TOP (2014) 22:844–851
DOI 10.1007/s11750-014-0340-1

DISCUSSION

Comments on: Support Vector Machines Maximizing Geometric Margins for Multi-class Classification

Yann Guermeur

Published online: 15 August 2014

1 Introduction

The article deals with multi-class discrimination with support vector machines (SVMs). The authors present multi-class SVMs (MSVMs) which they have introduced in recent years: multiobjective MSVMs (MMSVMs). Those machines are based on the same functional class as that of the standard MSVMs (Guermeur 2012). They differ in the nature of the learning problem, which is no longer a standard optimization problem (convex quadratic programming problem), but a multiobjective optimization problem (taking the form of a second-order cone programming problem). The aim is to maximize exactly all geometric margins, so as to improve generalization performance. This performance is assessed empirically, through experiments performed on data sets from the UCI benchmark repository. In our comments, we make use of the latest results of the statistical theory of large margin multi-category classifiers to study the connection between the (width of the) geometric margins and the generalization performance.

The organization of these comments is as follows. Section 2 discusses the characteristics of the *all-together* (AT) MSVMs. Section 3 is devoted to the theoretical study of the generalization performance of these machines and the MMSVMs. At last, we discuss in Sect. 4 the options available to bridge the gap between theory and practice.

2 On the standard MSVMs

In the introduction of the article, the authors provide us with four references for the MSVM implementing the AT method (Bredensteiner and Bennett 1999; Guermeur

This comment refers to the invited paper available at doi:10.1007/s11750-014-0338-8.

Y. Guermeur (✉)
LORIA-CNRS, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex, France
e-mail: Yann.Guermeur@loria.fr

2002; Vapnik 1998; Weston and Watkins 1998). At least three models can be added to this list: that of Crammer and Singer (2001), Lee et al. (2004) and its quadratic loss extension, the M-SVM^2 (Guermeur and Monfrini 2011). It is noteworthy that the first two papers can be found in the bibliography. The model of Lee and her co-authors appears especially interesting in the framework of this study focusing on generalization performance since it is historically the first one to be Fisher consistent (Lee et al. 2004; Zhang 2004; Tewari and Bartlett 2007). As for the M-SVM^2, its model selection can be performed by minimizing over the regularization path a radius-margin bound (Bonidal 2013). On the other hand, it is well known that the models of Weston and Watkins, Vapnik, and Bredensteiner and Bennett are identical. The point is that for all the standard MSVMs, we have:

$$\sum_{p \in M} w^p = 0 \tag{1}$$

and consequently

$$\begin{aligned}
\sum_{p<q} ||w^p - w^q||^2 &= \frac{1}{2} \sum_{p \in M} \sum_{q \in M} ||w^p - w^q||^2 \\
&= \frac{1}{2} \sum_{p \in M} \sum_{q \in M} \left\{ ||w^p||^2 + ||w^q||^2 - 2\langle w^p, w^q \rangle \right\} \\
&= m \sum_{p \in M} ||w^p||^2 - \sum_{p \in M} \left\langle w^p, \sum_{q \in M} w^q \right\rangle \\
&= m \sum_{p \in M} ||w^p||^2.
\end{aligned}$$

Thus, it springs from (1) that the penalizer of all the standard MSVMs can take the form $\sum_{p \in M} \|w^p\|^2$. For these machines, we do not need to make the hypothesis that the description space $\mathcal{X}$ is a subset of $\mathbb{R}^n$. The hypotheses regarding this space and the set $M = [\![1, m]\!]$ of the categories are those which are at the basis of the statistical theory of pattern recognition, more precisely *agnostic learning* (Kearns et al. 1992). We assume that $(\mathcal{X}, \mathcal{A})$ and $(M, \mathcal{B})$ are measurable spaces and the link between descriptions and categories can be characterized by an unknown probability measure P on the measurable space $(\mathcal{X} \times M, \mathcal{A} \otimes \mathcal{B})$. Obviously, the statistical properties of the MMSVMs should be studied in the same framework.

3 Dependence of the guaranteed risks on the geometric margins

In the framework of pattern recognition, irrespective of the class of functions involved, all the guaranteed risks can be written as a sum of two terms: a sample-based estimate of performance and a control term which is an increasing function of the *capacity* of the class (see for instance, Vapnik 1998). In the case of large margin multi-category classifiers, the central capacity measure is a covering number. The nature of this number

varies as a function of the pathway followed to derive the bound. We now discuss the characteristics of the bounds available, focusing on their dependence on the sample size l and the number of categories m. In the specific case when the classifier is an MSVM or an MMSVM, we establish the way the covering numbers of interest can be upper bounded as a function of restrictions imposed on the corresponding functional class, restrictions precisely related to the width of the geometric margins. This calls for the introduction of standard definitions, starting with margin operators.

Definition 1 (Δ operator, Definition 6 in Guermeur 2007) Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $\mathbb{R}^m$. Define Δ as an operator on $\mathcal{F}$ such that:

$$\begin{aligned} \Delta : \mathcal{F} &\longrightarrow \Delta\mathcal{F} \\ f &\mapsto \Delta f = \left((\Delta f)_p\right)_{p\in M} \end{aligned}$$

$$\forall t \in \mathcal{T}, \quad \Delta f(t) = \frac{1}{2}\left(f_p(t) - \max_{q\in M\setminus\{p\}} f_q(t)\right)_{p\in M}.$$

Definition 2 (Δ^* operator, Definition 7 in Guermeur 2007) Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $\mathbb{R}^m$. Define Δ^* as an operator on $\mathcal{F}$ such that:

$$\begin{aligned} \Delta^* : \mathcal{F} &\longrightarrow \Delta^*\mathcal{F} \\ f &\mapsto \Delta^* f = \left((\Delta^* f)_p\right)_{p\in M} \end{aligned}$$

$$\forall t \in \mathcal{T}, \quad \Delta^* f(t) = \left(\left(2\mathbb{1}_{\{p\in \operatorname{argmax}_{q\in M} f_q(t)\}} - 1\right)\max_{q\in M}(\Delta f)_q(t)\right)_{p\in M}.$$

Definition 3 (Classes of functions $\mathcal{F}_\mathcal{G}$, $\mathcal{F}^*_\mathcal{G}$, and $\mathcal{F}^\#_\mathcal{G}$) Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ into $[-M_\mathcal{G}, M_\mathcal{G}]^m$. For all g in $\mathcal{G}$, the functions f_g and f^*_g from $\mathcal{X}\times M$ into $[-M_\mathcal{G}, M_\mathcal{G}]$ are defined by:

$$\forall (x,p) \in \mathcal{X}\times M, \quad \begin{cases} f_g(x,p) = (\Delta g)_p(x) \\ f^*_g(x,p) = (\Delta^* g)_p(x) \end{cases}.$$

Then, the classes $\mathcal{F}_\mathcal{G}$ and $\mathcal{F}^*_\mathcal{G}$ are defined as follows:

$$\mathcal{F}_\mathcal{G} = \left\{f_g : g \in \mathcal{G}\right\}, \quad \mathcal{F}^*_\mathcal{G} = \left\{f^*_g : g \in \mathcal{G}\right\}.$$

In the sequel, $\Delta^\#$ is used in place of Δ or Δ^* in the formulas that hold true for both margin operators. We define accordingly $\mathcal{F}^\#_\mathcal{G} = \left\{f^\#_g : g \in \mathcal{G}\right\}$.

Definition 4 [ϵ-cover, ϵ-net and covering numbers (Kolmogorov and Tihomirov 1961)] Let (E, ρ) be a metric or pseudo-metric space, $E' \subset E$ and $\epsilon \in \mathbb{R}_+^*$. An *$\epsilon$-cover* of E' is a coverage of E' with open balls of radius ϵ the centers of which belong to E. These centers form an *ϵ -net* of E'. A *proper ϵ -net* of E' is an *ϵ-net* of E' included in E'. If E' has an ϵ-net of finite cardinality, then its *covering number* $\mathcal{N}(\epsilon, E', \rho)$ is the smallest cardinality of its ϵ-nets:

$$\mathcal{N}(\epsilon, E', \rho) = \min\left\{\left|E''\right| : \left(E'' \subset E\right) \wedge \left(\forall e \in E',\ \rho\left(e, E''\right) < \epsilon\right)\right\}.$$

If there is no such finite net, then the covering number is defined to be infinite. $\mathcal{N}^{(p)}(\epsilon, E', \rho)$ will designate a covering number of E' obtained by considering proper ϵ-nets only. In the finite case, we have thus:

$$\mathcal{N}^{(p)}(\epsilon, E', \rho) = \min\left\{\left|E''\right| : \left(E'' \subset E'\right) \wedge \left(\forall e \in E',\ \rho\left(e, E''\right) < \epsilon\right)\right\}.$$

The definition of a covering number thus involves the specification of a (pseudo-) metric. We will make use of two of them.

Definition 5 (Pseudo-distance $d_{\mathcal{F},\mathbf{t}_n,\infty}$) Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $\mathbb{R}^m$. For $n \in \mathbb{N}^*$, let $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$. Then, the empirical pseudo-metric $d_{\mathcal{F},\mathbf{t}_n,\infty}$ on $\mathcal{F}$ is defined as follows:

$$\forall\left(f, f'\right) \in \mathcal{F}^2,\quad d_{\mathcal{F},\mathbf{t}_n,\infty}\left(f, f'\right) = \max_{1 \leqslant i \leqslant n} \left\|f\left(t_i\right) - f'\left(t_i\right)\right\|_\infty.$$

Definition 6 (Pseudo-distance $d_{\mathcal{F},\mathbf{t}_n,2}$) Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{T}$. For $n \in \mathbb{N}^*$, let $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$. Then, the empirical pseudo-metric $d_{\mathcal{F},\mathbf{t}_n,2}$ on $\mathcal{F}$ is defined as follows:

$$\forall\left(f, f'\right) \in \mathcal{F}^2,\quad d_{\mathcal{F},\mathbf{t}_n,2}\left(f, f'\right) = \left\|f - f'\right\|_{L_2(\mu_{\mathbf{t}_n})} = \left(\frac{1}{n}\sum_{i=1}^{n}\left(f\left(t_i\right) - f'\left(t_i\right)\right)^2\right)^{\frac{1}{2}}$$

where $\mu_{\mathbf{t}_n}$ denotes the uniform probability measure on $\{t_i : 1 \leqslant i \leqslant n\}$.

The standard way to derive an upper bound on a covering number consists in establishing a generalized Sauer–Shelah lemma (Alon et al. 1997; Mendelson and Vershynin 2003) involving an extension of the Vapnik–Chervonenkis (VC) dimension (Vapnik and Chervonenkis 1971). In Guermeur (2007), it was proved that the extensions characterizing the learnability of large margin multi-category classifiers are the γ–Ψ-dimensions.

Definition 7 (γ–Ψ-dimensions, Definition 28 in Guermeur 2007) Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $\mathbb{R}^m$ and $\Delta^\#$ a margin operator. Let Ψ be a family of mappings ψ from M into $\{-1, 1, *\}$. For $\gamma \in \mathbb{R}_+^*$, a subset $s_{\mathcal{T}^n} = \{t_i : 1 \leqslant i \leqslant n\}$ of $\mathcal{T}$ is said to be *γ–Ψ -shattered* (*Ψ -shattered with margin γ*) by $\left(\mathcal{F}, \Delta^\#\right)$ if there is a mapping

$\psi^n = \left(\psi^{(i)}\right)_{1 \leqslant i \leqslant n}$ in Ψ^n and a vector $\mathbf{b}_n = (b_i)_{1 \leqslant i \leqslant n}$ in $\mathbb{R}^n$ such that, for each vector $\mathbf{k}_n = (k_i)_{1 \leqslant i \leqslant n}$ in $\{-1, 1\}^n$, there is a function $f_{\mathbf{k}_n}$ in $\mathcal{F}$ satisfying

$$\forall i \in [\![1, n]\!], \quad \begin{cases} \text{if } k_i = 1, \ \exists p_1 : \psi^{(i)}(p_1) = 1 \ \wedge \ \left(\Delta^{\#} f_{\mathbf{k}_n}\right)_{p_1}(t_i) - b_i \geqslant \gamma \\ \text{if } k_i = -1, \ \exists p_2 : \psi^{(i)}(p_2) = -1 \ \wedge \ \left(\Delta^{\#} f_{\mathbf{k}_n}\right)_{p_2}(t_i) + b_i \geqslant \gamma \end{cases}.$$

The *γ–Ψ -dimension*, or *Ψ -dimension with margin γ*, of $\left(\mathcal{F}, \Delta^{\#}\right)$, denoted by γ–Ψ-dim$\left(\mathcal{F}, \Delta^{\#}\right)$, is the maximal cardinality of a subset of $\mathcal{X}$ γ–Ψ-shattered by $\left(\mathcal{F}, \Delta^{\#}\right)$, if this cardinality is finite. If no such maximum exists, $\left(\mathcal{F}, \Delta^{\#}\right)$ is said to have infinite γ–Ψ-dimension.

These dimensions can be seen either as scale-sensitive extensions of the Ψ-dimensions (Ben-David et al. 1995), or multivariate extensions of the fat-shattering dimension (Kearns and Schapire 1994). One of them appears easier to handle due to its connection with the one-against-one decomposition scheme, the extension of the Natarajan dimension (Natarajan 1989) (margin Natarajan dimension, denoted γ-N-dim).

For a m-category classifier computing a class of functions $\mathcal{G}$ from $\mathcal{X}$ into $\mathbb{R}^m$, Theorem 22 in Guermeur (2007), an extension of Corollary 9 in Bartlett (1998) and Theorem 4.1 in Vapnik (1998), provides us with a guaranteed risk whose control term grows as the square root of the logarithm of $\mathcal{N}^{(p)}\left(\epsilon, \Delta^{\#}\mathcal{G}, 2l\right)$, the supremum over $\mathcal{X}^{2l}$ of $\mathcal{N}^{(p)}\left(\epsilon, \Delta^{\#}\mathcal{G}, d_{\Delta^{\#}\mathcal{G}, \mathbf{x}_{2l}, \infty}\right)$. This covering number (with Δ^* as margin operator) can be bounded from above by means of a generalized Sauer–Shelah lemma involving the margin Natarajan dimension of $(\mathcal{G}, \Delta)$ (Lemma 39 in Guermeur 2007). Thus, characterizing the connection between the generalization performance of a MSVM (or a MMSVM) and its geometric margins can boil down to deriving an upper bound on its margin Natarajan dimension in terms of those margins. This is precisely what we get with the following theorem, a straightforward multi-class extension of Theorem 4.6 in Bartlett and Shawe-Taylor (1999):

Theorem 1 (Margin Natarajan dimension of the MSVMs, Theorem 48 in Guermeur 2007) *Let κ be a real-valued positive type function (kernel) (Berlinet and Thomas-Agnan 2004) and let $\mathbf{H}_\kappa$ be the corresponding reproducing kernel Hilbert space (RKHS) (Berlinet and Thomas-Agnan 2004). Let $\mathbf{H}_{\kappa,m}$ be the RKHS of $\mathbb{R}^m$-valued functions (Wahba 1992) at the basis of a m-category MSVM (MMSVM) with kernel κ (Guermeur 2012). Let us assume that the image of $\mathcal{X}$ by the reproducing kernel map is included in the closed ball of radius $\Lambda_{\mathcal{X}}$ about the origin in $\mathbf{H}_\kappa$. Let $\bar{\mathbf{H}}_{\kappa,m}$ be the restriction of $\mathbf{H}_{\kappa,m}$ characterized by:*

$$\sup_{\mathbf{w}=(w^p)_{p \in M} \in \bar{\mathbf{H}}_{\kappa,m}} \frac{1}{2} \max_{(p,q) \in M^2} \left\| w^p - w^q \right\|_{\mathbf{H}_\kappa} \leqslant \Lambda_{\mathbf{H}_{\kappa,m}}.$$

Then, for any positive real value γ, the following bound holds true:

$$\gamma\text{-}N\text{-}dim\left(\bar{\mathbf{H}}_{\kappa,m}, \Delta\right) \leqslant \binom{m}{2} \left(\frac{\Lambda_{\mathbf{H}_{\kappa,m}} \Lambda_{\mathcal{X}}}{\gamma}\right)^2.$$

The bound sketched above is not utterly satisfactory due to its suboptimal dependence on the sample size l. Indeed, its control term decreases with l as a $O\left(\frac{\ln(l)}{\sqrt{l}}\right)$. The optimal convergence rate, $\frac{1}{\sqrt{l}}$, can be obtained by following a more direct path involving a different capacity measure: the Rademacher average (Bartlett et al. 2005). To the best of our knowledge, the first result of this kind is Corollary 8.1 in Mohri et al. (2012). This bound is not utterly satisfactory either since its control term grows quadratically with m. The reason for this drawback basically rests in the fact that even in the case of kernel machines, the Rademacher averages associated with multivariate models cannot be bounded as straightforwardly as those associated with univariate models. The property used by Mohri and his co-authors to adapt the bi-class line of reasoning to the multi-class case, i.e., cope with this difficulty, appears in the proof of Theorem 8.1 in Mohri et al. (2012). It is the sub-additivity of the supremum. The quadratic dependence can be seen as an artifact of this choice. To make the best of both worlds, so as to optimize both dependences (on l and m), we propose a hybrid approach. In short, it consists in following the proof of Theorem 8.1 in Mohri et al. (2012) up to the point where the Rademacher average appears, and then apply Dudley's integral inequality (see for instance Theorem 11.17 in Ledoux and Talagrand 1991), to switch back to a covering number. The corresponding covering number is $\mathcal{N}(\epsilon, \mathcal{F}_{\mathcal{G}}^{\#}, l)$, the supremum over $(\mathcal{X} \times M)^l$ of $\mathcal{N}(\epsilon, \mathcal{F}_{\mathcal{G}}^{\#}, d_{\mathcal{F}_{\mathcal{G}}^{\#}, \mathbf{z}_l, 2})$, with $\mathbf{z}_l = ((x_i, y_i))_{i \in I} \in (\mathcal{X} \times M)^l$. Bounding from above this covering number as a function of the margin Natarajan dimension of $(\mathcal{G}, \Delta)$ remains an open problem. The only solution available so far to make use of a generalized Sauer–Shelah lemma consists in treating separately the classes of functions to which the component functions of the model of interest belong. To that end, one can apply the following lemma, whose proof raises no difficulty.

Lemma 1 *Let $\mathcal{G} = \prod_{p=1}^{m} \mathcal{G}_p$ be a class of functions from $\mathcal{X}$ into $\left[-M_{\mathcal{G}}, M_{\mathcal{G}}\right]^m$. Then,*

$$\mathcal{N}\left(\frac{m}{\sqrt{2}}\epsilon, \mathcal{F}_{\mathcal{G}}, l\right) \leqslant \prod_{p=1}^{m} \mathcal{N}\left(\epsilon, \mathcal{G}_p, l\right).$$

To bound from above the covering numbers of the classes $\mathcal{G}_p$ in terms of a generalized VC dimension, one can use a variant of Theorem 1 in Mendelson and Vershynin (2003). Then, one can easily verify that the convergence rate obtained is (at worst) $\sqrt{\frac{\ln(l)}{l}}$ ("halfway" between that of the two previous bounds), while the control term only grows as the square root of m. To finish the derivation of the guaranteed risk, it remains to bound m fat-shattering dimensions. Turning back to the case of the MSVMs and MMSVMs, this can be done by means of a result already mentioned: Theorem 4.6 in Bartlett and Shawe-Taylor (1999). The problem is that the resulting bound cannot be used to provide a theoretical justification to the MMSVMs. Indeed, by handling independently the component functions of the classifier, we have lost most of the connection between the control term of the guaranteed risk and the geometric margins (whose definition involves the difference of two component functions).

4 Discussion

Tatsumi and Tanino have introduced multi-class support vector machines which are based on a principle in full accordance with the intuition borrowed from the bi-class case: a direct maximization of the geometric margins. The experimental evidence they provide is very promising. However, strange as it may seem, the statistical theory of large margin multi-category classifiers still fails to fully justify their choices. This justification could come as the byproduct of the derivation of sharper bounds on the risk. We conjecture that a bound exhibiting the optimal convergence rate with a control term growing only as the square root of the number of categories could be obtained from an appropriate implementation of the generic chaining method (Talagrand 2005).

Acknowledgments The author would like to thank E. Didiot for carefully reading this manuscript.

References

Alon N, Ben-David S, Cesa-Bianchi N, Haussler D (1997) Scale-sensitive dimensions, uniform convergence, and learnability. J ACM 44(4):615–631

Bartlett P (1998) The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. IEEE Trans Inf Theory 44(2):525–536

Bartlett P, Shawe-Taylor J (1999) Generalization performance of support vector machines and other pattern classifiers. In: Schölkopf B, Burges C, Smola A (eds) Advances in kernel methods—support vector learning, chap 4. The MIT Press, Cambridge, pp 43–54

Bartlett P, Bousquet O, Mendelson S (2005) Local Rademacher complexities. Ann Stat 33(4):1497–1537

Ben-David S, Cesa-Bianchi N, Haussler D, Long P (1995) Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. J Comput Syst Sci 50(1):74–86

Berlinet A, Thomas-Agnan C (2004) Reproducing kernel hilbert spaces in probability and statistics. Kluwer, Boston

Bonidal R (2013) Sélection de modèle par chemin de régularisation pour les machines à vecteurs support à coût quadratique. Ph.D. thesis, Université de Lorraine

Bredensteiner E, Bennett K (1999) Multicategory classification by support vector machines. Computat Optim Appl 12(1/3):53–79

Crammer K, Singer Y (2001) On the algorithmic implementation of multiclass kernel-based vector machines. J Mach Learn Res 2:265–292

Guermeur Y (2002) Combining discriminant models with new multi-class SVMs. Pattern Anal Appl 5(2):168–179

Guermeur Y (2007) VC theory of large margin multi-category classifiers. J Mach Learn Res 8:2551–2594

Guermeur Y (2012) A generic model of multi-class support vector machine. Int J Intell Inf Database Syst 6(6):555–577

Guermeur Y, Monfrini E (2011) A quadratic loss multi-class SVM for which a radius-margin bound applies. Informatica 22(1):73–96

Kearns M, Schapire R (1994) Efficient distribution-free learning of probabilistic concepts. J Comput Syst Sci 48(3):464–497

Kearns M, Schapire R, Sellie L (1992) Toward efficient agnostic learning. In: COLT'92, pp 341–352

Kolmogorov A, Tihomirov V (1961) ϵ-entropy and ϵ-capacity of sets in functional spaces. Am Math Soc Transl Ser 2(17):277–364

Ledoux M, Talagrand M (1991) Probability in Banach spaces: isoperimetry and processes. Springer, Berlin

Lee Y, Lin Y, Wahba G (2004) Multicategory support vector machines: theory and application to the classification of microarray data and satellite radiance data. J Am Stat Assoc 99(465):67–81

Mendelson S, Vershynin R (2003) Entropy and the combinatorial dimension. Invent Math 152:37–55

Mohri M, Rotamizadeh A, Talwalkar A (2012) Foundations of machine learning. The MIT Press, Cambridge

Natarajan B (1989) On learning sets and functions. Mach Learn 4(1):67–97

Talagrand M (2005) The generic chaining: upper and lower bounds of stochastic processes. Springer, Berlin

Tewari A, Bartlett P (2007) On the consistency of multiclass classification methods. J Mach Learn Res 8:1007–1025

Vapnik V (1998) Stat Learn Theory. Wiley, New York

Vapnik V, Chervonenkis A (1971) On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab Appl XVI (2):264–280

Wahba G (1992) Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In: Casdagli M, Eubank S (eds) Nonlinear modeling and forecasting, SFI studies in the sciences of complexity, vol XII, pp 95–112

Weston J, Watkins C (1998) Multi-class support vector machines. Tech. Rep. CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science (1998)

Zhang T (2004) Statistical analysis of some multi-category large margin classification methods. J Mach Learn Res 5:1225–1251