ELSEVIER

Pattern Recognition Letters 22 (2001) 407–412

Pattern Recognition Letters

www.elsevier.nl/locate/patrec

An improved bound on the finite-sample risk of the nearest neighbor rule

Richard Nock [a,*], Marc Sebban [b]

[a] Département Scientifique Interfacultaire, Université des Antilles-Guyane, Campus universitaire de Schoelcher, 97233 Schoelcher, France
[b] Département de Sciences Juridiques et Economiques, Université des Antilles-Guyane, Campus universitaire de Fouillole,
97159 Pointe-à-Pitre, France

Received 2 March 2000; received in revised form 3 August 2000

Abstract

This paper extends a previous risk study of the well-known nearest neighbor (NN) rule with fixed and finite reference samples. Our result is competitive with some previously obtained in fairly restrictive and complex settings, and beats these in general cases. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Nearest neighbor; Risk bounds; Finite sample; Fixed sample

1. Introduction

The nearest neighbor (NN) rule is one of the simplest and oldest non-parametric classification techniques. It uses a set of observations $\boldsymbol{S}$ from a metric space $\boldsymbol{X}$ to classify members of $\boldsymbol{X}$ into one of c classes. For each $x \in \boldsymbol{X}$, it chooses the element $x' \in \boldsymbol{S}$ which is the nearest to x and gives the same class to x as that of x' (an arbitrary tie-breaking rule is assumed when more than one point are at minimal distance). Historically, the first appearance of a similar classification rule occurred in (Fix and Hodges, 1951). Since then, much work has been done for the theoretical risk study of the NN rule or variants (Cover, 1968; Cover and Hart, 1967; Drakopoulos, 1995; Okamoto and Yugami, 1996; Snapp and Venkatesh, 1998; Venkatesh et al., 1992).

The first result (Cover and Hart, 1967) showed that, as the reference sample's size goes towards infinity, modulo some smoothness and independence constraints, the risk of the NN rule, $P_{\boldsymbol{S}}$, is upperbounded by $2P^* - cP^{*^2}/(c-1)$, where P^* is Bayes optimal risk. Apart from the upperbound itself, this result is important because it allows to evaluate the difficulty of a pattern recognition problem by computing bounds for P^* (Cover, 1968; Drakopoulos, 1995). Later results relaxed the infinite size constraint on $\boldsymbol{S}$, and studied the expectation of the NN risk when samples of equal size are drawn, as for example (Cover, 1968; Venkatesh et al., 1992).

A recent result (Snapp and Venkatesh, 1998) even focused on generalizing bounds to the k-NN rule, in which the k nearest points vote to give a class. Recently too, the study was relaxed to give

[*] Corresponding author. Fax: +596-72-73-62.
E-mail addresses: rnock@univ-ag.fr, nock@lirmm.fr, Richard.Nock@martinique.univ-ag.fr (R. Nock), msebban@univ-ag.fr, sebban@univ-lyon2.fr (M. Sebban).

PII: S0167-8655(00)00133-1

risk upperbounds for finite and *fixed* reference samples $\boldsymbol{S}$ for problems with unrestricted number of classes (Drakopoulos, 1995). This work constitutes the starting point of our work.

Drakopoulos (1995) carried out his study by putting continuity assumptions to bound the variations of $\boldsymbol{X}$; namely, Hölder continuity was assumed regarding the likelihood functions. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be metric spaces upon which metrics $d_{\boldsymbol{A}}$ and $d_{\boldsymbol{B}}$ are defined, then a function $f : \boldsymbol{A} \rightarrow \boldsymbol{B}$ is Hölder continuous iff

$$\exists \alpha > 0, K > 0 : \forall x, y \in \boldsymbol{A},$$
$$d_{\boldsymbol{B}}(f(x), f(y)) \leqslant K d_{\boldsymbol{A}}(x, y)^{\alpha}.$$

Using this hypothesis, Drakopoulos (1995) proves that the risk of the NN rule is upperbounded by $2P^* - cP^{*^2}/(c-1) + K_{\boldsymbol{S}}$, where $K_{\boldsymbol{S}}$ is the maximal variation (assuming Hölder continuity) of the likelihood functions between one point of $\boldsymbol{X}$ and its NN in $\boldsymbol{S}$. This term can be thought of as a *penalty factor* due to the finiteness of $\boldsymbol{S}$. In a second theorem, Drakopoulos (1995) strengthens the result and obtains a smaller upperbound for $P_{\boldsymbol{S}}$, but at the expense of a very restrictive hypothesis completing Hölder continuity. This hypothesis expresses that on any point of $\boldsymbol{X}$, the overall variation of the likelihood functions over all classes (using L_2 norm), between x and its NN in $\boldsymbol{S}$, is upperbounded by $K'(c) \times g(\sup_{y \in \boldsymbol{X}} P^*(y))$. Here $K'(c)$ is increasing and converges to 2, but $g(\cdot)$ is a decreasing function which converges to zero as the maximal *local* Bayes error, $\sup_{y \in \boldsymbol{X}} P^*(y)$, attain its maximal possible value, $(c-1)/c$. By means of words, if there exists in $\boldsymbol{X}$ one point x for which Bayes rule does only a little better than a simple coin toss, then for *any* point y of $\boldsymbol{X}$, the conditional class probabilities between y and its NN in $\boldsymbol{S}$, potentially very far from y, are constrained to be practically the same. In spite of this limitation, Theorem 2 of Drakopoulos (1995) is interesting because it shows that, modulo the assumption, the penalty factor due to the finiteness of $\boldsymbol{S}$ vanishes as P^* increases, another key factor when studying the difficulty of a pattern recognition problem.

Our aim in that paper is to exhibit a stronger behavior (that is, a smaller upperbound) in a more general setting, using weaker and simpler hypotheses. Our first hypothesis is weaker than the first one of Drakopoulos (1995). Our second assumption is much less restrictive with respect to the second one of Drakopoulos (1995). Informally, it states that there exists on average over $\boldsymbol{X}$ a positive correlation between the likelihood functions of certain classes. After having presented our main result, we propose an application of the theorem in the two-classes case. In that framework, the result is even stronger than in the general case, since the second hypothesis disappears. The vanishing property of the penalty factor is therefore proven under the weakest restriction scheme, yet it generalizes all corresponding results of Drakopoulos (1995).

2. Definitions and related theorems

Most of our notations follow those of Drakopoulos (1995). Let

$$\boldsymbol{S} = \{x_1, x_2, \ldots, x_{|\boldsymbol{S}|}\}$$

be a finite sample set over a metric space $\boldsymbol{X}$ upon which a metric is defined; here, $|\cdot|$ denotes the cardinality. Assume that each x_i is labeled with one of c classes $\theta_1, \theta_2, \ldots, \theta_c$. Define variables X, Θ_X, that take values over $\boldsymbol{X}$ and $\{\theta_1, \theta_2, \ldots, \theta_c\}$, respectively. Similarly to Drakopoulos (1995), the results we present rely implicitly on the fact that the pairs (observation, class) of $\boldsymbol{S}$ are independently identically distributed according to the distribution (X, Θ_X). Throughout the paper, we make use of the following notations:

$$\forall i \in \{1, 2, \ldots, c\},$$
$$\begin{cases} a_i(x) \triangleq Pr(\Theta_X = \theta_i \mid X = x), \\ b_i(x) \triangleq Pr(\Theta_X = \theta_i \mid X = \arg\min_{y \in \boldsymbol{S}} \mathrm{d}(x, y)). \end{cases}$$

In order to achieve our results, we make the following hypothesis, which explicitly bounds the variations of class-conditional probabilities.

(H) There exists positive increasing functions $f_i(.) : \mathbb{R}^+ \rightarrow [0, 1]$ such that

$$\forall i \in \{1,2,\ldots,c\},$$
$$|a_i(x) - b_i(x)| \leqslant f_i(\mathrm{d}(x, \arg\min_{y\in \boldsymbol{S}} \mathrm{d}(x,y))),$$

Drakopoulos (1995) puts $f_i(a) \overset{\Delta}{=} K_i a^{\alpha_i}$ for some $K_i > 0, \alpha_i > 0$. This states Hölder continuity as presented in the introduction. Remark that (H) belongs indeed to the weakest hypotheses one could consider, since the choice $f_i(\cdot) \overset{\Delta}{=} 1$ boils down to removing the constraint on class i. For the sake of readability, we make use of the shorthand

$$\forall i \in \{1,2,\ldots,c\}, \delta_i(x) \overset{\Delta}{=} f_i(\mathrm{d}(x, \arg\min_{y\in \boldsymbol{S}} \mathrm{d}(x,y))).$$

Let $\vec{\delta}(x)$ denote the corresponding c-components vector for all $\delta_i(x)$. $\forall x \in \boldsymbol{X}$, $P_*(x)$ denotes the Bayesian error on x and $P_{\boldsymbol{S}}(x)$ denotes the error of the NN rule using $\boldsymbol{S}$ on x. P_* and $P_{\boldsymbol{S}}$ are the corresponding errors over the whole $\boldsymbol{X}$. We are now ready to state the first result of Drakopoulos (1995). For the sake of comparison, the theorem is stated locally on every x.

Theorem 1 (Drakopoulos, 1995). *Suppose* (H) *satisfied assuming Hölder continuity.* $\forall x \in \boldsymbol{X}$, *we have*

$$P_*(x) \leqslant P_{\boldsymbol{S}}(x) \leqslant 2P_*(x) - \frac{c}{c-1}P_*(x)^2 + \max_{i\in\{1,2,\ldots,c\}} \delta_i(x).$$

An overall upperbound on $P_{\boldsymbol{S}}$ can be easily obtained by taking the expectations using the probability density function $p(x)$ of X over X. Drakopoulos (1995) obtains the upperbound

$$P_{\boldsymbol{S}} \leqslant 2P_* - \frac{c}{c-1}P_*^2 + \sup_{x\in \boldsymbol{X}} \max_{i\in\{1,2,\ldots,c\}} \delta_i(x). \quad (1)$$

Note the degradation of the penalizing factor depending on $\vec{\delta}(x)$, to ensure easy integration. At the expense of a strong constraint, Drakopoulos (1995) was able to prove a different upperbound, which integrates Bayes optimal risk in the penalizing factor.

(H′) Hypothesis (H) is satisfied and $\forall x \in \boldsymbol{X}$,

$$\sqrt{\sum_{i=1}^{c}\left(\sup_{x\in \boldsymbol{S}} K_i \mathrm{d}(x, \arg\min_{y\in \boldsymbol{S}} \mathrm{d}(x,y))^{\alpha_i}\right)^2} \leqslant 2\sqrt{\frac{c-1}{c}}\left(1 - \frac{c}{c-1}P_*(x)\right).$$

For the sake of simplicity, define

$$L = \sqrt{\sum_{i=1}^{c}\left(\sup_{x\in \boldsymbol{S}} K_i \mathrm{d}(x, \arg\min_{y\in \boldsymbol{S}} \mathrm{d}(x,y))^{\alpha_i}\right)^2}.$$

With the help of (H′), Drakopoulos proves the following theorem

Theorem 2 (Drakopoulos, 1995). *Suppose* (H′) *satisfied assuming Hölder continuity.* $\forall x \in \boldsymbol{X}$, *we have*

$$P_*(x) \leqslant P_{\boldsymbol{S}}(x) \leqslant 2P_*(x) - \frac{c}{c-1}P_*(x)^2 + L\sqrt{\frac{c-1}{c}}\left(1 - \frac{c}{c-1}P_*(x)\right).$$

Integration over X brings the same bounds with the dependencies on x removed. A careful look at the proof of Theorem 2 (Drakopoulos, 1995) even allows to replace for the local risk on every $x \in \boldsymbol{X}$ the quantity L by the smaller one

$$L'(x) = \sqrt{\sum_{i=1}^{c}\left(K_i \mathrm{d}(x, \arg\min_{y\in \boldsymbol{S}} \mathrm{d}(x,y))^{\alpha_i}\right)^2}.$$

Even with that refined bound, it is worthwhile remarking that (H′) is highly restrictive. In particular, on a point x of $\boldsymbol{X}$ where Bayes rule performs only slightly better than a simple coin toss, the variations of class-conditional probabilities are constrained to be almost zero between x and its NN in S, a point which can be very far from x. More generally, the constraint on $Pr(\Theta_X \mid X)$ is very strong when studied over the possible samplings of S. Indeed, the satisfaction of (H′) over all (or many) samplings implies that if the problem admits a point $x \in \boldsymbol{X}$ for which $P_*(x) \approx (c-1)/c$, then *all* points $y \in \boldsymbol{X}$ satisfy $P_*(y) \approx (c-1)/c$. In contrast, the single satisfaction of (H) over all possible samplings of S does not suffer the local influence of some points of $\boldsymbol{X}$, and the constraints it puts over $Pr(\Theta_X \mid X)$ can be reduced provided $\boldsymbol{X}$ has a reasonable finite diameter.

3. Improved results

We begin with some useful definitions for this section. We denote the Bayesian class of x as $m_x = \arg\max_{i\in\{1,2,\ldots,c\}} a_i(x)$. The vectors $\vec{a}_{\backslash m_x}(x)$, $\vec{b}_{\backslash m_x}(x)$ and $\vec{\delta}_{\backslash m_x}(x)$ are the $c-1$ components vectors derived respectively from $\vec{a}(x)$, $\vec{b}(x)$ and $\vec{\delta}(x)$ to which component m_x is removed. The following are general definitions, for some arbitrary vectors $\vec{d}(x)$ and $\vec{e}(x)$ having the same dimension, c, and whose components are referred to as $d_i(x)$ and $e_i(x)$, respectively ($\forall i \in \{1,2,\ldots,c\}$).

$$\boldsymbol{E}(\vec{d}(x)) = \frac{1}{c}\sum_{i=1}^{c} d_i(x), \tag{2}$$

$$\boldsymbol{V}(\vec{d}(x)) = \sum_{i=1}^{c}\left(d_i(x) - \boldsymbol{E}(\vec{d}(x))\right)^2, \tag{3}$$

$$\begin{aligned}\mathrm{Cov}(\vec{d}(x),\vec{e}(x)) = \sum_{i=1}^{c}[(d_i(x) - \boldsymbol{E}(\vec{d}(x)))\\ \times (e_i(x) - \boldsymbol{E}(\vec{e}(x)))].\end{aligned} \tag{4}$$

Hypothesis (H*) is the following one.

$$(\mathrm{H}^*) \qquad \int_{\boldsymbol{X}} \mathrm{Cov}(\vec{a}_{\backslash m_x}(x), \vec{b}_{\backslash m_x}(x)) p(x)\,\mathrm{d}x \geqslant 0.$$

Contrasting with (H$'$), which is used to prove Theorem 2, (H*) has four advantages. First, the constraint is not a local constraint expressed on every $x \in \boldsymbol{X}$. Second, it relies actually on the average behavior over $\boldsymbol{X}$ of certain functions. Third, this constraint does not concern the Bayes classes for each point of $\boldsymbol{X}$, but the other, "minor" classes. Fourth, it seems reasonable to think that, as $\boldsymbol{S}$ grows, on average, the points of $\boldsymbol{X}$ will come close enough to their neighbor in $\boldsymbol{S}$ so that the class conditional probabilities will not fluctuate that much between them, leading to a situation in which (H*) is satisfied. We are now ready to state our main result.

Theorem 3. *Suppose* (H) *and* (H*) *satisfied. We have*

$$P_* \leqslant P_{\boldsymbol{S}} \leqslant 2P_* - \frac{c}{c-1}P_*^2 + \sup_{x\in\boldsymbol{X}} \delta_{m_x}(x)\left(1 - \frac{cP^*}{c-1}\right).$$

Proof. Proving $P_* \leqslant P_{\boldsymbol{S}}$ is trivial. We prove the right inequality. We have

$$P_{\boldsymbol{S}}(x) = 1 - \vec{a}(x)\cdot\vec{b}(x) \tag{5}$$

$$\begin{aligned} &= 1 - (1 - P_*(x))\left(1 - \sum_{i=1,i\neq m_x}^{c} b_i(x)\right)\\ &\quad - \sum_{i=1,i\neq m_x}^{c} a_i(x)b_i(x) \end{aligned} \tag{6}$$

$$= P_*(x) + \sum_{i=1,i\neq m_x}^{c} b_i(x)(1 - P_*(x) - a_i(x)) \tag{7}$$

$$\begin{aligned} &= P_*(x) + \sum_{i=1,i\neq m_x}^{c} a_i(x)(1 - P_*(x) - a_i(x))\\ &\quad + \sum_{i=1,i\neq m_x}^{c} (b_i(x) - a_i(x))(1 - P_*(x) - a_i(x))\\ &= 2P_*(x) - \frac{c}{c-1}P_*(x)^2 - \boldsymbol{V}(\vec{a}_{\backslash m_x}(x))\\ &\quad + \sum_{i=1,i\neq m_x}^{c} (b_i(x) - a_i(x))(1 - P_*(x) - a_i(x)). \end{aligned} \tag{8}$$

We have made use of the following relationships: in (6) $\sum_{i=1}^{c} a_i(x) = \sum_{i=1}^{c} b_i(x) = 1$ and $a_{m_x}(x) = 1 - P_*(x)$, and in (8) we have

$$\begin{aligned} &\sum_{i=1,i\neq m_x}^{c} a_i(x)(1 - P_*(x) - a_i(x))\\ &\quad = P_*(x) - P_*(x)^2 - \sum_{i=1,i\neq m_x}^{c} a_i(x)^2\\ &\quad = P_*(x) - P_*(x)^2 - \boldsymbol{V}(\vec{a}_{\backslash m_x}(x)) - \frac{P_*(x)^2}{c-1}\\ &\quad = P_*(x) - \frac{c}{c-1}P_*(x)^2 - \boldsymbol{V}(\vec{a}_{\backslash m_x}(x)). \end{aligned}$$

We upperbound the factor

$$\sum_{i=1,i\neq m_x}^{c} (b_i(x) - a_i(x))(1 - P_*(x) - a_i(x)) - \boldsymbol{V}(\vec{a}_{\setminus m_x}(x)).$$

First, remark that the quantity

$$(1 - P_*(x)) \sum_{i=1,i\neq m_x}^{c} (b_i(x) - a_i(x))$$
$$= -(1 - P_*(x))(b_{m_x}(x) - a_{m_x}(x)),$$

this due to the fact that $\sum_{i=1}^{c} a_i(x) = \sum_{i=1}^{c} b_i(x) = 1$. Now, we have

$$- \sum_{i=1,i\neq m_x}^{c} (b_i(x) - a_i(x))a_i(x) - \boldsymbol{V}(\vec{a}_{\setminus m_x}(x))$$
$$= - \sum_{i=1,i\neq m_x}^{c} a_i(x)b_i(x) + \frac{P_*(x)^2}{c-1}.$$

The quantity $\mathrm{Cov}(\vec{a}_{\setminus m_x}(x), \vec{b}_{\setminus m_x}(x))$ is also equal to

$$\sum_{i=1,i\neq m_x}^{c} a_i(x)b_i(x) - \frac{1}{c-1} \sum_{i=1,i\neq m_x}^{c} a_i(x) \sum_{j=1,j\neq m_x}^{c} b_j(x),$$

which is also

$$\sum_{i=1,i\neq m_x}^{c} a_i(x)b_i(x) - \frac{P_*(x)^2}{c-1} + \frac{P_*(x)(b_{m_x}(x) - a_{m_x}(x))}{c-1}.$$

We get

$$P_{\boldsymbol{S}}(x) = 2P_*(x) - \frac{c}{c-1} P_*(x)^2 - (b_{m_x}(x) - a_{m_x}(x)) \times \left(1 - \frac{cP^*(x)}{c-1}\right) - \mathrm{Cov}(\vec{a}_{\setminus m_x}(x), \vec{b}_{\setminus m_x}(x)).$$

Integrating over X, using the fact that

$$0 \leqslant \mathrm{var}(P^*(x)) = \int_X P^*(x)^2 p(x)\mathrm{d}x - P^{*^2},$$

and using the fact that $-(b_{m_x}(x) - a_{m_x}(x)) \leqslant |b_{m_x}(x) - a_{m_x}(x)| \leqslant \delta_{m_x}(x)$, we get

$$P_{\boldsymbol{S}} \leqslant 2P_* - \frac{c}{c-1} P_*^2 + \sup_{x\in \boldsymbol{X}} \delta_{m_x}(x) \left(1 - \frac{cP^*}{c-1}\right) - \int_X \mathrm{Cov}(\vec{a}_{\setminus m_x}(x), \vec{b}_{\setminus m_x}(x))p(x)\, \mathrm{d}x.$$

There remains to use hypothesis (H*) to get the desired upperbound. □

When the number of classes increases, this bound becomes better and better with respect to Theorem 2. Ultimately, our penalty $\sup_{x\in \boldsymbol{X}} \delta_{m_x}(x)$ becomes negligible with respect to the quantity $L\sqrt{(c-1)/c}$. On the other side, when $c = 2$, remark that we have $\int_{\boldsymbol{X}} \mathrm{Cov}(\vec{a}_{\setminus m_x}(x), \vec{b}_{\setminus m_x}(x))p(x)\mathrm{d}x = 0$, since the function is 0 everywhere. In that case, we can state the following theorem, which extends all theorems of Drakopoulos (1995) while using only (H).

Corollary 4. *Suppose* (H) *satisfied. Whenever* $c = 2$, *we have*

$$P_* \leqslant P_{\boldsymbol{S}} \leqslant 2P_* - 2P_*^2 + \sup_{x\in \boldsymbol{X}} \delta_{m_x}(x)(1 - 2P^*).$$

For example, the theorem of Drakopoulos (1995) which is proven under the weakest hypothesis among all his results ((H) restricted to Hölder continuity) would only lead to the upperbound

$$P_{\boldsymbol{S}} \leqslant 2P_* - 2P_*^2 + \sup_{x\in \boldsymbol{X}} \max_{i\in\{1,2\}} \delta_i(x).$$

This upperbound is larger than that of Corollary 4 for two reasons. First, its penalty term, $\sup_{x\in \boldsymbol{X}} \max_{i\in\{1,2\}} \delta_i(x)$, is not decreasing as a function of P^*. Second, the dependence on $\vec{\delta}$ uses its maximal component, instead of the (eventually smaller) single component of Bayes class.

Fig. 1 shows a synthetic and pathologic example of a two-classes problem on which our bounds considerably outperform those of Drakopoulos (1995). Here, X takes values over the interval $[a, b]$, and the observations are one dimensional. Note that the class-conditional densities satisfy Hölder continuity, with $\alpha_1 = \alpha_2 \geqslant 1$. Suppose that ϵ is very

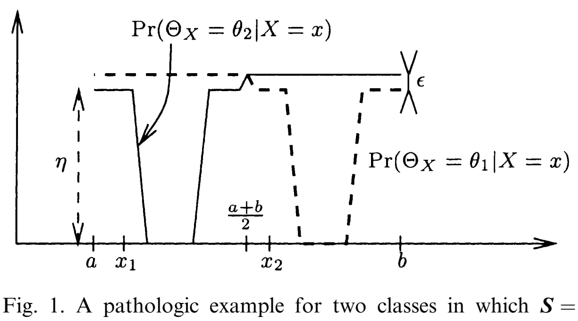

Fig. 1. A pathologic example for two classes in which $\boldsymbol{S} = \{x_1, x_2\}$.

small compared to η. In that case, the bounds of Drakopoulos (1995) in in Eq. (1) gives the penalty term

$$\sup_{x \in \boldsymbol{X}} \max_{i \in \{1,2\}} \delta_i(x) \approx \eta.$$

This term is obtained for the points on the immediate right side of x_1, where $Pr(\Theta_X = \theta_2 \mid X = x)$ approaches zero. Corollary 4 gives the much smaller bound

$$\sup_{x \in \boldsymbol{X}} \delta_{m_x}(x)(1 - 2P^*) \approx \epsilon(1 - 2P^*),$$

which is even smaller than ϵ. This term is obtained for the points $\leqslant (a+b)/2$ for which x_2 is the NN. Suppose now that $P_* = 1/8$, in which case in Eq. (1) gives approximately $P_{\boldsymbol{S}} \leqslant 7/32 + \eta$ whereas our bound in Corollary 4 gives the much better upperbound $P_{\boldsymbol{S}} \leqslant 7/32 + 3\epsilon/4$.

To finish with this illustrative example, suppose that $\boldsymbol{S}$ contains only x_1. In that case, the penalty term of Drakopoulos (1995) approaches $\eta + \epsilon$ (obtained for the points where $Pr(\Theta_X = \theta_1 \mid X = x)$ approaches zero), whereas ours remains the same.

4. Conclusion

In this paper, we have provided new results about the NN risk in the case where the reference sample is fixed and finite. Our bounds (Theorem 3 and Corollary 4), as well as those of Drakopoulos (1995) show that the NN risk in the fixed an finite case can be expressed by the sum of the infinite sample risk plus a penalty term. The theoretical interest of our results is to provide better upperbounds for the the NN risk as opposed to Drakopoulos (1995), while using weaker (and sometimes the weakest) hypotheses. Apart from this general consideration, we think our bounds also provide an interesting glimpse into the way the penalty term behaves, as they show that its influence can be reduced on hard problems, for which P_* tends to be high. Another interesting feature of our results is that they hold in a general and practical setting as opposed to Cover (1968), because in experimental works the reference set $\boldsymbol{S}$ is fixed and usually of more or less restricted size. There is however a price to pay to cast the results into this new setting. This is the knowledge of a bound on the variations of class-conditional probabilities in the domain not covered by $\boldsymbol{S}$, but it appears that this is not a difficulty really hard to bypass. Indeed, conventional statistical analyses such as inferential statistics heavily rely on distributional assumptions such as normality, thereby leading in our case to computable bounds for the penalty term. Modulo this analytical step, we think our bounds can be of higher practical potential than those of Cover (1968) and Drakopoulos (1995) to evaluate the real difficulty of a pattern recognition problem by computing bounds for P_*. This is an important problem which early motivated the obtention of risk bounds for the NN rule Cover (1968, 1995), and eventually contributed to its widespread and use.

References

Cover, T., 1968. Rates of convergence for nearest neighbor procedures. In: Proc. First Hawaii Conf. on System Sciences, pp. 413–415.

Cover, T., Hart, P.E., 1967. Nearest Neighbor pattern classification. IEEE Trans. Infomation Theory, 21–27.

Drakopoulos, J.A., 1995. Bounds on the classification error of the nearest neighbor rule. In: Proc. 12th International Conference on Machine Learning, pp. 203–208.

Fix, E., Hodges, J.L., 1951. Discrimatory analysis nonparametric discrimination. Technical Report TR-21-49-004, Rept 4, USAF School of Aviation Medicine, Randolph Field TX.

Okamoto, S., Yugami, N., 1996. Theoretical analysis of the nearest neighbor classifier in noisy domains. In: Proc. 13th International Conference on Machine Learning, pp. 355–363.

Snapp, R.R., Venkatesh, S.S., 1998. Asymptotic derivation of the finite-sample risk of the k nearest neighbor classifier. Technical Report UVM-CS-1998-0101, University of Vermont, Burlington.

Venkatesh, S.S., Snapp, R.R., Psaltis, D., 1992. BELLMAN STRIKES AGAIN! the growth rate of sample complexity with dimension for the nearest neighbor classifier. In: Proc. 5th International Conference on Computational Learning Theory, pp. 93–102.