Nan Li and Hao Helen Zhang

Sparse Learning with Non-convex Penalty in Multi-classification

Nan Li[1]; Hao Helen Zhang[2]

[1] Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital, Memphis, Tennessee, U.S.A.
[2] Department of Mathematics, University of Arizona, Tucson, Arizona, U.S.A.

Abstract:
Multi-classification is commonly encountered in data science practice, and it has broad applications in many areas such as biology, medicine, and engineering. In multiclass problems, variable selection is much more challenging than in binary classification or regression problems. In addition to estimating multiple discriminant functions for separating different classes, we need to decide which variables are important for each individual discriminant function as well as for the whole set of functions. In this paper, we address the multi-classification variable selection problem by proposing a new form of penalty, supSCAD, which first groups all the coefficients of the same variable associated with all the discriminant functions altogether and then imposes the SCAD penalty on the supnorm of each group. We apply the new penalty to both soft and hard classification and develop two new procedures: the supSCAD multinomial logistic regression and the supSCAD multi-category support vector machine. Our theoretical results show that, with a proper choice of the tuning parameter, the supSCAD multinomial logistic regression can identify the underlying sparse model consistently and enjoys oracle properties even when the dimension of predictors goes to infinity. Based on the local linear and quadratic approximation to the non-concave SCAD and nonlinear multinomial log-likelihood function, we show that the new procedures can be implemented efficiently by solving a series of linear or quadratic programming problems. Performance of the new methods is illustrated by simulation studies and real data analysis of the Small Round Blue Cell Tumors and the Semeion Handwritten Digit data sets.
Keywords:
logistic regression; SCAD; supnorm; SVM; variable selection

1. Introduction:

Multiclass classification has broad applications in practice such as handwritten zip code digit recognition and cancer classification based on DNA microarray data (Hastie et al., 2009). Usually, a large number of variables are collected but some of them are uninformative in prediction. Therefore, it is essential to identify important variables in order to increase both classification accuracy and model interpretability.

This paper is motivated by precision medicine in cancer research where one goal is to extract important information from omics data, such as genomics, transcriptomics, or proteomics and classify tumors into different cancer subtypes in order to provide optimal treatment. Since the number of genes is usually much larger than the sample size, it is critical to select "signature" genes which can characterize cancer subtypes and have strong prediction power. This work is motivated by the classification of small round blue cell tumors in childhood (Khan et al., 2001) using a small set of important genes. We propose and study a new class of learning methods for joint multiclass classification and variable selection.

Based on the composite of supnorm and SCAD function, we propose a new form of penalty called the supSCAD penalty to achieve group sparsity for multiclass problems. What makes the supSCAD penalty attractive is its ability to remove noise covariates, i.e., those covariates not contributing to discriminating different classes. One motivating example is the multi-type cancer classification using genes. Typically, only a small subset of "important" genes are needed to classify cancer into different subtypes, and the rest of genes are either redundant or non-informative. The supSCAD penalty is designed to enforce group-wise parsimony in all the coefficients associated with one variable without directly penalizing individual coefficients, and therefore the estimated coefficients are less biased than other methods such as the group LASSO. The new penalty demonstrates competitive performance for both multinomial logistics regression and multi-category support vector machine, and enjoys nice theoretical properties, even if the data dimension diverges. An efficient algorithm is developed by combining the difference convex algorithm (DCA; Wu and Liu (2009)) and the local linear approximation (Zou and Li, 2008).

2. Methodology:

Consider a *K*-class problem with the input vector $x \in R^d$ and the output $y \in \{1, \dots, K\}$. For linear classification rules, there are $(d + 1)$ coefficients associated with each decision function, including the intercept term. Altogether, all the coefficients associated with the *K* decision functions can be expressed as a $K \times (d + 1)$ coefficient matrix. The *j*th column of the matrix, expressed as $\beta_{(j)} = (\beta_{1j}, \dots, \beta_{Kj})^T$, consists of *K* coefficients associated with x_j, where $j = 0, 1, \dots, d$ and x_0 is the intercept. The *k*th row $\beta_k = (\beta_{k0}, \beta_{k1}, \dots, \beta_{kd})$ consists of $(d + 1)$ coefficients characterizing the decision function f_k, where $k = 1, \dots, K$. For the variable selection purpose, we treat the elements in $\beta_{(j)}$ as a group. Define the supnorm of $\beta_{(j)}$ as $\|\beta_{(j)}\|_\infty = \max_{k=1,\dots,K} |\beta_{kj}|$, where the importance of x_j is directly controlled by its largest absolute element. If $\|\beta_{(j)}\|_\infty = 0$, then all the *K* coefficients associated with x_j are set to zero. Otherwise, if x_j is important with a positive supnorm, then no penalty is imposed on the remaining elements. To borrow desired sparsity and oracle property of SCAD penalty (Fan and Li, 2001), we propose supSCAD penalty with the following form

$$J_\lambda(\|\zeta\|_\infty) = \begin{cases} \lambda\|\zeta\|_\infty & if\ \|\zeta\|_\infty \le \lambda \\ -\frac{(\|\zeta\|_\infty^2 - 2a\lambda\|\zeta\|_\infty + \lambda^2)}{2(a-1)} & if\ \lambda < \|\zeta\|_\infty \le a\lambda \\ \frac{(a+1)\lambda^2}{2} & if\ \|\zeta\|_\infty > a\lambda \end{cases}, \text{ where}$$

$\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \dots, \zeta_K)^T$, $\|\zeta\|_\infty = \max_{i=1,\dots,K} |\zeta_i|$, $a > 2$, and $\lambda > 0$ is the tuning parameter.

Applied supSCAD to soft classification method, we have supSCAD multinomial logistic regression as

$$\min_\beta -\left\{\left\{\sum_{i=1}^n \left[\sum_{k=1}^K I(y_i = k)\beta_k^T x_i - \log\left(\sum_{k=1}^K \exp(\beta_l^T x_i)\right)\right\}\right\} + n\sum_{j=1}^d J_\lambda\left(\|\beta_{(j)}\|_\infty\right)$$

$$subject\ to \sum_{k=1}^K \beta_{kj} = 0,\ \ j = 0, 1, \dots, d.$$

We show that supSCAD multinomial logistic regression estimator is root-(n/d_n) consistent, where the subscript in d_n is used to emphasize its dependence on sample size n. The computation can be decomposed into two loops. The outer loop approximates the negative multinomial log-likelihood by its second order Taylor expression. Under the approximated log-likelihood, the inner loop solves the non-convex supSCAD function by DCA (Le Thi Hoai and Tao, 1997) or LLA (Zou and Li, 2008) techniques.

Applied supSCAD to hard classification method, we have supSCAD multi-category support vector machine (supSCAD MSVM) with the following form

$$\min_{\beta_0,\beta} \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} I(y_i \neq k)\left[\beta_{k0} + \beta_k^T x_i + 1\right]_+ + \sum_{j=1}^{d} J_\lambda\left(\left\|\beta_{(j)}\right\|_\infty\right)$$

$$subject\ to \sum\nolimits_{k=1}^{K} \beta_{k0} = 0, and \sum_{k=1}^{K} \beta_{kj} = 0, \qquad j = 1, \ldots, d.$$

Additionally, we have supSCAD multi-category proximal support vector machine (supSCAD MPSVM) as

$$\min_{\beta_0,\beta} \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} I(y_i \neq k)\left[\beta_{k0} + \beta_k^T x_i + 1\right]^2 + \sum_{j=1}^{d} J_\lambda\left(\left\|\beta_{(j)}\right\|_\infty\right)$$

$$subject\ to \sum\nolimits_{k=1}^{K} \beta_{k0} = 0, and \sum_{k=1}^{K} \beta_{kj} = 0, \qquad j = 1, \ldots, d.$$

In the same scheme of applying DCA or LLA techniques, supSCAD MSVM/MPSVM can be solved through a series of linear or quadratic programming problems.

3. Result:

We evaluate our proposed supSCAD estimator in terms of variable selection, class prediction, and probability estimation, under three simulation experiments and two real examples. The three simulation experiment settings are: (1) a four-class linear classification example with two “strong” important variables and two “weak” important variables for all the classes; (2) a four-class linear classification example with different important variables across classes; (3) a three-class high-dimensional example with two important variables for all the classes and 198 noise variables.

For comparison, we consider four soft classifiers and six hard classifiers. Four soft classification methods are L1 logistic regression (L1 LR), group-L1 logistic regression (GroupL1 LR), supSCAD logistic regression (supSCAD LR), and the composite MCP logistic regression (comp-MCP LR). Six hard classification methods are the standard MSVM (L2 MSVM), L1 MSVM (L1 MSVM), supSCAD MSVM/MPSVM, and the composite MCP MSVM/MPSVM. All simulations are conducted using Matlab (MATLAB, 2014) and Tomlab (Holmström et al., 2010), an optimization environment within Matlab.

For each simulation experiment, we generate a training data set and a test data set from the same distribution. We use the training data to train the classifiers and select the best tuning parameter from a series of λ values, and use the test data to evaluate performance of the estimated classifiers. A total of 100 simulations are conducted under each setting. Every classifier is evaluated in terms of its prediction and variable selection accuracy. For a soft classifier, we also examine its performance in estimating the conditional class probabilities.

Overall supSCAD LR achieves best performance in all three experiments, including the second experiment which has different important variables across classes and violates the supSCAD underlying assumption. The supSCAD focuses on the group-level selection and it does not imposes penalty at the individual level. By design, this second example has important variables with zero coefficients for certain classes, which favours the bi-level selection procedures such as the comp-MCP most. However, supSCAD LR still gives satisfactory results.

In real data analysis, Small Round Blue Cell Tumors is a 4-class problem, consisting 63 training samples and 20 independent testing samples. After filtering and standardizing, we rank the genes by their marginal separation power in the training set. We then select the top 100 and bottom 100 genes as the prediction covariates to feed to various classification methods. Leave-one-out cross validation is used to select the optimal tuning parameter, and then the trained models are used to predict the class labels for 20 test samples. Results are summarized and compared in the first row of Table 1. It is observed that the LR-based methods have overall better classification accuracy than the SVM-based methods for this data set, and the L1 LR, supSCAD LR, and the comp-MCP LR methods all have the test error zero.

Semeion Handwritten Digit Data consists of 1593 handwritten digits (0–9), each of which was scanned and stretched in a square box 16 × 16 in a gray scale of 256 values (i.e. 256 attributes). Then each pixel of each image was scaled into a Boolean (1/0) value of a fixed threshold. The whole data set is roughly equally distributed among the 10 classes. We split the data into six groups with roughly equal size and class distribution, with one group used for testing and the remaining five groups used for training. We perform 5-fold CV to select the best tuning parameter. The classification results are the second row of Table 1, and the supSCAD LR gives the lowest classification error 0.11.

Table1: Classification results for two real datasets.

Method	**L1 LR**	**GroupL1 LR**	**supSCAD LR**	**comp-MCP LR**	**L2 MSVM**
SRBCT	0	0.05	0	0	0
Semeion	0.37	0.35	0.11	0.12	0.20
Method	**L1 MSVM**	**supSCAD MSVM**	**supSCAD MPSVM**	**comp-MCP MSVM**	**comp-MCP MPSVM**
SRBCT	0.05	0.10	0.05	0.05	0.05
Semeion	0.27	0.25	0.19	0.27	0.19

4. Discussion and Conclusion:

This newly proposed penalty, supSCAD, enhances sparse learning in multi-classification by retaining the merits from both SCAD and supnorm penalties. It can incorporate the natural group effects of the coefficients associated with the same covariate to construct more parsimonious classifiers with desired oracle properties.

To tackle the numerical challenge of non-differentiability and non-convexity of the objective function, we have proposed an efficient iterative algorithm based on the LLA or DCA. For multiclass probability estimation, supSCAD is applied to multinomial logistic regression to conduct variable selection and conditional probability estimation simultaneously. An optimization procedure involving quadratic approximation to the multinomial loglikelihood function and nested DCA/LLA for the supSCAD penalty is developed and evaluated by numeric experiments. We further extend the penalty to the multi-category SVM framework and develop supSCAD MSVM/MPSVM, which demonstrate competitive performance compared to other regularized MSVM in simulated and real data examples.

The major underlying assumption of the supSCAD penalty is that the covariates across different classes can be naturally grouped for each predictor. Though this assumption may not hold for very complex problems, we find that it can still achieve reasonably good results even when the assumption is violated. For further improvement, our supSCAD penalty can be easily extended to incorporate within-group sparsity structure by imposing additional penalty on individual coefficients, such as LASSO or adaptive LASSO penalty, e.g., $\sum_{j=1}^{d} J_{\lambda}(\|\beta_{(j)}\|_{\infty}) + \lambda_c \sum_{k=1}^{K}\sum_{j=1}^{d} |\beta_{kj}|$. However, choosing the extra tuning parameter λc may cause additional computation cost. Its theoretical and computational properties are interesting for investigation in future work.

References:

1. Fan J, Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348–1360.
2. Hastie T, Tibshirani R, Friedman J (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.
3. Holmström K, Göran AO, Edvall MM (2010). User's Guide for Tomlab 7.
4. Khan J,Wei JS, Ringner M, Saal LH, Ladanyi M,Westermann F, et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6): 673–679.
5. Le Thi Hoai A, Tao PD (1997). Solving a class of linearly constrained indefinite quadratic problems by dc algorithms. *Journal of Global Optimization*, 11(3): 253–285.
6. MATLAB (2014). *version 8.3 (R2014a)*. The MathWorks Inc., Natick, Massachusetts.
7. Wu Y, Liu Y (2009). Variable selection in quantile regression. *Statistica Sinica*, 19(2): 801–817.
8. Zou H, Li R (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4): 1509.