Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 52-20 (2019) 73–78

Consensus and Reliability: The Case of Two Binary Classifiers ⋆

A.T.J.R. Cobbenhagen * **A. Carè** ** **M.C. Campi** **
F.A. Ramponi ** **W.P.M.H. Heemels** *

* *Dept. of Mechanical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands (e-mail: a.t.j.r.cobbenhagen@tue.nl).*
** *Dept. of Information Engineering, University of Brescia, Italy.*

Abstract: In this paper we consider the problem of estimating the probability of misclassification when consensus is achieved between two binary classifiers that are trained on the same training set. Firstly, it is shown that, under consensus, the probability of misclassification compares favourably with that of the best of the two classifiers. Secondly, we provide accurate, and yet simple to compute, estimates of the probability of consensus and the probability of misclassification under consensus. This paper provides a theoretical basis for these estimates and demonstrates their accuracy by simulation results on a synthetic data set and on a medical data set for breast cancer cell classification.

Keywords: Consensus, Classifiers, Multi-agent, Machine learning, Optimisation.

1. INTRODUCTION

1.1 Problem description and main contributions

In this paper we consider the situation where two binary classifiers are constructed using the same training set. We are interested in the probability of misclassification when the two classifiers agree (consensus). The motivation for this study is the empirical evidence that, under consensus, a higher probability of correct classification is achieved.

The main contributions of this paper consists of two novel results on this probability of misclassification given consensus. Both results are valid irrespective of the distribution by which samples are drawn.

Firstly, we present a novel theorem (Theorem 2) that compares the probability of misclassification under consensus with the probability of misclassification of the best classifier. From this theorem, two corollaries are derived that rigorously quantify the probability of misclassification for the case at hand on the basis of certain empirical indicators.

Secondly, for a certain family of classification algorithms, we present simple-to-compute estimators of the probability that the two classifiers are both wrong and the probability that they are in consensus. These estimators are also tested on numerical simulations.

⋆ Roy Cobbenhagen and Maurice Heemels were supported by "Toeslag voor Topconsortia voor Kennis en Innovatie" (TKI HTSM) from the Ministry of Economic Affairs, the Netherlands.

A. Carè, M.C. Campi and F.A. Ramponi were supported by the H&W 2015 program of the University of Brescia under the project "Classificazione della fibrillazione ventricolare a supporto della decisione terapeutica" (CLAFITE).

1.2 Previous work

Research on the subject of combining classifiers is vast: an overview can be found in, e.g., Kittler et al. (1998); Kittler (1998); Džeroski and Ženko (2004). Applications can be found in Petrakos et al. (2001) and, in Kuncheva et al. (2000), statistical tests on multiple dependent classifiers are described. Our approach is radically different from these works as we do not require additional statistical tests, and our estimators can be computed from structural properties (i.e., the number of "support points"). Our investigation narrows the gap between these multi-classifier studies in the machine learning community and the *scenario approach* (Calafiore and Campi (2006)) from the optimisation and control community following up the previous contributions Campi (2010); Margellos et al. (2015); Manganini et al. (2015); Baronio et al. (2017).

1.3 Preliminaries and notation

Let $y : \mathbb{R}^n \to \{0, 1\}$ be a random mapping from a vector of n features x to a label $y \in \{0, 1\}$. Hence, to a given x the mapping assigns a probability that $y = 0$ and that $y = 1$. The objective of *supervised classification* is to construct a function $\hat{y} : \mathbb{R}^n \to \{0, 1\}$ of x such that $\hat{y}(x) = y(x)$ with high probability, where $\hat{y}$ is constructed using a collection of N, previously recorded, data points, called the "training set". We call $\hat{y}$ a *binary classifier*. In this paper we denote the training set by $\tau_N := \{(x_1, y_1), ..., (x_N, y_N)\}$ and assume that the data points (x_i, y_i) in the training set τ_N are independent and identically distributed (i.i.d.) according to a probability measure $\mathbb{P}$ over $\Delta = \mathbb{R}^n \times \{0, 1\}$. It is assumed that the marginal probability of x over $\mathbb{R}^n$ admits density. No other knowledge regarding $\mathbb{P}$ is assumed.

2405-8963
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2019.12.129

When given a new feature vector $x \in \mathbb{R}^n$, the classifier provides a prediction $\hat{y}(x)$ of the corresponding label. The probability of misclassification of a classifier $\hat{y}$ is

$$V = \mathbb{P}\{\hat{y}(x) \neq y(x)\}\,.^{1}$$

Let $\hat{y}_A : \mathbb{R}^n \to \{0,1\}$ and $\hat{y}_B : \mathbb{R}^n \to \{0,1\}$ denote two classifiers called A and B, respectively. It is assumed that these classifiers are trained on the same training set of size N. To improve readability, we will use "consensus" as a shorthand for "$\hat{y}_A(x) = \hat{y}_B(x)$" and "$A$ wrong" ("B wrong") for "$\hat{y}_A(x) \neq y(x)$" ("$\hat{y}_B(x) \neq y(x)$"), and similarly for the "right" case. We use the following notations throughout the whole paper

$$\begin{aligned}
V_A &= \mathbb{P}\{A \text{ wrong}\}\\
V_B &= \mathbb{P}\{B \text{ wrong}\}\\
V_{A\cap B} &= \mathbb{P}\{A \text{ wrong} \wedge B \text{ wrong}\}\\
V_{A\cup B} &= \mathbb{P}\{A \text{ wrong} \vee B \text{ wrong}\}\\
\alpha &= \mathbb{P}\{\text{consensus}\}\\
&= \mathbb{P}\{(A \text{ wrong} \wedge B \text{ wrong}) \vee (A \text{ right} \wedge B \text{ right})\}\\
V_{ag} &= \frac{V_{A\cap B}}{\alpha} = \mathbb{P}\{A \text{ wrong} \wedge B \text{ wrong} \mid \text{consensus}\}\\
V_{best} &= \min\{V_A, V_B\}\\
V_{worst} &= \max\{V_A, V_B\}.
\end{aligned}$$

1.4 Structure of the paper

The remainder of the paper is structured as follows. Section 2 presents a novel result that compares the probability of misclassification conditioned on agreement (i.e., V_{ag}) with the probability of misclassification of the best classifier. Section 3 presents a new estimator for V_{ag} and some ancillary theoretical results. A demonstration by simulation of the accuracy of the estimator is presented in Section 4. The paper ends with conclusions in Section 5.

2. MOTIVATION: GETTING CLOSE TO THE BEST PERFORMANCE

The main motivation to investigate V_{ag} is the experimental evidence that, in general, V_{ag} is much smaller than V_A and V_B. This evidence is in part supported by Theorem 2, which shows that, even though V_{ag} can be worse than V_{best}, it is still relatively close to the performance of the best classifier.

2.1 Probability of consensus

Before we discuss the main result (Theorem 2), we establish the following lemma on the probability that the two classifiers agree (i.e., they are in consensus).

Lemma 1. (Probability of consensus).

$$\begin{aligned}
\alpha = 1 - V_{A\cup B} + V_{A\cap B} &= 1 - V_A - V_B + 2V_{A\cap B}\\
&= 1 + V_A + V_B - 2V_{A\cup B}.
\end{aligned}$$

[1] Note that V is a random variable on Δ^N, and V can also be interpreted as a conditional probability:

$$V = \mathbb{P}^{N+1}\{\hat{y}(x) \neq y(x) \mid \tau_N\}\,,$$

where

$$\mathbb{P}^{N+1} = \underbrace{\mathbb{P} \times \cdots \times \mathbb{P}}_{N+1 \text{ times}}$$

is the product probability since the samples are i.i.d.

Proof. Using the definition of the probability of consensus, we obtain

$$\begin{aligned}
\alpha &= \mathbb{P}\{\text{consensus}\}\\
&= \mathbb{P}\{(A \text{ right} \wedge B \text{ right}) \vee (A \text{ wrong} \wedge B \text{ wrong})\}\\
&= \mathbb{P}\{A \text{ right} \wedge B \text{ right}\} + \mathbb{P}\{A \text{ wrong} \wedge B \text{ wrong}\}\\
&= 1 - \mathbb{P}\{A \text{ wrong} \vee B \text{ wrong}\} + \mathbb{P}\{A \text{ wrong} \wedge B \text{ wrong}\}\\
&= 1 - V_{A\cup B} + V_{A\cap B}.
\end{aligned}$$

By the inclusion-exclusion principle,

$$V_{A\cup B} = V_A + V_B - V_{A\cap B}, \tag{1}$$

which can be substituted to obtain the latter two equalities of the claim. □

2.2 'Stay with the best' theorem

Using the result of Lemma 1, we can prove the following theorem.

Theorem 2. If $V_A + V_B < 1$, then

$$V_{ag} \leq \frac{V_{best}}{1 + V_{best} - V_{worst}}. \tag{2}$$

Proof. Lemma 1 implies that

$$V_{ag} = \frac{V_{A\cap B}}{1 - V_A - V_B + 2V_{A\cap B}}. \tag{3}$$

Combining (3) with the assumption $V_A + V_B < 1$, we get $V_{ag} < \frac{1}{2}$. Hence, it holds true that

$$\frac{\partial V_{ag}}{\partial V_{A\cap B}} = \frac{1}{\alpha}(1 - 2V_{ag}) > 0, \tag{4}$$

that is, V_{ag} is an increasing function of $V_{A\cap B}$ so that we can upper bound V_{ag} by substituting the maximum value that $V_{A\cap B}$ can take. Since $V_{A\cap B}$ is the probability that both A and B are wrong, it is upper bounded by the minimum of V_A and V_B. Therefore, it holds that

$$\begin{aligned}
V_{ag} &\leq \frac{\min\{V_A, V_B\}}{1 - V_A - V_B + 2\min\{V_A, V_B\}}\\
&= \frac{\min\{V_A, V_B\}}{1 + \min\{V_A, V_B\} - \max\{V_A, V_B\}}.
\end{aligned}$$ □

Remark 3. Inequality (2) becomes an equality, namely when the worst classifier is wrong every time the best classifier is wrong ($V_{A\cap B} = V_{best}$ in (3)).

Remark 4. The condition $V_A + V_B < 1$ cannot be removed; however, it is very mild because any practically useful classifier classifies with a probability of error smaller than 50%.

The interpretation of Theorem 2 is that one achieves a probability of misclassification close to that of the classifier that performs better. For example, suppose that $V_{best} = 0.01$ and the other classifier is ten times worse, i.e., $V_{worst} = 0.10$. Then, according to Theorem 2, $V_{ag} \leq 0.011$, i.e., V_{ag} is much closer to the probability of misclassification of the best classifier than to that of the worst classifier. This is achieved by abstaining from classifying in case of disagreement, which normally occurs in feature regions that are difficult to classify (e.g., regions where y takes on value 0 or 1 with an evenly split probability).

2.3 Data-dependent applications of Theorem 2

In practice, the true values of V_A and V_B are unknown. However, (upper) bounds on the probability of misclassification are sometimes available in the spirit of the so-called

Probably-Approximately-Correct (PAC) learning. In notable situations, these bounds can be obtained from the training set, i.e., without resorting to any extra validation or testing data, see e.g., Graepel et al. (2005); Carè et al. (2018); Campi and Garatti (2018); Carè et al. (2019). We formalize the eventuality that such data-dependent bounds are available by the following assumption.

Assumption 5. There exist data-dependent bounds such that, for each of the classifiers $j \in \{A, B\}$, it holds that

$$\mathbb{P}^N\{V_j \leq \epsilon_j(\tau_N)\} \geq 1 - \beta_j,$$

where $\epsilon_j(\tau_N) \in (0, 1]$ denotes the upper bound on the probability of misclassification and $1 - \beta_j \in (0, 1)$ is the confidence with which the upper bound holds. In cases of interest, β_j is a very small value.

We can use these data-dependent bounds to leverage Theorem 2. We first provide a deterministic result in Corollary 6 and then the probabilistic, data-dependent counterpart in Corollary 7.

Corollary 6. Let $\epsilon_A, \epsilon_B \geq 0$ such that $\epsilon_A + \epsilon_B < 1$ and define $\epsilon_{\max} = \max\{\epsilon_A, \epsilon_B\}, \epsilon_{\min} = \min\{\epsilon_A, \epsilon_B\}$. Under the condition that $V_A \leq \epsilon_A$ and $V_B \leq \epsilon_B$, it holds that

(i) $V_{ag} \leq V_{best} \dfrac{1}{1 - \epsilon_{\max}}$,

(ii) $V_{ag} \leq \dfrac{\epsilon_{\min}}{1 + \epsilon_{\min} - \epsilon_{\max}}$.

Proof. We have $V_{worst} \leq \epsilon_{\max}$ by assumption and $V_{best} \geq 0$ is trivially true. Using these two inequalities to bound from below the denominator of (2) yields the first inequality.

In order to prove the second inequality, we again use $V_{worst} \leq \epsilon_{\max}$ in (2) and observe that $V_{best}/(V_{best} + 1 - \epsilon_{\max})$ is an increasing function of V_{best} since $\epsilon_{\max} < 1$. In order to bound from above this expression we therefore substitute the largest possible value of V_{best}, which is $\epsilon_{\min}$. □

The following Corollary 7 justifies the usage of the data-dependent bounds $\epsilon_A(\tau_N), \epsilon_B(\tau_N)$ to draw conclusions about V_{ag} by showing that it is a rare event (of probability at most $\beta_A + \beta_B$) that one observes that the condition $\epsilon_A(\tau_N) + \epsilon_B(\tau_N) < 1$ is satisfied and yet the conclusions of Corollary 6 are not correct.

Corollary 7. Let $\epsilon_A(\tau_N), \epsilon_B(\tau_N)$ be the bounds in Assumption 5. Define $\epsilon_{\max}(\tau_N) = \max\{\epsilon_A(\tau_N), \epsilon_B(\tau_N)\}$, $\epsilon_{\min}(\tau_N) = \min\{\epsilon_A(\tau_N), \epsilon_B(\tau_N)\}$ and introduce the (bad) event $\mathcal{B} = \left\{V_{ag} > V_{best} \frac{1}{1-\epsilon_{\max}(\tau_N)} \ \vee \ V_{ag} > \frac{\epsilon_{\min}(\tau_N)}{1+\epsilon_{\min}(\tau_N)-\epsilon_{\max}(\tau_N)}\right\}$. Then, it holds that

$$\mathbb{P}^N\{\epsilon_A(\tau_N) + \epsilon_B(\tau_N) < 1 \wedge \mathcal{B}\} \leq \beta_A + \beta_B.$$

Proof. By Corollary 6, it holds that, for any τ_N, $\epsilon_A(\tau_N) + \epsilon_A(\tau_N) < 1 \wedge \mathcal{B} \implies V_A > \epsilon_A(\tau_N) \vee V_B > \epsilon_B(\tau_N)$. Thus,

$$\begin{aligned}&\mathbb{P}^N\{\epsilon_A(\tau_N) + \epsilon_B(\tau_N) < 1 \wedge \mathcal{B}\}\\ &\quad\leq \mathbb{P}^N\{V_A > \epsilon_A(\tau_N)\} + \mathbb{P}^N\{V_B > \epsilon_B(\tau_N)\}\\ &\quad\leq \beta_A + \beta_B.\end{aligned}$$ □

Theorem 2 and Corollaries 6 and 7 offer worst-case guarantees that hold true in full generality. On the other hand, simulation evidence shows that in many situations conditioning on agreement will improve the probability of misclassification well beyond worst case. In the next section, we venture beyond the results in Corollaries 6 and 7 and try to lay the foundations of a new theory for an accurate estimate of the actual probability of misclassification under consensus. Our study here is preliminary and is meant to offer new avenues for further investigations.

3. PRACTICAL ESTIMATORS FOR V_{ag}

In this section, we assume that the classifiers A and B are constructed by means of two algorithms that fit the theoretical framework of Campi (2010); Carè et al. (2018). As a consequence, the obtained classifiers can be characterized by their "support points" (the notion of "support point" is analogue to that of "support constraint" in the theory of the scenario approach, see e.g., Campi and Garatti (2008, 2018)), defined as follows.

Definition 8. (Support point, support set). A data point in the training set is a support point for a classifier if and only if the removal of that data point from the training set followed by retraining yields a different classifier. The support set of a classifier is the set of its support points.

The following fact of the theory in Campi (2010); Carè et al. (2018) is crucial in what follows.

Fact 9. A data point $(x_i, y_i) \in \tau_N$ is a support point if and only if the classifier trained on $\tau_N \setminus \{(x_i, y_i)\}$ misclassifies (x_i, y_i). [2]

We denote by S_A^N (S_B^N) the support sets of classifier A (B) trained on τ_N. For reasons that will be clear soon, we have used the superscript N as a reminder of the size of the traning set. We will also use the shorthands $k_A^N = |S_A^N|$, $k_B^N = |S_B^N|$. It is a well-known fact that there is a strong relation between the cardinality of the support set and the probability of misclassification. In particular, it holds that (see e.g. Calafiore (2009))

$$\mathbb{E}_N\{V_A\} = \frac{\mathbb{E}_{N+1}\{k_A^{N+1}\}}{N+1} \tag{5}$$

(likewise for classifier B), where the expectation on the left-hand side, with respect to τ_N, is taken on the probability of misclassification V_A discussed throughout the paper, while the expectation on the right-hand side is with respect to a larger training set $\tau_{N+1} = \{(x_1, y_1), ..., (x_{N+1}, y_{N+1})\} \in \Delta^{N+1}$ and is taken on the number of support points of the classifier trained on τ_{N+1}. This shows that $k_A^{N+1}/(N+1)$ is a reasonable estimator of V_A. In the same spirit, we define

$$k_{A\cup B}^{N+1} = |S_A^{N+1} \cup S_B^{N+1}|$$

and show that

$$\mathbb{E}_N\{V_{A\cup B}\} = \frac{\mathbb{E}_{N+1}\{k_{A\cup B}^{N+1}\}}{N+1}. \tag{6}$$

Proof. Let A^* and B^* denote two classifiers trained on the enlarged training set τ_{N+1} and let A_i and B_i denote the classifiers trained on the N data points in $\tau_{N+1} \setminus \{(x_i, y_i)\}$.

[2] Recall that, in order for this fact to hold, the special initialization point in the constructions of Campi (2010); Carè et al. (2018) must not be counted as belonging to the training set.

Hence, in particular, A_{N+1} and B_{N+1} can be understood as the classifiers A and B, trained on τ_N, that have been discussed throughout the paper. For brevity, let us denote by δ_i a data point (x_i, y_i). Fact 9 ensures that

$$\mathbf{1}\{\delta_i \text{ s.p. for } A^*\} = \mathbf{1}\{A_i \text{ wrong on } \delta_i\}, \tag{7}$$

where "s.p." stands for "support point" and $\mathbf{1}\{\cdot\}$ is the indicator function (likewise for B).

Every possible ordering of $N+1$ data points is equally likely to occur because of the i.i.d. assumption. This allows us to state the following identity (for more information, see a similar derivation in Calafiore (2009)):

$$\begin{aligned}
&\sum_{i=1}^{N+1} \int_{\Delta^{N+1}} \mathbf{1}\{A_i \text{ or } B_i \text{ wrong on } \delta_i\} \mathbb{P}(\mathrm{d}\delta_i) \\
&\qquad \mathbb{P}^N(\mathrm{d}\delta_1, ..., \mathrm{d}\delta_{i-1}, \mathrm{d}\delta_{i+1}, ..., \mathrm{d}\delta_{N+1}) \\
&= (N+1) \int_{\Delta^{N+1}} \mathbf{1}\{A_{N+1} \text{ or } B_{N+1} \text{ wrong on } \delta_{N+1}\} \\
&\qquad \mathbb{P}(\mathrm{d}\delta_{N+1})\, \mathbb{P}^N(\mathrm{d}\delta_1, ..., \mathrm{d}\delta_N)\,. \qquad (8)
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}_N\{V_{A\cup B}\} &= \mathbb{E}_N\{\mathbb{P}\{A \text{ or } B \text{ wrong on } \delta_{N+1}\}\} \\
&= \int_{\Delta^N} \int_{\Delta} \mathbf{1}\{A_{N+1} \text{ or } B_{N+1} \text{ wrong on } \delta_{N+1}\} \\
&\qquad \mathbb{P}(\mathrm{d}\delta_{N+1})\, \mathbb{P}^N(\mathrm{d}\delta_1, ..., \mathrm{d}\delta_N) \\
&= \int_{\Delta^{N+1}} \mathbf{1}\{A_{N+1} \text{ or } B_{N+1} \text{ wrong on } \delta_{N+1}\} \\
&\qquad \mathbb{P}^{N+1}(\mathrm{d}\delta_1, ..., \mathrm{d}\delta_{N+1}) \\
[(8)] &= \tfrac{1}{N+1} \int_{\Delta^{N+1}} \left(\sum_{i=1}^{N+1} \mathbf{1}\{A_i \text{ or } B_i \text{ wrong on } \delta_i\} \right) \\
&\qquad \mathbb{P}^{N+1}(\mathrm{d}\delta_1, ..., \mathrm{d}\delta_{N+1}) \\
[(7)] &= \tfrac{1}{N+1} \int_{\Delta^{N+1}} \left(\sum_{i=1}^{N+1} \mathbf{1}\{\delta_i \text{ s.p. for } A^* \text{ and/or } B^*\} \right) \\
&\qquad \mathbb{P}^{N+1}(\mathrm{d}\delta_1, ..., \mathrm{d}\delta_{N+1}) \\
&= \tfrac{1}{N+1} \int_{\Delta^{N+1}} \left| S_A^{N+1} \cup S_B^{N+1} \right| \mathbb{P}^{N+1}(\mathrm{d}\delta_1, ..., \mathrm{d}\delta_{N+1}) \\
&= \frac{\mathbb{E}_{N+1}\{k_{A\cup B}^{N+1}\}}{N+1}.
\end{aligned}$$

□

We conclude that $k_{A\cup B}^{N+1}/(N+1)$ is a reasonable estimator of $V_{A\cup B}$.

Taking expectation on both sides of (1), we obtain

$$\mathbb{E}_N\{V_{A\cap B}\} = \mathbb{E}_N\{V_A\} + \mathbb{E}_N\{V_B\} - \mathbb{E}_N\{V_{A\cup B}\}\,. \tag{9}$$

Defining

$$k_{A\cap B}^{N+1} = |S_A^{N+1} \cap S_B^{N+1}|,$$

and noting that $k_{A\cap B}^{N+1} = k_A^{N+1} + k_B^{N+1} - k_{A\cup B}^{N+1}$, substitution of the right-hand sides of (5) and (6) into the right-hand side of (9) leads to the conclusion that

$$\mathbb{E}_N\{V_{A\cap B}\} = \frac{\mathbb{E}_{N+1}\{k_{A\cap B}^{N+1}\}}{N+1}, \tag{10}$$

which shows that $k_{A\cap B}^{N+1}/(N+1)$ is a reasonable estimator of $V_{A\cap B}$.

Finally, $1 - \frac{k_{A\cup B}^{N+1} - k_{A\cap B}^{N+1}}{N+1}$ is obtained as a reasonable estimator of α by recalling that $\mathbb{E}_N\{\alpha\} = 1 + \mathbb{E}_N\{V_A\} + \mathbb{E}_N\{V_B\} - 2\,\mathbb{E}_N\{V_{A\cup B}\}$ in view of Lemma 1.

The issue with the estimators obtained so far is that they are based on the enlarged τ_{N+1} and not on the available τ_N. To fill this gap, we rely on a well-educated guess (heuristic): the number of support points of a classifier trained on τ_N is expected to be close to the number of support points when an additional training point is considered.

The reasoning behind this heuristic assumption is that, when N is large and V_A is reasonably low, an $(N+1)$-th data point is unlikely to be misclassified, so that the number of support points is unlikely to change when the training set is enlarged with this point. On the other hand, if an $(N+1)$-th data point is misclassified, then the new number of support points could, at least in principle, take on any value between 1 and $N+1$. However, we conjecture that, most of the times, this value will be close to the previous one. In other terms, we conjecture that $k_j^N \approx k_j^{N+1}$ for all $j \in \{A, B, A\cup B, A\cap B\}$.

We are now in the position to propose the following estimators (of α, $V_{A\cap B}$ and V_{ag}, respectively):

$$\hat{\alpha} = 1 - \frac{k_{A\cup B}^N - k_{A\cap B}^N}{N+1}, \tag{11a}$$

$$\hat{V}_{A\cap B} = \frac{k_{A\cap B}^N}{N+1}, \tag{11b}$$

$$\hat{V}_{ag} = \frac{k_{A\cap B}^N}{N+1+k_{A\cap B}^N - k_{A\cup B}^N}, \tag{11c}$$

where

$$\begin{aligned}
k_{A\cup B}^N &= |S_A^N \cup S_B^N|, \\
k_{A\cap B}^N &= |S_A^N \cap S_B^N|.
\end{aligned}$$

3.1 Additional remarks

A remarkable fact is that the probability $V_{A\cup B}$ actually corresponds to the probability of misclassification of a classifier that fits the framework of Campi (2010); Carè et al. (2018). Such a classifier can be called the "union classifier" and is defined as follows. If A and B agree, the union classifier classifies according to the value of this agreement. In the case where A and B disagree, one of them must be right and the other must be wrong. In this case, the union classifier outputs a ternary value and hence it is deliberately wrong. The union classifier is therefore right if both A and B are right, otherwise it is wrong, and hence we can claim that the probability of misclassification of the union classifier is $V_{A\cup B}$. The "union classifier" has support set equal to $S_A^N \cup S_B^N$.

Starting from relation (10), one could be tempted to define the corresponding "intersection classifier", in a similar fashion as the "union classifier" above, and to study it in the light of Campi (2010); Carè et al. (2018). However, in general, the set of points $S_A^N \cap S_B^N$ is not a support set of the "intersection classifier" according to the theory of Campi (2010); Carè et al. (2018). This can be shown by noting that the removal of a point in $S_A^N \setminus S_B^N$ from the training set followed by retraining yields a different "intersection classifier".

4. SIMULATION RESULTS

4.1 Synthetic data set

In the following simulations, the classifiers are *Guaranteed Error Machines* (GEM), see Campi (2010); Carè et al. (2018). Following Carè et al. (2018), the regions of the GEM classifiers are restricted to (hyper)spheres as opposed to more general quadrics as in the original GEM algorithm.

In this section, the following synthetic problem is considered. There are $n = 2$ features, $x^{(1)} \in [0,1]$ and $x^{(2)} \in [0,1]$, mapped to a label $y \in \{0,1\}$ according to the function

$$y(x) = \begin{cases} 1, & \text{if } x^{(2)} \geq \left(x^{(1)} - \frac{1}{2}\right) \cos\left(25x^{(1)}\right) + \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

The marginal distribution of $\mathbb{P}$ with respect to x is the uniform distribution over $[0,1]^2$.

In order to demonstrate the accuracy of the estimators, the following simulation experiment was performed. In total, 100 data sets of size $N = 1000$ were generated and on each data set two GEM classifiers were trained. Classifier A always had $(0.5, 1)$ as the starting point (with label "1") and classifier B always had $(0.5, 0)$ as the starting point (with label "0"). Figure 1 shows a training set and Figure 2 shows the two GEM classifiers trained on it.

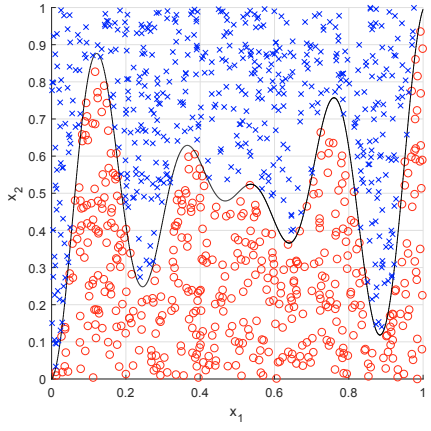

Fig. 1. Sample training set of size $N = 1000$. Circles indicate values with label "0" and crosses indicate values with label "1". The solid line indicates the division according to the function (12).

The number of support points of each classifier and the number of support points in common were computed for each pair of classifiers for all 100 training sets. In order to compute the "true" probability of agreement and probability of both being wrong, a Monte Carlo simulation of $2 \cdot 10^5$ samples was performed. Figure 3 shows the result for the estimator $\hat{V}_{ag}$ versus the Monte Carlo estimate. It can be seen that these values are in good agreement. Quantitatively, the mean difference between these two values is 0.0020, the mean absolute difference is 0.0064 and the largest absolute deviation is 0.0211.

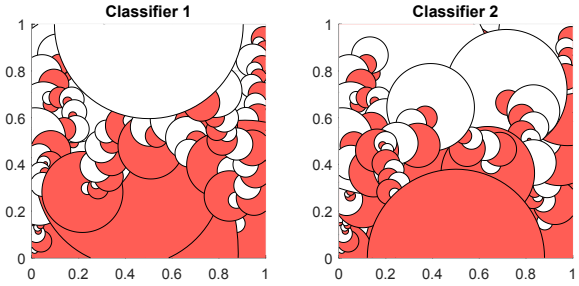

Fig. 2. Two GEMs trained on the training set of Figure 1. The coloured regions are classified as "0" by the classifiers, "1" otherwise. By a Monte Carlo simulation of $2 \cdot 10^5$ samples it was determined that $V_A \simeq 0.098$, $V_B \simeq 0.10$, $V_{A \cap B} \simeq 0.052$ and $\alpha \simeq 0.90$, which yields $V_{ag} \simeq 0.057$.

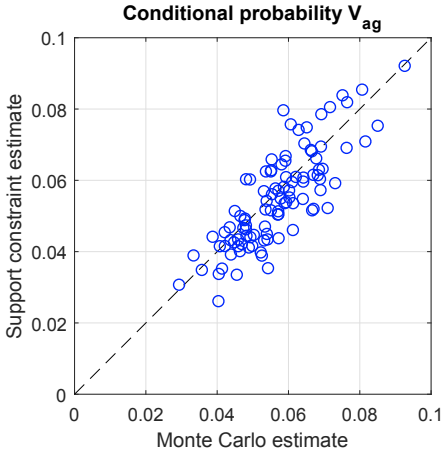

Fig. 3. The estimator $\hat{V}_{ag}$ versus the Monte Carlo estimate of V_{ag}.

4.2 Medical data set

We applied the results developed in the paper to the well-known diagnostic BreastW data set, obtained from Dua and Graff (2017). This data set can be used to train a classifier that predicts whether a particular breast tissue cell nucleus corresponds to a malignant or benign cell. The data set consists of 569 data points (357 benign, 212 malignant) with $n = 30$ features.

In this simulation example we construct two GEMs according to the original GEM algorithm as described in Campi (2010), which can construct quadrics. [3] We randomly selected two data entries to be the starting point for each of the two GEMs. These starting points were subsequently removed from the training set. From the remaining 567 points we randomly selected $N = 397$ data points (approximately 70%) to train the GEMs. The remaining 170 points were used for validation.

We obtained two GEMs with $k_A^N = 27$, $k_B^N = 17$, $k_{A \cap B}^N = 6$, $k_{A \cup B}^N = 38$. The comparison between the validation and the estimators from this paper is displayed in Table 1.

[3] The complexity parameter of GEM was set so as to achieve complete classification.

Table 1. Simulation results for the medical data set.

Probability	Fraction in validation	Estimator
V_A	0.0706	0.0678
V_B	0.0529	0.0427
$V_{A\cap B}$	0.0176	0.0151
α	0.9118	0.9196
V_{ag}	0.0194	0.0164

5. CONCLUSION AND FUTURE WORK

This paper established a worst case result showing that the probability of being wrong under consensus cannot be much worse than the probability of misclassification of the best classifier. Subsequently, inspired by the theory of the scenario approach, we proposed a practical estimator for the actual probability that two classifiers agree and are both wrong. The strength of the results were demonstrated on a synthetic data set and on a real-life medical data set.

The results of this paper are offered as a preliminary to future work. Firstly, it is interesting to study extensions of these results to general multi-agent/multi-classifier settings. In case of a large number of classifiers, unanimity may occur only with a low probability. This leads to the second direction of study: extension of the results to majority voting and other multi-agent decision schemes. Although a lot of research has been done on majority voting, the results of this paper can provide probabilistic guarantees and may shed a new light on multi-agent decision schemes. A third direction of future research is related to the exploitation of the wait-and-judge techniques presented in Campi and Garatti (2018); Carè et al. (2019) to the classification algorithms that we have considered in Section 3. The wait-and-judge approach can be used to improve the evaluation of the performance of classifiers under consensus. Finally, in this paper we provided estimators for certain key random quantities. We motivated their usage by a theoretical analysis accompanied by a heuristic. A fully rigorous analysis is currently ongoing research and could also be relevant to the important topic of leave-one-out stability (Evgeniou et al. (2004)).

REFERENCES

Baronio, F., Baronio, M., Campi, M., Carè, A., Garatti, S., and Perone, G. (2017). Ventricular defibrillation: Classification with G.E.M. and a roadmap for future investigations. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, 2718–2723. doi:10.1109/CDC.2017.8264054.

Calafiore, G.C. (2009). On the expected probability of constraint violation in sampled convex programs. *J. Optim. Theory Appl.*, 143(2), 405–412. doi:10.1007/s10957-009-9579-3.

Calafiore, G.C. and Campi, M.C. (2006). The scenario approach to robust control design. *IEEE Trans. Automat. Contr.*, 51(5), 742–753. doi:10.1109/TAC.2006.875041.

Campi, M.C. (2010). Classification with guaranteed probability of error. *Mach. Learn.*, 80, 63–84. doi:10.1007/s10994-010-5183-x.

Campi, M.C. and Garatti, S. (2008). The exact feasibility of randomized solutions of uncertain convex programs. *SIAM J. Optim.*, 19(3), 1211–1230. doi:10.1137/07069821X.

Campi, M.C. and Garatti, S. (2018). Wait-and-judge scenario optimization. *Math. Program.*, 167(1), 155–189. doi:10.1007/s10107-016-1056-9.

Carè, A., Garatti, S., and Campi, M.C. (2019). The wait-and-judge scenario approach applied to antenna array design. *Comput. Manag. Sci.* doi:10.1007/s10287-019-00345-5.

Carè, A., Ramponi, F.A., and Campi, M.C. (2018). A new classification algorithm with guaranteed sensitivity and specificity for medical applications. *IEEE Control Syst. Lett.*, 2(3), 393–398. doi:10.1109/LCSYS.2018.2840427.

Dua, D. and Graff, C. (2017). UCI Machine Learning Repository. URL `http://archive.ics.uci.edu/ml`.

Džeroski, S. and Ženko, B. (2004). Is Combining Classifiers with Stacking Better than Selecting the Best One? *Mach. Learn.*, 54(3), 255–273. doi:10.1023/B:MACH.0000015881.36452.6e.

Evgeniou, T., Pontil, M., and Elisseeff, A. (2004). Leave One Out Error, Stability, and Generalization of Voting Combinations of Classifiers. *Mach. Learn.*, 55(1), 71–97. doi:10.1023/B:MACH.0000019805.88351.60.

Graepel, T., Herbrich, R., and Shawe-Taylor, J. (2005). PAC-Bayesian Compression Bounds on the Prediction Error of Learning Algorithms for Classification. *Mach. Learn.*, 59(1-2), 55–76. doi:10.1007/s10994-005-0462-7.

Kittler, J. (1998). Combining classifiers: A theoretical framework. *Pattern Anal. Appl.*, 1(1), 18–27. doi:10.1007/BF01238023.

Kittler, J., Hatef, M., Duin, R., and Matas, J. (1998). On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(3), 226–239. doi:10.1109/34.667881.

Kuncheva, L., Whitaker, C., Shipp, C., and Duin, R. (2000). Is independence good for combining classifiers? In *Proc. 15th Int. Conf. Pattern Recognition. ICPR-2000*, volume 2, 168–171. IEEE, Barcelona, Spain. doi:10.1109/ICPR.2000.906041.

Manganini, G., Falsone, A., and Prandini, M. (2015). A majority voting classifier with probabilistic guarantees. In *2015 IEEE Conf. Control Appl.*, 1084–1089. IEEE, Sydney, Australia. doi:10.1109/CCA.2015.7320757.

Margellos, K., Prandini, M., and Lygeros, J. (2015). On the connection between compression learning and scenario based single-stage and cascading optimization problems. *IEEE Trans. Automat. Contr.*, 60(10), 2716–2721. doi:10.1109/TAC.2015.2394874.

Petrakos, M., Benediktsson, J.A., and Kanellopoulos, I. (2001). The effect of classifier agreement on the accuracy of the combined classifier in decision level fusion. *IEEE Trans. Geosci. Remote Sens.*, 39(11), 2539–2546. doi:10.1109/36.964992.