DE GRUYTER OPEN

JAISCR, 2013, Vol. 3, No. 3, pp. 229– 242

DOI 10.2478/jaiscr-2014-0016

A SURVEY OF OLD AND NEW RESULTS FOR THE TEST ERROR ESTIMATION OF A CLASSIFIER

Davide Anguita, Luca Ghelardoni, Alessandro Ghio and Sandro Ridella

Department of Biophysical and Electronic Engineering, University of Genova,
Via Opera Pia 11A, 16145 Genova, Italy

Abstract

The estimation of the generalization error of a trained classifier by means of a test set is one of the oldest problems in pattern recognition and machine learning. Despite this problem has been addressed for several decades, it seems that the last word has not been written yet, because new proposals continue to appear in the literature. Our objective is to survey and compare old and new techniques, in terms of quality of the estimation, easiness of use, and rigorousness of the approach, so to understand if the new proposals represent an effective improvement on old ones.

1 Introduction

The main objective of supervised learning methods, like *Neural Networks* (*NNs*) [1] and *Support Vector Machine* (*SVM*) [2], consists in estimating an input-output relationship by exploiting a set of patterns, named the *training set*. In general, the probability distribution originating the data is unknown and additional *a priori* information is seldom available, therefore this relationship must be inferred exclusively from the available data.

A classifier is considered to be reliable and effective if it does not simply *memorize* the training samples, but if it can *learn* from them: in other words, a “good” trained model must be able to capture the underlying phenomenon characterizing the observed patterns and correctly predict the labels of new and previously unobserved inputs, if originated from the same distribution of the training set. This capacity is defined as the *generalization ability* of a classifier, while the misclassification rate on unseen patterns is known as the *generalization error* rate.

Obviously, if we knew the probability distribution of the data, we could exactly compute the generalization error. However, in practice, statistical estimates are computed, instead, which consist in upper bounding the generalization error, in probability, according to some user–defined confidence. This approach is not only of theoretical interest, but of paramount importance in many application fields, such as, for example, forensic statistics [3].

The use of statistical upper bounds for measuring the generalization ability of a classifier has been extensively addressed, in recent years, by the Machine Learning community [4, 5, 6, 7]. Two main categories of methods exist to face this problem: *in-sample* and *out-of-sample* techniques [8, 9]. *In-sample* procedures, such as the Maximal Discrepancy and the Rademacher Complexity [4, 10], compute an estimation of the generalization error without exploiting a separate test set: in fact, the training set is used for both building the model and estimating the misclassification rate. Though providing new results and insights on the problem of error estimation and on the classification algorithms themselves, these methods aim at covering widely general cases and then result to be often inapplicable in practice [11]. In this work, instead, we focus on the more widespread *out-of-sample* methods [9, 12], where the generalization error is estimated by exploiting a separate *test set*, whose samples are independent of the training data. Many statistical results are available in the literature for this purpose: the objective of this paper is to survey some old and new ones and compare their performance under different conditions (e.g. by varying the amount of test data), in order to draw on and expand the analysis performed in [13].

Note that the test set approach can also be extended to resampling techniques [14], as shown in the Appendix A of this paper in the case of the well-known *K-fold Cross Validation* (*KCV*) [15], where the procedure of splitting the original dataset into a training and a test set is iterated several times. As this and other resampling methods extract only a limited amount of samples, to be used as a test set, the identification of effective methods (i.e. tight generalization error bounds) becomes quite important in these cases as well.

1.1 The Generalization Error of a Classifier

In this work we focus our attention on binary classifiers, as the extension of the results to the multi-class case is straightforward and it would unnecessarily complicate our analysis. Let us consider a test set of n i.i.d. patterns $\mathcal{X} = \{(x_i, y_i)\}$, $i = 1, ..., n$, originated from the same unknown distribution $P(x, y)$, which generated the training data, where $x_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$. Let $\hat{y} = f(x)$ be a classifier such that $f : \mathbb{R}^d \to \mathcal{Y}_f \subseteq \mathbb{R}$. We can define the empirical error of f on the test set $\mathcal{X}$ as[1]:

$$\hat{L}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(\hat{y}_i, y_i), \tag{1}$$

where we introduced the *loss function* $\ell : \mathcal{Y}_f \times \{\pm 1\} \to \mathcal{L} \subseteq [0, 1]$, which measures the discrepancy between the true label and the response of the classifier. In classification problems, where we are often interested in simply counting the number of misclassifications, a binary (e.g. *hard*) loss function, characterized by $\mathcal{L} = \{0, 1\}$, can be used:

$$\ell_H(\hat{y}, y) = \begin{cases} 1 & \text{if} \quad \text{sign}(\hat{y}) \neq y \\ 0 & \text{if} \quad \text{sign}(\hat{y}) = y \end{cases} \tag{2}$$

where

$$\text{sign}(y) = \begin{cases} 1 & \text{if} \quad y > 0 \\ -1 & \text{if} \quad y \leq 0. \end{cases} \tag{3}$$

As an alternative to the hard loss, we can exploit Lipschitz continuous *soft loss* functions when, instead of predicting a label, we are interested in estimating a posterior class probability (i.e. $\mathcal{L} = [0, 1]$). Two possible examples are the *logistic loss* [17]

$$\ell_{log}(\hat{y}, y) = \frac{e^{-\alpha y \hat{y}}}{1 + e^{-\alpha y \hat{y}}}, \tag{4}$$

where $\alpha > 0$ and the *linear soft loss*:

$$\ell_S(\hat{y}, y) = \begin{cases} 1 & \text{if} \quad \hat{y}y < -1 \\ \frac{1-\hat{y}y}{2} & \text{if} \quad -1 \leq \hat{y}y \leq 1 \\ 0 & \text{if} \quad \hat{y}y > 1 \end{cases} \tag{5}$$

which is a piecewise linear approximation of the former.

Given a fixed loss function and exploiting the information on the test set (e.g. the empirical error $\hat{L}_n(f)$), we are interested in finding a reliable estimation of the generalization error:

$$L(f) = \mathbb{E}\{\ell(\hat{y}, y)\}, \tag{6}$$

which cannot be explicitly computed as the probability distribution of the data is unknown. As it is well-known that $\hat{L}_n(f)$ usually underestimates $L(f)$, a theoretically rigorous approach for estimating the generalization capability of f is to compute an upper-bound of the error

$$L(f) \leq \hat{L}(f) \triangleq \hat{L}_n(f) + \Delta(f, \mathcal{X}, \delta), \tag{7}$$

which holds with a *coverage probability*

$$\mathcal{C}(L(f), n) = Pr\{L(f) \leq \hat{L}(f)\} = (1 - \delta), \tag{8}$$

and where δ is a user-defined confidence level. In Eq. (7), the additional term $\Delta(f, \mathcal{X}, \delta)$ represents the *unluckiness factor* of the bound, that is an estimation of the difference between the empirical and the generalization error, which depends on the particular classifier, the available test data, and the chosen confidence level. Usually, $\Delta(f, \mathcal{X}, \delta)$ decreases as the cardinality n of $\mathcal{X}$ increases, but can be quite large for small datasets, so we are interested in very tight estimations of this quantity.

In the next sections, we will analyze old and new bounds, under different conditions (e.g., by varying the number of patterns n and by using different loss functions $\ell(\cdot, \cdot)$), in order to identify the approach which gives the tightest $\Delta(f, \mathcal{X}, \delta)$. We will distinguish two different scenarios, related to

[1] In this work, we adopt the notation used in [4] and [16].

the previously mentioned loss functions. In Section 2, we will focus the attention on the binary loss function, which gives rise to an error distributed according to a binomial distribution, while in Section 3 the use of continuous real-valued loss functions will be analyzed. Finally, some experimental results will be presented in Section 4, where the behavior of the different bounds will be shown. To simplify the notation, we will omit the explicit dependency of both the empirical and the generalization error from the classifier f because, in this framework, the classifier is fixed after the training phase (i.e. L is used, instead of $L(f)$).

1.2 Preliminary Definitions

In this section we define some important quantities, which will be exploited in the following for estimating the generalization error bounds: the *variance* σ^2 and the *sample variance* s^2 of the empirical error.

Based on the definition of the empirical error of Eq. (1), the variance σ^2 is defined as:

$$\begin{aligned}\sigma^2 &= \mathbb{E}\{\hat{L}_n^2\} - L^2 \\ &= \mathbb{E}\{\hat{L}_n^2\} - \mathbb{E}\{\hat{L}_n\}^2. \end{aligned} \tag{9}$$

Analogously, we can also introduce the sample variance, computed according to the available text set $\mathcal{X}$ as:

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{n}[\ell(\hat{y}_i, y_i) - \hat{L}_n]^2 \\ &= \frac{1}{n}\left\{\left[\sum_{i=1}^{n}(\ell(\hat{y}_i, y_i))^2\right] - n\hat{L}_n^2\right\}. \end{aligned} \tag{10}$$

In the following sections, we will further deepen the analysis of these quantities, by exploiting the peculiarities of the different loss functions considered in the two scenarios.

2 Scenario 1 - Binary Loss Function

In the first scenario, we focus the attention on random variables that can assume only two different values, therefore the number of misclassifications is distributed according to a binomial distribution.

By taking into account the quantities defined in Section 1.2, it is worth noting that, in this scenario:

$$\mathbb{E}\{\hat{L}_n^2\} = \mathbb{E}\{\hat{L}_n\} = L. \tag{11}$$

As a straightforward consequence, Eq. (9) can be written as

$$\sigma^2 = L(1-L). \tag{12}$$

Analogously, by taking into account the sample variance of Eq. (10), it is possible to note that

$$\sum_{i=1}^{n}(\ell(\hat{y}_i, y_i))^2 = \sum_{i=1}^{n}(\ell(\hat{y}_i, y_i)) \tag{13}$$

and, by consequence,

$$s^2 = \hat{L}_n(1-\hat{L}_n). \tag{14}$$

2.1 Normal Approximation

A simple formula for the estimation of the confidence interval can be derived by the application of the central limit theorem and, consequently, from the approximation of the binomial by a normal distribution [19]. Under this hypothesis, the inequality, which represents the upper-bound of the confidence interval, can be written as:

$$\frac{L-\hat{L}_n}{\sqrt{\frac{\sigma^2}{n}}} \leq z_{1-\delta}, \tag{15}$$

where n is the sample size, σ^2 is the variance and $z_{1-\delta}$ denotes the $(1-\delta)$-th percentile of the standard normal distribution.

As the variance is unknown, we can approximate σ^2 by exploiting the definition of sample variance of the empirical error, according to Eq. (14). Then, an explicit upper-bound for the generalization error formulation can be obtained:

$$L \leq \hat{L}_n + z_{1-\delta}\sqrt{\frac{\hat{L}_n\left(1-\hat{L}_n\right)}{n}}, \tag{16}$$

or, equivalently:

$$\Delta(f, \mathcal{X}, \delta) = z_{1-\delta}\sqrt{\frac{\hat{L}_n\left(1-\hat{L}_n\right)}{n}}, \tag{17}$$

according to Eq. (7).

Unfortunately, as the binomial distribution could noticeably differ from the normal approximation, especially when the number of test patterns n is small, and the sample variance could differ from the true one, the bound is not rigorous and its coverage probability could fall well below the user-defined confidence level $(1-\delta)$.

2.2 Wilson Score Approach

We can avoid approximating σ^2 with s^2 thank to the Wilson score approach [20], which exploits Eq. (12), resulting in the following inequality:

$$\frac{L-\hat{L}_n}{\sqrt{\frac{L(1-L)}{n}}} \leq z_{1-\delta}. \tag{18}$$

The explicit formulation for the generalization error can be computed by solving the previous inequality respect to L:

$$L \leq \frac{\hat{L}_n+\frac{1}{2n}z^2_{1-\delta}}{1+\frac{1}{n}z^2_{1-\delta}}+\frac{z_{1-\delta}\sqrt{\frac{1}{4n^2}z^2_{1-\delta}+\frac{\hat{L}_n\left(1-\hat{L}_n\right)}{n}}}{1+\frac{1}{n}z^2_{1-\delta}}, \tag{19}$$

and we can explicitly compute the unluckiness factor, obtaining:

$$\Delta(f,\mathcal{X},\delta)=\frac{(1-2\hat{L}_n)A+B}{2(1+A)}, \tag{20}$$

where A and B are constants such that

$$A = \frac{1}{n}z^2_{1-\delta} \tag{21}$$

$$B = \sqrt{\frac{1}{n^2}z^2_{1-\delta}+\frac{4\hat{L}_n(1-\hat{L}_n)}{n}}. \tag{22}$$

Note that, analogously to the Normal Approximation presented in Section 2.1, the value of the Wilson Score bound is not rigorous, due to the approximation of the binomial distribution with the normal one.

2.3 Clopper-Pearson Bound

We can avoid the normal approximation, by exploiting the exact Clopper-Pearson approach [21]. This is obtained by expressing the probability that $p=n\hat{L}_n$ (or less) misclassifications are obtained, when classifying n samples, for which the generalization error equals L, as [22]:

$$\begin{aligned} Bin(n,p,L) &= Pr\left(\sum_{i=1}^{n}\hat{y}_i \leq p\right) = \\ &= \sum_{j=0}^{p}\binom{n}{j}L^j(1-L)^{n-j}. \end{aligned} \tag{23}$$

In order to derive the bound for the generalization error in the form of Eq. (7), we can consider the inverse of the binomial tail, i.e.:

$$\hat{L}=\overline{Bin}(n,p,\delta) \tag{24}$$

where

$$\overline{Bin}(n,p,\delta) \equiv \max_{\hat{L}}\left\{\hat{L}: Bin(n,p,\hat{L}) \geq \delta\right\}, \tag{25}$$

which can be solved numerically. By exploiting the previous relation, we can formulate the Clopper-Pearson bound for the generalization error as

$$L \leq \overline{Bin}(n,n\hat{L}_n,\delta), \tag{26}$$

which rigorously holds with probability $(1-\delta)$. Alternatively, the unluckiness factor can be simply derived from Eq. (26):

$$\Delta(f,\mathcal{X},\delta)=\overline{Bin}(n,n\hat{L}_n,\delta)-\hat{L}_n. \tag{27}$$

3 Scenario 2 - Bounded Real-Valued Loss Function

In this section, we focus our attention on the analysis of random variables, which can assume any value within the interval[2] $[0,1]$. The bounds presented here can be exploited when the soft loss functions $\ell_S(\cdot,\cdot)$ of Eq. (5) or $\ell_{log}(\cdot,\cdot)$ of Eq. (4) are used for measuring the generalization error of a classifier. It is worth observing that the bounds presented in this section can also be exploited when a binary loss function is used, at least in principle. However, the Clopper-Pearson is obviously superior in this case, because it takes in account the additional information given by modelling the errors using a binomial distribution. The current scenario, instead, asks for nonparametric methods because the distribution of the errors is continuous and bounded but unknown.

Let us consider again the quantities introduced in Section 1.2. Eq. (11) can be reformulated as:

$$\mathbb{E}\{\hat{L}_n^2\} \leq \mathbb{E}\{\hat{L}_n\}=L. \tag{28}$$

Then, Eq. (9) becomes

$$\sigma^2 \leq L(1-L). \tag{29}$$

[2] Obviously, a generalization of these results to any bounded interval $[a,b]$ is trivially obtained by a simple rescaling argument.

Analogously, it is possible to note that

$$\sum_{i=1}^{n}(\ell(\hat{y}_i,y_i))^2 \leq \sum_{i=1}^{n}(\ell(\hat{y}_i,y_i)) \tag{30}$$

so to reformulate Eq. (10) as

$$s^2 \leq \hat{L}_n(1-\hat{L}_n). \tag{31}$$

3.1 Chebyshev Bound

The Chebyshev inequality represents the generalization of the Wilson Score approach to the case of continuous bounded loss functions [23]:

$$(L-\hat{L}_n)^2 \leq \frac{\sigma^2}{n\delta} \tag{32}$$

but, differently from Eq. (19), this bound is rigorous, i.e. it holds with a coverage probability at least equal to $(1-\delta)$.

By applying Eq. (29) we get:

$$(L-\hat{L}_n)^2 \leq \frac{L(1-L)}{n\delta}. \tag{33}$$

Then, we can solve Eq. (33) with respect to L and obtain a rigorous and explicit upper bound for the generalization error. In particular, the unluckiness factor assumes the same formulation as in Eq. (20) (as expected, because the Chebyshev bound is a generalization of the Wilson Score interval to the case of continuous bounded soft loss functions) but, in this case, A and B are constants such that

$$A = \frac{1}{\delta n} \tag{34}$$

$$B = \sqrt{A(A-4\hat{L}_n^2+4\hat{L}_n)}. \tag{35}$$

3.2 Guttman Bound

A less–known bound was found by Guttman [24], which uses the actual sample variance in addition to the true variance value and, in general, is tighter than Eq. (32). With probability $(1-\delta)$:

$$(L-\hat{L}_n)^2 \leq \frac{s^2}{n-1}+\frac{1}{\sqrt{\delta}}\sigma^2\sqrt{\frac{2}{n(n-1)}}. \tag{36}$$

By applying the upper-bound of Eq. (31) and solving with respect to L, we can obtain the explicit Guttman bound for the generalization error, which obviously still holds with a coverage probability equal to $(1-\delta)$. The unluckiness factor can be written in the same form as Eq. (20), where the two constants assume the following values:

$$A = \frac{1}{\sqrt{\delta}}\sqrt{\frac{2}{n(n-1)}} \tag{37}$$

$$\mathrm{B} = \sqrt{A\left(4\hat{L}_n-4\hat{L}_n^2+\tfrac{4s^2}{n-1}+A\right)+\tfrac{4s^2}{n-1}}. \tag{38}$$

3.3 Bernstein Bound

A useful bound for estimating the generalization ability of a classifier, when a real-valued indicator is used for the error, is the Bernstein bound [25]. The estimated error can be expressed as:

$$L \leq \hat{L}_n+\sigma\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}+\frac{\log\left(\frac{1}{\delta}\right)}{3n}. \tag{39}$$

As, once again, σ is usually unavailable in practice, we can exploit the upper-bound of Eq. (31) and write:

$$L \leq \hat{L}_n+\sqrt{L(1-L)}\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}+\frac{\log\left(\frac{1}{\delta}\right)}{3n}, \tag{40}$$

whose solution can be found through a numerical procedure.

3.4 Maurer-Pontil Bound

Because of the application of the upper-bound of Eq. (31) in Eq. (39), the estimation of the generalization error through the Bernstein approach requires a numerical procedure for computing the unluckiness factor $\Delta(f,\mathcal{X},\delta)$. Recently, a new bound has been proposed, which overcomes this problem by exploiting the sample variance instead of the true one [26, 27]:

$$L \leq \hat{L}_n+s\sqrt{\frac{2\log\left(\frac{2}{\delta}\right)}{n}}+\frac{7\log\left(\frac{2}{\delta}\right)}{3(n-1)}. \tag{41}$$

This last inequality can be considered an empirical formulation of the Bernstein bound, which is still rigorous but explicit and, therefore, easy to compute.

3.5 Chernoff Bound

If we want to avoid the numerical procedure for computing the upper-bound of the generalization error L, as an alternative to the Maurer-Pontil bound, the estimation through the Chernoff approach can be used instead [28]. In this case, only the information concerning the empirical error rate is exploited and the unluckiness factor can be expressed as:

$$\Delta(f,\mathcal{X},\delta) = \sqrt{\frac{2\hat{L}_n \log\left(\frac{1}{\delta}\right)}{n}} + \frac{2\log\left(\frac{1}{\delta}\right)}{n}. \tag{42}$$

As the Maurer-Pontil bound, the Chernoff bound is rigorous, holding with probability $(1-\delta)$.

3.6 Hoeffding Bound

One of the oldest bounds that can be used for estimating the generalization ability of a classifier is the Hoeffding bound [29], which is still object of investigation [30, 31] and is not widespread, because its formulation is not easy to deal with. The bound is

$$Pr\left\{L \geq \hat{L}_n + \Delta\right\} \leq \left[\left(\frac{1-L}{1-L+\Delta}\right)^{1-L+\Delta}\left(\frac{L}{L-\Delta}\right)^{L-\Delta}\right]^n, \tag{43}$$

where the dependency of Δ from the classifier, the data and the confidence value has been omitted for the sake of simplicity. If we consider the worst–case scenario, $L = \hat{L}_n + \Delta$, we can avoid L to appear in the right term of the inequality and rewrite Eq. (43) as:

$$Pr\left\{L \geq \hat{L}_n + \Delta\right\} \leq \left[\left(\frac{1-\hat{L}_n-\Delta}{1-\hat{L}_n}\right)^{1-\hat{L}_n}\left(\frac{\hat{L}_n+\Delta}{\hat{L}_n}\right)^{\hat{L}_n}\right]^n. \tag{44}$$

By setting the right side of Eq. (44) equal to δ, it is possible to find a bound in the form of Eq. (7), characterized by $\mathcal{C}(L,n) = (1-\delta)$, as we did in the previous cases. Unfortunately, this last equality does not have a closed–form solution and must be solved numerically. Furthermore, note that a solution can be found only when $\hat{L}_n \in (0,1)$. The case $\hat{L}_n = 0$ can be computed[3] as the limit $\hat{L}_n \to 0$ to obtain:

$$Pr\left\{L - \hat{L}_n \geq \Delta\right\} \leq (1-\Delta)^n. \tag{45}$$

By setting $\delta = (1-\Delta)^n$, we obtain the explicit value of the unluckiness factor for the Hoeffding bound when no patterns are misclassified by f in the test set:

$$L \leq \Delta(f,\mathcal{X},\delta) = 1 - \sqrt[n]{\delta}. \tag{46}$$

Note that, in the current and past literature, Eq. (44) is seldom used, while a more elegant, but much looser, formulation [29] is usually cited:

$$Pr\left\{L \geq \hat{L}_n + \Delta\right\} \leq e^{-2n\Delta^2} \tag{47}$$

By setting $\delta = e^{-2n\Delta^2}$, we obtain the following explicit formulation:

$$L \leq \hat{L}_n + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}}, \tag{48}$$

which is characterized by a coverage probability equal to $(1-\delta)$ and represents the most widely applied Hoeffding's result.

4 Experimental Results

In this section, we are interested in comparing the tightness of the bounds, which we previously presented and which are summarized in Table 1. Our target is to provide useful insights on the test error estimation in both presented scenarios, by estimating the generalization error $L(f)$ according to Eq. (7).

4.1 Scenario 1 - Results

As a first step, we analyze the three generalization error bounds fixing the confidence level and the empirical misclassification rate, while varying the number of patterns in the test set. For this purpose, let us suppose to set $\delta = 5\%$ and let us vary the cardinality of the test set $\mathcal{X}$ in the range $n \in [10, 200]$, in accordance with the experiments performed elsewhere in the literature (e.g. see [19]). Large datasets are not taken in account for two main reasons:

[3] The value of the bound for $\hat{L}_n \to 1$ can be analogously obtained but it is obviously not interesting as it corresponds to a classifier that exhibits a 100% error on the test set.

Table 1. Bounds introduced in Sections 2 and 3: the last column indicates the short name, that will be used for the experimental results.

Scenario 1		
Bound	**Eq.**	**Short name**
Normal approximation	(16)	NOR
Wilson score approach	(20), (22), (22)	WIL
Clopper-Pearson bound	(26)	CP
Scenario 2		
Bound	**Eq.**	**Short name**
Chebyshev bound	(20), (34), (35)	CHE
Guttman bound	(20), (37), (38)	GUT
Bernstein bound	(40)	BER
Maurer-Pontil bound	(41)	MAU
Chernoff bound	(42)	CRF
Tighter Hoeffding bound	(44), (45)	T_HOE
Conventional Hoeffding bound	(48)	HOE

- as n increases, the bounds tend to be characterized by a similar performance;
- as we noted in the introduction, these bounds can also be exploited when applying the KCV procedure (see also Appendix A), where only small subsets of the dataset are available for testing purposes (e.g. 10% of the original dataset [12]).

Figs. 1 and 2 show the estimations $\hat{L}$ (see Eq. (7)) for the three bounds in the case of $\hat{L}_n = 25\%$ and $\hat{L}_n = 0\%$, respectively. We note that:

- WIL and CP give similar $\hat{L}$ values, even if it is worth observing that WIL is not rigorous and, then, there is no guarantee about the coverage probability of the bound;
- NOR always outperforms WIL and CP, but, besides being non-rigorous, its error estimation is useless when $\hat{L}_n$ is small, as predicted by theory and as clearly shown in Fig. 2 (where $\hat{L} = 0\%$ $\forall n$ using NOR).

Now we fix $n = 10$ and $n = 200$ (the extreme values we used for the previous experiments) and we plot the trend of $\hat{L}$ by varying $\hat{L}_n$ in the range of interest for classification problems $[0\%, 50\%]$. We observe that:

- when n is small (Fig. 3), CP results to be sometimes looser than the non-rigorous approaches NOR and WIL;
- as n increases (Fig. 4), the three bounds tend to predict similar values of L, as expected by the theory (i.e. the binomial distribution can be safely approximated by a normal distribution).

As CP is sometimes loose, it is interesting to compute the true coverage probability for the three bounds [19], so to verify if the looseness is justified by rigorousness requirements. Different trends of $\mathcal{C}(L,n)$ are shown:

- Figs. 5 and 6 present the coverage probability, when n is set, respectively, to 10 and 200 and L is varied in the range $[0\%, 50\%]$. CP is the only guaranteed bound, as expected, while both WIL and NOR, despite giving tighter estimates for L, are characterized by a coverage probability which sometimes falls below the nominal confidence ($1 - \delta = 95\%$), even when n is increased;
- Figs. 7 and 8 present the coverage probability, when L is respectively set[4] to 1% and 25% and n is varied in the range $[10, 200]$. As in the previous experiments, CP represents the only guaranteed bound, even if WIL results to be a valuable alternative, albeit non–rigorous, when either the

[4] The case $L = 0\%$ is avoided because it is easy to verify that, for NOR, WIL and CP, $\mathcal{C}(0\%, n) = 100\%$, $\forall n$.

true L is small or the nominal confidence boundary can be violated in practice.

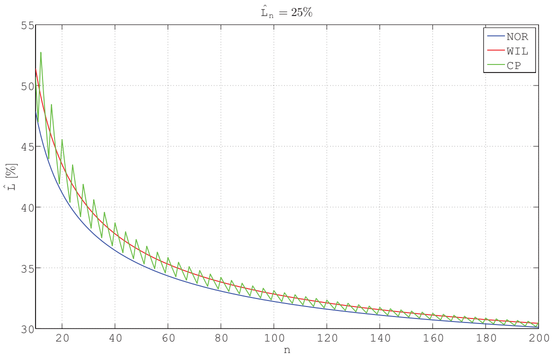

Figure 1. Scenario 1 - Trend of $\hat{L}$ ($\hat{L}_n = 25\%$).

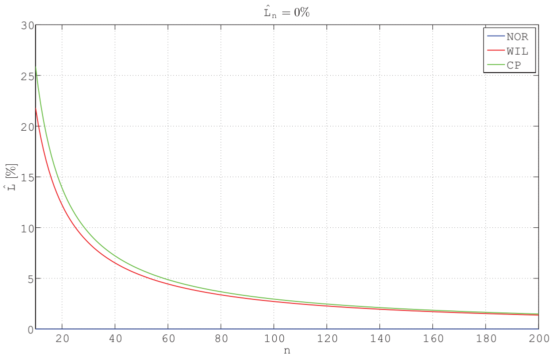

Figure 2. Scenario 1 - Trend of $\hat{L}$ ($\hat{L}_n = 0\%$).

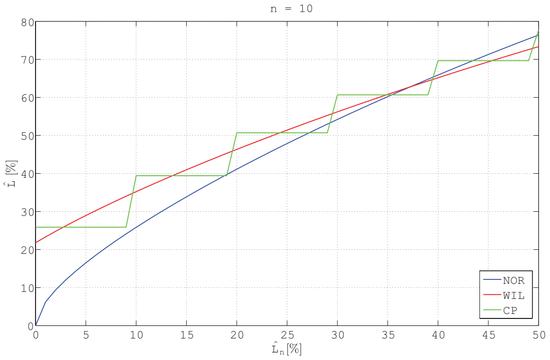

Figure 3. Scenario 1 - Trend of $\hat{L}$ ($n = 10$).

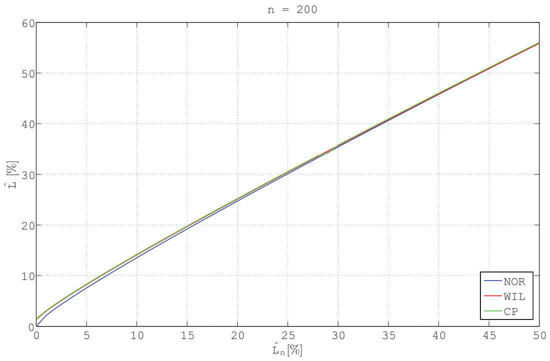

Figure 4. Scenario 1 - Trend of $\hat{L}$ ($n = 200$).

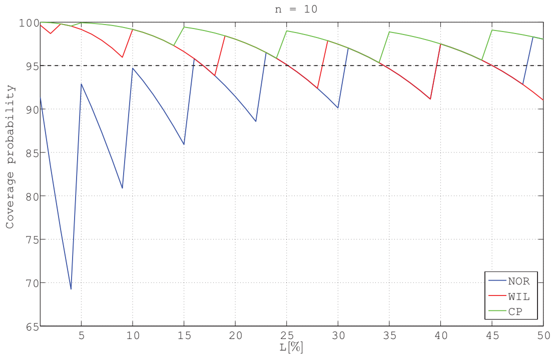

Figure 5. Scenario 1 - Coverage probability varying L ($n = 10$).

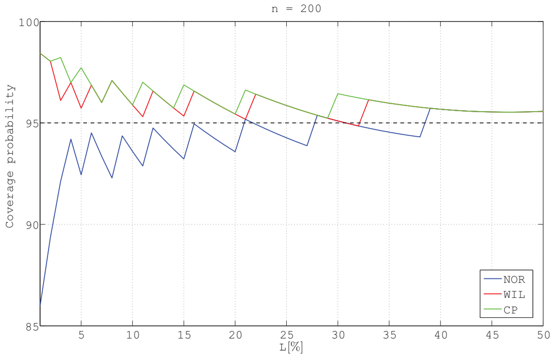

Figure 6. Scenario 1 - Coverage probability varying L ($n = 200$).

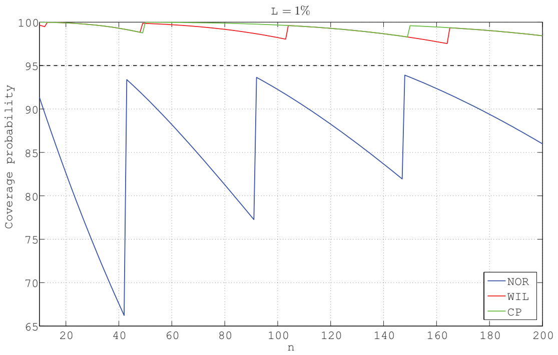

Figure 7. Scenario 1 - Coverage probability varying n ($L = 1\%$).

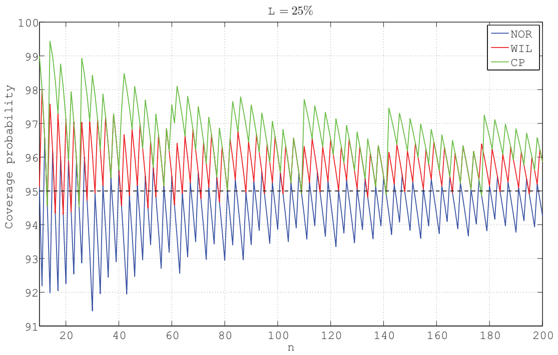

Figure 8. Scenario 1 - Coverage probability varying n ($L = 25\%$).

4.2 Scenario 2 - Results

We analyze now the second scenario, presented in Section 3, where a real-valued bounded loss function (e.g., $\ell_S(\cdot,\cdot)$ or $\ell_{log}(\cdot,\cdot)$) is used. Our objective is to verify the tightness of the bounds, computed using the seven possible choices: GUT, CHE, HOE, T_HOE, BER, MAU and CRF, which are all rigorous approaches. In the computation of GUT and MAU, we have to fix the sample variance s^2. For this purpose, as a real-valued continuous loss function is used, we can exploit Eq. (31) and set the sample variance to the worst case value $s^2 = \hat{L}_n(1 - \hat{L}_n)$.

We plot the trend of the bounds by varying, alternatively, n and $\hat{L}_n$. In particular:

- Fig. 9 shows the predicted generalization error values when we fix $\hat{L}_n = 25\%$ and we let n vary in the range of interest. T_HOE allows to find the tightest estimation, but it requires a numerical procedure for finding the value of $\hat{L}$. As a valuable alternative, the conventional Hoeffding bound (HOE) is explicit, is trivially computable and, in these cases, guarantees an acceptable performance;
- Fig. 10 presents the trend of $\hat{L}$ by varying $n \in [10, 200]$ and fixing $\hat{L}_n = 0\%$: as shown in Section 3.6, the explicit formulation of Eq. (46) must be used for T_HOE, which still guarantees the optimal performance and, in this case, does not need any numerical procedure;
- Figs. 11 and 12 present the values of $\hat{L}$ when n is respectively set to 10 and 200 and $\hat{L}_n$ is varied in the range $[0\%, 50\%]$. As in the previous experiments, T_HOE provides the tightest estimations for every value of n and $\hat{L}_n$. If the user wants to avoid a numerical procedure for finding $\hat{L}$ when $\hat{L}_n > 0\%$, CRF (especially when n is large and $\hat{L}_n$ is small) and HOE can be effectively used, instead. BER could be a valuable alternative, but it requires a numerical procedure as well.

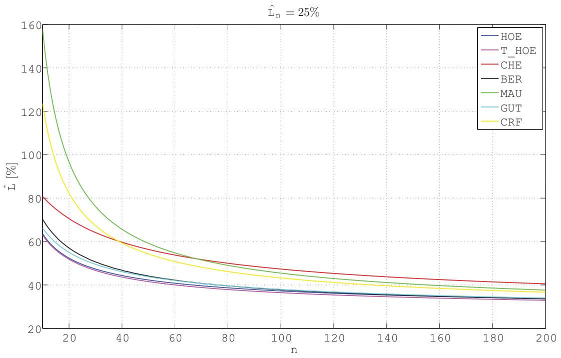

Figure 9. Scenario 2 - Trend of $\hat{L}$ ($\hat{L}_n = 25\%$).

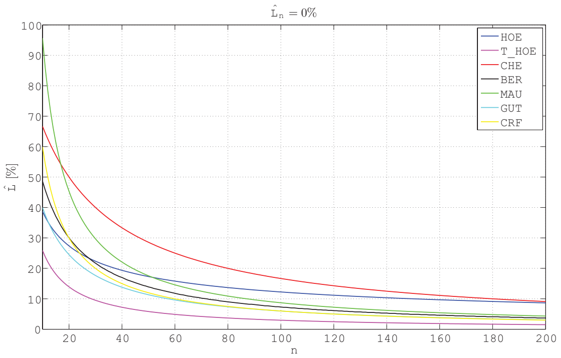

Figure 10. Scenario 2 - Trend of $\hat{L}$ ($\hat{L}_n = 0\%$).

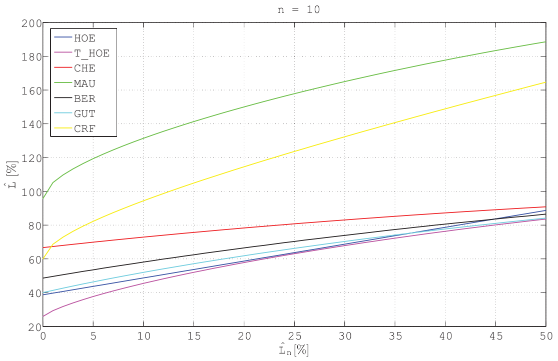

Figure 11. Scenario 2 - Trend of $\hat{L}$ ($n = 10$).

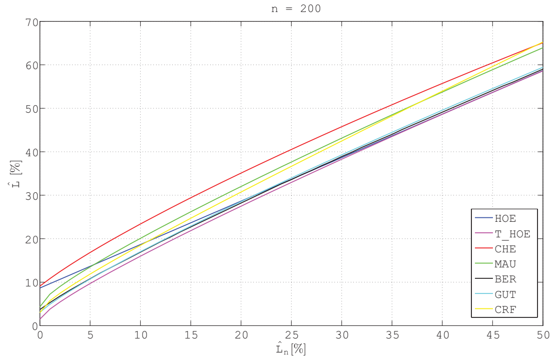

Figure 12. Scenario 2 - Trend of $\hat{L}$ ($n = 200$).

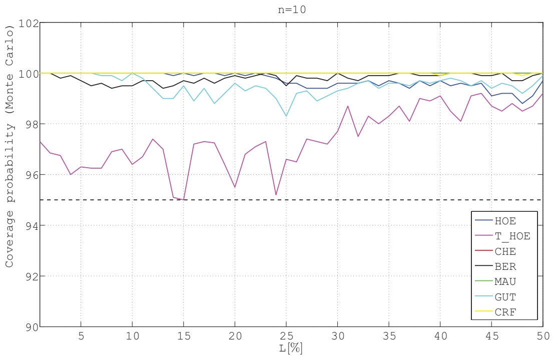

Figure 13. Scenario 2 - Experimental $\mathcal{C}(L,n)$ varying L ($n = 10$).

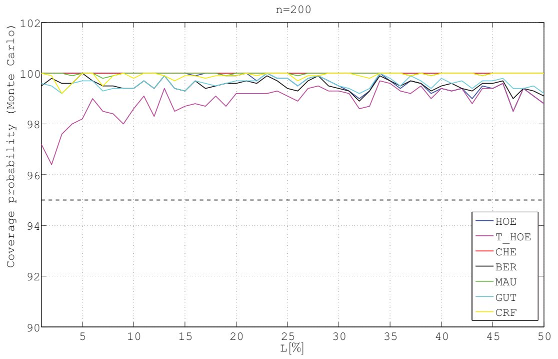

Figure 14. Scenario 2 - Experimental $\mathcal{C}(L,n)$ varying L ($n = 200$).

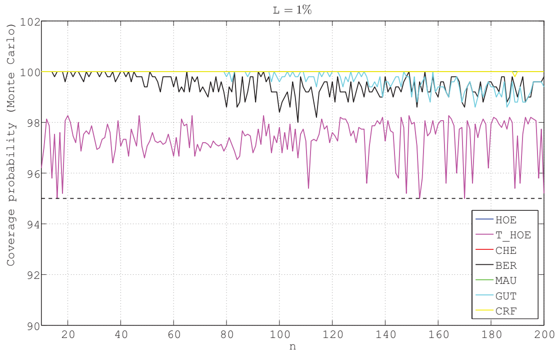

Figure 15. Scenario 2 - Experimental $\mathcal{C}(L,n)$ varying n ($L = 1\%$).

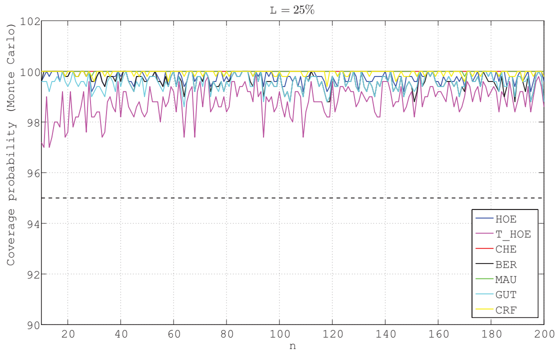

Figure 16. Scenario 2 - Experimental $\mathcal{C}(L,n)$ varying n ($L = 25\%$).

Our last tests address the estimation of the coverage probability of the bounds mentioned before. For this purpose, we use a Monte Carlo method for computing $\mathcal{C}(L,n)$, where the following procedure is repeated τ times for every (L,n) pair:

- we create a set of n error values, sampled from a normal distribution with mean equal to L and variance equal to $L(1-L)$, which is the worst case of Eq. (29);
- we compute the estimation of the generalization error $\hat{L}$ (see Eq. (7)) using GUT, CHE, HOE, T_HOE, BER, MAU and CRF;
- we check if $L \leq \hat{L}$.

In our experiments, we set $\tau = 1000$. The obtained results are shown in Figs. 13, 14, 15 and 16:

- Figs. 13 and 14 present the coverage probability, obtained with the described Monte Carlo procedure, when L is varied and n is set to 10 and 200, respectively. As we can see, all the values of $\mathcal{C}(L,n)$ are above the nominal confidence, as expected by theory. The bound characterized by a coverage probability closer to 95% is T_HOE, which also provides the tightest estimate of L (see, for example, Fig. 12);
- Figs. 15 and 16 show the values of $\mathcal{C}(L,n)$ when n is varied and L is set to 0% and 25%, respectively. For these plots, we can draw the same conclusions as in the cases of Figs. 13 and 14.

4.3 Using Scenario 2 Bounds in Scenario 1 - Results

In the previous sections, two scenarios have been separately analyzed: as a first issue, the case of hard loss binary functions has been taken into account; subsequently, we focused on continuous soft loss functions. However, as remarked in Section 3, the bounds designed for soft loss functions can be exploited in the case of binary random variables as well.

Thus, we consider a binary hard loss function for evaluating the generalization error and take into account the best performing rigorous approaches, according to the results of Sections 4.1 and 4.2 (CP and T_HOE). Moreover, we also contemplate WIL, as it represents a valuable alternative when a rigorous bound is not needed. We compare the three trends by alternatively varying n and $\hat{L}_n$:

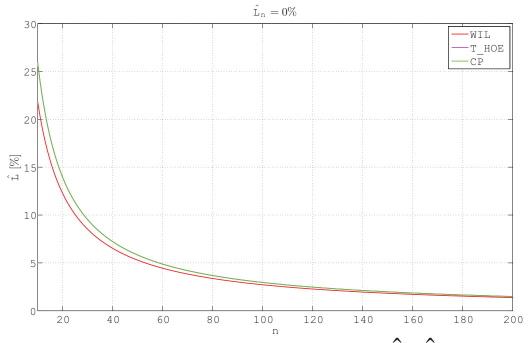

Figure 17. CP vs. T_HOE - Trend of $\hat{L}$ ($\hat{L}_n = 0\%$).

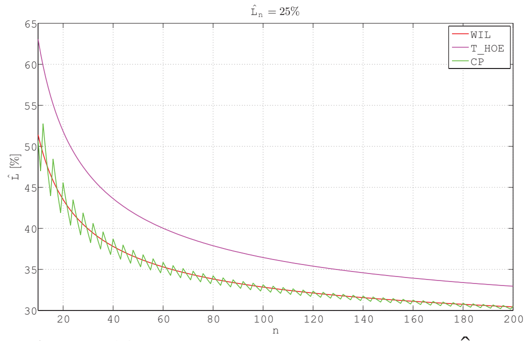

Figure 18. CP vs. T_HOE - Trend of $\hat{L}$ ($\hat{L}_n = 25\%$).

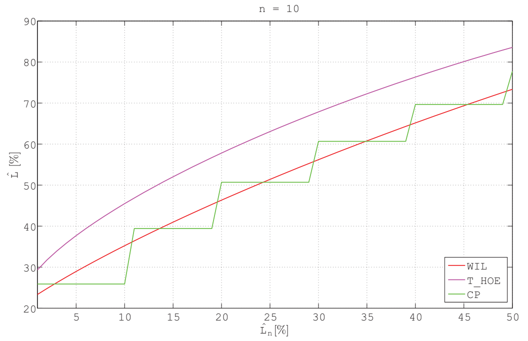

Figure 19. CP vs. T_HOE - Trend of $\hat{L}$ ($n = 10$).

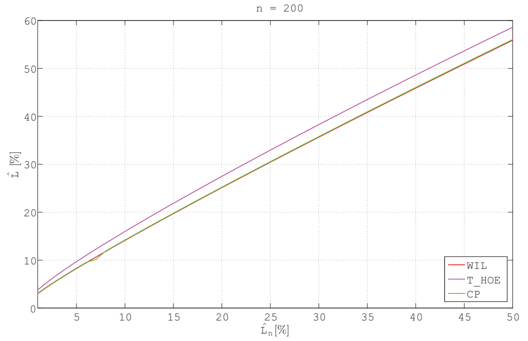

Figure 20. CP vs. T_HOE - Trend of $\hat{L}$ ($n = 200$).

- Figs. 17 and 18 show the CP, WIL and T_HOE bounds, when the empirical error is equal to 0% and 25%, respectively, and n is varied. It is worth noting that CP and T_HOE performs identically when $\hat{L}_n = 0\%$; on the contrary, CP becomes tighter than T_HOE as $\hat{L}_n$ increases, since the Clopper-Pearson approach takes into account the actual binomial distribution of the error. The performance of the WIL approach are comparable to CP: as also highlighted in Section 4.1, WIL sometimes outperforms CP, but its coverage probability could fall below the nominal value $(1-\delta)$, at least when n is small;

- The results obtained by varying $\hat{L}_n$ and by fixing n to 10 and 200, respectively, are shown in Figs. 19 and 20. As in the previous analysis, CP results to be the tightest rigorous approach, though, as n increases, the gap between CP and T_HOE decreases. WIL, as in the previous cases, represents a non-rigorous, but tight, alternative.

5 Conclusions

In this survey, we presented several approaches to estimate the generalization error of a classifier, when a test set is available. In particular, we took into account two scenarios, according to the usual approaches to classification problems: in the first one, we considered a loss function which can assume only two binary values; in the second one, we supposed to use a real-valued function, bounded in the interval $[0,1]$. In order to compare the different alternatives, we performed different tests by varying the number of patterns and the empirical error; the actual coverage probability for the bounds was also computed.

Concerning the first scenario, we can observe that:

- if we need a guaranteed and rigorous bound, the Clopper-Pearson approach is the only theoretically justified method for estimating the generalization error L. Unfortunately, as clearly shown by Figs. 7 and 8, Clopper-Pearson can be conservative, as the coverage probability is larger than the nominal confidence value, causing the looseness of $\hat{L}$;

- the Normal Approximation approach, besides being non-rigorous, is not effective in giving good generalization estimates in practice;

- the Wilson Score bound represents a good compromise for error estimation. However it must be used only when a rigorous approach is not strictly required, as its coverage probability falls, even if rarely, below the nominal confidence.

By analyzing the bounds we presented for the second scenario, we can sketch the following conclusions:

- all the bounds we analyzed in this framework are rigorous;

- generally speaking, the Hoeffding approach of Eq. (44) guarantees the tightest estimate, but no closed-form expression can be found for this bound and a numerical procedure for finding the value of the generalization error must be implemented. If a 'paper and pencil' approach is desired, the Guttman and the conventional Hoeffding bounds of Section 3.2 and Eq. (48), respectively, represent valuable alternatives. The Bernstein bound is characterized by an appealing performance, but it requires a numerical procedure for estimating the generalization error as well.

As the bounds for soft losses can be exploited for binary losses as well, we compared the Clopper-Pearson inequality (i.e. the only rigorous bound analyzed for Scenario 1), with the best performing method of Section 3 (i.e. the tighter Hoeffding formulation of Eq. (44)). The Wilson Score approach of Section 2.2 has also been included in this analysis, as it could be exploited, when a rigorous approach is not required by the application. The results clearly confirm what expected by theory: in general, the Clopper-Pearson bound is noticeably tighter than the Hoeffding one because the former method takes into account the actual binomial distribution of the error. Moreover, if rigorousness is not an issue, the Wilson Score bound is sometimes even tighter than Clopper-Pearson. Then, if not required by a specific problem, using the bounds for the Scenario 2 in the framework of Scenario 1 is not recommended.

As a final remark, we can safely claim that the methods appeared in the literature more than fifty years ago are still the best ones. The only caveat is that, due to the widespread use of computing tools since those days, the elegance of the closed–form formulations should be finally abandoned, in favor to their implicit, but tighter, counterparts, which needs a numerical solution.

A Extension to the case of K-Fold Cross Validation (KCV)

Let us suppose that the set of n d-dimensional data $\mathcal{X}$ must be used for both training the classifier and estimating the generalization error (e.g. when n is small, a subset of data cannot be allocated only for testing purposes). As an alternative to the conventional test set approaches, previously considered in this work, *resampling techniques* can be used instead. In particular, in this Appendix, we focus our attention on the well-known *K-Fold Cross Validation* (*KCV*) approach [14, 15] and we show how the results presented in the previous sections, targeted to the test set method, can be easily extended to this resampling case.

The KCV consists in dividing the available set $\mathcal{X}$ in k parts, each one consisting of n/k samples: $(k-1)$ parts are used, in turn, as a training set and the remaining one is used as a test set. The error performed by the trained model on the test set can be reliably used for estimating $L(f)$, the true generalization error, because it has not been exploited for training the model.

We can define the empirical error on the j-th test set as:

$$\hat{L}_{n/k}(f_j) = \frac{k}{n}\sum_{i=1}^{n/k} \ell(\hat{y}_i, y_i) \qquad (49)$$

where f_j is the classifier trained on the remaining $\frac{k-1}{k}n$ samples, and $\ell(\cdot,\cdot)$ is a (either hard or soft) loss function. Then, an unspecific test set bound can be applied to estimate the generalization error, according to Eq. (7):

$$L(f) \leq \hat{L}_{n/k}(f_j) + \Delta(f_j, \mathcal{X}_j^{test}, \delta), \qquad (50)$$

where the unluckiness factor depends on the classifier found using the j-th training set, on the confidence level δ, and on the j-th test set $\mathcal{X}_j^{test}$.

If we randomly pick up one of the trained model f_j to classify a new point, it is possible to show [32] that the performance of the model will be bounded by

$$L(f) \leq \frac{1}{k}\sum_{j=1}^{k} \left\{\hat{L}_{n/k}(f_j) + \Delta(f_j, \mathcal{X}_j^{test}, \delta)\right\}, \qquad (51)$$

which holds with a coverage probability greater than or, at least, equal to $(1-\delta)$. Therefore, all the analysis performed in this survey can be applied to this case as well.

References

[1] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[2] C. Cortes, V. Vapnik, "Support–Vector Networks", *Machine Learning*, Vol. 20, 1995, pp. 273–297.

[3] L. Roberge, S. B. Long, D. B. Burnham, "Data Warehouses and Data Mining tools for the legal profession: using information technology to raise the standard of practice", *Syracuse Law Review*, vol. 52, 2002, pp. 1281–1292.

[4] P. Bartlett, S. Boucheron, G. Lugosi, "Model Selection and Error Estimation", *Machine Learning*, Vol. 48, 2002, pp. 85–113.

[5] L. Devroye, L. Gyorfi, G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer Verlag, 1996.

[6] V.N. Vapnik, A.Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities", *Theory of Probability and its Applications*, Vol. 16, 1971, pp. 264-280.

[7] O. Gascuel, G. Caraux, "Distribution-free performance bounds with the resubstitution error estimate", *Pattern Recognition Letters*, Vol. 13, 1992, pp. 757–764.

[8] K. Duan, S. S. Keerthy, A. Poo, "Evaluation of simple performance measures for tuning SVM parameters", *Neurocomputing*, vol. 51, 2003, pp. 41–59.

[9] D. Anguita, A. Boni, S. Ridella, F. Rivieccio, D. Sterpi, "Theoretical and practical model selection methods for Support Vector classifiers", In: "*Support Vector Machines: Theory and Applications*", L. Wang, Springer, 2005.

[10] D. Anguita, A. Ghio, S. Ridella, "Maximal Discrepancy for Support Vector Machines", *Neurocomputing*, vol. 74, 2011, pp. 1436–1443.

[11] C.J.C. Burges, "A tutorial on Support Vector Machines for classification", *Data Mining and Knowledge Discovery*, vol. 2, 1998, pp. 121-167.

[12] C.W. Su, C.C. Chang, C.J. Lin, “A practical guide to Support Vector classification”, Technical report, Dept. of Computer Science, National Taiwan University, 2003.

[13] D. Anguita, L. Ghelardoni, A. Ghio, S. Ridella, “Test error bounds for classifiers: A survey of old and new results”, In: *Proc. of the 2011 IEEE Symposium on Foundations of Computational Intelligence (FOCI)*, Paris, France, 2011, pp. 80-87.

[14] D. Anguita, A. Ghio, S. Ridella, D. Sterpi, “K–Fold Cross Validation for Error Rate Estimate in Support Vector Machines”, In: *Proc. of the Int. Conf. on Data Mining (DMIN’09)*, 2009, pp. 291–297.

[15] M. Anthony, S. B. Holden, “Cross–Validation for binary classification by real–valued functions: theoretical analysis”, In: *Proc. of the 11th Conf. on Computational Learning Theory*, 1998, pp. 218–229.

[16] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, 2000.

[17] J.C. Platt, “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods”, *Advances in Large Margin Classifiers*, 1999, pp. 61–74.

[18] P. Bartlett, A. Tewari, “Sparseness vs Estimating Conditional Probabilities: Some Asymptotic Results”, *Journal of Machine Learning Research*, Vol. 8, 2007, pp. 775–790.

[19] L.D. Brown, T.T. Cai, A. DasGupta, “Interval estimation for a binomial proportion”, *Statistical Science*, Vol. 16, 2001, pp. 101–133.

[20] E.B. Wilson, “Probable inference, the law of succession, and statistical inference”, *Journal of the American Statistical Association*, Vol. 22, 1927, pp. 209–212.

[21] C.J. Clopper, E.S. Pearson, “The use of confidence intervals for fiducial limits illustrated in the case of the binomial”, *Biometrika*, Vol. 26, 1934, pp. 404-413.

[22] J. Langford, “Tutorial on practical prediction theory for classification”, *Journal of Machine Learning Research*, Vol. 6, 2005, pp. 273-306.

[23] A. Papoulis, *Probability, Random Variables, and Stochastic Processes, 3rd ed.*, McGraw-Hill, 1991.

[24] L. Guttman, “A distribution-free confidence interval for the mean”, *The Annals of Mathematical Statistics*, Vol. 19, 1948, pp. 410–413.

[25] S.N. Bernstein, *Sobranie sochinenie, Tom IV. Teoriya veroyatnostei, Matematicheskaya statistika (1911-1946) (Collected works, Vol. IV. The theory of probability, Mathematical statistics (1911-1946))*, Izdat. Nauka, 1964.

[26] J.Y. Audibert, R. Munos, C. Szepesvari, “Exploration-exploitation trade-off using variance estimates in multi-armed bandits”, *Theoretical Computer Science*, Vol. 410, 2009, pp. 1876–1902.

[27] A. Maurer, M. Pontil, “Empirical Bernstein Bounds and Sample Variance Penalization”, In: *Proc. of the Int. Conference on Learning Theory (COLT)*, 2009.

[28] Y. Mansour, D. McAllester, ”Generalization Bounds for Decision Trees”, In: *Proc. of the Thirteenth Annual Conference on Computational Learning Theory*, 2000, pp. 69-74.

[29] W. Hoeffding, “Probability inequality for sum of bounded random variables”, *Journal of the American Statistical Association*, Vol. 58, 1963, pp. 13–30.

[30] V. Bentkus, “An Inequality for Large Deviation Probabilities of Sums of Bounded i.i.d. Random Variables”, *Lithuanian Mathematical Journal*, Vol. 41, 2001, pp. 112–119.

[31] V. Bentkus, “On Hoeffdings inequalities”, *Annals of Probability*, Vol. 32, 2004, pp. 1650–1673.

[32] A. Blum, A. Kalai, J. Langford, “Beating the HoldOut: Bounds for Kfold and Progressive CrossValidation”, *Proc. of the Conference on Computational Learning Theory*, 1999, pp. 203-208.