Journal of Multivariate Analysis 166 (2018) 282–299

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Adaptively weighted large-margin angle-based classifiers

Sheng Fu [a,b], Sanguo Zhang [a,b], Yufeng Liu [c,d,e,f,g,*]

[a] School of Mathematical Science, University of the Chinese Academy of Sciences, Beijing 100049, China
[b] Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China
[c] Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, NC 27599, USA
[d] Department of Genetics, University of North Carolina at Chapel Hill, NC 27599, USA
[e] Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599, USA
[f] Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, NC 27599, USA
[g] Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, NC 27599, USA

ARTICLE INFO

Article history:
Received 20 June 2017
Available online 15 March 2018

AMS subject classifications:
62H30

Keywords:
Fisher consistency
Multicategory classification
Robustness
SVM
Weighted learning

ABSTRACT

Large-margin classifiers are powerful techniques for classification problems. Although binary large-margin classifiers are heavily studied, multicategory problems are more complicated and challenging. A common approach is to construct k different decision functions for a k-class problem with a sum-to-zero constraint. However, such a constraint can be inefficient. Moreover, many large-margin classifiers can be sensitive to outliers in the training sample. In this article, we use the angle-based classification framework to avoid the explicit sum-to-zero constraint, and we propose two adaptively weighted large-margin classification techniques. Our new methods are Fisher consistent and more robust against outliers under suitable conditions. Numerical experiments further indicate that our methods give competitive and stable performance when compared with existing approaches.

1. Introduction

Classification is an archetypal example of supervised learning problem with many diverse applications, e.g., to handwritten digit recognition, disease diagnosis, and fraud detection. Many classification methods are available in the literature, including logistic regression, Fisher linear discriminant analysis (LDA), support vector machine (SVM), tree-based methods, and many more; see [2,11] for comprehensive reviews.

Large-margin classifiers have attracted a lot of attention in recent years. The SVM is an appealing and powerful method which enjoys great success in various situations [3,5,6,8,22]. It was originally designed for two-class problems. The basic idea is to search for a single classification function that maximally separates the two classes. One appealing property of the SVM is that the resulting classification function only depends on a subset of the training data. The kernel trick implicitly maps the training data from the original input space into a more flexible, possibly high-dimensional, feature space, which enables the SVM to handle high-dimensional data and nonlinear classification problems. These three characteristics – maximum margin principle, support vector property, and kernel trick – enable the SVM to be successful and explain its popularity.

Since it is frequent to have more than two classes in practice, extensions dealing with multicategory classification problems are necessary and challenging. Two general approaches are used to handle k-class classification problems:

* Correspondence to: Department of Statistics and Operations Research, CB3260, University of North Carolina, Chapel Hill, NC 27599, USA.
E-mail addresses: fusheng13@mails.ucas.ac.cn (S. Fu), sgzhang@ucas.ac.cn (S. Zhang), yfliu@email.unc.edu (Y. Liu).

https://doi.org/10.1016/j.jmva.2018.03.004
0047-259X/

(i) Reduce the multicategory problem through a sequence of binary classifiers, e.g., using a one-against-one [22] or a one-against-rest [9] strategy. Both of them have obvious drawbacks. The one-against-one approach evaluates all possible pairwise classifiers and then combines $k(k-1)/2$ individual binary classifiers; these classifiers may suffer from the tie-in-vote problem. The one-against-rest approach trains k classifiers to separate each class from the combined remaining classes, and often leads to imbalanced training data [13,14].

(ii) Consider all classes simultaneously [4,22,23]. This type of method uses a k-tuple function vector $\boldsymbol{f}$ which maps the covariates to $\mathbb{R}^k$. The prediction rule assigns the class label to the category that corresponds to the maximum element within the function vector. To reduce the parameter space and ensure the uniqueness of optimal solution, a sum-to-zero constraint is often employed on the decision vector $\boldsymbol{f}$. Many existing Multicategory SVM (MSVM) methods fall within this framework; see, e.g., [7,13,17,21].

Despite their great success, the SVM methods still suffer from some drawbacks in certain situations. In particular, the SVM uses the unbounded hinge loss and consequently the resulting classifiers may be affected by points far away from their own classes, namely "outliers" in the training data. To overcome these problems, Shen et al. [20] and Liu et al. [15] proposed ψ-learning for binary and multicategory problems, respectively. Wu and Liu [24,25] applied the truncated hinge loss to obtain robust multicategory SVMs. However, the ψ-learning and robust truncated MSVMs both follow the non-convex optimization framework. Wu and Liu [26] proposed a weighted large-margin classification technique to gain robustness, which can be more computationally efficient than methods in [15,20,24,25].

For binary classification, the two classes can be separated by a single decision function. Analogously, it suffices to use $k-1$ decision functions for k-class problems. Recently, Zhang and Liu [29] and Zhang et al. [30] proposed a new angle-based large-margin classification framework. They used a regular k-vertex simplex structure centered at the origin in $\mathbb{R}^{k-1}$. Each vertex stands for one class and is a $(k-1)$-dimensional vector. The classification function vector $\boldsymbol{f}$ maps the covariates of a given instance to $\mathbb{R}^{k-1}$. The prediction rule assigns a new instance to the category whose corresponding vertex vector has the smallest angle with respect to the mapped classification function vector $\boldsymbol{f}$.

Although multicategory ψ-learning [15] and robust MSVMs [24,25] are less sensitive to outliers, there are still drawbacks. Because the ψ loss and the truncated hinge loss are non-convex, the corresponding optimization problems are more complex than the convex quadratic programming or linear programming for typical MSVMs. Furthermore, these methods need to estimate k functions with the sum-to-zero constraint.

In this paper, we first provide two types of general loss functions under the angle-based framework and study their properties. To ensure robustness while reducing computational complexity, we apply weighted learning techniques to two multicategory large-margin angle-based classification formulations without the sum-to-zero constraint. Our methods have several attractive properties. First, the angle-based classification strategy avoids the sum-to-zero constraint, which makes the formulation more compact and efficient. Second, the adaptively weighted methods we propose are robust and computationally effective, given that the objective function is convex. Third, the new formulations are Fisher consistent under suitable conditions. Finally, our numerical experiments demonstrate that the proposed methods offer competitive classification prediction accuracy compared to various simultaneous multicategory classifiers.

The rest of the paper is organized as follows. In Section 2, we briefly review several existing MSVM methods, and then introduce the angle-based classification framework with both convex and non-convex loss functions. Some consistency results for general loss functions are provided as well. In Section 3, we propose convex weighted learning techniques to achieve robustness and discuss their connections with non-convex robust methods. The results of our numerical studies showing the effectiveness of the new methods are reported in Section 4. Some discussion is given in Section 5. All proofs and some details are included in Appendix A.

2. Multicategory large-margin classification

Let $P(\boldsymbol{X}, Y)$ be some unknown probability distribution of $(\boldsymbol{X}, Y)$, where $\boldsymbol{X}$ is a vector of predictors, and Y is a label. Consider a k-category classification problem with $k \geq 2$. Suppose the training set is $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ from $P(\boldsymbol{X}, Y)$, where $\boldsymbol{x}_i$ is a p-dimensional covariate, and $y_i \in \{1, \ldots, k\}$ represents the corresponding label. Our target is to build a classification model based on the training data, which can be used to predict the class label of the new observed data. In this section, we first review the simultaneous large-margin classifiers in Section 2.1, and then introduce the angle-based classification framework with general loss functions in Section 2.2. Robust non-convex loss functions are provided in Section 2.3. Some statistical properties of the angle-based methods are given in Section 2.4.

2.1. Multicategory classification with k functions

Define a decision function $\boldsymbol{f} = (f_1(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x}))^\top \in \mathbb{R}^k$ in the vector form, where each component corresponds to one class. The prediction rule is $\hat{y} = \operatorname{argmax}_{j\in\{1,\ldots,k\}} f_j(\boldsymbol{x})$ for any new data point $\boldsymbol{x}$. Then, the corresponding optimization of large-margin classifiers can typically be written as

$$\min_{\boldsymbol{f}\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell\{\boldsymbol{f}(\boldsymbol{x}_i), y_i\} + \lambda J(\boldsymbol{f}), \quad \text{subject to} \sum_{j=1}^{k} f_j(\boldsymbol{x}) = 0 \text{ for all } \boldsymbol{x},$$