**, 564–583.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society, Series B* **58**, 267–288.

Wang, J., Shen, X. and Liu, Y. (2007). Probability estimation for large-margin classifiers. *Biometrika* **95**, 149–167.

Wu, Y., Zhang, H. H. and Liu, Y. (2010). Robust model-free multiclass probability estimation. *Journal of the American Statistical Association* **105**, 424–436.

Zhang, H. H., Liu, Y., Wu, Y. and Zhu, J. (2008). Variable selection for the multicategory svm via adaptive sup-norm regularization. *Electronic Journal of Statistics* **2**, 149–167.

