TEST (2016) 25:44–46
DOI 10.1007/s11749-015-0474-y

DISCUSSION

Comments on: Probability enhanced effective dimension reduction for classifying sparse functional data

Chong Zhang[1] · **Yufeng Liu**[2,3,4,5]

Published online: 25 January 2016

The authors are to be congratulated for their solid contribution in providing a powerful method that handles classification and dimension reduction problems with functional data sets. This type of problems has drawn much attention in the literature, and is known to be difficult due to the complex structure of the corresponding data sets. In particular, many existing dimension reduction methods ignore the relationship between predictors and labels, and perform dimension reduction only using the covariates. Such procedures can be suboptimal and may lead to unstable results, especially when the predictors are sparsely observed. The proposed PEFCS method integrates the observed labels in the dimension reduction step by estimating class-conditional probabilities, and is shown to enjoy more competitive and robust performance in numerical examples.

This interesting paper leads to many promising research directions. For example, class-conditional probability (we denote it by $P_j(\hat{X}_i) = \text{pr}(Y = j \mid \hat{X}_i)$) estimation

This comment refers to the invited paper available at: doi:10.1007/s11749-015-0470-2.

✉ Yufeng Liu
yfliu@email.unc.edu

Chong Zhang
chong.zhang@uwaterloo.ca

1 Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada

2 Department of Statistics and Operations Research,
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

3 Department of Genetics, University of North Carolina at Chapel Hill,
Chapel Hill, NC, USA

4 Department of Biostatistics, University of North Carolina at Chapel Hill,
Chapel Hill, NC, USA

5 Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill,
Chapel Hill, NC, USA

is a crucial step in the PEFCS method. In the literature, it is known that classification methods can be grouped into two main categories: *soft* and *hard* classifiers (Wahba 2002; Liu et al. 2011). Soft classifiers directly estimate class-conditional probabilities, which further leads to classification rules. Typical examples of soft classification include Fisher's LDA and logistic regression. In contrast, hard classifiers bypass direct estimation of probabilities and focus on classification boundary estimation. Typical examples of hard classification include the support vector machine (SVM, Boser et al. 1992; Cortes and Vapnik 1995) and ψ-learning (Shen et al. 2003). In Liu et al. (2011), it was observed that the classification performance of various classifiers depends heavily on the underlying distribution of (X, Y). The authors use the hinge loss for the SVM in this paper. Therefore, a possible generalization of the proposed technique is to employ a more general loss function in their equation (5) for probability estimation, instead of using the weighted SVM. We will briefly discuss the idea below.

Consider the optimization problem

$$\min_{g \in \mathcal{F}_K} \sum_{i=1}^{n} \ell \left\{ Y_i g \left(\hat{X}_i \right) \right\} + \lambda \|g\|^2_{\mathcal{F}_K}, \tag{1}$$

where $\ell(\cdot)$ is a differentiable loss function for a soft classifier. One can verify that $P_{+1}(\hat{X}_i)$ can be estimated using $\ell'\{-\hat{g}(\hat{X}_i)\}/\ell'\{\hat{g}(\hat{X}_i)\}$ (Liu et al. 2011). For standard classification where the predictors are scalars or vectors, Liu et al. (2011) pointed out that when the underlying class-conditional probability, as a function of the predictors, is relatively smooth, soft classifiers tend to perform better than the hard ones. Moreover, the transition behavior from soft to hard classifiers were thoroughly investigated using the large-margin unified machine family proposed by Liu et al. (2011). For functional data classification, the comparison between soft and hard classifiers and the corresponding transition behavior are largely unknown, and further exploration in this direction can be very interesting.

Another potential research topic is to extend the PEFCS methodology to handle multicategory problems. In this case, the construction of slices in the EDR method becomes more involved. In particular, when $Y \in \{+1, -1\}$, only one direction of the EDR space can be recovered, because of the existence of homogeneity in learning problems with binary responses. To overcome this difficulty, Shin et al. (2014) proposed to construct slices based on $P_{+1}(\hat{X}_i)$. In multicategory classification, estimation of the class-conditional probabilities becomes more complex, as one needs to estimate a probability vector $\{P_1(\hat{X}_i), P_2(\hat{X}_i), \ldots, P_k(\hat{X}_i)\}$. Furthermore, how to construct $S_{(P_1, P_2, \ldots, P_k)|X}$ remains unclear. Therefore, it can be interesting and challenging to develop new methodology in this future research direction. Next, we provide one possible way to generalize the PEFCS methodology for multicategory problems.

For margin-based classification, when the number of classes is three or larger, one classification function $g(\cdot)$ is not enough to discriminate all classes. To overcome this difficulty, a common approach in the literature is to use k functions for k classes, and impose a sum-to-zero constraint on the k functions to reduce the parameter space and to ensure some theoretical properties such as Fisher consistency. Recently, Zhang and Liu (2014) suggested that using k functions and the sum-to-zero constraint can

be inefficient and suboptimal, and proposed the angle-based large margin classifiers for multicategory classification. In particular, consider a simplex $\boldsymbol{W}$ with k vertices $\{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_k\}$ in a $(k-1)$-dimensional space, such that

$$\boldsymbol{W}_j = \begin{cases} (k-1)^{-1/2}\mathbf{1}_{k-1}, & j = 1, \\ -\left(1+k^{1/2}\right)/\{(k-1)^{3/2}\}\mathbf{1}_{k-1} + \{k/(k-1)\}^{1/2}\,\boldsymbol{e}_{j-1}, & 2 \le j \le k, \end{cases}$$

where $\mathbf{1}_{k-1}$ is a vector of 1's with length $k-1$, and $\boldsymbol{e}_j \in \mathbb{R}^{k-1}$ is a vector with the jth element 1 and 0 elsewhere. In angle-based classification, one uses a $(k-1)$-dimensional classification function $\boldsymbol{f} = (f_1, \ldots, f_{k-1})^T$, which maps $\boldsymbol{x}$ to $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^{k-1}$, where $\boldsymbol{x}$ is the predictor vector. Observe that $\boldsymbol{f}$ introduces k angles with respect to $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_k$, namely, $\angle(\boldsymbol{f}, W_j)$; $j = 1, \ldots, k$. The prediction rule is based on which angle is the smallest. In particular, $\hat{y}(\boldsymbol{x}) = \operatorname{argmin}_{j\in\{1,\ldots,k\}} \angle(\boldsymbol{f}, W_j)$, where $\hat{y}(\boldsymbol{x})$ is the predicted label for $\boldsymbol{x}$. Based on the observation that

$$\operatorname*{argmin}_{j\in\{1,\ldots,k\}} \angle(\boldsymbol{f}, W_j) = \operatorname*{argmax}_{j\in\{1,\ldots,k\}} \langle \boldsymbol{f}, \boldsymbol{W}_j \rangle,$$

Zhang and Liu (2014) proposed the following optimization problem for the angle-based classifier

$$\min \sum_{i=1}^{n} \ell\{\langle \boldsymbol{W}_{y_i}, \boldsymbol{f}(\boldsymbol{x}_i)\rangle\} + \lambda J(\boldsymbol{f}), \tag{2}$$

where $\ell(\cdot)$ is a binary margin-based surrogate loss function, $J(\boldsymbol{f})$ is a penalty on $\boldsymbol{f}$ to prevent overfitting, and λ is a tuning parameter to balance the goodness of fit and the model complexity. One advantage of the angle-based classifier is that it is free of the commonly used sum-to-zero constraint, hence it can be more efficient for learning with big data sets. Thus, generalization of the PEFCS method in the angle-based framework should be feasible and promising.

Acknowledgments The authors were supported in part by National Science and Engineering Research Council of Canada (NSERC) (Zhang), US NSF Grant DMS-1407241 (Liu) and NIH/NCI Grant R01 CA-149569 (Liu), and National Natural Science Foundation of China (NSFC 61472475) (Liu).

References

Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Haussler D (ed) Proceedings of the 5th annual workshop on computational learning theory, COLT '92, Association for Computing Machinery, New York, NY, U.S.A., pp 144–152. doi:10.1145/130385.130401

Cortes C, Vapnik VN (1995) Support vector networks. Mach Learn 20:273–297

Liu Y, Zhang HH, Wu Y (2011) Soft or hard classification? Large margin unified machines. J Am Stat Assoc 106:166–177

Shen X, Tseng GC, Zhang X, Wong WH (2003) On ψ-learning. J Am Stat Assoc 98:724–734

Shin SJ, Wu Y, Zhang HH, Liu Y (2014) Probability-enhanced sufficient dimension reduction for binary classification. Biometrics 70(3):546–555

Wahba G (2002) Soft and hard classification by reproducing kernel Hilbert space methods. Proc Natl Acad Sci 99(26):16,524–16,530

Zhang C, Liu Y (2014) Multicategory angle-based large-margin classification. Biometrika 101(3):625–640