Noname manuscript No.
(will be inserted by the editor)

Comments on: Support vector machines maximizing geometric margins for multi-classification

Yoonkyung Lee

Received: date / Accepted: date

The paper provides an overview of the multi-objective multi-class support vector machine (MMSVM) based on a series of research articles written by the authors and their collaborators. As an extension of the binary SVM, the MMSVM takes an all-together approach to classification with multiple categories. Along with the description of the original multi-objective optimization problem for the method, the paper lays out computational strategies for its implementation and further approximation of the solutions to the initial problem via alternative single-objective second-order cone programming problems. The progression from the initial problem to various alternative formulations is methodically presented with summary of the relation between the two sets of corresponding solutions.

Geometric Margin vs. Functional Margin

The MMSVM is primarily motivated by the observation that the geometric margin and "functional margin" could be different when they are evaluated for a pair of estimated discriminant hyperplanes from other multi-class SVMs. Note that there are two different definitions of the functional margin for a linear discriminant function, $f(x) = w^\top x + b$, in the literature. One is $y_i f(x_i)$, individually defined for each instance (x_i, y_i) and the other is $1/\|w\|$, a half of the distance between the two hyperplanes given by $w^\top x + b = \pm 1$ in the binary case. The latter is the definition used in the paper. In the linearly separable case, the functional margin coincides with the geometric margin, that is, the minimal distance of training data to the discriminant hyperplane ($\min_i |w^\top x_i + b|/\|w\|$). The main thrust of the series of the papers reviewed by Tatsumi and

Lee's research was supported in part by National Science Foundation grant DMS-12-09194.

Y. Lee
Department of Statistics, The Ohio State University, Columbus, OH 43210, USA
E-mail: yklee@stat.osu.edu

Tanino is that simultaneous maximization of the geometric margins as opposed to maximization of the functional margins as in the existing multi-class SVMs would produce better generalization.

Margin Maximization as a Form of Regularization

However, Vapnik's theoretical argument for maximization of margin for better generalization lies in the result (see Theorem 10.3 in Vapnik (1998)) that the upper bound of the V-C dimension of the class of δ-margin normalized separating hyperplanes is inversely related to δ^2. This, in turn, suggests simultaneous maximization of the margin and minimization of the training error to attain a smaller upper bound of the generalization error as stated in the Corollary of the aforementioned theorem. In my view, the original Vapnik's arguments rest on the functional margin ($1/\|w\|$) rather than the geometric margin. The latter is properly defined and coincides with the former only in the limited separable case, while the "functional margin" through the norm of w is well-defined as a measure of the complexity of a discriminant function in general. It is well known that maximization of margin or equivalently minimization of $\|w\|^2$ is a form of regularization. This idea of stabilizing solutions to ill-posed problems with a regularizer originates from Tikhonov in solving an integral operator equation in mathematics, and a similar idea can be found in statistics, for instance, in ridge regression for stabilizing the variance of least squares estimators with the l_2 norm of the regression coefficients as a regularizer. For the reasons, I am not convinced of the significance of the distinction between the two kinds of margin in extending the SVM to the multi-category case as emphasized in the paper, and unsure whether it would bring substantial differences in practice.

Separable vs. Non-Separable Case

The paper seems to be focused almost exclusively on the separable case (except for the review of the binary SVM) where the geometric margins for a pair of classes can be meaningfully specified. However, it is not clear how the notion of geometric margin is extended to the non-separable case. When it comes to practical applications, I believe, separable problems are exceptions rather than the norm. Consequently, the authors' statement that "the large margin guarantees the generalization ability of the SVM" has to be modified in most settings. A large margin is only one side of the equation to ensure high classification accuracy. For better generalization, it is paramount to strike a balance between the empirical risk and the complexity of a classifier. The trade-off comes in the form of a choice of the regularization parameter or the cost parameter C in SVM. This critical aspect does not seem to be stressed adequately in the paper. Moreover, the focus on the separable case inevitably

limits the scope of comparisons as evident in the numerical experiment in Section 5.3. Only three out of eight benchmark problems had feasible solutions for comparison.

Turning Focus from Margin (or Penalty) to Loss

As in many of the existing extensions of the binary SVM, the multi-class extension in the paper takes an operational/computational point of view of the loss function (e.g. hinge loss, $(1 - yf(x))_+$ for the binary case) that it is an entity only implicitly defined through the inequality constraints for data. By contrast, there are other extensions primarily motivated by a statistical viewpoint, in particular, how to devise a loss function in the multi-category case so that the resulting discriminant functions are consistent with the Bayes decision rule. For example, Lee et al (2004) proposed an extension of the hinge loss that is properly classification-calibrated and thus ensures the Bayes risk consistency, after noting that the one-versus-rest approach with the binary SVM is not consistent nor is the commonly used multi-class extension in Vapnik (1998). More comprehensive take on this theoretical aspect of multi-category classification and conditions for proper loss functions can be found in Zhang (2004) and Tewari and Bartlett (2007).

Differential Penalties

With Bayes risk consistency being an asymptotic and rather minimal property, employing a consistent extension alone would not directly translate to near optimal performance, especially when the sample size is small to modest. In a nutshell, the key to the optimal performance is how to approximate the ideal partition of the input space given by the Bayes decision rule using data. Depending on the geometric characteristics of the partition, conceivably, the complexity of the discriminant functions $f_j(x) = w_j^\top x + b_j$ that induce the classification boundaries (as measured by $\|w_j\|^2$) would vary across classes. As the number of classes increases, it is likely that configuration of the classes over the input space would become more complex, warranting differences in the complexity of f_j. This increasing complexity of the partition also explains partially why various combination approaches might produce better performance than the direct approach in spite of the lack of consistency. I suspect that simultaneous maximization of the geometric margins might have the net effect of allowing for differential penalties on w_j, and the observed benefits of the multi-objective approach in some applications might be attributed to its flexibility in the size of w_j.

Computation and Empirical Validation

In addition to mathematical learning-theoretic justification and statistical sense of optimality, one cannot overlook the importance of computational efficiency, scalability and ease of implementation for a learning algorithm. For multi-category classification problems, the combination approach may seem more appealing than the direct approach in this regard. Tatsumi and Tanino's paper puts great emphasis on the computational aspect of the optimization for the MMSVM. The initial multi-objective formulation (M1) is relaxed and progressively transformed to a single-objective second-order cone programming problem (εM2) with the aid of the ε constraint method and additional constraints. The transformation requires a choice of a pair of classes (r, s) and ε constants for (εM) and another constant c_{rs} for (εM2) a priori. Presumably, the quality of the solutions from the transformed version would hinge on the choice of the parameters, but there is little discussion about how to specify them and how much impact incorrect specification has on error rates. Some numerical examples in the paper indicate fixing the parameters by using the solution to a version of multi-class SVM maximizing the functional margins. Given the additional computational complexity, its observed gain in accuracy over benchmark problems seems too incremental to render a convincing case for the method.

Acknowledgements This discussion brings me back to my own dissertational work under the guidance of Grace Wahba in the early 2000's. The work has evolved into many ideas in different forms and shapes over the years. I am grateful to Grace for her inspirations and wish her very happy 80th birthday this year (2014) along with all of her former and present students and colleagues.

References

Lee Y, Lin Y, Wahba G (2004) Multicategory Support Vector Machines, theory, and application to the classification of microarray data and satellite radiance data. Journal of the American Statistical Association 99:67–81

Tewari A, Bartlett P (2007) On the Consistency of Multiclass Classification Methods. Journal of Machine Learning Research 8:1007–1025

Vapnik V (1998) Statistical Learning Theory. Wiley, New York

Zhang T (2004) Statistical analysis of some multi-category large margin classification methods. Journal of Machine Learning Research 5:1225–1251