Research Article

Efficacious End User Measures—Part 1: Relative Class Size and End User Problem Domains

E. Earl Eiland and Lorie M. Liebrock

Computer Science and Engineering Department, New Mexico Institute of Mining and Technology, 801 Leroy Place, Socorro, NM 87801, USA

Correspondence should be addressed to E. Earl Eiland; eee@nmt.edu

Received 29 June 2012; Accepted 28 October 2012

Academic Editor: Konstantinos Lefkimmiatis

Copyright © 2013 E. Earl Eiland and Lorie M. Liebrock. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Biological and medical endeavors are beginning to realize the benefits of artificial intelligence and machine learning. However, classification, prediction, and diagnostic (CPD) errors can cause significant losses, even loss of life. Hence, end users are best served when they have performance information relevant to their needs; this paper’s focus. Relative class size (rCS) is commonly recognized as a confounding factor in CPD evaluation. Unfortunately, rCS-invariant measures are not easily mapped to end user conditions. We determine a cause of rCS invariance, joint probability table (JPT) normalization. JPT normalization means that more end user efficacious measures can be used without sacrificing invariance. An important revelation is that without data normalization, the Matthews correlation coefficient (MCC) and information coefficient (IC) are not relative class size invariants; this is a potential source of confusion, as we found not all reports using MCC or IC normalized their data. We derive MCC rCS-invariant expression. JPT normalization can be extended to allow JPT rCS to be set to any desired value (JPT tuning). This makes sensitivity analysis feasible, a benefit to both applied researchers and practitioners (end users). We apply our findings to two published CPD studies to illustrate how end users benefit.

1. Introduction

Biological compounds and systems can be complex, making them difficult to analyze and challenging to understand. This has slowed applying biological and medical advances in the field. Recently, artificial intelligence and machine learning, being particularly effective classification, prediction and diagnostic (CPD) tools, have sped applied research and product development. CPD can be described as the act of comparing observations to models, then deciding whether or not the observations fit the model. Based on some predetermined criterion or criteria, a decision is made regarding class membership (\(x \in A\) or \(x \notin A\)). In many domains, class affiliation is not the end result, rather it is used to determine subsequent activities. Examples include medical diagnoses, bioinformatics, intrusion detection, information retrieval, and patent classification. The list is virtually endless. Incorrect CPD output can lead to frustration, financial loss, and even death; correct CPD output is important. Hence, a number of CPD algorithms have been developed and the field continues to be active.

Characterizing CPD effectiveness, then, is necessary. For example, CPD tool developers need to know how their particular modification affects CPD performance, and practitioners want to make informed choices between CPD options before deploying a tool in the field. Jamain and Hand, summarizing their results in a classifier meta-analysis, comment:

"The real question a user generally wants to answer is "which classification methods [are] best for me to use on my problem with my data . . ." [1]."

This question has not been addressed in studies we have read. Indeed, Jamain and Hand generalize the sentiment of R.P.W Duin’s comment regarding comparing automated, heavily parametrized classifiers.

"It is difficult to compare these types of classifiers in a fair and objective way [2]."
Seemingly, the research community has viewed the end user’s need as too complex to address. Thus, for the most part, researchers have focused on addressing their own needs. End user issues, when discussed, have been constrained to specific problem domains. It might be fair to state that each end user’s need is, in some way, unique. However, that does not mean that the apparent complexities faced by end users cannot be identified and managed. Ideally, a means of satisfying end user needs without sacrificing researcher needs will emerge. At a minimum, it should be possible for end users to be enlightened regarding measure suitability (which measures best quantify how a CPD will impact their situation). This paper is a first step in identifying a general structure of CPD problems faced by end users and using that structure to identify CPD evaluation measures and tools relevant to end users. To the extent that research studies present CPD performance information by which end users can estimate impact in their situation, the studies provide improved service to the end user.

Our primary focus is on summary statistics. In the current context, summary statistics are formulae that take measurement suite elements as input and output a single value which represents the target CPD’s overall quality. However, because multiple values are condensed into a single value, information is lost. To the extent essential information is retained, the summary statistic can prove useful for CPD evaluation. A key characteristic of summary statistics is that they are not monotonic; they have optima. Useful summary statistic optima indicate overall classifier quality. Ideally, these summary statistics also quantify some aspect of classifier output efficacious to end users. End users can directly use such values to estimate how the CPD will impact their situation.

As presented by Hand [3], measurement theory distinguishes between two entity attribute types: intrinsic, those that are part of an entity’s definition (e.g., density or mass) and extrinsic, those that are expressions of the entity’s interaction with the environment (e.g., weight). Attributes such as density and weight can be quantified, so we can also talk about intrinsic and extrinsic measures. When reported in joint probability tables (JPTs), CPD output is partitioned into four distinct categories: \( T_+ \), \( F_+ \), \( T_- \), and \( T_- \). After any dataset has been tested, the final object count in each category is influenced by the environmental factors rCS and boundary (B). \( rCS = Y/Y \). B is an n element vector that defines a “surface” that encloses one class, for example, “class A.” In every case, there will also be an optimum boundary \( B' \). All elements outside that surface are in class \( \overline{A} \), rather than class “A.” Because \( T_+ \), \( F_+ \), \( F_- \), and \( T_- \) are sensitive to rCS and B, they are extrinsic measures.

1.1. Nomenclature. Although this paper applies well-established stochastic concepts, not all discussions use the same terminology. To avoid confusion, we define our lexicon for quantities measured (each being the size of the defined set):

- \( T_+ \): correctly identified events in class A, the “class of interest” (if such a class exists);
- \( F_+ \): class \( \overline{A} \) events incorrectly flagged as class A;
- \( F_- \): class A events incorrectly flagged as class \( \overline{A} \);
- \( Z \): events flagged as class A;
- \( \overline{Z} \): events flagged as class \( \overline{A} \);
- \( Y \): actual class A events in the data set;
- \( \overline{Y} \): actual class \( \overline{A} \) events in the data set;
- \( N \): the data set;

These values are often presented in an JPT as in Table 1. When appropriate, these symbols will also be used to represent populations. Context will determine whether a quantity or a population is being referenced.

End users are interested in how a process will function in their environment, so they need measures sensitive to extrinsic factors. From a purely academic perspective, the goal for many researchers is to characterize the CPD process independent of extrinsic factors; thus, they want intrinsic measures. Presumably because of the immediacy of the need, significant progress has been made in identifying and characterizing intrinsic measures. We are interested in extrinsic measures useful for end users; little attention has been paid to their needs.

Because of the disparity between researcher and end user needs, providing for end user needs requires careful consideration. A researcher is interested solely in CPD performance; effects caused by external factors must be accounted for, if not eliminated. In contrast, end users need to incorporate external factors, not compensate for or eliminate them. Thus, in order to have research reports that are readily applicable by end users may require providing values that hold little relevance for researchers. We propose an “end user efficacious” measure suite and a means by which end users can tailor research results to their specific environment.

This study builds on Sokolova et al. and other CPD summary statistic characterization studies [4–14]. A challenge categorical problem evaluators face, when comparing to CPD results reported by others, is adjusting for data set effects. One of the major data set issues is that test sets used may well have different rCSs, with different applicability and/or utility. This can cloud results. As an example, we ran a CPD on two test sets drawn from the same class populations. Since both the class source populations and CPD were the same for each test, one would expect statistically indistinguishable output. However, since JPT categories are extrinsic, the anticipated similarity may be masked. The only difference

### Table 1: Values in the lexicon are often organized into a joint probability table (JPT), such as this.

<table>
<thead>
<tr>
<th>Actual target classification</th>
<th>( T_+ )</th>
<th>( F_+ )</th>
<th>( T_- )</th>
<th>( F_- )</th>
<th>( Z )</th>
<th>( \overline{Z} )</th>
<th>( Y )</th>
<th>( \overline{Y} )</th>
<th>( N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test result</td>
<td>Positive</td>
<td>( T_+ )</td>
<td>( F_+ )</td>
<td>( T_- )</td>
<td>( F_- )</td>
<td>( Z )</td>
<td>( \overline{Z} )</td>
<td>( Y )</td>
<td>( \overline{Y} )</td>
</tr>
<tr>
<td>Totals</td>
<td>( Y )</td>
<td>( \overline{Y} )</td>
<td>( N )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
between the test outputs shown in Table 2 is one test set has a relative class size of 9:1 (rCS = 9) and the other a relative class size of 1:1 (rCS = 1). The CPD performs equally well in each test; however, rCS introduces a bias in the JPTs that makes the CPD performance equality difficult to recognize. When rCS = 1, there are twice as many F− observations as F+. However, when rCS = 9, the ratio between F− and F+ goes to 18:1! Without knowledge of the test sets used, an observer could well conclude that these were two significantly different CPDs, with significantly different applicability and/or efficacy. This is an obvious problem for researchers, thus significant effort has been applied to mitigate it; a selection of rCS invariant measures are available:

(i) the Youden index [15];
(ii) two related measures, diagnostic odds ratio (DOR) [16] and diagnostic power (DP) [17] (DP = (3/π) log(DOR));
(iii) the Matthews correlation coefficient (MCC) [18];
(iv) the receiver (or relative) operating characteristic (ROC) area under the curve (AUC) [19, 20];
(v) information theoretic measures such as the information coefficient (IC) [21, 22].

Youden was addressing the rCS’s biasing effect in 1950; thus, the problem has been known for well over half a century, yet reports regarding mitigation are still entering the literature [9, 23]. In the works reviewed, consideration of end user problem environment was tightly constrained and the view of the data virtually unrestricted. We invert these criteria; first identifying problem interactions with rCS (a broad view of end user needs), then viewing the data such that it addresses the question posed by Jamain and Hand (a constrained view of the data).

rCS is generally confounding in the research environment; this is presumably also true for some end users. However, for other end users, rCS may be important for their problem. For these end users, basing decisions on rCS invariant measures may be misleading. Hence, we start by asking two questions:

(i) “Is rCS important for all end user CPD problem domains?”
(ii) “If not, what characteristics define when to incorporate rCS?”

Consider relating these two questions to a pair of real-world problems. A less effective CPD used with a rheumatoid arthritis test could lead to either more people than necessary being treated, or fewer. Likewise, a poorly selected intrusion detection boundary could cause an IT system to have excessive errors (false alarms or missed attacks).

In order to consider the two questions posed above, we will use a statistical nomenclature to describe a supervised CPD test bed. Viewed from a statistical perspective, observations on the dataset processed provide estimates of the underlying (class) population probabilities (rCS is the odds expression of that probability). Observations in the test system input are drawn from the specific populations (because the source class populations are known for each observation, a “ground truth” exists). The source population relative class sizes can be represented as a probability, for example, the probability that a randomly selected input will be a member of class A. This is the leading probability (Pleading), the probability before the inputs interact with the defined process. In the examples stated, class A members would consist of RA-positive individuals and malicious information system activity. In the field, “ground truth” for any particular individual cannot be known prior to being processed (otherwise there would be no need for evaluation). However, in the test scenario being described, ground truth is known for each test set member. Since the source class is known for each CPD input element, input uncertainty does not affect any individual CPD output.

The balance of this paper is organized as follows. Section 2 considers relative class size. Section 3 describes the research protocol used in this study. Section 4 discusses efficacious measures for end users and considers existing summary statistics. Section 5 considers a cause of rCS invariance in measures and implications thereof. This is followed in Section 6 where we present two examples using the proposed format and tool. The main body of this paper closes with a summary of our findings and presents future work in Section 7. Four appendices with equation derivations and additional JPT normalization details wrap up the paper.

2. Relative Class Size

Abstractly described, test set elements interact with the defined process. This interaction “modifies” the elements (perhaps only by adding a tag indicating strength of the match with a model), leading to a test for class A membership. The probability that a randomly selected output will be detected as a member of class A is the subsequent probability (Psubsequent). Psubsequent describes the state of the data stream after interacting with the defined process and is the combined result of the input mix (quantified as a probability (Pleading = Y/N) or an odds ratio (rCS = Y/N)) and the defined process. The defined process contributes its own uncertainty (Pevent) to the observed output. Thus, the test system can be described by the equation Psubsequent = f(Pleading, Pevent). The CDP test set model is illustrated in Figure 1.

Tying the test set model to the examples, Psubsequent consists of the patient’s RA diagnosis and the stream of
intrusion detector classifications. $P_{\text{event}}$ for the RA diagnosis consists of the strength of the match between the compound assayed and RA, test quality and the boundary used to determine class membership (diseased, not diseased). Similarly, $P_{\text{event}}$ for the intrusion detection example consists of the appropriateness of the model that represents the malicious activity, the reliability of tags defining the activity, and the algorithm (or perhaps rule set) used to make malicious/non-malicious determination.

In the CPD test system described, $\hat{P}_{\text{leading}}$ is a characteristic of the input test dataset; hence, it is always fixed.8,9 It is, in fact, related to rCS:

$$\hat{P}_{\text{leading}} = 1 - \frac{rCS}{1 + rCS}. \quad (1)$$

We can now restate the original question within our framework: "are there problem domains where $P_{\text{leading}}$ is important rather than confounding?"  

One such situation could arise where individual results are significant only to the extent to which they contribute to a cumulative result. Consider setting intrusion detection boundaries. The end user is interested in limiting the impact of intrusions and intrusion prevention. The impact is cumulative, with each evaluation activity contributing. In this case, relative class size (expressed as $P_{\text{leading}}$ or rCS) is important. If the end user were to base its boundary on $P_{\text{event}}$, by using a rCS invariant measure (a measure that could not reflect the end user's estimate of their attack rate), there would likely be either excessive false alarm processing costs or excessive expenses due to missed attacks. Cases of this type, where each individual outcome contributes to a cumulative result, require knowledge of both $P_{\text{leading}}$ and $P_{\text{event}}$.

Are there conditions in which $P_{\text{leading}}$ rather than being essential, might instead cause errors? We suggest that one such situation is when individual results are important and cumulative results are not. Consider a person tested for rheumatoid arthritis (RA). Depending upon the physician's office ordering the test, the frequency of RA+ tests could vary considerably.8 If each office set test boundaries to minimize their respective error rates, there would be a range of test scores that would be classified differently by different offices. Clearly, both diagnoses cannot be correct; a person cannot be simultaneously RA+ and RA-. In this case, considering the physician's rCS-based $P_{\text{leading}}$ does not minimize the error for the patient.

With regard to rCS, we see that while basic research benefits from rCS invariant measures, these measures are not suitable for all end users; rCS invariance will be confounding for some end users. For these end users, any specific environment can have any of (literally) an infinite number of rCS values. Indeed, an end user's expected rCS can vary over time. Thus, a “one size fits all” solution will not be particularly efficacious. Our goal to provide for end user rCS needs, thus, resolves into two tasks:

(i) identify both rCS-sensitive and rCS-invariant measures that are efficacious for end users;

(ii) identify a means by which end users can tailor reported CPD results to reflect performance for their expected rCS.
3. Research Protocol

Although in many problem domains, populations tend to be normally distributed, this is not universal. In order to avoid limiting the applicability of our results, we use analytic procedures that are insensitive to distribution. To preserve generality, our analysis is strictly nonparametric; medians are used instead of means and quantiles are used instead of standard deviations. We also execute our tests with the Monte Carlo method, a nonparametric analytical tool often used when problem complexity (in our case, potential end user problem complexity) is not amenable to mathematical analysis.

CPD evaluation studies can be partitioned into two groups: those that use "real-world" data and those that use simulated data. Characterizing CPD evaluation measures requires observing how the measures respond as CPD output varies. Real-world data, such as those available in repositories, for example, the UCI Machine Learning Repository, provide the opportunity to test against a wide variety of complex data types [24]. However, observing the effect of incremental changes on real-world data is difficult at best. For our purpose, we use simulated CPD output. Although any distribution could be used, we assert normality when generating datasets. All data sets used in this study were generated such that the classes were normally distributed \((N(\mu, \sigma^2))\); \(\mu\) is the distribution mean and \(\sigma\) is the standard deviation. The figures displayed were based on four hundred datasets consisting of two hundred thousand randomly drawn observations from two source populations: positive = \(N(1.0, 0.0225)\) and negative = \(N(2.0, 0.0625)\). Separate tests were run with datasets having rCSs of

\[
2^0:1, 2^1:1, 2^2:1, \ldots, 2^{13}:1.
\]

A total of 5,600 independent data sets were used in this study.

For each summary statistic evaluated, we observed how the reported metric was affected by rCS versus boundary versus metric output. The 3D results are presented as contour plots. Because the measure values are asymptotic to one (thus nonlinear), we use the median of the four hundred runs for each test case; means are not valid for non-linear scales. It is impractical to present confidence intervals on 3D data, but on the 2D graphs in Appendix D, the ninety percent confidence interval (90% CI) is displayed for select test series. To illustrate, the 90% CI is indicated by the vertical lines at each rCS tested in Figure 2; the horizontal line indicates the median.

This protocol provides the flexibility and repeatability necessary for analysis, yet abides by the constraints necessary for analysis of less tractable problem domains with difficult problem environments (e.g., complex CPD input and output distributions).

4. End User Efficacy

End users have two activities: CPD selection and CPD application. Regardless of any end user problem distinctions, these two activities address common interests:

(i) process application is concerned with the accuracy of the CPD for both possible outcomes:

(1) "given that the test is positive, to what extent can the result be relied upon?"

(2) "given that the test is negative, to what extent can the result be relied upon?"

Mathematically, this can be expressed as a conditional probability, or a conditional odds. These values are monotonic, so difficult to use for optimum boundary identification;

(ii) process selection needs to choose the CPD with the best expected accuracy ("given the set of choices for CPD, which CPD will provide the best results and to what extent can its results be relied upon?"). Summary statistic output (based on the two monotonic measures) can inform end users for process selection.

The end user efficacious measures differ for the two rCS problem types; so they are further discussed in the following sections.

4.1. When rCS Is Important. Measure efficacy depends upon whether or not the impact on the end user is cumulative. The two CPD application questions can be expressed mathematically as

(i) "given that the test is positive, to what extent can the result be relied upon?" \(\Omega(T_+ | Z) = T_+ / F_+\);

(ii) "given that the test is negative, to what extent can the result be relied upon?" \(\Omega(T_- | \bar{Z}) = T_- / F_-\).

Proportions, being asymptotic to one, are not ratio measures [25] and, thus, have limited utility. We use odds ratios instead.
Effect of rCS and boundary on “TAR”

Figure 3: Being the sum of the observed correct classifications, TAR is a good measure for evaluating rCS sensitive CPDs. It is significant that the dashed line, indicating the optimum boundary, is not vertical; this shows that TAR is rCS sensitive.

The effect is cumulative (or additive); so the two conditional expressions, instead of being a measure suite, provide ancillary information. For CPD selection, end users will be interested in the proportion of the input stream that can be expected to be correct. Thus, a measure such as the total accuracy rate (TAR)

\[ \text{TAR} = \frac{T_+ + T_-}{T_+ + T_- + F_+ + F_-} \]  

represents the per element expected accuracy. TAR, being asymptotic to one, is not a ratio measure; therefore, averaging is not a valid operation. The total accuracy odds ratio (TOR) would be a better choice:

\[ \text{TOR} = \frac{T_+ + T_-}{F_+ + F_-}. \]  

Proportions, such as TAR and odds ratios such as TOR, are alternate expressions of the same CPD output. In fact, odds ratios can be transformed into proportions using (1).

Figure 3 shows TAR, as boundary and rCS vary. The TAR contour appears to vary little and be relatively constant over a wide boundary range. The optimum boundary (shown on the graph as the black dashed line) intersects the x-axis at around 1.45 and slopes toward 1.0. Additionally, the contour around the optimum boundary flattens as rCS increases. We can also see the optimum boundary and the reported accuracy rate both change as rCS varies. TAR is sensitive to rCS; thus, it is useful for problem domains where cumulative effects are important.

There are two other commonly seen rCS-sensitive summary statistics: F-score and Matthews correlation coefficient (MCC). Another measure, information coefficient (IC); is becoming more prevalent, so we consider it as well.  

4.1. F-Score. F-score is the complement of a summary statistic proposed by van Rijsbergen [26]. The measure suite for F-score is recall and precision; van Rijsbergen's measure is based on information retrieval performance criteria put forth by Cleverdon [27]. Cleverdon's criteria address practitioner needs in information retrieval. Recall quantifies a CPD's completeness (the probability that the desired observations in the database are correctly identified). Precision quantifies the probability that undesired observations are mistakenly labeled as desired. For the information retrieval domain, these data seem to be what end users need to know. F-score is now being seen in other problem domains.

Recall and precision correspond to the conditional probabilities \( P(T_+ \mid Y) \) and \( P(T_+ \mid Z) \) (also known as “True positive rate” (TPR) and “positive predictive value” (PPV)). In the problem domain within which they were introduced (information retrieval), these measures quantify how well an CPD relates an object to a concept, such as selecting a document based on keywords. F-score is defined as

\[ F_\beta = \frac{(1 + \beta^2) \text{(precision)} \text{(recall)}}{(\beta^2 \text{(precision)} + \text{recall})}, \]  

where \( \beta \) is the relative weight of precision and recall:

\[ \beta = \frac{\text{importance of precision}}{\text{importance of recall}}. \]  

If precision and recall have equal weights, then \( F_\beta = F_1 \), which is the harmonic mean of precision and recall:

\[ F_1 = 2 \frac{\text{(precision)} \cdot \text{(recall)}}{\text{precision + recall}}. \]  

Since precision and recall are conditional probabilities, we can convert the F-score equation into JPT values. After substitution and rearranging terms,

\[ F_1 = \frac{T_+}{T_+ + F_+/2 + F_-/2}. \]  

The derivation is provided in Appendix A. Notably, using the harmonic mean results in \( F_1 \) not being equally weighted in the denominator. While this may be suitable for the information retrieval domain and some others, it is hardly universal.

In contrast to TAR, F-score's rCS sensitivity varies, depending upon the class monitored. The effect can be seen in Figures 4(a) and 4(b). Interestingly, both TAR and F-score are well-accepted measures. This may indicate the existence of another CPD problem structure element. Analysis of this possibility is postponed for later consideration.

4.1.2. Matthews Correlation Coefficient. The Matthews correlation coefficient (MCC) is a more recent measure, introduced by Matthews [18]. MCC is the application of Pearsons correlation coefficient to CPD evaluation. In a subsequent classifier measure survey, Baldi et al. restated the measure in the form commonly seen today [21]:

\[ \text{MCC} = \frac{(T_+ \cdot T_-) - (F_+ \cdot F_-)}{\sqrt{Y \cdot Y - Z \cdot Z}}. \]
Although not mentioned explicitly in Matthews’ work, Baldi et al. note that the equation requires normalized distributions:

\[
\frac{a - \bar{a}}{\sigma_A}, \quad \text{where } a \in \{A\},
\]

where \(A = \{a_1, a_2, a_3, \ldots, a_s\}\) is the class of the input dataset and \(s = |A|\), \(\bar{a}\) is the mean of \(A\), \(\bar{A}\) is treated in the same manner. One effect of distribution normalization is class size equalization: rCS = 1 (this is discussed in Section 5). The expressions used to generate normalized JPTs are shown in Table 3 the class size equalization is indicated by the ones in the “normalized totals” row. To demonstrate Matthews initial intent to use normalized distributions, we recalculated Matthews et al’s original results using both actual and normalized JPTs. The values using normalized JPTs matched Matthews results; the values using actual JPT values varied from Matthews reported values by approximately a factor of twenty. Thus, MCC, when introduced, was intended to be calculated on normalized JPTs.

The JPT normalization prerequisite applied by Matthews et al. and noted by Baldi et al. seems to have been lost, although the belief that MCC is rCS-invariant persists [28–34]. As a consequence, Baldi et al.’s equation is sometimes applied without first normalizing the JPTs. In two of these reports, Cannon et al. [28] and Mirceva et al. [34] include JPTs. Upon recalculating their results, we determined that the values presented were based on nonnormalized JPTs. In both cases, there were substantial differences between the results on normalized and nonnormalized JPTs. In the Cannon et al. results, the difference affected not only the values, but also the process rankings. Using normalized JPTs, MOLPRINT, the process ranked last, moved into the upper fifty percent of processes tested. Having rankings of processes substantially change due to such changes could result in selection of a suboptimal process for use in real-world settings.

Figure 5 shows how normalization impacts the rCS sensitivity for MCC. Figure 5(a) shows MCC’s response when the raw JPT values are used. The peak boundary (indicated on the graph by the dashed line) shifts as rCS varies and the value decreases as rCS increases (indicated by the sloping dashed line that intersects the contours). In contrast, Figure 5(b) shows that with JPT normalization, the peak boundary and calculated optimum MCC value are fixed (indicated by a vertical dashed line and contours that do not intersect the peak boundary). Exactly the same data sets were used for both graphs; the only difference is the presence or absence of JPT normalization.

Comparing MCC results is complicated by the fact that the published reports we surveyed did not identify whether or not the MCC values reported were on normalized JPTs or not. As seen in Figures 5(a) and 5(b), comparing results across tests where rCS is not normalized could lead to errors. Since rCS affects the optimum boundary when raw JPTs are used, a simple correction of reported values by recalculating MCC on normalized JPTs will most probably be for a suboptimal boundary; thus, the corrected MCC value will also be suboptimal. In this section, we characterized MCC using nonnormalized data. Appendix C derives an rCS-invariant MCC expression.
As explained by Baldi et al., IC is the mutual information (I) normalized by the entropy in ground truth (H); I is the mutual information contained in ground truth regarding the test set S(Y ∪ Ȳ) and the CPD prediction of ground truth, as contained in Z ∪ Z̄:

$$IC = \frac{I(Y \cup Ȳ, Z \cup Z̄)}{H(Y \cup Ȳ)}.$$  \hfill (11)

Expressing I and H in terms of JPT categories,

$$I(Y \cup Ȳ, Z \cup Z̄) = -H\left(\frac{T_+}{N}, \frac{F_+}{N}, \frac{F_-}{N}, \frac{T_-}{N}\right)$$

$$- \frac{T_+}{N} \log\left(|Y| \ast |Z|\right) - \frac{F_+}{N} \log\left(|Ȳ| \ast |Z|\right)$$

$$- \frac{F_-}{N} \log\left(|Y| \ast |Z̄|\right) - \frac{T_-}{N} \log\left(|Ȳ| \ast |Z̄|\right),$$ \hfill (12)

where

$$H\left(\frac{T_+}{N}, \frac{F_+}{N}, \frac{F_-}{N}, \frac{T_-}{N}\right) = -\frac{T_+}{N} \log\frac{T_+}{N} - \frac{F_+}{N} \log\frac{F_+}{N} - \frac{F_-}{N} \log\frac{F_-}{N} - \frac{T_-}{N} \log\frac{T_-}{N}.$$ \hfill (13)

Information-theory-based measures are gaining traction in the literature [35–43]. Some of these reports indicate the belief that the measures are rCS-invariant [38, 40, 43]. Solis and Rackovsky [41] note that their particular information theoretic measure may not be rCS-invariant. The belief that information theoretic measures are rCS-invariant comes from the fact that information theory applies to probability density functions, which are always normalized (rCS = 1) [45, 46]. Unless JPTs are normalized prior to use, IC and related measures cannot be guaranteed to be rCS-invariant.

Like other measures, IC compares target CPD output to an CPD using random classification. However, it differs in that IC is based on the entropy existing in the test set and CPD output. If the input and output are the same, then IC = 1; if the output of the process is equivalent to that of random selection, then IC = 0. A side effect of IC’s use of logs is increased computational complexity. All of the other measures evaluated have a complexity of O(N), IC is O(N^2). This may limit IC’s utility for large data sets. IC’s computational complexity did affect our analysis. Had we calculated IC on the two hundred thousand element test sets used for the other measures, it would have taken approximately six months. Consequently, we tested IC on twenty thousand element test sets. In Figure 6, we can see that the peak boundary shifts as rCS increases; thus, IC is not rCS-invariant. As with the other rCS-sensitive measures, JPT normalization can confer rCS invariance.

Relative to end user interests, it is unclear what MCC quantifies and under what context the value will be relevant; MCC’s end user efficacy is limited to optimum boundary identification.

MCC’s complexity makes determining an underlying measure suite difficult. This detail will be addressed in the future.

4.1.3. Mutual Information Coefficient. Rosi and Sander introduced an information-theory-based measure into the literature in 1993 [22]. It was subsequently included in a measure comparison by Baldi et al. [21]. Since then, it has gained some traction in biological literature [35–43] and has been seen in network management literature [44]. The measure is sometimes called the information coefficient or mutual information coefficient; we use the acronym IC.
When rCS is confounding, the conditional probability (or the harmonic mean of unequal set sizes) do not exist; so the arithmetic mean of the conditional odds on normalized JPTs is appropriate:

\[
\text{EPA} = \frac{(T_+ Y) / (F_+ Y) + (T_- Y) / (F_- Y)}{2}.
\]

To our knowledge, this end user summary statistic is not found in the literature. We apply this summary statistic in the meta-analysis in Section 6.

As noted in the introduction, rCS-invariant summary statistics are already in use. We review three commonly seen rCS-invariant summary statistics:

(i) the Youden index [15];
(ii) two related measures: diagnostic odds ratio (DOR) [16] and diagnostic power (DP) [17] (\(\text{DP} = (\sqrt{3}/\pi) \log(\text{DOR})\));
(iii) the receiver (or relative) operating characteristic area under the curve (AUC) [19, 20].

Two other summary statistics, the Matthews correlation coefficient (MCC) [18] and mutual information coefficient (IC) [21] are commonly held to be rCS-invariant, but in fact are not. They were discussed in Section 4.1.

4.3.1. Youden Index. The Youden index (traditionally represented by \(J\)) was proposed in 1950 and is seen in medical diagnostic studies [15]. There are a number of expressions of \(J\). The original is

\[
J = \frac{1}{2} \left[ \frac{T_+ - F_+}{T_+ + F_+} + \frac{T_- - F_-}{T_- + F_-} \right].
\]

Perhaps a more common representation is

\[
J = \text{sensitivity} + \text{specificity} - 1,
\]

where

\[
\text{sensitivity} = \frac{T_+}{\bar{Y}}, \quad \text{specificity} = \frac{T_-}{\bar{Y}}.
\]

Further, sensitivity is also known as the true positive rate (TPR) and specificity is the complement of the false positive rate; specificity = 1 − FPR = 1 − (\(F_+ / \bar{Y}\)). Hence, an even simpler (and better, according to the minimum description length principal) definition would be

\[
J = \text{TPR} - \text{FPR}.
\]

In this form, the Youden index can be taken to be a summary statistic of the measure suite \{TPR, FPR\}.

\(J\) is special in that \(J = 0\) indicates an CPD with an output equal to that of tossing a fair coin. \(J = 1\) with a perfect CPD...
and \( J = -1 \) for an CPD that classifies everything. As noted in their respective literature bases, \( J \) shares a characteristic with AUC, in that it is insensitive to rCS. On our source populations, the optimum boundary is approximately 1.4. This can be seen in Figure 7. There is an issue with end user efficacy, however. \( J \) quantifies the spread between the TPR and FPR. This information has little bearing on the “pretest” question posed at the beginning of this section.

4.3.2. Diagnostic Odds Ratio (DOR) and Discriminant Power (DP). Two related measures are the diagnostic odds ratio (DOR) [16] and discriminant power (DP) [17]. DOR is defined as

\[
\text{DOR} = \frac{T_+/F_-}{F_+/T_-},
\]

where \( T_+/F_- \) is true positive odds (TPO) and \( F_+/T_- \) is false positive odds (FPO). After simplification,

\[
\text{DOR} = \frac{T_+T_-}{F_+F_-}.
\]

Discriminant power is defined as

\[
\text{DP} = \frac{\sqrt{3}}{\pi} \log (X + \log W),
\]

where

\[
X = \frac{\text{sensitivity}}{1 - \text{sensitivity}}; \quad Y = \frac{\text{specificity}}{1 - \text{specificity}}.
\]

Recasting the equation, we get

\[
\text{DP} = \frac{\sqrt{3}}{\pi} \log \left( \frac{T_+T_-}{F_+F_-} \right).
\]

The derivation is provided in Appendix B. Comparing the two measures, we see that

\[
\text{DP} = \frac{\sqrt{3}}{\pi} \log (\text{DOR}).
\]

DOR and DP are found in medical research. Interestingly, \( \text{DP} = -\infty \) and \( \text{DOR} = 0 \) when either \( T_+ = 0 \) or \( T_- = 0 \), both need not equal zero. Similarly, \( \text{DP} = \infty \) and \( \text{DOR} = 0 \) when either \( F_+ = 0 \) or \( F_- = 0 \), both need not equal zero. Hence, an CPD can classify some observations correctly (total Accuracy > 0), yet have \( \text{DP} = -\infty \) and \( \text{DOR} = 0 \). This is counterintuitive, since one would expect \( \text{DP} = -\infty \) and \( \text{DOR} = 0 \) to indicate a totally ineffective CPD and DP = \( \infty \) and DOR = \( \infty \) to indicate a perfect CPD, rather than something in between. Since \( T_+ \) and \( T_- \) are (statistically) independent,\(^{14}\) (as are \( F_+ \) and \( F_- \)), the DP and DOR could, in a probabilistic sense, be interpreted as the odds that, given two random observations, one will be classified \( T_+ \) and the other \( T_- \) (one will be classified \( F_+ \) and the other \( F_- \)). While the question seems similar, the fact that the DOR and DP optimum boundaries are different from the other inherently rCS invariant measures tested suggests that the two questions are significantly different. Perhaps this is because the DP and DOR treat the problem as a multiplicative function; we identify the problem as an additive function. This value would seem to be directly relevant in niche CPD scenarios, but not to general CPD problem types.

In medical studies, when the event tested for (\( T_+ \)) has a low probability, DOR approximates relative risk: the rate at which the event was observed in group A versus the rate observed in group B. This is valuable information. However, when applied in the more general CPD domain, there is a problem. In any specific CPD task, the category of interest may not have a sufficiently low probability \( T_+ \); thus, the approximation may not always be acceptably close.

Unfortunately, DOR and DP have a challenging sensitivity to boundary; the optimum boundary is indicated by \( \min(\text{DOR}) \) (or \( \min(\text{DP}) \)). Thus, for any test run, the boundary with the smallest \( T_+ \) and \( T_- \) relative to \( F_+ \) and \( F_- \) gives the best accuracy. Not only is this counterintuitive, but also a potential error source. The problem originates from the fact that the greater the \( \min(\text{DOR}) \) (or \( \min(\text{DP}) \)), the better the results. Thus, if the boundary used to partition the test output is not at \( \min(\text{DOR}) \) (or \( \min(\text{DP}) \)), the results may appear better than they really are.\(^ {15} \) Most observations regarding DOR apply to DP as well. For example, \( \text{DP} = 0 \) when \( T_+ T_- = F_+ F_- \).

One important characteristic of DOR and DP is that they are rCS invariant. An important difference between DOR/DP and the other rCS-invariant measures is that their optimum boundaries, although constant in our tests, are offset from the “minimum error boundary.” These effects can be seen in Figure 8.\(^ {16} \) Since DOR and DP are minima, they follow a valley in the contour graph, instead of a ridge. Also contrary to the other measures, DOR decreases when the absolute class size effect becomes noticeable. This means that the contours are closed, instead of open as seen for the other measures. DOR’s vertical optimum boundary line and constant value
4.3.3. Receiver Operating Characteristic Area under the Curve (AUC). ROC has a solid history. Swets campaigned diligently to establish it as the evaluation criterion of choice [20, 47, 48]. The $\{TPR, FPR\}$ measurement suite is the basis for the AUC summary statistic. The title originates from the fact that it is the area under a “ROC curve,” a curve defined by false positive (FPR = $F_i / \overline{Y}$) and true positive (TPR = $T_i / \overline{Y}$) rates. These values are calculated from JPTs of CPD output for a number of boundaries across the observed range, then graphed as the ROC curve [19, 49]. The literature describes the ROC curve (AUC) as being rCS-invariant as well as boundary-invariant. Because it is boundary-invariant, AUC is a popular tool in our present research environment. However, AUC has been criticized on theoretical terms recently [50, 51].

In contrast to the other summary statistics reviewed, AUC is generally accompanied by the ROC curve (indeed, the ROC curve may be presented without providing AUC). To a person skilled in the art, the ROC curve provides a great deal more information regarding CPD performance than does the single value AUC summary statistic (this is, of course, true for any measure suite, since consolidation of multiple values into a single summary statistic value means that information is lost).

Compared to our end user focused criteria, ROC-AUC, being boundary-invariant, is not useful for boundary identification. Nor is it efficacious for end users.

There are numerous ROC-AUC variants [52, 53]. Vanderlooy and Hüllermeier determined in their comparison, that despite intuitive appeal, none of the variants confer any CPD selection improvement. From the end user perspective, since the underlying measure units remain the same, they all have the same limited efficacy.

Figure 9 shows the optimum boundary versus rCS for the proposed normalized PPV, NPV average and existing summary statistics, normalized MCC, Youden index, and DOR. As can be seen in the figure, DOR peaks at a different boundary than the other rCS-invariant measures tested, and (excepting DOR) the optimum boundary is relatively stable until rCS $> 2^6$, after which the detected optimum boundary starts dropping rapidly. For our context, a key finding from Figure 9 is that not only does DOR have weak end user efficacy, it also should not be used to identify the optimum boundary.

4.4. Findings regarding rCS-Invariant End User Problems. For problems requiring rCS invariance, we find that

(i) end users need three values. For CPD selection, the expected total predictive accuracy (a summary statistic) $EPA = ((T_i / \overline{Y})/(F_i / \overline{Y}) + (T_i / \overline{Y})/(F_i / \overline{Y}) / 2$ is important. When the CPD is used in the field, the summary statistic value has no meaning. Instead, end users need the information provided by the two measure suite elements; the positive predictive value odds ratio $PPV = (T_i / \overline{Y})/(F_i / \overline{Y})$ and the negative predictive value odds ratio $NPV = (T_i / \overline{Y})/(F_i / \overline{Y})$. 

(see in Figure 8(b)) indicate that DOR (and hence, DP) is rCS-invariant. Because of this boundary bias, they may not be useful for selecting boundaries. For example, in our test environment, TAR at the common optimum boundary is 0.994, TAR at DOR optimum boundary is 0.958; the difference is significant at the 95% confidence level.
Effect of rCS on the optimum boundaries for some rCS-invariant summary statistics

Figure 9: Other than DOR/DP, all of the rCS-invariant measures tested identified the same optimum boundary.

(ii) Although many commonly seen summary statistics can be used to identify the optimum boundary, as seen in this study, not all can (e.g., ROC-AUC and DOR). Figure 9 shows how the optimum boundary identified by DOR differs from that identified by the other summary statistics tested.

(iii) Of the rCS sensitive summary statistics evaluated, only the EPA output answers the end user’s CPD selection question. The others may be useful for niche problems, but provide little useful information for the “common” end user.

5. JPT Normalization

In statistical circles, standardizing distributions is a well-established technique. One effect of standardization is that the area under the probability density function (pdf) equals 1. This simplifies pdf analysis, since the area of any segment of the area under the curve can be interpreted as the probability of an event occurring within that segment. Similarly, distribution standardization facilitates pdf comparisons. Since the CPD analysis domain considers processes with overlapping pdfs, it intersects with the pdf comparison domain, but is neither a superset nor a subset. Where appropriate, distribution standardization is a useful tool.

In CPD analysis, distribution standardization takes the form of JPT normalization. Table 4 shows a JPT displaying “raw” data—actual category cardinality. After normalization, the class totals (bottom row in Table 5) are one. Thus, JPT normalization seems to be a cause for rCS-invariance in measures. As such, it provides a benefit to end users with rCS invariant problems: any JPT-based CPD evaluation measure will have rCS invariant output, if the input JPTs are normalized. (Illustrated in Figure 10).

Although any measure can be rCS-invariant when the JPTs are normalized, some measures have emerged which have intrinsic rCS invariance. These inherently rCS invariant measures all have \{TPV, FPV\} (ratios that normalize the JPTs) as measure suites, thus rather than being counter examples, they provide empirical evidence that JPT normalization is the root cause for rCS invariance in measures; proof is beyond the scope of this paper. An overview of commonly seen rCS invariant measures is provided in Appendix D.

There is also a benefit for end users with rCS-sensitive problems. Statisticians use distribution standardization to mitigate rCS effects; however, the process is reversible. JPTs with rCS = 1 can be “tuned” to any desired rCS simply by multiplying one class by a constant \(c\) so that \(c \frac{Y}{Y'}\) equals the desired value. Thus, an end user with an rCS-sensitive problem can adjust reported results to fit their need. JPT tuning also allows end users to execute sensitivity analyses and estimate how the CPD will perform in their environment, over the expected rCS range. These insights are applied to a real-world problem in Section 6; a comparison of two RA diagnostic tests and an intrusion detection problem.

However, the optimum boundary is rCS-dependant; thus, the tool is not complete. To apply to all end users, results for all possible optimum boundaries would need to be provided. This is impractical, if not impossible, for CPD test reports to include. As illustrated in Endnote 15, the tuned JPTs will indicate trends, but cannot be considered definitive. Nonetheless, JPT tuning extends JPT normalization in a way we have not previously seen in the literature and provides end users with a useful capability.

6. Examples

In this section, we use our proposed end user efficacious data analysis on two real-world problems, a meta-analysis comparing two medical diagnostic tests for rheumatoid arthritis (RA) by [54] and data from a cyber security masquerading study by [55].
Advances in Artificial Intelligence

6.1 A Meta-Analysis of Rheumatoid Arthritis Diagnostic Tests.

The meta-analysis is quite thorough and accounts for many potential variations between studies. Three hundred and two relevant studies were found; eighty-six satisfied the rigorous inclusion criteria. The team concludes that one test is better than the other, however, does so without using a summary statistic. Our reanalysis adds the three recommended measures identified in Section 4.3.

The study uses two measures: positive likelihood ratio (LR⁺) and negative likelihood ratio (LR⁻). Given a typical test using supervised inputs (where ground truth is known), these two values are efficacious for researchers. They are less efficacious where end users have only the CPD output and ground truth is unknown. As we note in Section 4.3, PPV and NPV are more relevant for end users.

On pooled data, the LR⁻ differences between the tests were statistically insignificant. However, the LR⁺ results were statistically significant. On the pooled data, the “anti-CCP” positive test results were more frequently correct than the “RF” diagnoses. The authors make one important point regarding rheumatoid arthritis treatment; it is harmful and costly to treat persons with false positive results. Hence, it is important to correctly diagnose negatives as well as positives: total expected predictive accuracy (EPA) is important for rheumatoid arthritis treatment. JPT normalization allows measurement of EPA that is rCS-invariant. Using normalized JPT data,

$$\text{EPA} = \frac{T_+/F_+ + T_-/F_-}{2}. \quad (25)$$

In our extension to Nishimura et al.’s report, we calculate the normalized EPA, PPV, and NPV on the pooled test data. Table 6 shows the original likelihood ratios reported by Nishimura et al. and EPA (the parenthesized range is the 95% confidence interval). LR⁺ and PPV show similar values, but LR⁻ is about one-sixth of the NPV; end users should be cautious when interpreting likelihood ratios. Comparing EPAs for each test and keeping in mind the end user context requires rCS invariance, the anti-CCP test correct diagnosis rate is a little more than twice the correct diagnoses rate of the RA test. This is true, even though, as can be seen in Table 7, the RF test actually more accurately detects RA’s presence.

The authors note that “the better accuracy of anti-CCP antibody was mainly due to its higher specificity.” In comparing the JPTs in Table 7, the anti-CCP pooled data
Table 7: These normalized JPTs of Nishimura et al.'s [54] pooled anti-CCP and RF test data were generated using their reported sensitivities and specificities. A person without RA is far less likely to be misdiagnosed than one with the disease, when the anti-CCP test is used.

<table>
<thead>
<tr>
<th>Actual RA condition</th>
<th>Diseased</th>
<th>Not diseased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-CC test result</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>0.67</td>
<td>0.05</td>
</tr>
<tr>
<td>Negative</td>
<td>0.33</td>
<td>0.95</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 8: The above two JPTs have been “tuned” to a population where the diseased population is one hundred times the undiseased population. In this environment, the cumulative results will cause the anti-CCP test to appear to outperform the RF test.

<table>
<thead>
<tr>
<th>Actual RA condition</th>
<th>Diseased</th>
<th>Not diseased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-CC test result</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>67</td>
<td>0.05</td>
</tr>
<tr>
<td>Negative</td>
<td>33</td>
<td>0.95</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>1</td>
</tr>
</tbody>
</table>

The cyber security problem domain is one where cumulative effects (e.g., processing false alarms \(F_{+}\)) are important. Consider an end user desiring to detect masquerading attacks, in which an attacker pretends to be an authorized user in order to gain access to a system. Determining the appropriate boundary for the detector is necessary in order to balance the effects of false alarms and missed attacks. This balance is subjective to the relative volume of normal and masquerade system activity; thus, rCS is important; end users will want to incorporate rCS.

Schonlau et al. simulate a masquerade attack by capturing UNIX commands resulting from specific users, then inserting UNIX commands generated by another user into the original command stream. They compare a number of detection algorithms. The best performing was based on data compression. The Bayesian classifier [55] they used did not perform as well. We compare the two algorithms using the seven summary statistics discussed earlier.

Schonlau et al. use ROC curves to compare their various detection algorithms. From a research perspective, this is appropriate, since ROC is invariant to rCS and does not require boundary selection. However, as noted in Section 4.3.3, ROC provides limited information to end users. To illustrate the effect of using an inappropriate measure type, we reanalyze one of the user command streams with both rCS-sensitive and rCS-invariant measures. Table II shows the results.

For all seven summary statistics considered, the classifier with the higher value is better. Clearly, regardless of the measure, the compression algorithm outperforms the Bayesian classifier. The summary statistic values and associated optimum thresholds, however, vary widely.

What do we learn about the two classifiers from the measure values? The IC measures information content; MCC measures covariance. Youden index and DOR/DP quantify more esoteric characteristics. All four measure classifier...
Table 9: The top two JPTs have been “tuned” to a population where the diseased population is one hundredth of the undiseased population. In this environment, the two tests appear statistically indistinguishable.

<table>
<thead>
<tr>
<th>Actual RA condition</th>
<th>Diseased</th>
<th>Not diseased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-CCP Positive</td>
<td>0.67</td>
<td>5</td>
</tr>
<tr>
<td>Negative</td>
<td>0.33</td>
<td>95</td>
</tr>
<tr>
<td>Totals</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>EPA</td>
<td>144 (136–150)</td>
<td></td>
</tr>
</tbody>
</table>

(a)

Table 10: This table shows the total accuracy ratios for RA tested populations of 100:1 (the rheumatologist) and 1:100 (the general practitioner).

<table>
<thead>
<tr>
<th>Test</th>
<th>Patient bases (diseased : undiseased)</th>
<th>Rheumatologist (100 :1)</th>
<th>General practitioner (1 :100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-CCP</td>
<td></td>
<td>670 (850–650)</td>
<td>144 (136–150)</td>
</tr>
<tr>
<td>RF</td>
<td></td>
<td>230 (250–212)</td>
<td>137 (131–143)</td>
</tr>
</tbody>
</table>

Table 11: These tables show the results for the compression-based classifier and the Bayesian classifier. The measures output on different scales and measure different characteristics; they cannot be directly compared. Because these are two different classifiers, the output ranges differ.

(a)

<table>
<thead>
<tr>
<th>Compression classifier</th>
<th>rCS-sensitive measures</th>
<th>rCS-invariant measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure</td>
<td>Value</td>
<td>Boundary</td>
</tr>
<tr>
<td>IC</td>
<td>0.528</td>
<td>0.800</td>
</tr>
<tr>
<td>TAR</td>
<td>0.930</td>
<td>0.800</td>
</tr>
<tr>
<td>MCC</td>
<td>0.786</td>
<td>0.600</td>
</tr>
<tr>
<td>$T_+ F$-score</td>
<td>0.829</td>
<td>0.800</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Bayesian classifier</th>
<th>rCS-sensitive measures</th>
<th>rCS-invariant measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure</td>
<td>Value</td>
<td>Boundary</td>
</tr>
<tr>
<td>IC</td>
<td>0.068</td>
<td>188</td>
</tr>
<tr>
<td>TAR</td>
<td>0.620</td>
<td>638</td>
</tr>
<tr>
<td>MCC</td>
<td>0.053</td>
<td>-228</td>
</tr>
<tr>
<td>$T_+ F$-score</td>
<td>0.543</td>
<td>-387</td>
</tr>
</tbody>
</table>

Table 12: This table shows how JPT tuning can assist end users in estimating how a CPD will work in their environment. An executive looking at TOR will see that there are 25 correct classifications for every incorrect in the expected operating range. The IT system administrator looking at $\Omega_F$-score will see that there will be thousands of errors for every correct $T_+$. Their decisions regarding the usefulness of this CPD may differ.

<table>
<thead>
<tr>
<th>rCS =</th>
<th>TOR</th>
<th>$\Omega_F$-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>78</td>
<td>7.1</td>
</tr>
<tr>
<td>3.7</td>
<td>13.3</td>
<td>4.9</td>
</tr>
<tr>
<td>1.000</td>
<td>19</td>
<td>0.042</td>
</tr>
<tr>
<td>10 K</td>
<td>25</td>
<td>0.0021</td>
</tr>
<tr>
<td>100 K</td>
<td>25</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

A corporate executive might be concerned about the effect all four categories could have on the enterprise’s performance; thus, TOR would be most relevant. Consider the TOR score of Schonlau et al.’s raw data in Table 12. The executive can tell that when rCS = 3.7, there will be over 13 correctly classified events for every incorrect classification. Based on these values, both persons might decide that performance is acceptable. JPT tuning, however, changes the picture considerably. The executive will see accuracy triple, but the system administrator will see a decrease in accuracy of more than three orders of magnitude. The executive and system administrator may now have different opinions.

Another problem with selecting an inappropriate summary statistic can be seen in Table 11. Not all measures have the same optimum threshold. An end user relying on an inappropriate summary statistic to determine a useful boundary for masquerade detection may be disappointed with their results.
After a classifier is selected, the two perspectives can lead to different system optimizations. When made available to end users, TAR and F-score values can help stakeholders such as executives and IT managers make more informed decisions. Since TAR/TOR and F-score/ΩF-score may have different optimum boundaries, practitioners and decision makers may benefit from having both values reported for each optimum boundary over an rCS range. That way, end users will have an appreciation of the tradeoff associated with selecting a particular solution.

7. Conclusion

This paper is a first step in identifying the structure of CPD problems faced by end users. Using that structure, we characterize how CPD evaluation measures are relevant to end users and identify end user relevant evaluation tools. To that end, we have defined rCS’s importance to end user utility, research reports should include information on the cumulative output or each individual CPD output, rCS values used are from normalized JPTs, then the four normalized values are based on normalized JPTs. If the values used are from normalized JPTs, then Y and \( \tilde{Y} \) both equal one, thus are unnecessary.

Appendices

A. Restating \( F_1 \) in Terms of JPT Values

As defined,

\[
F_1 = 2 \frac{(\text{precision})(\text{recall})}{\text{precision} + \text{recall}},
\]

where

\[
\text{precision} = \frac{T_+}{T_+ + F_-},
\]
\[
\text{recall} = \frac{T_+}{T_+ + F_-}.
\]

Substituting, we have

\[
F_1 = 2 \frac{(T_+ / (T_+ + F_-)) (T_+ / (T_+ + F_-))}{(T_+ / (T_+ + F_-) + (T_+ / (T_+ + F_-))}.
\]

Multiplying and creating common denominators,

\[
F_1 = \frac{2T_+^2}{(T_+ + F_-) (T_+ + F_-)} / ((T_+ + F_-) (T_+ + F_-)).
\]

Multiplying numerator and denominator by \((T_+ + F_-)\),

\[
F_1 = \frac{2T_+}{2T_+ + F_-} = \frac{T_+}{T_+ (F_+/2) + (F_-/2)}.
\]

B. Restating DP in Terms of JPT Values

\[
\text{DP} = \frac{\sqrt{3}}{\pi} (\log X + \log W),
\]

where

\[
X = \frac{\text{sensitivity}}{1 - \text{sensitivity}},
\]
\[
Y = \frac{\text{specificity}}{1 - \text{specificity}},
\]

sensitivity = \( \frac{T_+}{Y} \),

\[
1 - \text{sensitivity} = \frac{F_-}{Y},
\]

specificity = \( \frac{T_-}{Y} \),

\[
1 - \text{specificity} = \frac{F_+}{Y}.
\]

Combining the logs, we get

\[
\text{DP} = \frac{\sqrt{3}}{\pi} (\log (XY)).
\]
Illustrated in Table 14. Using Table 14 definitions, any JPT can be defined in terms of the TPR and FPR. This is the AUC and Youden index incorporate JPT normalization. Section 3, Table 1.

Multiplying top and bottom by \( Y_2 \), we are left with

\[
DP = \frac{\sqrt{3}}{\pi} \log \left( \frac{T_p Y \; T_p Y}{F_p Y \; F_p Y} \right). \tag{B.5}
\]

Then, substituting for \( X \) and \( Y \),

\[
DP = \frac{\sqrt{3}}{\pi} \left( \log \left( \frac{\text{specificity}}{1 - \text{specificity}} \right) \frac{\text{sensitivity}}{1 - \text{sensitivity}} \right). \tag{B.4}
\]

C. Derivation of Normalized MCC Equation

An important side note is that MCC, as commonly calculated,

\[
\text{MCC} = \frac{(T_p T_n) - (F_p F_n)}{\sqrt{Y_1 Z_2 Y_2 Z_2}} \tag{C.1}
\]

is not rCS-invariant as is sometimes reported [28–34]; it must use normalized JPTs (as in Table 13). Substituting the normalized JPT values in (C.1) and collecting terms, the rCS-invariant MCC is

\[
\text{Normalized MCC} = \frac{(T_p T_n) - (F_p F_n)}{\sqrt{Y_1 (Y_1 + Y_2) Z_2 (Z_1 + Z_2)}}. \tag{C.2}
\]

Equation (C.2) can be used in lieu of normalizing JPTs prior to calculating MCC.

D. Measures with Intrinsic rCS Invariance

Although the AUC, Youden index and DOR/DP are distinctly different measures; they all have one key similarity: normalized input. The AUC and Youden index both are (TPR, FPR), and since TPR and FPR are conditional probabilities \( P(T_+ \mid Y) \) and \( P(F_- \mid Y) \), likewise, TNR and FNR are conditional probabilities \( P(T_- \mid Y) \) and \( P(F_+ \mid Y) \). If, in the JPT, we replace \( T_+ \) by TPR, \( F_+ \) by FPR, \( T_- \) by TNR, \( F_- \) by FNR, then the marginal totals \( Y \) are replaced by 1s and \( N \) becomes 2. This is shown in Table 15. Since the two marginal totals representing class size are equal, this process compensates for rCS: the CPD output JPTs have been normalized. In this paper, calculations and discussion using the rCS invariant JPT form shown in Table 14 will refer to "normalized" versions. Any discussions not referring to "normalization" are of measures using the "raw" JPT form as presented in Section 3, Table 1.

Regardless of the actual test set rCSs, the input values for the AUC and Youden index incorporate JPT normalization. Although not as evident, this is also true for DOR and DP. Any JPT can be defined in terms of the TPR and FPR. This is illustrated in Table 14. Using Table 14 definitions,

\[
\text{DOR} = \frac{(C_p \text{TPR}) (C_\pi (1 - \text{FPR}))}{(C_\pi (1 - \text{TPR}) (C_\pi \text{FPR})}. \tag{D.1}
\]

Thus, we find that DOR and DP are based on normalized JPTs as well.

From the literature, we see that MCC is rCS-invariant when calculated on normalized JPTs. Presumably, other rCS-sensitive measures will be rCS-invariant when calculated on normalized JPTs as well. We tested this hypothesis by calculating accuracy, F-score, and MCC values on normalized versions. Figure II displays the peak Accuracy rate and F-score on normalized JPTs and compares them to the output of the established rCS-invariant measures, AUC, DOR, and Youden index (DP, being just a log expression of DOR was left out.) The graphs are provided solely to compare their response to rCS. Any conclusions from Figure II beyond that must be made with care.

Figure II brings out some interesting points.

(i) Confidence interval response to rCS seems to fall into two categories. All of the normalized measures (including AUC, Youden index, DOR, and DP) have relatively stable CIs below rCS = 26. Above rCS = 26, there is an observable trend away from the stable value. This is due to a well-known issue with absolute sample size related to the strong law of large numbers. In our tests, the problem became statistically significant when the smaller sample had less than four hundred members.
Figure 11: All of the normalized summary statistics exhibit rCS invariance. All of the lines vary from the horizontal, indicating that rCS invariance weakens when $rCS > 2^6$. This is a well-known absolute sample size issue. In our tests, the problem became statistically significant when $|A| < 400$. 

(a) Peak normalized accuracy rate (90% CI). It strongly resembles the Figures 11(b), 11(c), and 11(d).

(b) Peak normalized F-score (90% CI). When $\beta = 1$, it strongly resembles Figures 11(a), 11(c) and 11(d).

(c) Peak normalized MCC (90% CI). It strongly resembles Figures 11(a), 11(b) and 11(d).

(d) Peak Youden index (90% CI). This measure turns out to be related to the normalized accuracy rate.

(e) Peak AUC (90% CI). It appears somewhat less sensitive to absolute sample size.

(f) Best DOR (90% CI). As rCS invariance weakens, the DOR value drops.
For normalized accuracy rate, Youden index, normalized $F$-score, and normalized MCC, the 90% confidence interval generally increases as rCS increases. Analyzing the CIs is difficult because (all but DOR) are measured on scales with an upper bound, their scales are not linear. The CI changes observed, however, are consistent with expectations. In general, as the positive class size decreases, normalization magnifies any changes in $T_p$ and $F_p$ far more than normalization of the negative class makes offsetting reductions (the positive class decreases by a factor of $2^{14}$, while the negative class increases by a factor of less than $2^1$).

(ii) Measure families have been found in the summary statistics evaluated. As discussed earlier, DOR and DP are related. The test also reveals a similarity between the normalized accuracy rate and Youden index:

\[ \text{Youden index} = TPR - FPR \]

\[ \text{norm accuracy rate} = \frac{TPR + 1 - FPR}{2} \]  
(D.3)

so that

\[ \text{norm accuracy rate} = \frac{\text{Youden Index} + 1}{2} \]  
(D.4)

Thus, we see that normalized accuracy rate and Youden index are related.

(iii) JPT normalization can inflate reported process accuracy. Each graph in Figure II exhibits rCS stability when rCS < $2^6$. However, when rCS > $2^6$, rCS invariance seems to weaken. This turns out to be a function of the absolute size of the smaller class and is a consequence of the strong law of large numbers. As class sample size decreases, its representation of the source population decreases. The problem is that as sample size decreases, distribution tails lose their definition. When a sample size is magnified by JPT normalization, the undefined tails do not reappear, thus causing the sample to represent a source population with a smaller variance. This means the class overlap is underrepresented. Since process accuracy is inversely related to class overlap, a reduction in estimated class overlap will result in process accuracy overestimation. In our tests, the difference became statistically significant when sample sizes fell below four hundred members.

Violating the strong law of large numbers also affects the optimum boundary. As the apparent source population variance decreases, the boundary shifts toward that class. This can be seen in all of the contour graphs. In order to increase rCS, our protocol decreases $|A|$. A has the lower mean; thus, as rCS increases, the calculated optimum boundary starts shifting toward $\mu_A$. In our tests, when rCS > $2^6$, the shift becomes statistically significant.

**Acknowledgments**

The authors are thankful for the insightful comments in the early days of this work by Dr. Lynda Ballou, Mathematics Department, NM, USA Institute of Mining and Technology, Socorro, New Mexico. They are also indebted to Dr. Andrew Barnes, Applied Statistics Lab, General Electric Global Research, Niskayuna, NY, USA for taking the time to discuss this work. The journal reviewer’s comments strengthened the paper considerably.

**Endnotes**

1. There are two levels of tool development. If development is “basic research,” then evaluation is application-agnostic. If it is “applied research,” then the focus is application-specific and evaluation needs tend to align with practitioner needs. For the purpose of this discussion, researchers do basic research; end users consist of practitioners and applied researchers.

2. CPD are a subset of the more general group of categorical problems. Our investigations apply to both.

3. The measure suite members quantify some particular aspect of CPD performance, thus providing greater CPD performance detail. These measure suite elements tend to be monotonic; thus, they are difficult to use individually to quantify overall CPD performance.

4. If the summary statistic only generates a single value, it is by definition $B^*$. $B^*$ can also be a range. If there are multiple $B^*$'s, they need not be continuous.

5. For example, measures used for CPD evaluation have been tested for invariance to various JPT perturbations [6]. Sokolova and Lapalme claim to be the first to comprehensively assess invariance to JPT perturbations; no boundary invariance assessments are known and only one summary statistic, AUC, claims boundary invariance. This study observes boundary effect on the metrics, but does not look for a basis for boundary invariance. This is left for future work.

6. The receiver operating characteristic originated in signal theory and gained its name from that problem domain. ROC, however, is now commonly used to analyze categorical data represented in JPTs. Although “receiver operating characteristic” may be the most commonly seen label in the literature, “relative operating characteristic,” being less domain specific, has been proposed as a more appropriate name.

7. End user knowledge regarding actual input population probabilities for their environment may vary. As pointed out later, such information may be either important or confounding.

8. The example in Section 6 presents a case with two physician's offices. One was a general practitioner, where patients tested for RA had an rCS of 0.01 (RA$_-/RA_+$). The other was a rheumatologist. In that office, patients tested for RA had an rCS of 100 (RA$_+/RA_-$).
9. MCC is often touted as being rCS-invariant. This, however, is only true in a special case. This and related implications are discussed in Section 5.

10. IC is also considered to be rCS-invariant. As with MCC, this is only true in a special case.

11. Or their complements, the probability that desired observations are incorrectly identified and the probability that desired observations are mistakenly labeled as undesired.

12. A Google scholar search for “Matthews correlation coefficient” turned up well over one thousand articles. The publications cited are but a small sample.

13. Using the results shown in Figures 5(a) and 5(b) as an example, if a test was run on a sample with rCS = 2\(^8\) on raw JPTs, the MCC = 0.33 and \(B^* = 1.1\). Recalculating MCC for the normalized JPT observed at rCS = 2\(^8\) and \(B^* = 1.1\), results in MCC < 0.69. However, the actual peak is MCC > 0.85.

14. Independence is a highly overloaded term. In this context, it means that any change to \(T\), will not affect \(T\).

15. DOR and DP are seen frequently in medical studies. In this problem domain, the inappropriate boundary risk may not always be present. The risk would exist in a study of heart attacks versus cholesterol levels; cholesterol level is a continuous variable. However, in a study of heart attacks versus family history, family history could be binary (a close relative died/did not die). In this type of test, boundary sensitivity is not an issue; care must be taken, however, in test design. Just by changing the test to a count (“how many close relatives died/did not die,” for instance) causes the problem to reappear.

16. DP and DOR are measured on different scales than the other summary statistics. In order to facilitate comparison, DOR was converted from an “odds ratio” type measure (bounded by \([0, \infty]\)) to a “probability” type measure (bounded by \([0, 1]\)). The relation between the two forms is

\[
\text{probability measure} = 1 - \frac{1}{\text{odds measure} + 1}. \quad \text{(D.5)}
\]

17. There is a similar measurement suite the “detection error tradeoff” (DET) [56]. DET plots the missed detection rate instead of the correct detection rate on the y axis. Since the two values are each other’s complement, comments herein regarding ROC apply equally to DET. Interestingly, DET is plotted using log scales. This is a real challenge for measures with a lower bound of zero.

18. Since all of the inherently rCS-invariant measures studies have [TPR, FPR] as measurement suites, the ROC curve could be presented for each of them as well.

19. We noticed a similar effect on the measure’s values. The values started becoming overly optimistic (once again, excepting DOR, the values of which dropped). The cause turned out to be a well-known issue with absolute sample size. The effect became significant when class A’s size fell below 400 elements.

20. Pdf standardization is confounding when evaluating a problem where rCS is important; thus, standardization is not appropriate for all problems. The set of all CPD problems is greater than the set of problems where pdf standardization is useful. Likewise, the set of all pdf comparisons includes problems with other than overlapping (or potentially overlapping) probability distribution functions, hence, the CPD problem domain intersects with the pdf comparison domain, but is neither a superset nor a subset.

21. This expression does not require that rCS = 1 initially. With the exception of \(Y\) or \(Y\) equaling zero, any JPT can be transformed (tuned) from one rCS to another.

22. There may be a solution to this deficiency; we will investigate this in future work.

23. As nicely summarized by [57], meta-analysis is a statistical technique for combining the findings from independent studies. Meta-analysis is most often used to assess the clinical effectiveness of healthcare interventions; it does this by combining data from two or more randomized control trials. Meta-analysis of trials provides a precise estimate of treatment effect, giving due weight to the size of the different studies included. The validity of the meta-analysis depends on the quality of the systematic review on which it is based. Good meta-analyses aim for complete coverage of all relevant studies, look for the presence of heterogeneity, and explore the robustness of the main findings using sensitivity analysis.

24. Anti-CCP refers to an assay using cyclic citrullinated peptide (CCP) to detect the anti-CCP antibody.

25. RF is an initialism referring to rheumatoid factor, an antibody used as a marker for RA.

26. On a single tailed test as used here, only one bound is relevant; thus, the bound indicates a 97.5% confidence.

27. In a CPD setting where rCS is important, JPT tuning enables a capability previously unavailable: sensitivity analysis. For the practitioner, this means that CPD performance can be evaluated for the expected rCS. Moreover, CPD performance can be identified over the rCS range the practitioner might expect.

28. The implications of the difference between TAR and \(T\), F-score will be addressed in future work.

29. All of the other measures are bound. In order to facilitate comparison, DOR was transformed from an “odds” format to the equivalent “probability” format.

References


