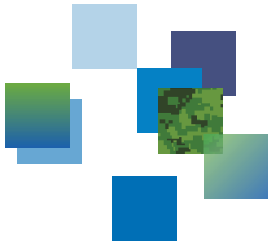
Defence Research and Development Canada
Recherche et développement pour la défense Canada

DRDC | RDDC



About reliability estimation of detectors and Bayesian reasoning

Anne-Laure Jousselme
DRDC – Valcartier Research Centre

Defence Research and Development Canada

Scientific Report
DRDC-RDDC-2016-R040
April 2016

Canada

About reliability estimation of detectors and Bayesian reasoning

Anne-Laure Jousselme
DRDC – Valcartier Research Centre

Defence Research and Development Canada
Scientific Report
DRDC-RDDC-2016-R040
April 2016

Table of contents

Table of contents i

Abstract ii

Significance for defence and security ii

Résumé iii

Importance pour la défense et la sécurité iii

Acknowledgements iv

1 Introduction 1

2 Classical Bayesian reasoning scheme 2

2.1 Likelihood functions $P(A|E)$ 3

2.2 Posterior probability $P(E|A)$ 4

2.3 Prior probability $P(E)$ 5

3 $P(A|E)$ or $P(E|A)$ as a reliability measure? 6

4 How to characterise a detector? 8

5 How to reduce the “false positive rate”? 9

6 Conclusions and discussion 12

References 14

Abstract

This brief note addresses the problem of performance estimation of detectors, with a specific emphasis on detectors for rare events. The discussion is framed in the Bayesian reasoning framework for estimating posterior probability $P(E|A)$ about events E based on observations A. The question addressed is which measure between $P(A|E)$ or $P(E|A)$ should be used as a requirement for performance of detectors. The two measures are first presented and their respective meaning discussed. Beyond the semantics associated to the term "reliability", the semantics of the mathematical quantities concerned is explained and the impact of this measure on detector's designers is discussed. Some ideas to improve performances of detectors for rare events are finally sketched, emphasising that the need for such detectors to be only components of a larger situation assessment system, so that correlations with other detectors or other pieces of information can be drawn, and that the decision maker can decide based on a more global picture of the situation.

Significance for defence and security

The reflection put forward in this work could be useful to any scientific work where performance criteria are to be used as requirements for detectors or classifiers design. More specifically, it recalls the need for scientists and operational people to understand the meaning and consequences of the choice of performance measures for detectors of rare events.

Résumé

Cette brève note aborde le problème de l'estimation de la performance de détecteurs, et la performance de détecteurs d'événements rares en particulier. La discussion utilise le cadre de raisonnement Bayésien qui permet d'estimer la probabilité *a posteriori* $P(E|A)$ de l'événement d'intérêt E étant donnée une observation A. La question posée est qui de la mesure $P(A|E)$ ou $P(E|A)$ devrait être utilisée pour la performance de détecteurs. Les deux mesures sont d'abord présentées et leurs significations respectives discutées. Au-delÃ de la sémantique associée au terme "fiabilité", la sémantique des quantités mathématiques concernées est expliquée et l'impact du choix d'une mesure ou de l'autre sur les concepteurs de détecteurs est discuté. Quelques idées pour améliorer les performances de détecteurs d'événements rares sont finalement esquissées, soulignant la nécessité pour ces détecteurs d'être seulement des composants d'un système d'évaluation de la situation, afin de mettre en corrélation leurs alarmes avec d'autres informations ou autres détections et que le preneur de décision puisse avoir une image globale de la situation.

Importance pour la défense et la sécurité

La réflexion proposée dans ce travail pourrait être utile à tout travail scientifique où les critères de performance sont utilisés comme exigences pour la conception de détecteurs ou de classifieurs. Plus particulièrement, elle rappelle le besoin pour les scientifiques et les opérationels de comprendre la signification et les conséquences du choix des mesures de performance de détecteurs d'événements rares.

Acknowledgements

The author would like to thank Jean-Robert Simard, Pierre Lahaie and Luc Pigeon for initiating the discussion and for the very fruitful exchanges about the topic which led to this document.

1 Introduction

The topic addressed in this document is the reliability assessment of detectors. More specifically, the question is which of the two quantities $P(E|A)$, the probability that the event of interest E occurred given that an alarm A has been issued by the detector, and $P(A|E)$, the probability that the detector sends an alarm given that the event E was indeed present, should be considered as a measure of reliability for a detector.

In the following, the term "reliability" will be used to refer to detector performances in detection[1]: We expect that a "reliable" detector will have both a high true positive rate $P(A|E)$ and low false alarm rate $P(A|\overline{E})$[2]. The true positive rate together with its counter-part the false alarm rate are the measures traditionally used in Bayesian reasoning schemes to quantify the reliability of a detector. The question is whether $P(E|A)$ would be a more adequate measure.

The question has been submitted by colleagues designing Chemical Biological Radiological Nuclear and Explosive (CBRNE) detectors. Traditionally, $P(A|E)$ together with $P(A|\overline{E})$ are the basic quantities used to characterise detectors' performance (understood here as reliability). End-users' requirements are then commonly expressed in terms of true positive and false alarm rates that detectors' designers attempt to meet. Unfortunately, despite good performances in laboratory it may happen that the detector does not behave as expected in an operational setting and provides more "false alarms" than the original specifications. To address this issue, using the alternative measure $P(E|A)$ is considered and the consequences of such a choice are explored in this document.

[1] Whether the term "reliability" is adequate for such a characterisation is not the purpose of this document.

[2] $\overline{E}$ is the absence of event E.

2 Classical Bayesian reasoning scheme

Let us consider the case of a detector $\mathcal{D}$ designed to detect some specific event of interest E and to trigger an alarm A whenever it indeed detects the event:

- $P(A|E)$ is the probability that the detector triggers (sends an alarm) given the event indeed occurred; and
- $P(E|A)$ is the probability that the event indeed occurs provided that the detector sent an alarm.

Usually, $P(A|E)$ is called the *likelihood* as it represents the prior likelihood that A occurs under the condition E. The other quantity $P(E|A)$ is the *posterior probability*, interpreted as the confidence the decision maker may have in the occurrence of E in light of evidence A. It is inferred from the likelihood and prior probabilities as given by Bayes' rule:

$$P(E|A) = \frac{P(A|E)P(E)}{P(A|E)P(E) + P(A|\overline{E})P(\overline{E})} \tag{1}$$

where $\overline{E}$ is the complement event of E, *i.e.*, "the event E did not occur". After an alarm occurred (from $\mathcal{D}$), the decision maker, based on the two values $P(E|A)$ and $P(\overline{E}|A)$, needs to decide either that E indeed occurred or that E did not occurred. A classical decision rule is the Maximum A Posteriori (MAP) rule which leads to decide the hypothesis with the highest posterior probability: decide E if $P(E|A) > P(\overline{E}|A)$ and decide $\overline{E}$ if $P(E|A) < P(\overline{E}|A)$. Other considerations such as the cost or risk maybe involved in a weighting sum of the posteriors. These costs are $C(E|\overline{E})$, $C(E|E)$, $C(\overline{E}|E)$ and $C(\overline{E}|\overline{E})$ expressing for instance, in the case of $C(E|\overline{E})$, the cost associated to deciding E while E did not occurred. In general, costs associated with correct decisions are null while costs associated with errors depend on the type of error: $C(E|\overline{E})$ may be lower than $C(\overline{E}|E)$ meaning that a "false alarm" is less damageable than a miss.

We distinguish between two steps in the detection process:

(1) the **design step**, where the detector is designed, the internal parameters are set (*e.g.*, threshold), and the performances are estimated under controlled experimental conditions, with accessible ground truth. Avoiding philosophical considerations about the accessibility to truth, "ground truth" is used here to refer to some "gold standard" we consider as "truth". In laboratory[3] tests, the designer knows under which condition the test is performed and thus which event E or $\overline{E}$ the detector is supposed to detect. The likelihoods are estimated at that step; and

[3]Here, the term "laboratory" is used to denote the place where the detector's design experiments take place, be it in-door or out-door.

(2) the **operational step**, where the detector is deployed and used to detect events E, under ill-defined conditions of use, with no access to ground truth. At this point, if a gold standard is to be defined, it will necessarily be different from the previous one. The operator (user) may act as a "ground truth producer", systematically comparing the alarms of the detector with his/her **own** estimation of the occurrence of E.

At both steps, the events A and E are defined. While the event A is the same at both the design and operational steps, the event E differs. During the design step, we have no doubt on the fact that E was present: the detector designer provoked the conditions E or that $\overline{E}$. However, during the operational step, we have no means to determine that E was present or not. In particular, what is called a "false alarm" at step (1) should not be called as such at step (2) since it is unknown which event E or $\overline{E}$ the detector was indeed detecting.

2.1 Likelihood functions $P(A|E)$

The conditional probability $P(A|E)$ is usually referred to as the *True Positive Rate* (TPR) and the conditional probability $P(A|\overline{E})$ to as the *False Positive Rate* or False Alarm Rate (FAR). These two quantities are the criteria traditionally used to characterise detectors' performances: Varying the decision threshold τ on the couple (FAR; TPR) of variables defines the so-called Receiver Operating Characteristic (ROC) curve. Any detector with performances belonging to the ROC curve is *optimal* in terms of both the true positive and false alarm rates. The selection of the operating point (one specific point on the ROC curve, thus one particular detector) is left to the user since he/she is the only one able to decide of this tradeoff based on his/her specific needs.

Under controlled laboratory experimental conditions, $P(A|E)$ is often estimated as the number of occurrences of A while E was indeed true. Other said, E is presented to the detector, and each alarm A from $\mathcal{D}$ contributes positively to $P(A|E)$:

$$\hat{P}(A|E) = \frac{\text{Nb of occurrences of } A \text{ when } E \text{ was present}}{\text{Nb of tests where } E \text{ was present}} \tag{2}$$

where $\hat{P}(A|E)$ is a frequentist estimation of $P(A|E)$. If $\mathcal{D}$ triggers each time E is present, then $\hat{P}(A|E) = 1$ which means that the detector perfectly detected E during the experimental tests. Obviously, this measure is not enough to characterise the detector's performance since a detector which always sends an alarm whatever E is present or not will also lead to $\hat{P}(A|E) = 1$. The true positive rate needs thus to be associated with the false alarm rate, itself estimated as:

$$\hat{P}(A|\overline{E}) = \frac{\text{Nb of occurrences of } A \text{ when } E \text{ was absent}}{\text{Nb of tests where } E \text{ was absent}} \tag{3}$$

where $\hat{P}(A|\overline{E})$ is the estimation of $P(A|\overline{E})$. The best detector is the one whose true positive rate is 1 and false alarm rate is 0.

These values $\hat{P}(A|E)$ and $\hat{P}(A|\overline{E})$ represent the detector's performances obtained under controlled experimental conditions. Rather, we should write $\hat{P}(A|E, C_l)$ and $\hat{P}(A|\overline{E}, C_l)$ where C_l represents the experimental conditions under which the estimations of TRP and FAR have been obtained. These values are further used as indicators of the detector's performance and possibly as a prediction quality factor.

> Later in operation, when a detector sends an alarm, $\hat{P}(A|E)$ tells us that during laboratory tests, each time E was presented to the detector it correctly identified it $\hat{P}(A|E)$ % of the time. Equivalently, $\hat{P}(A|\overline{E})$ tells us that during laboratory tests, each time E was not presented to the detector wrongly sent an alarm $\hat{P}(A|\overline{E})$ % of the time.

2.2 Posterior probability $P(E|A)$

What the decision maker is interested in though, is not exactly the past performances of the detector (as represented by $P(A|E)$ and $P(A|\overline{E})$) but rather the posterior probability $P(E|A)$ which is the probability that E indeed occurred given that the detector sent an alarm. The posterior probability, as computed by Bayes' rule, has thus a different meaning than $P(A|E)$: It expresses the belief or confidence one may have in the occurrence of E **after** an alarm occurred.

As Bayes' formula (1) states, $P(E|A)$ is an update of the prior probability of occurrence of the event $P(E)$, in light of a new alarm. We notice as well that for fixed past performances $P(A|E)$ and $P(A|\overline{E})$, our confidence in the occurrence of E after an alarm will be high if $P(E)$ was initially high and low if $P(E)$ was low.

That means that despite a very good detector ($P(A|E)$ high, $P(A|\overline{E})$ low), the posterior probability that the event E indeed occurred given an alarm maybe very low. This phenomenon is well documented in the medical community especially and is known under the name "false positive paradox" (see for instance [1]). This "paradox" occurs when the prior probability of the event we want to detect $P(E)$ is far less than the true positive rate $P(A|E)$. The conclusion is that it is not because a detector is very reliable (according to laboratory tests) that the chance that E occurs, $P(E|A)$ will be very high.

For instance, if we use the values $P(A|E) = 0.99$ (very high true positive rate) and $P(A|\overline{E}) = 0.001$ (null false positive rate) together with a prior $P(E) = 0.0001$ in Bayes' rule (1), the probability that E indeed occurred after receiving a positive

alarm is still very low:

$$P(E|A) = \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.001 \times 0.9999} = 0.0901 \tag{4}$$

In the case of CBRNE detectors where base rates are usually very low (if available), this means that Bayesian reasoning leads to a very low probability of the event E even if the detector triggered.

$P(E|A)$ represents an updated value of the prior probability $P(E)$: Before receiving any alarm, our belief in the occurrence of E is $P(E)$. Without any other evidence (or information), it will not be changed. As soon as an alarm is sent, our belief in E is updated according to Bayes' rule. Then, if the detector is reliable (according to past performances represented by the likelihood functions $P(A|E)$ and $P(A|\overline{E})$), then our belief in E is increased.

Thus, the reading of a low $P(E|A)$ should not be that the detector does not perform well, but rather that our belief in E still still low despite an alarm. However, this value should be compared to:

- the prior $P(E)$: If $P(E|A) > P(E)$, then our belief in E increased and if $P(E|A) < P(E)$, then our belief in E decreased; and
- $P(\overline{E}|A)$ which is the probability that E did not occur while an alarm was received: If $P(E|A) > P(\overline{E}|A)$, then E is more likely than $\overline{E}$. After receiving an alarm, the non-occurrence of the event may still be more likely that its occurrence, simply because our initial belief in $\overline{E}$ was much higher. $P(E)$ acts as a normalisation constant.

2.3 Prior probability $P(E)$

The prior probability of the event E represents the decision maker's subjective uncertainty about the occurrence of E without any additional information. This prior may be estimated by different means:

- The most common case is a frequentist estimation based on past observed occurrences of the event ("base rate"). We can notice that, if the event has never been observed in the past, its prior probability is 0, which would not be valid in Bayes' rule;
- It could be provided by the decision maker at the beginning of the experiment as a subjective assessment of his/her initial belief. In the case the he/she has not reason to believe more in E than in $\overline{E}$, then $P(E)$ would be set to $\frac{1}{2}$; and
- It could be provided by an external source (say some intelligence).

As we saw above, this prior probability has a **high impact** on the value of the posterior probability, especially for extreme values, either very low or very high $P(E)$. In such cases, the update as computed by Bayes' rule would provide a very slight change in the belief of the decision maker after an alarm is indeed sent. How to elicit the prior for Bayesian reasoning depends critically on the nature of the problem at hand. But as noticed in [2], the prior probabilities "may be influenced by base rates, but need not be the same".

3 $P(A|E)$ or $P(E|A)$ as a reliability measure?

It may be argued that $P(E|A)$ is a better measure of "reliability" than the traditional likelihood measure $P(A|E)$. This argument has been for instance put forward by Levin in [3] as a reply to Tversky and Kahneman's study[4] on the poor human ability at inductive reasoning [5]. Levin argues that measuring reliability by $P(E|A)$ rather than by $P(A|E)$ is consistent with people intuitive estimation. Levin's argument of using $P(E|A)$ as a measure of reliability is however shattered by a deeper analysis of Bayesian reasoning and the meaning of the different quantities involved, by Sherry in [6].

Let us recall Tversky and Kahneman's example of blue and green cabs [5] reused in [3, 6][5]:

> Tversky and Kahneman's cabs [5]:
> *A cab was involved in a hit-and-run accident at night: Two cab companies, the Green and the Blue, operate in the city. You are given the following data:*
>
> *(i) 85% of the cabs in the city are Green and 15% are Blue.*
>
> *(ii) A witness identified the cab as a Blue cab. The court tested his ability to identify cabs under the appropriate visibility conditions. When presented with a sample of cabs (half of which were Blue and half of which were Green) the witness made correct identifications in 80% of the cases and erred in 20% of the cases.*
>
> *(iii) A witness identified the cab as a Blue cab. The court tested his ability to identify cabs under the appropriate visibility conditions. When presented with a sample of cabs (half of which were Blue and half of*

[4] Tversky and Kahneman's study addressed in particular the problem of "base rate fallacy" reported in the literature as an error in human thinking according to which base rates are neglected in favor of specific information (see for instance [4]).

[5] The analogy with our general setting is straightforward: The witness is the detector, thus $W_B = A$, and the event of interest is "the car is blue", thus $E = B$.

which were Green) the witness made correct identifications in 80% of the cases and erred in 20% of the cases.

Question: What is the probability that the cab involved in the accident was Blue rather than Green?

Let us denote (as in [6]) by B the event "the car involved was blue" and W_B the event "the witness testifies it is a blue cab". We have thus $P(W_B|B) = P(W_G|G) = 0.8$, $P(W_B|G) = P(W_G|B) = 0.2$, $P(B) = 0.15$ and $P(G) = 0.85$, which introduced in Bayes' theorem gives:

$$P(B|W_B) = \frac{P(W_B|B)P(B)}{P(W_B|B)P(B) + P(W_B|G)P(G)} = \frac{0.8 \times 0.15}{0.8 \times 0.15 + 0.2 \times 0.85} = 0.41 \tag{5}$$

The expression (5) is an instantiation of Bayes' rule and, as discussed in Section 2.2, $P(B|W_B)$ (corresponding to $P(E|A)$ in our detector example) is an update of the probability of B ("the cab involved in the accident was blue") in light of new evidence ("the witness testifies the car was blue"). Before any evidence, the belief of the jurors (the decision makers) in B is 0.15, the prior probability as estimated by statistics of the colors of the town's cabs $P(B)$. Then, a new evidence occurs (under the form of a testimony, or an alarm) and the new probability of B becomes $P(B|W_B) = 0.41$. It is thus increased because (1) the witness testified that the car was blue (as represented by $P(W_B)$) and (2) the witness had previously been estimated to be reliable to 0.8, meaning thus that his testimony should be positively considered (as represented by $P(W_B|B)$).

The main argument of Sherry against Levin's proposal is that because it is unknown whether B or G (E or $\overline{E}$ in our case) was indeed true during the night of the accident (*i.e.*, when the detector actually sends an alarm), the measure of $P(E|A)$ is difficult to exploit without further evidence.

$P(E|A)$ can be taken as a measure of reliability, as Sherry agrees [6], but with a distinct meaning than $P(A|E)$ however: $P(E|A)$ tells "how reliable the testimony is when [the jurors] do not know whether the witness was attempting to identify a green or a blue cab when he testifies". Translated into our detector example, it means that $P(E|A)$ measures how "reliable" the detector is when the decision maker **does not know** which situation the detector was attempting to detect, E or $\overline{E}$. It differs from $P(A|E)$ in the sense that $P(E|A)$ relates to a particular event E, while $P(A|E)$ relates to events of type E in general (under the specified experimental conditions).

The problem of assessing $P(E|A)$ is that "it must be made without [the decision maker] knowing the [event the detector] is trying to identify" [6].

We see several consequences of assessing detectors' reliability using $P(E|A)$ rather than $P(A|E)$:

(1) $P(E|A)$ cannot be assessed by tests because we do not know the ground truth (concurring with Sherry's argument). To respect the meaning of the conditional bar $|$, we should consider A as being the condition of experimentation, thus observe alarms and see if E was present or not upon alarms. Although we can control the environment in laboratory tests (provide E or $\overline{E}$) and observe how the detector reacts (A or $\overline{A}$), the reverse operation is not trivial: How to control alarms and imagine they have an impact on E?

(2) $P(E|A)$ should be assessed through Bayes' rule. In that case, there is an issue in having the performance of the detector depending in the prior probability of the event it attempts to detect. In particular, introducing $P(E)$ in the computation of the detector's reliability comes to considerably reduce the "performances" of detectors of rare events, and this, regardless the (possibly very good) detector's individual laboratory performances. Because the event it aims at detecting is very improbable, its performance will be very low. This approach is thus very disputable.

(3) There is semantic issue in interpreting $P(E|A)$ as a reliability measure. According to Bayesian reasoning, $P(E|A)$ represents our belief in the occurrence of E once A has been observed. Although indeed there is no restriction of naming it a "reliability" measure, doing so may create some confusion for both (1) the decision maker who may think that the detector is not good, while it is and (2) the detector designer who, if advocating for a Bayesian reasoning, may not understand the decision maker's requirements in terms of detection performances. Anyhow, warning the decision maker or end-user about the respective meanings of both values $P(A|E)$ and $P(E|A)$ is of great importance. The semantic confusion between the conditional posterior and prior is known in the literature as "the inverse fallacy" or "conditional probability fallacy" (see for instance [7]).

4 How to characterise a detector?

The question of how to reasonably characterise rare events detector's performance remains. As alternatives to true positive and false alarm rates, detector performances are often characterised in terms of *sensitivity* and *specificity* derived from the former ones. Other aggregated measures are also often used such as the *accuracy*, the *precision*, the *F-measure* or the *area under* (the ROC) *curve* (AUC) (See for instance [8]). Which measure to use for characterising detectors' (or classifiers') performances is still an open question widely addressed in the literature: For instance in [9] some measures are surveyed; a procedure and measure properties for rare events detectors

evaluation is proposed in [10]; in [11] a new measure is proposed; the problem of detection in imprecise environments (when the target operating conditions are difficult to specify precisely) is addressed in [12], proposing an approach which combines ROC analysis, decision analysis and computational geometry; in [13], several measures of performances are analysed for threat detection included human factor considerations using computer based training.

Besides the measures, the experimental conditions should be specified together with the performances and should cover a wide range of the environmental conditions hopefully including the one under which the detector will be deployed. A series of ROC curves (and deduced measures) should thus be provided.

5 How to reduce the "false positive rate"?

The detector $\mathcal{D}$ may be improved in different ways: Using different subsets of features, extracting new features, using different modalities, combining different types of detectors, etc.

The fusion of several detectors is a good approach as it has been proved that their diversity generally improves detection performances (see for instance [14] about multiple classifier systems). Let us consider the case of a detection system with two detectors $\mathcal{D}_1$ and $\mathcal{D}_2$. Each detector provides its own alarm, A_1 and A_2 respectively and, the decision about the presence or absence of E is based on these two sources. How to combine the outputs of $\mathcal{D}_1$ and $\mathcal{D}_2$ is widely addressed within the fusion and pattern recognition communities, to name only these two. Intuitively, we imagine that if two *independent* detectors trigger simultaneously, it would be more likely that E is present than if a single one triggers:

$$P(E|A_1, A_2) > P(E|A_1) \tag{6}$$

Bayes' rule is indeed one means (among others) to combine these two pieces of information. The posterior probability $P(E|A_1, A_2)$ represents the subjective belief of the decision maker given that two distinct and independent[6] alarms occurred. According

[6] If the detectors are independent, $P(A_1, A_2) = P(A_1)P(A_2)$.

to Bayes' rule, this posterior is:

$$\begin{aligned} P(E|A_1, A_2) &= \frac{P(A_1, A_2|E)P(E)}{P(A_1, A_2)} \\ &= \frac{P(A_1|E)P(A_2|E)P(E)}{P(A_1)P(A_2)} \\ &= \frac{P(A_1|E)P(E)}{P(A_1)} \times \frac{P(A_2|E)}{P(A_2)} \\ &= P(E|A_1) \times \frac{P(A_2|E)}{P(A_2)} \end{aligned} \tag{7}$$

Thus, if we denote by $P(E|A_1)$ our belief in the occurrence of E after A_1 only is sent, our belief after both A_1 and A_2 are sent is:

- either increased by a factor $\frac{P(A_2|E)}{P(A_2)}$ if $P(A_2|E) > P(A_2)$, that is if $\mathcal{D}_2$'s reliability (as estimated by the true positive rate) is higher that the total probability of A_2; and
- or decreased by a factor $\frac{P(A_2|E)}{P(A_2)}$ if $P(A_2|E) < P(A_2)$, that is if $\mathcal{D}_2$'s reliability (as estimated by the true positive rate) is higher that the total probability of A_2.

Based on the semantic distinction explained in Section 3, $P(E|A_1, A_2)$ should not be confounded by the "reliability" of the (combined) detector.

However, regarding the global performances (TPR and FAR) of the combined independent detectors, the probability that the two alarms occur concurrently upon the presence of E depends on the way the global alarm, say A_{12}, is defined. Let us consider two cases:

$A_{12} = A_1$ AND A_2 In this case, an alarm is generated by the combined detector $\mathcal{D}_{12}$ if **the two** detectors $\mathcal{D}_1$ and $\mathcal{D}_2$ trigger. Also, no alarm will be generated if **only one** triggers. Then, we have:

$$\begin{cases} A_{12} = A_1 \text{ AND } A_2 \\ \overline{A_{12}} = \left\{ (A_1 \text{ AND } \overline{A_2}) \text{ OR } (\overline{A_1} \text{ AND } A_1) \text{ OR } (\overline{A_1} \text{ AND } \overline{A_2}) \right\} \end{cases}$$

Thus[7]:

$$\begin{aligned} P(A_{12}|E) &= P(A_1, A_2|E) = P(A_1|E)P(A_2|E) \\ P(A_{12}|\overline{E}) &= P(A_1, A_2|\overline{E}) = P(A_1|\overline{E})P(A_2|\overline{E}) \end{aligned} \tag{8}$$

[7]The comma below denotes the logical AND.

Hence, compared to a single detector (say $\mathcal{D}_1$ for instance), the FAR is lower but the TPR is also lower:

$$\begin{aligned} P(A_{12}|E) &< P(A_1|E) \\ P(A_{12}|\overline{E}) &< P(A_1|\overline{E}) \end{aligned} \tag{9}$$

$A_{12} = A_1$ OR A_2 In this case, an alarm is generated by the combined detector $\mathcal{D}_{12}$ if **one of the two** detectors $\mathcal{D}_1$ and $\mathcal{D}_2$ triggers. Also, no alarm will be generated only if **the two do not** trigger. Then, we have:

$$\begin{cases} A_{12} = A_1 \text{ OR } A_2 = \left\{(A_1 \text{ AND } A_2) \text{ OR } (A_1 \text{ AND } \overline{A_2}) \text{ OR } (\overline{A_1} \text{ AND } A_1)\right\} \\ \overline{A_{12}} = \overline{A_1} \text{ AND } \overline{A_2} \end{cases}$$

Thus:

$$\begin{aligned} P(A_{12}|E) &= P(A_1, A_2|E) + P(\overline{A_1}, A_2|E) + P(A_1, \overline{A_2}|E) \\ P(A_{12}|\overline{E}) &= P(A_1, A_2|\overline{E}) + P(\overline{A_1}, A_2|\overline{E}) + P(A_1, \overline{A_2}|\overline{E}) \end{aligned} \tag{10}$$

Hence, compared to a single detector (again $\mathcal{D}_1$), the TPR is higher but the FAR is also higher:

$$\begin{aligned} P(A_{12}|E) &> P(A_1|E) \\ P(A_{12}|\overline{E}) &> P(A_1|\overline{E}) \end{aligned} \tag{11}$$

The two combination rules presented above are just two trivial ones. A plethora of alternative rules and techniques have been (and still are) developed in the literature. Some examples are given in [15, 16] to name only these two references. In general, the performance of a detector combining the outputs of $\mathcal{D}_1$ and $\mathcal{D}_2$ is not trivial. That means that there is no trivial function f such that:

$$P(A_{12}|E) = f(P(A_1|E), P(A_2|E)) \tag{12}$$

The global performance depends on the combination method for the two detectors.

Interesting approaches for combining detectors or sources of information in general are also developed within the framework of evidence theory (also known as Dempster-Shafer theory [17, 18]). In particular, a discounting operation on information provided by unreliable sources is available which aims at minimising the impact of less reliable sources (while still keeping them in the process) and maximising the one of highly reliable sources (see for instance [19, 20, 21]).

6 Conclusions and discussion

This brief addressed the problem of measuring reliability of detectors, and the impact of such a measure on the detector's designers. The discussion was framed in the Bayesian reasoning framework for estimating posterior probability about events based on observations. Beyond the semantics associated to the term "reliability", the semantics of the mathematical objects concerned has been explained and some aspects have been highlighted. In particular, we should not confound:

- the likelihood function $P(A|E)$ which has been obtained under controlled experimental conditions, when the ground truth was available and which can reasonably be considered as a measure of the detector's reliability regarding its ability in providing correct detections; with
- the posterior probability, $P(E|A)$ which expresses the confidence (or belief) of the decision maker in the occurrence of E given that (or posterior to) (1) the detector sent an alarm AND (2) the prior probability of the occurrence of E, $P(E)$. $P(E|A)$ is thus the update of the prior estimation of $P(E)$ in light of a new evidence A. It may be very low especially in the case of the detection of rare events ("false positive paradox"). Such a low value should not lead to the conclusion that the detector was poor in fact, but should rather lead to a mitigation of the reaction in case of alarms.

Using $P(E|A)$ as a reliability or performance criterion to be satisfied by rare events detectors would (1) put strong constraints on the detectors' designers who would need to meet performance criteria beyond their control and thus, hardly reachable. Also, (2) it would induce some semantic confusion for decision makers who may not understand the meaning of the desirable value they provide.

The use of Bayes' rule to estimate the decision maker's subjective belief and update prior probabilities on light of new evidence in the case of rare events detection is debatable, especially because this approach is very sensitive to the prior probabilities. The estimation of the latter is problematic if one wants to avoid the false positive paradox. Moreover, estimating priors through base rates is irrelevant in that particular case because these would be very low. The estimation of the detectors' performance on the output of Bayes' rule suffers from the same drawback.

In order to meet both detector designers' and decision makers' considerations, some approaches could be a conjunction of:

1. a specification of performance curves or measures covering a wide range of experimental conditions;
2. a judicious training of the users about the meaning of the alarms and how they should be considered; and

3. the design of higher-level decision support systems integrating a variety of sources of information together with adequate measures of performance including decision maker needs and technical specifications [22, 23, 24].

About the last item, when decisions have a high cost (of miss or of useless intervention), the decision maker will base his/her decision on a series of pieces of information from multiple sources, possibly one of them only being a detector of event E. Other sources of information are for instance, other detectors for event E based on the same or distinct modalities, several detectors for other events $E_1, \ldots, E_n$, related to event E, contextual information such as a global level of threat, intelligence or expert reports possibly linked to E, background knowledge on similar past events (as recorded for instance in databases), etc. As providing support to the decision maker, these different sources are then fused (combined) to obtain a picture as clear, as precise, as timely, etc, as possible of the situation. Some challenges in reaching that state of situation awareness are the uncertainty representation and management of possible conflict between sources outputs. This includes (but not only) how to judiciously consider the reliability (estimated through past experiments) of the different sources so that they bring their best to the fusion process: Even unreliable sources may bring very relevant and critical information to a global fusion system, for instance if they measure some parameters unavailable to the other sources. The information fusion community has been addressing these issues for years.

How to adequately characterise an individual detector is still an open research question which deserves attention, especially in the case of rare events. This numerical assessment of detector performances is critical as it may serve several purposes such as setting criteria to be meet by detectors' designers', but also as an estimation of the detectors' reliability to be further reintroduced in some fusion process. The validation of rare events detectors in operation is delicate (and may not be necessarily desirable) as it would require observing (without any doubt and in real operation) the event of interest, and even a high number of the kind.

References

[1] M. Hanrahan, "The false positive paradox – Bayes' theorem." www.mthanrahan.com/2013/04/the-false-positive-paradox-bayes-theorem.html, April 2013.

[2] J. J. Koehler, "The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges," *Bahavioral and brain sciences*, vol. 19, no. 1, pp. 1–53, 1996.

[3] M. Levin, "A misuse of Bayes's theorem," *Informal Logic*, vol. 19, no. 1, pp. 63–66, 1999.

[4] M. Bar-Hillel, "The base-rate fallacy in probability judgments," *Acta Psychologica*, vol. 44, pp. 211–233, 1980.

[5] A. Tversky and D. Kahneman, "Causal schemata in judgments under uncertainty," Tech. Rep. PTR-1060-77-10, Defense Advanced Research Project Agency, 1977.

[6] D. Sherry, "Bayes's theorem and reliability: A reply to Levin," *Informal Logic*, vol. 25, no. 2, pp. 167–177, 2005.

[7] G. Villejoubert and D. Mandel, "The inverse fallacy: An account of deviations from Bayes' theorem and the additivity principle," *Memory & Cognition*, vol. 30, no. 2, pp. 171–178, 2002.

[8] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[9] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, pp. 427–437, 2009.

[10] M. Sokolova, K. El Emam, S. Chowdhury, E. Neri, S. Rose, and E. Jonker, "Evaluation of rare event detection," in *Advances in Artificial Intelligence*, vol. 6085 of *Lecture Notes in Computer Science*, pp. 379–383, Springer, 2010.

[11] D. J. Hand, "Measuring classifier performance: a coherent alternative to the area under the ROC curve," *Machine Learning*, vol. 77, pp. 103–123, 2009.

[12] F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Machine Learning*, vol. 42, pp. 203–231, 2001.

[13] F. Hofer and A. Schwaninger, "Reliable and valid measures of threat detection performance in X-ray screening," in *Proc. of the IEEE ICCST*, 2004.

[14] L. I. Kuncheva, *Combining Pattern Classifiers – Methods and algorithms*. Wiley, 2004.

[15] F. Roli, G. Giacinto, and G. Vernazza, "Methods for designing multiple classifier systems," in *Proc. of the second int. workshop on Multiple Classifier Systems* (F. Roli and J. Kiltter, eds.), vol. 2096 of *LNCS*, pp. 78–87, 2001.

[16] D. M. J. Tax and R. P. W. Duin, "Combining one-class classifiers," *Lecture Notes in Computer Science*, vol. 2096, pp. 299–308, 2001.

[17] A. Dempster, "Upper and lower probabilities induced by multivalued mapping," *The Annals of Mathematical Statistics*, vol. 38, pp. 325–339, April 1967.

[18] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[19] G. Rogova and V. Nimier, "Reliability in information fusion: Literature survey," in *Proceedings of the 7^{th} Annual Conference on Information Fusion* (ISIF, ed.), (Stockholm, Sweden), pp. 1158–1165, 2004.

[20] H. Guo, W. Shi, and Y. Deng, "Evaluating sensor reliability in classification problems based on evidence theory," *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, vol. 36, pp. 970–981, October 2006.

[21] A. Martin, A.-L. Jousselme, and C. Osswald, "Conflict measure for the discounting operation on belief functions," in *Proceedings of the 11^{th} Annual Conference on Information Fusion*, (Cologne, Germany), 2008.

[22] J. Llinas, "Assessing the performance of multisensor fusion processes," in *Hanbook of Multisensor Data Fusion*, ch. 20, CRC Press LLC, 1st ed., 2001.

[23] J. Salerno, "Measuring situation assessment performance through the activities of interest score," in *Proc. of the 11th Int. conf. on Information Fusion*, 2008.

[24] E. Blasch, R. Breton, and P. Valin, "Information fusion measures of effectiveness (MOE) for decision support," in *Proc. of SPIE, Signal Processing, Sensor Fusion, and Target Recognition XX*, vol. 8050, 2011.

This page intentionally left blank.

DOCUMENT CONTROL DATA

(Security markings for the title, abstract and indexing annotation must be entered when the document is Classified or Designated.)

1. ORIGINATOR (The name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's report, or tasking agency, are entered in section 8.)

DRDC – Valcartier Research Centre
2459 de la Bravoure Road, Québec QC G3J 1X5, Canada

2a. SECURITY MARKING (Overall security marking of the document, including supplemental markings if applicable.)

UNCLASSIFIED

2b. CONTROLLED GOODS

(NON-CONTROLLED GOODS)
DMC A
REVIEW: GCEC APRIL 2011

3. TITLE (The complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C or U) in parentheses after the title.)

About reliability estimation of detectors and Bayesian reasoning

4. AUTHORS (Last name, followed by initials – ranks, titles, etc. not to be used.)

Jousselme, A.-L.

5. DATE OF PUBLICATION (Month and year of publication of document.)	6a. NO. OF PAGES (Total containing information. Include Annexes, Appendices, etc.)	6b. NO. OF REFS (Total cited in document.)
April 2016	24	24

7. DESCRIPTIVE NOTES (The category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Scientific Report

8. SPONSORING ACTIVITY (The name of the department project office or laboratory sponsoring the research and development – include address.)

DRDC – Valcartier Research Centre
2459 de la Bravoure Road, Québec QC G3J 1X5, Canada

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)	9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)
06cb	

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)	10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)
DRDC-RDDC-2016-R040	

11. DOCUMENT AVAILABILITY (Any limitations on further dissemination of the document, other than those imposed by security classification.)

(X) Unlimited distribution
( ) Defence departments and defence contractors; further distribution only as approved
( ) Defence departments and Canadian defence contractors; further distribution only as approved
( ) Government departments and agencies; further distribution only as approved
( ) Defence departments; further distribution only as approved
( ) Other (please specify):

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11)) is possible, a wider announcement audience may be selected.)

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

This brief note addresses the problem of performance estimation of detectors, with a specific emphasis on detectors for rare events. The discussion is framed in the Bayesian reasoning framework for estimating posterior probability $P(E|A)$ about events E based on observations A. The question addressed is which measure between $P(A|E)$ or $P(E|A)$ should be used as a requirement for performance of detectors. The two measures are first presented and their respective meaning discussed. Beyond the semantics associated to the term “reliability”, the semantics of the mathematical quantities concerned is explained and the impact of this measure on detector’s designers is discussed. Some ideas to improve performances of detectors for rare events are finally sketched, emphasising that the need for such detectors to be only components of a larger situation assessment system, so that correlations with other detectors or other pieces of information can be drawn, and that the decision maker can decide based on a more global picture of the situation.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus. e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

Detection; Performances; ROC curve; Reliability; False alarms; Fusion

DRDC | RDDC

SCIENCE, TECHNOLOGY AND KNOWLEDGE FOR CANADA'S DEFENCE AND SECURITY | SCIENCE, TECHNOLOGIE ET SAVOIR POUR LA DÉFENSE ET LA SÉCURITÉ DU CANADA

www.drdc-rddc.gc.ca