Generating Actionable Information for Classifier Assessment

E. Earl Eiland, New Mexico Institute of Mining and Technology
Lorie M. Liebrock, Ph.D., New Mexico Institute of Mining and Technology

Abstract. DoD systems are becoming increasingly automated and classifiers are essential to their operation. Classifier assessment, then, is also essential, not only during tool selection, but also for maintaining deployed systems. Classifier evaluation is, however, not trivial: there are numerous classifier evaluation measures used. How does one select a measure, and once selected, what do the values mean? This article addresses what constitutes actionable information regarding a classifier's effect on mission capability, maps some commonly seen measures to these needs, and recommends a means of maximizing the actionable information generated in classifier evaluations.

1. Introduction

DoD systems are becoming increasingly automated. Although human supervision is important, more and more operational activities are coming under computer control. Classification is a key element in these control systems. Incorrect classification can have devastating effects. Assessing classifiers is an important activity, not only during tool selection but also at deployment and during operation. Practitioners need actionable information.

Imagine you are responsible for choosing the target selection software for a system such as Iron Dome. Some missiles may impact in desert or agricultural areas, causing little damage. Conversely, some missiles may impact hospitals or power or water plants, causing severe damage and/or casualties. How does one generate actionable information for such a mission? And how can that information be mapped to the software selection task? Unfortunately, classifier evaluation is not a trivial task. One challenge is that there are numerous measures used and their relevance to mission capability is unclear. This article will apply actual industrial control system intrusion detection system (ICS IDS) test results to a hypothetical mission. We will identify what constitutes actionable information for mission capability, define two specific classification mission types, and review some commonly seen classifier evaluation measures ("total accuracy rate"[1] (TAR), F *score* [1] and the "*Youden index*" (J)[2]) for applicability, then recommend a measure selection protocol. ICS IDS utility will be measured relative to the mission. The practitioner's takeaway from this article will be an understanding of classifier evaluation and an evaluation protocol for generating actionable information.

In this discussion, classifier test results are presented in the "de facto" standard format, the joint probability table (JPT) (also called confusion matrix, error matrix and contingency table). Table 1 shows the JPT and defines the variables used for various test category counts.

		Actual classification (ground truth)		
		Y	$\overline{Y}$	*Totals* ↓
Test Result	$+ : s_i \in \{Z\}$	t_+ (malicious)	f_+ (false alarm)	$\lvert Z \rvert = t_+ + f_+$
	$- : s_i \in \{\overline{Z}\}$	f_- (missed attack)	t_- (normal)	$\lvert \overline{Z} \rvert = f_- + t_-$
Totals		$\lvert Y \rvert = t_+ + f_-$	$\lvert \overline{Y} \rvert = f_+ + t_-$	$\lvert S \rvert = \lvert Y \rvert + \lvert \overline{Y} \rvert = \lvert Z \rvert + \lvert \overline{Z} \rvert$

Table 1: Category Cardinalities Shown in a Joint Probability Table.

		Actual classification (ground truth)		
		Y	$\overline{Y}$	*Totals* ↓
Test Result	$+ : s_i \in \{Z\}$	177	0	177
	$- : s_i \in / \{\overline{Z}\}$	23	310000	310023
Totals		200	310000	310200

Table 2: Observed Test Results, Organized into a Joint Probability Table.

	Measure value	
Measure	NO GO	GO
Accuracy	0.9993	**0.9999**
F1-score	0.009852	**0.9389**
Youden Index	0.004978	**0.8850**

Table 3: The bolded values for each of the three measures support a "GO" decision. However, the measure values vary greatly. What do these values indicate to decision makers about the extent to which the target IDS will mitigate the success of an adversary to cripple their mission?

2. Use Case

Our task is to make a "GO/NO GO" recommendation on deploying an advanced ICS anomaly detector on a power system provisioning a mission-critical service. "GO" indicates deployment is justified; "NO GO" indicates deployment is not justified: the existing situation (no anomaly detection) is favorable. Table 2 summarizes our evaluation's test outcome. (These results are an actual test result of an advanced anomaly detector in a recent ICS IDS development project. [3] Test conditions set the significant digits to four). At first glance, the detector results look compelling; there are few missed attacks and no false alarms. However, decision makers are looking at the big picture: how is the mission affected? For this discussion, we assume the mission effect is measured in U.S. dollars; the units could be anything relevant (e.g., pounds, kilowatts, gallons treated). We want to compare implications of both the "GO" and "NO GO" options. Table 3 shows the results for three measures in the literature.

For all three measures, higher is better, so the measures all support a "GO" decision. However, how great will that benefit be? Unfortunately, these measures cannot provide that information.

3. Actionable Information Considerations for Measuring Mission Capability

For classifier selection, deployment, and field checking, practitioners must quantify performance. We have distilled their needs to four questions which correlate directly to rows in Table 4.

3.1 What Is the Classifier's Impact on My Mission?

To address category impacts, consider what the measure quantifies and how that relates to the mission capability. This question is actually a bit simplistic, because there are two different mission types: one where the relative difference in class sizes is important and one where the relative class size ($ratio_+ = |Y| / |\overline{Y}|$) is confounding. [4]

Our intrusion detection use case includes both types of instances. High-impact events may have devastating, perhaps irrecoverable, effects on the mission; even a single occurrence of such an event must be avoided. In this case, the relative frequency of such attacks is irrelevant. Indeed, the occurrence of these events may be sufficiently infrequent so that a distribution cannot be reasonably estimated.

At the other end of the spectrum, some events may be mere nuisances with little effect on the mission. The risk of "nuisance level" events may be sufficiently frequent that the impact is cumulative: $ratio_+$ is important. Commonly seen measures do not reflect this risk well. Question one in Table 4 shows three measures' ability to quantify impact.

For this use case, we also define an intermediate event category. This impact category's missional effect is significant, but not irrecoverable. We also stipulate that a relative attack frequency can be reasonably estimated. Intermediate impact events will have different impact parameters, but can use the same measure as the nuisance events.

Impact is problem specific and can affect each JPT category differently. In this paper, we refer to a vector of the expected per-event impacts: $I = (\iota T_+, \iota F_+, \iota F_-, \iota T_-)$.

3.2 What Is the Boundary that Provides the Optimum Impact?

Practitioners need to know what boundary settings produce the best results. Only a boundary (B) sensitive measure can provide this information. One mission's optimum boundary (B*) may not be optimum for another. In the intrusion detection use case, the three impact levels could each have independent optimum boundaries. This information could affect the detector's deployment strategy.

A measure useful to practitioners must be sensitive to the same factors and to the same degree as practitioners are to their respective mission capabilities. With regard to utility, the practitioner's context is defined by the importance, or impact, of elements from each JPT category in that context. A small change to any element of **I** will generate corresponding changes in a compliant measure.

A useful measure is sensitive to ***I***, allowing practitioners to tune the measure's output to match their mission. Commonly seen measures do facilitate boundary selection; however, they do not quantify impact, so there is no guarantee that their B^*s are optimum for impact. [5] Boundary optimization in Table 4 shows three measures' utility for optimum boundary selection.

3.3 How Sensitive is the Impact to Boundary Selection?

Classifiers are often used in situations where the practitioner has imperfect knowledge of the environment; they may not be able to determine the optimum boundary with certainty. In these cases, a practitioner's boundary selection may be suboptimal. They will benefit from knowing how a slightly suboptimal boundary will affect the mission outcome. Only a boundary-sensitive measure can provide this information. Although many commonly seen measures do have Bs, their values cannot be mapped to impact.

Indeed, different measures will have different curvatures on the same test data; they cannot be used to assess boundary sensitivity. Boundary sensitivity in Table 4 shows three measures' ability to quantify boundary sensitivity.

3.4 How Can I Measure Classifier Performance in the Field?

Practitioners may have imperfect mission and context knowledge, resulting in unexpected classifier performance, or mission conditions may shift, resulting in performance drift. Either situation may justify re-tuning a classifier after deployment; a classifier's field performance must be monitorable. For our intrusion detection use case, the events flagged as malicious can be analyzed, thereby determining ground truth (t_+, f_+), but all that is known about missed attacks(t_-) is discovered by other means. Evaluating a short-range missile interception system, such as Israel's Iron Dome, has the same problem: ground truth (t_-,f_-) is known for ignored missiles[2], but not for the intercepted missiles (Z in this case). In these two cases, field data is incomplete; however, practitioners can rely on knowing Z and $\overline{Z}$. With these two values, practitioners can compare actual and predicted results. Unexpected variations trigger deeper analysis. A generally useful measure, then, must be calculable from classifier output: Z and $\overline{Z}$. Commonly seen measures are not amenable to the practitioner's need for this perspective. Perspective in Table 4 shows three measures' utility for field checks.

	Measure		
Practitioner's need	TAR	Fβ -score	Youden
1: category impacts (Section 3.1)	No	Partial	No
2: boundary optimization (Section 3.2)	Yes*	Yes*	Yes*
3: boundary sensitivity (Section 3.3)	No	No	No
4: perspective (Section 3.4)	No	No	No

Table 4: In the general case, there are four characteristics a measure needs to well-inform practitioners on classifier performance. They are posed as questions in Sections 3.1-3.4; none of the measures considered satisfy all four practitioner's questions. Hence, the measures provide practitioners little actionable information.

**The measure does not quantify the classifier's impact on the practitioner's use case. The B* is optimizing another characteristic, so there is no guarantee that the B* identified is correct for impact.*

4. The Utility of Classifier Evaluation Measures for Practitioners

Table 4 summarizes *Total Accuracy Rate, Fβ-score* and *Youden Index* utility for practitioners. Each measures' ability to answer the four questions posed in Section 3 were evaluated. Key findings were:

- Most are ordinal scale measures and none were ratio scale measures. Answering the questions requires ratio scale values, so none could

answer practitioner questions one through three.

- Each measure's characteristics were, at best, suitable for niche problems. None of the questions were answered for the general case.
- None included mission capability impact. (Category impact in Table 4.)
- Excepting total accuracy rate, the summary measures reviewed define classifier quality equal to a fair coin as "zero." However, from the practitioner's perspective, a fair coin is just another classification algorithm: randomly tagging output as Z or $\overline{Z}$ with probability 0.5 may not have a zero effect on mission capability. A meaningful zero for research may not be meaningful for any particular mission. Thus, the measures could not answer questions one through three.

Additional measure characterizations and in-depth analysis are available in Eiland and Liebrock [5, 6].

4.1 Two Evaluation Measures Valuable for Practitioners

As noted in Section 3, based on relative class size importance, there are two classification mission types, and the existing measures do not factor in expected event impacts. Here we show that there are two measures which include *I*. The vector of JPT category impacts are expressed as statistical expectations (expected individual element impact) by category or class.

For missions such as detecting low- and medium-impact ICS attacks and short-range missile defense, impact is cumulative and ratio$_+$ is an important factor. For this mission type,

$$\iota_I = \iota_{T_+}\frac{t_+}{|S|} + \iota_{F_+}\frac{f_+}{|S|} + \iota_{F_-}\frac{f_-}{|S|} + \iota_{T_-}\frac{t_-}{|S|}$$

is an appropriate measure. ι can also be expressed on Z and $\overline{Z}$, the outputs actually observed by the practitioner:

$$\iota_I = \iota_Z\frac{|Z|}{|S|} + \iota_{\overline{Z}}\frac{|\overline{Z}|}{|S|}.$$

An in-depth discussion and analysis of ι is in the online technical report [6].

For events such as detecting ICS attacks with potentially high impact, each event's impact is independent and ratio$_+$ is confounding; normalized JPTs are used. Normalization mathematically balances relative class size (*ratio*$_+ = 1$), thus it mitigates any skew resulting from ratio$_+$. To facilitate comparison with non-normalized JPTs, the sum of all categories is kept at one ($|S| = 1$) and the individual input class values (Y and $\overline{Y}$) add up to 0.5.

For missions where ratio+ is confounding, the expected impact is

$$\iota_\sigma = \frac{1}{2}\left(\frac{\iota_{T_+}t_{+n}}{|Z_n|} + \frac{\iota_{F_+}f_{+n}}{|Z_n|} + \frac{\iota_{T_-}t_{-n}}{|\overline{Z_n}|} + \frac{\iota_{F_-}f_{-n}}{|\overline{Z_n}|}\right).$$

ι and ισ are suitable for many classification problems and are extensible to classification problems with more than two classes. For ισ, the JPT must be normalized, then each category must be conditioned by its classification tag. Then, the impact-adjusted values can be summed up.

An in-depth discussion and analysis of ι and ισ is in the online technical report [6].

4.1.1 Applying ι and ισ to Practitioner's Questions

How well do the impact measures ι and ισ meet the practitioner's needs?

1. **What is the classifier's impact on my mission capability?** The four impacts specified in *I* quantify the expected mission impact of each output type. These are part of the impact equations, so the measure values quantify mission impact. Also, the impact measures are ratio-scale.
2. **What is the boundary that provides the optimum impact?** The impact measures are boundary sensitive, so the optimum expected impact can be mapped to a value or values in the boundary continuum.
3. **How sensitive is the impact to boundary selection?** Since the impact measures are boundary sensitive, the change in expected impact can be calculated relative to the variation in the selected boundary from the true optimu
4. **How can I assess classifier performance in the field?** Both impact measures can be stated as functions of the classifier's output, Z and $\overline{Z}$. Practitioners can compare these values to the actual (observed) impact values, thereby determining if the classifier is performing as expected.

Both ι and ισ satisfy the four criteria for quantifying a classifier's impact on the practitioner's mission capability.

5. Measure Comparison Protocol

Section 2 describes the ICS use case; this section illustrates applying the new measures to it and the actionable information generated. An essential initial step to using ι and ισ is defining *I*. Using ι and ισ does require additional mission-environment-specific information, the vector *I*. Table 5 lists the impact (*I*) vectors for this use case. The values for each category are based on the following assumptions:

- Legitimate events processed as legitimate (T_): The enterprise realizes a $100 benefit from every legitimate event.
- Legitimate events processed as malicious (F+): Although there is some loss incurred, the enterprise still realizes a $99 benefit from every false alarm event.
- Malicious events processed as malicious (T+): The enterprise avoids the potential negative effects but experiences a small detection and response cost per event.
- Malicious events processed as legitimate (F_): The enterprise realizes a cost from every missed attack. For this use case, a specific event's impact class is mapped directly to the expected ιF_. High-impact events generate a major loss; low-impact events generate a small reduction in per-event income.

These values are for illustration purposes; each practitioner must determine values for their situation.

Another initial step is measure selection. ι is appropriate when the mission results affected by the classifier are cumulative and ground truth for the input population is known. If either of these conditions does not exist, then ισ must be used.

Once *I* has been defined and the measure selected, then the evaluation consists of inserting the JPT data and *I* into the selected measure (ι or ισ) and comparing the values. The *B* with the *JPT* that generates the best impact value is B^*.

	Impact value			
Impact Class	ιT_+	ιF_+	ιF_-	ιT_-
Low	$ – 0.001000	$99.00	$90.00	$100.00
Medium	$ – 0.001000	$99.00	$ – 0.001000	$100.00
High	$ – 0.001000	$99.00	$ – 10,000	$100.00

Table 5: Impact values for this use case. For clarity, impact classes are tied to missed attacks (ιF_-). All other vector values are kept constant. The table reflects the test's precision to four significant digits.

	Measure value	
Measure	NO GO	GO
Accuracy	0.9993	**0.9999**
F1-score	0.009852	**0.9389**
Youden Index	0.004978	**0.885**
Low impact (u)	**$99.99**	$99.94
Medium impact (u)	**$99.93**	**$99.93**
High impact ($\iota\sigma$)	–$4937	**–$941.7**

Table 6: The bolded values for each measure indicate the decision supported. Whereas the common measures all support a "GO" decision, the impact measures show that the decision is not so clear-cut. A "GO" decision is only strongly supported for high-impact events. The impact measures provide more actionable information to practitioners.

Table 6 shows the values using the new measures. Do the proposed measures, $\iota\mu$ and $\iota\sigma$, provide more actionable information? All three common measures support a "GO" decision, but those measures do not reflect how the ICS IDS affects the practitioner's mission capability. The impact measures, however, are quantified in relevant units (for this example, we use dollars per event). Decision makers need only compare the predicted net income for deploying the ICS IDS ("GO") to the predicted net income without the ICS IDS ("NO GO"). The impact measures show that a "GO" decision is only strongly supported for high impact events. For medium impact events, the correct decision is debatable and the ICS IDS is contraindicated for low impact events. Practitioners receive this insight from $\iota\mu$ or $\iota\sigma$, but not the other measures. Given the information in Table 5, one might suspect that the ICS IDS would not be cost effective for low impact events, but $\iota\mu$ or $\iota\sigma$ quantify the expected result. Further, $\iota\mu$ shows that the "GO" decision is strongly supported for high impact events. However, the impact value is still strongly negative, indicating the enterprise still has substantial residual risk. Decision makers may want to implement additional mitigations.

6. User Reaction

The initial work was inspired by anecdotal stories and informal discussions with practitioners. However, a group of practitioners was surveyed to substantiate the stories. The group size was too small for a rigorous statistical analysis. However, two conclusions were possible: the more experience practitioners had with classifier evaluation, the less satisfied they were with the existing measures, and after exposure to $\iota\mu$ and $\iota\sigma$ and viewing a use case applying them, Accuracy, F_1-score and Youden index, the respondents overwhelmingly felt that $\iota\mu$ and $\iota\sigma$ provided substantial additional actionable information and the extra effort needed to define *I* was justified.

7. Conclusion

Historically, practitioners have had difficulty assessing classifier effect on their mission capability. This paper identifies actionable information desired by practitioners, evaluates three commonly seen measures for mission relevance, and recommends changes that provide better quality (more actionable) information. Testing demonstrated the increase in actionable information received by practitioners.

Questions expressing information useful to practitioners were identified:

1. What is the classifier's impact on my mission?
2. What is the boundary that provides the optimum impact?
3. How sensitive is the impact to boundary selection?
4. How can I measure classifier performance in the field?

Commonly seen classifier evaluation measures fail to answer these questions. To illustrate this point, values using three measures — Total Accuracy Rate, F1-score and Youden index — are presented and discussed. Two measures were identified that do answer the questions: $\iota\mu$ and $\iota\sigma$.

Generally, the intent of publishing classifier performance data is to inform a broad readership, including practitioners. Practitioners would be better served if classifier performance reports provided data that enabled them to calculate their own mission-relevant impacts.

One suggestion is for researchers to publish $\iota\mu$ and $\iota\sigma$ with balanced *I*, on normalized JPT(B^*), and the normalized JPT(B) for a range of boundaries. Researchers can use the published $\iota\mu$ and $\iota\sigma$ values, and practitioners have the values necessary to calculate actionable information for their specific missions.

The four-step process for practitioners is:

1. Identify the expected impacts and define *I*. If the problem is complex, like the intrusion detection example, then multiple *I* s may be needed.
2. Identify the appropriate measure, $\iota\mu$ or $\iota\sigma$. Various aspects of a complex problem may need to be addressed separately, and the same measure may not be appropriate for all aspects.
3. If $\iota\mu$ is appropriate, then use JPT tuning to compensate for the mission domain's ratio_+, then condition the published JPTs by Z and $\overline{Z}$.
4. Calculate the selected measure. The B with the JPT that generates the best impact value is B^*. B^* will provide the best results for the target classifier. In the example given, the best impact is the maximum. This, however, is problem-dependent. In some cases, the best result may be a minimum.

With these values, practitioners can also determine the classifier output's *B* sensitivity.

NOTES

1. TAR has been in use so long, its source is not found cited.
2. Ground truth is also known for missed missiles, but that is a separate matter.

REFERENCES

1. Van Rijsbergen, C.J. (1979.) "Information Retrieval." Available online at http://www.dcs.gla.ac.uk/Keith/Preface.html
2. Youden, W.J. (1950.) "Index for rating diagnostic tests." Cancer, 3:32–35.
3. Eiland, E. Earl & Stride, Eric. (June 2016.) "Unpublished data set." Emerging Technologies Directorate. root9B, LLC.
4. Japkowicz, Nathalie & Shah, Mohak. (2014.) "Evaluating Learning Algorithms: A Classification Perspective." Cambridge University Press. New York, N.Y.
5. Eiland, E. Earl & Liebrock, Lorie M. (2013.) "Efficacious end user measures; Part 1: Relative class size and end user problem domains." Adv. in Artif. Intell. 2013:2:2–2:23.
6. Eiland, E. Earl. & Liebrock, Lorie M. (2015.) "Efficacious end user measures for discriminant analysis tools." Technical report, New Mexico Institute of Mining and Technology.

ABOUT THE AUTHORS

E. Earl Eiland is a senior cyber security engineer with root9B, LLC. In his current position, he has lead development of a novel control system anomaly detector and is rolling out a company-wide innovation program. He holds an M.S. in computer science from New Mexico Tech and is completing a computer science doctorate at New Mexico Tech. His interests include risk-based classifier evaluation, machine intelligence applications and control system cybersecurity.
earl.eiland@root9b.com
505.453.2605
root9B, LLC
310 South Saint Mary's Street, Suite 1600
San Antonio, TX 78205

Lorie M. Liebrock is Dean of Graduate Studies and a professor of computer science and engineering at New Mexico Tech. Her research integrates enterprise cybersecurity, foundations of computer science, computer forensics, information assurance, parallel processing, and visualization for complex problems. Dr. Liebrock holds M.S. and Ph.D. degrees in computer science from Rice University and B.S. and M.S. degrees in computer science from Michigan Technological University. She is a member of ACM and IEEE.
liebrock@nmt.edu