A COHERENT CLASSIFIER/PREDICTION/DIAGNOSTIC PROBLEM FRAMEWORK AND RELEVANT SUMMARY STATISTICS[1]

by

E. Earl Eiland

Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

New Mexico Institute of Mining and Technology
Socorro, New Mexico
August, 2017

[1]Portions of this dissertation have been previously published in Open Access journals. The publisher's Terms of Service are included in Appendix G.

Executing a doctoral program and dissertation requires two things: fortitude to begin the journey and persistence to carry through to the end. Were it not for my parents and grandparents raising me to follow my passion and not fear the unknown, I would not have stepped on this path.

Likewise, my wife graciously endured untold hours of my being (mentally and/or physically) "in my cave". Without her patience and encouragement, I might never have completed this.

Truly, these are the giants upon whose shoulders I stood to achieve this goal. My heartfelt gratitude and love goes to each and every one of them.

E. Earl Eiland
New Mexico Institute of Mining and Technology
August, 2017

It is better to design performance measures according to what one actually wants in the environment than according to one thinks the agent should behave.

Stuart Russell and Peter Norvig

ABSTRACT

Classification is a ubiquitous decision activity. Regardless of whether it is predicting the future, e.g., a weather forecast, determining an existing state, e.g., a medical diagnosis, or some other activity, classifier outputs drive future actions. Because of their importance, classifier research and development is an active field.

Regardless of whether one is a classifier developer or an end user, evaluating and comparing classifier output quality is important. Intuitively, classifier evaluation may seem simple, however, it is not. There is a plethora of classifier summary statistics and new summary statistics seem to surface regularly. Summary statistic users appear not to be satisfied with the existing summary statistics. For end users, many existing summary statistics do not provide actionable information. This dissertation addresses the end user's quandary.

The work consists of four parts:

1. Considering eight summary statistics with regard to their purpose (what questions do they quantitatively answer) and efficacy (as defined by measurement theory).

2. Characterizing the classification problem from the end user's perspective and identifying four axioms for end user efficacious classifier evaluation summary statistics.

3. Applying the axia and measurement theory to evaluate eight summary statistics and create two compliant (end user efficacious) summary statistics.

4. Using the compliant summary statistics to show the actionable information they generate.

By applying the recommendations in this dissertation, both end users and researchers benefit. Researchers have summary statistic selection and classifier evaluation protocols that generate the most usable information. End users can also generate information that facilitates tool selection and optimal deployment, if classifier test reports provide the necessary information.

Keywords: classifier evaluation; efficacious summary statistic axioms; summary statistics selection; end user efficacy

ACKNOWLEDGMENTS

The value of this work was greatly improved by input some very patient people. First and foremost is my research adviser, Lorie M. Liebrock, with whom I had innumerable discussions. I am also grateful for the valuable comments received from my research committee. During the formative stage of this study, Lynda Ballou, Mathematics Department, New Mexico Institute of Mining and Technology, Socorro, New Mexico and Andrew Barnes, Statistics Laboratory, General Electric Global Research, Niskayuna, New York, both graciously listened to and critiqued my musings. Their insights helped define a fruitful path.

To adapt a well known adage,
"It takes a village to raise a Doctor of Philosophy".

This dissertation was typeset with LaTeX[2] by the author.

[2]The LaTeX document preparation system was developed by Leslie Lamport as a special version of Donald Knuth's TeX program for computer typesetting. TeX is a trademark of the American Mathematical Society. The LaTeX macro package for the New Mexico Institute of Mining and Technology dissertation format was written for the Tech Computer Center by John W. Shipman.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LEXICON

$\boldsymbol{A}$: The source population from which Y was drawn.

$\overline{\boldsymbol{A}}$: The source population from which $\overline{Y}$ was drawn.

$\boldsymbol{B}$: Boundary. B^* is the optimum boundary.

Bias : A systematic difference that favors one group over another or weights one group more than another. For multiplicative summary statistics, $\iota_{T_+} = \iota_{F_+} = \iota_{F_-} = \iota_{T_-}$ defines unbiased impacts. For additive summary statistics, $|\iota_{T_+}| = |\iota_{F_+}| = |\iota_{F_-}| = |\iota_{T_-}|$ defines unbiased impacts.

Boundary : The classification tool input descriptor values used to partition S into Z and $\overline{Z}$.

CI : Confidence interval, the range within which a measurement will occur some specified percent of the time (X% CI).

Class imbalance : see relative class size.

Confounding factor : also called a confounding variable; an extraneous experimental input, for which if not controlled or accounted, will skew test results.

Confusion matrix : see joint probability table

Contingency table : see joint probability table

CPD : Classification, prediction, diagnosis; three common uses for classifiers.

Dependent variable : also called the response variable; a test output. Generally used along with "Independent variable".

Error matrix : see joint probability table

Expectation : In this context, a statistical value. This study's evaluations are non-parametric, so use median as a centrality measure. End users defining problem specific I will select the approach appropriate for their target data.

Explanatory variable : an experimental input which effects test output. Generally used along with "Response variable".

$\boldsymbol{F_+}$: False positive, class $\overline{A}$ events incorrectly flagged as class A. (Sometimes called a Type I error.)

$\boldsymbol{F_-}$: False negative, class A events incorrectly flagged as class $\overline{A}$. (Sometimes called as Type II error.)

Frequency table : see joint probability table

$\boldsymbol{\iota_{T_+}} = \sum_{s \in T_+} \iota_s / t_+$,

$\boldsymbol{\iota_{F_+}} = \sum_{s \in F_+} \iota_s / f_+$,

$\boldsymbol{\iota_{F_-}} = \sum_{s \in F_-} \iota_s / f_-$,

$\boldsymbol{\iota_{T_-}} = \sum_{s \in T_-} \iota_s / t_-$,

$\boldsymbol{\iota_Z} = \sum_{s \in Z} \iota_s / |Z| = \iota_{T_+} t_+ / |Z| + \iota_{F_+} f_+ / |Z|$,

$\boldsymbol{\iota_{\overline{Z}}} = \sum_{s \in \overline{Z}} \iota_s / |\overline{Z}| = \iota_{F_-} f_- / |\overline{Z}| + \iota_{T_-} t_- / |\overline{Z}|$,

$\boldsymbol{\iota_Y} = \sum_{s \in Y} \iota_s / |Y| = \iota_{T_+} t_+ / |Y| + \iota_{F_-} f_- / |Y|$ and

$\boldsymbol{\iota_{\overline{Y}}} = \sum_{s \in \overline{Y}} \iota_s / |\overline{Y}| = \iota_{F_+} f_+ / |\overline{Y}| + \iota_{T_-} t_- / |\overline{Y}|$.

$\boldsymbol{I}$: $I = (\iota_{T_+}, \iota_{F_+}, \iota_{F_-}, \iota_{T_-})$

Impact : A measure of CPD effect on a system's output.

Independent variable : see Explanatory variable; Generally used with "Dependent variable".

Joint probability table : A table presenting the cardinalities T_+, F_+, F_- and T_-. This is further described in Section 0.1 and illustrated in Table 1. JPT category cardinalities are lower case: $t_+ = |T_+|$, $t_- = |T_-|$, $f_+ = |F_+|$ and $f_- = |F_-|$.

JPT : joint probability table

pdf : probability distribution function

Proportion : see relative class size

rCS : relative class size

RA : rheumatoid arthritis

Relative class size : $ratio_+ = |Y| / |\overline{Y}|$ [42]

Independent variable : see Explanatory variable. Generally use along with "Dependent variable".

$\boldsymbol{S}$: The uncategorized data set.

Summary statistic : *summary statistic* $= f(I, JPT(B))$. Because multiple values are summarized into a single value, information is lost. To the extent essential information is retained, a summary statistic is useful. A key characteristic of summary statistics is when they are plotted against the boundary, they are not monotonic; they have optima. In this study's context, useful summary statistic optima indicate overall classifier utility for the boundary or boundaries and performance parameters at which this optimum utility was observed.

SS : summary statistic

$\boldsymbol{T_+}$: Correctly identified events in class A, the "class of interest" (if such a class exists).

$\boldsymbol{T_-}$: Correctly identified events of class $\overline{A}$, the other class.

$\boldsymbol{Y}$: Actual class A events in the data set.

$\boldsymbol{\overline{Y}}$: Actual class $\overline{A}$ events in the data set.

$\boldsymbol{Z}$: Events flagged as class A.

$\boldsymbol{\overline{Z}}$: Events flagged as class $\overline{A}$.

0.1 CPD input data

Although this paper applies well established stochastic concepts, not all discussions use the same terminology. To avoid confusion, we define our lexicon for test set categories:

		Actual classification (ground truth)		
		Y	$\overline{Y}$	*Totals* $\downarrow$
Test Result	$+ : s_i \in \{Z\}$	t_+	f_+	$\lvert Z\rvert = t_+ + f_+$
	$- : s_i \notin \{Z\}$	f_-	t_-	$\lvert\overline{Z}\rvert = f_- + t_-$
	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 are often organized into a joint probability table, such as shown here.

Frequently, these counts are presented as proportions of $|S|$ as shown in Table 2. The differences between Tables 1 and 2 are that the cell entries in Table 1 are integers, with the total of all four categories equaling $|S|$, whereas the JPT category proportions shown in Table 2 are rational numbers that sum up to one. Additionally, the proportional values represent the probability that for a given

Actual classification				
		Y	$\overline{Y}$	*Totals* $\downarrow$
Test Result	$+ : s_i \in \{Z\}$	$\frac{t_+}{\|S\|}$	$\frac{f_+}{\|S\|}$	$\frac{t_+ + f_+}{\|S\|} = \frac{\|Z\|}{\|S\|}$
	$- : s_i \notin \{Z\} \equiv s_i \in \{\overline{Z}\}$	$\frac{f_-}{\|S\|}$	$\frac{t_-}{\|S\|}$	$\frac{f_- + t_-}{\|S\|} = \frac{\|\overline{Z}\|}{\|S\|}$
	Proportional totals	$\frac{\|Y\|}{\|S\|}$	$\frac{\|\overline{Y}\|}{\|S\|}$	1

Table 2: The values in this JPT are proportions.

relative class size ($ratio_+ = \overline{Y}/Y$)), any randomly selected CPD output will be a member of that particular JPT category.

The CPD under test, configured with boundary vector (B, a "surface" that partitions the problem space), bins S into Z and $\overline{Z}$. The JPT(B) bin counts are snapshots of classifier labeling versus ground truth at B.

Although for supervised tests, the data set ground truth is known, it only represents the source population. Statistical methods are used to account for variation between the data sets and the source population. Confidence intervals indicate the quality of the resulting source population characterization. The nature of the source population is characterized by the class pdfs and their associated $ratio_+$.

0.2 Contour graphs

I use graphs to assess measure response to selected problem domain characteristics. A contour graph is used to show how measure output (contour lines on the graph, representing the Z-axis) is affected by $ratio_+$ (Y-axis) and boundary (B) (X-axis). This graph also shows the optimum boundary (B^*) versus $ratio_+$ as a dotted line. Figure 1 illustrates the graph type.

Another contour graph shows how the measure value versus boundary varies as class separation improves. Figure 2 is an example. The X-axis indicates the ordered sample index at which the measure was calculated. Decreasing the distribution overlap causes the true boundary ranges to change, so true boundary values are confounding. Index values provide an easy means of normalizing the test ranges. The Y-axis indicates the measure value recorded.

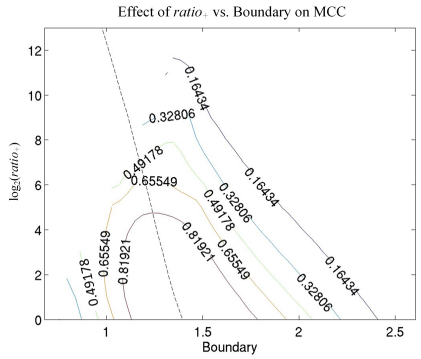


Figure 1: This contour graph shows how a measure is affected by boundary and relative class size ($ratio_+$). The dotted line indicates the optimum boundary (B^*). (This graph is for the Matthews Correlation Coefficient [*MCC*]).

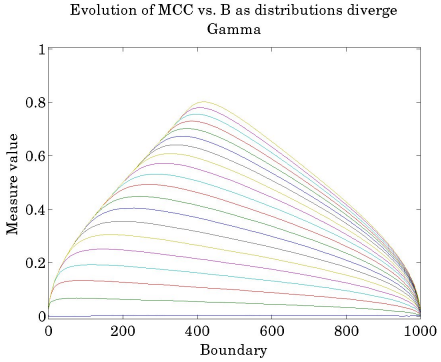


Figure 2: As class overlap decreases, classification improves. Measure responses, however vary. This graph shows one such test result. The bottom-most line represents the measure results for two identical gamma distributions that fully overlap. As the distributions diverge, causing the overlap to shrink, the test data classification improves. The MCC measure indicates this with increasing values. In this particular test, the best classification is skewed to the left side of each curve.

PREFACE

How did I end up studying "how to evaluate classifiers?" One might have expected work on a next generation cyber security tool, advanced memory management or the like. The most abstract answer might be a refusal to accept the status quo. My Master's thesis research proposed and tested a novel anomaly detection algorithm for cyber attack detection. Creating and executing the algorithm went smoothly. The challenge, was measuring classifier quality.

My literature review revealed many opinions regarding classifier assessment. Furthermore, none of the summary statistics seemed to quantify effect on an end user. I found no compelling reason for classifier summary statistic selection in the literature, but I did need to be able to compare my algorithm to others. As a matter of expedience, I selected one of the most common summary statistics, receiver operating characteristic area under the curve (*ROC-AUC*).

That might have ended my interest in classifier assessment, had I not faced the problem again at General Electric Global Research (GEGR). My first project was maturing my anomaly detector as an deployable cyber security intrusion detector. Once again, I had to revert to the default summary statistic, *ROC-AUC*. Now, however, I more strongly missed the end user's relevance. Compounding the issue, my original dissertation topic was going to examine potential enhancements to my thesis and results at GEGR. The lack of an efficacious end user summary statistic was like a stone in my shoe. Unless I radically changed my field of study, the assessment problem still loomed.

Finding an end user efficacious classifier evaluation summary statistic became a concurrent research project. As I learned about classifier assessment, or encountered a new summary statistic, I revisited my search. Many research fields specified class imbalance (aka relative class size [$ratio_+$]) invariance, but did not agree on a summary statistic. My efforts to define a $ratio_+$ invariant end user efficacious classifier evaluation summary statistic were not successful. My musings, however, did lead to a "EUREKA" moment – while considering $ratio_+$ invariance, I realized that some summary statistics were inherently $ratio_+$ invariant, but $ratio_+$ invariance is not a summary statistic characteristic. It can be conferred on any summary statistic using joint probability table (JPT) values by normalizing the JPT! Now, any JPT-based summary statistic that generated actionable information for end users was valid. The constraint that seemed to make the problem insoluble had been removed. With success seemingly on the horizon and the result applicable across a wide and diverse problem domain range, I convinced my research adviser to shift dissertation topics.

The rest, as the saying goes, is history.

This dissertation is accepted on behalf of the faculty of the Institute by the following committee:

Lorie M. Liebrock, Academic and Research Advisor

Scott Evans, Member

Subhashish Mazumdar, Member

Dongwan Shin, Member

Andrew Sung, Member

I release this document to the New Mexico Institute of Mining and Technology.

E. Earl Eiland Date

CHAPTER 1

INTRODUCTION

Some problem domains are analyzed using classifier output. Often, class affiliation is not the end result. Rather it drives subsequent activities. Examples include medical diagnoses determining the presence or absence of disease; information retrieval gathering relevant material; information system activity being flagged as malicious; determining insurance or credit terms etc. The list is virtually endless. (Because of classification's ubiquity, I use the initialism *CPD* for classification, prediction and diagnosis.) Incorrect classification can lead to frustration, financial loss and even death. Correct classification is important, hence, a number of CPD algorithms have been developed and the field remains active.

Characterizing CPD tool effectiveness, then, is important. For example, CPD tool developers need to know how their particular modification affects CPD performance; end users want actionable information upon which to choose between CPD options and to optimally deploy a tool in the field. The term "performance" can refer to a variety of operational parameters, including time factors (e. g., output latency), computation complexity and memory requirements. This research defines performance and effectiveness as characters of CPD output quality. The goal is to generate "actionable information", informing users on how a CPD tool will affect their target outcome, relative to the CPD's operational settings. I hypothesize that actionable information is generated by a subset of possible performance summary statistics.

At first glance, quantifying classifier effectiveness might seem intuitive; however, the consensus appears to be that classifier evaluation and comparison is actually difficult. If paper publication rate is an indicator, then summary statistic selection is still an open issue [6, 10, 21, 26, 35, 41, 78, 79]. There is a seemingly steady stream of publications characterizing summary statistics. Additionally, papers proposing new summary statistics appear regularly in the literature.

Of the myriad of proposed summary statistics, only a few have gained traction. Perhaps the most commonly seen summary statistics are *Total Accuracy Rate*, the *Receiver Operating Characteristic Area Under the Curve*, F_β*-score*, *Youden Index*, two related summary statistics, *Diagnostic Odds Ratio* and *Diagnostic Power*, *Mathews Correlation Coefficient* and *Mutual Information Coefficient*. These eight summary statistics will be used as the foundation for summary statistic evaluation. Summary statistic descriptions are in Chapter 5.

1.1 CPD end users are underserved stakeholders

All of these summary statistics condense multiple views of CPD quality into one value. However, because multiple values are summarized into a single value, information is lost. To the extent essential information is retained, a summary statistic can prove useful for CPD evaluation. A key characteristic of summary statistics is when they are plotted against the boundary, they are not monotonic; they have optima[1]. Useful summary statistic optima indicate overall classifier utility for the boundary or boundaries and performance parameters at which this optimum utility was observed. Ideally, these summary statistics also quantify some efficacious aspect of classifier output for end users. Efficacious summary statistic values enable end users to directly estimate the CPD's effect on their situation.

The disorganized state of CPD summary statistic selection is not new. Sokolova et al. comments "...the [summary statistics] in use now do not fully meet the needs of learning problems in which the classes are *equally important* and where *several algorithms are compared*" [78]. It seems reasonable that, if a researcher is not comfortable with the known summary statistics, that he/she may be inclined to develop one that suits.

End users want to know how a specific CPD tool will impact their problem. An informative summary statistic for any stakeholder must be sensitive to relevant problem domain characteristics and insensitive to irrelevant (confounding) characteristics. CPD tool stakeholders can be partitioned into three groups:

- **Basic researchers** focus on developing new CPD algorithms. This group expects that an effective new CPD algorithm will be useful in many problem domains, so their evaluations need to be application agnostic; specific problem domain characteristics are, in fact, confounding. Examples of basic researchers in the CPD context are the persons that introduced CPD techniques such as k nearest neighbor [16], neural net [60] and support vector machine [90].
- **Applied researchers** use CPD algorithms on specific problem domains to create tools useful for that domain; specific problem domain characteristics are important. Examples of tools incorporating CPD algorithms are anomaly-based intrusion detectors for cyber security [74, 80] and document classifiers for enterprise information retrieval systems [2]. In addition to independent applications, another applied research output is reusable programming language libraries (code libraries). Such libraries can be deployed by end users to create custom solutions. In this context, the focus is on the CPD tool; data sets are used to develop and test the tool.

[1] A common summary statistic, the receiver operating characteristic area under the curve, is boundary invariant. Since it neither increases nor decreases, it is monotonic. Since there is only a single value, it is also the optimum.

- **Practitioners** deploy CPD tools to solve specific problems in their domain. Domain-specific characteristics are important, as well as operational aspects such as impact sensitivity to class boundary settings (, the CPD tool setting that determines to which class each observation is allocated). Code libraries may be a frequently used method for CPD tool usage in other fields of study such as medicine [64], molecular biology [15, 61], finance [1], etc. In this context, the emphasis is the opposite of the other two groups; the CPD tool is used to evaluate the data, rather than the data used to evaluate the tool. Hence, investigators applying CPD methods are end users.

- **End users** encompass both practitioners and applied researchers, as defined above.

I consider what summary statistic characteristics are informative for end users. The researcher definitions align with more general definitions published by the National Science Board [62].

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 . . .' [40].

In the broader field of artificial intelligence, Russell and Norvig express a similar sentiment:

> As a general rule, it is better to design performance measures according to what one actually wants in the environment, rather than according to how one thinks the agents should behave [71].

In published studies read, specific proposed summary statistics may be mapped to specific problem domains. However, identifying a general means by which end users can quantify CPD effectiveness in their particular setting has not been addressed. 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 [23].

Seemingly, the research community has viewed the variety of end user needs as too complex and diverse 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. To the extent that research studies present CPD performance information by which end users can estimate CPD impact (how the CPD affects the end user) in their situation, the studies provide improved service to the end user.

In reviewing the literature, generally, new summary statistics are introduced in studies requiring a CPD, rather than in studies in which CPD evaluation is the focus. Investigators must allocate project assets to create the necessary deliverables, hence, limited assets are available to address summary statistic efficacy. In my literature review, I found no common understanding of what constitutes a good CPD evaluation summary statistic; this seemed glaring in its absence. I encountered no basis for, or foundation upon which to develop CPD evaluation summary statistics. In short, I found no "good summary statistic" criteria for CPD evaluation. One of the key contributions of this dissertation is establishing four axioms for end user efficacious CPD summary statistics.

1.2 End user summary statistic efficacy considerations

My interest is addressing end user interests. Regarding CPD selection and deployment, end users have three questions:

1) What is the CPD's impact on my problem? To answer this question, I consider what the summary statistic quantifies and how that relates to end users.

2) What is the boundary that provides the optimum impact? Clearly, only a boundary (B) sensitive summary statistic can provide this information. It is possible a summary statistic that does not quantify CPD impact may still share an optimum boundary (B^*) with one that does.

3) How sensitive is the impact to boundary selection? Clearly, only a boundary sensitive summary statistic can provide this information. Any two summary statistics with common optimum boundaries and boundary sensitivities are likely answering the same question, but (possibly) with different units of measure. This study will look for such occurrences.

Measurement theory provides additional insight on summary statistic efficacy: numbers are used in different ways. These uses constrain their information content and hence, their utility. I use the scale-type definitions proposed by Stevens [83]. Stevens defined four scale types, nominal, ordinal, interval and ratio. Ratio scales have the least functional constraints, so summary statistics using ratio scales are the most information rich. Ratio scales have two unique and readily identifiable characteristics:

They have meaningful zeros. In this dissertation's context, a meaningful zero is relative to the unit measured and the end user's problem. For example, if dollars profit is being measured, then zero means there is no profit. If energy produced is being measured, then zero means no energy was produced.

They have a "standard unit". Abstractly, this means that $m_x + 1 = m_{x+1}$. One implication of having a standard unit is that there is no upper bound and the lower bound can be either zero or negative infinity. In this context, a CPD could potentially have a negative impact on an end user, so the most widely useful summary statistic's range must be $(-\infty, \infty)$.

Reflecting on the end user interests, only a ratio scale summary statistic will satisfy question 1 above. A recurring topic in studies using CPDs is class imbalance ($ratio_+$). Accordingly, I evaluate summary statistics regarding i) the three questions posed above, ii) scale type and iii) sensitivity to the factors $ratio_+$ and pdf.

"Extending the body of knowledge", must cover new intellectual territory, validating other's work is not sufficient. Hence, a literature review is necessary; Chapter 2 covers this. The research protocol is described in Chapter 3. Understanding and clearly stating the problem must come prior to proposing a solution. The results of this process is covered in two chapters. Chapter 4 defines the summary statistic evaluation framework and identifies the necessary dependent variables. the Chapter 5 reviews the state-of-the-art and ends with a gap analysis and problem statement. The CPD problem domain is characterized in Chapter 6. Before a solution can be crafted, success criteria must be defined. To this end, Chapter 7 reviews the summary statistic criteria seen in other study fields. Those insights lead to defining four axioms for end user efficacious CPD summary statistics in Chapter 8. Two axiom compliant summary statistics are developed in Chapter 9. Using technical readiness levels, the summary statistics are mapped to prospective stakeholders in Chapter F. Examples of the enhanced efficacy of axiom compliant summary statistics are in Chapter 10. User study results are reported in Chapter 11. Conclusions and future work are in Chapter 12. The dissertation closes with References and Appendices.

CHAPTER 2

RELATED WORK

As noted in Chapter 1, this study focuses on CPD evaluation summary statistics that address end user interests. What constitutes related work? Certainly not the myriad of research projects that use CPD summary statistics. It is surprising how many projects "tweak" established summary statistics, potentially making them relevant. However, these enhanced summary statistics almost never gain a following. Some have; the summary statistics evaluated in Chapter 5 were selected based on their acceptance as "de facto" standards in some problem domains. Peer acceptance was the selection criterion for inclusion in this study.

This study is rare in that first the CPD problem was defined, then relevant summary statistic characteristics were defined. Only after these problem constraints were established was any actual summary statistic investigation executed. Of the prior work reviewed, only two investigators applied a similar strategy:

- Swets, noting a similarity between signal detection problems and medical diagnoses, suggested the receiver operating characteristic *ROC-AUC* be applied [84, 85, 86]. It has since been applied to many CPD problem types.
- Van Rijsbergen proposed a summary statistic that is the complement of F_β-*score* [88]. Van Rijisbergen applied measurement theory to information retrieval system evaluation, taking as his inputs select performance criteria put forth by Cleverdon [14]. The summary statistic is now used in many problem domains.

The *ROC-AUC* and F_β-*score* may be the most broadly accepted CPD summary statistics I saw. However, neither can be transformed into values which end users can use to predict actual impact in their situation.

The majority of the works on CPD evaluation were based on analyzing existing summary statistics and either inferring some insight that could reduce the confusion, or demonstrate the value of some proposed summary statistic characteristic. Examples include:

- Sokolova, et al., propose that for rare events, either T_+ or T_- is most important, and summary statistics should be selected based on their *Convergence* characteristic, relative to their sensitivity to the important class [77]. Their

work does not address whether or not a summary statistic provides actionable information; their work, although similar, addresses another issue.

- Martens, et al., propose two characteristics for a model underlying a CPD. Comprehensibility relates to the ease with which an end user can understand the model, justifiability relates to how well the model aligns with existing domain knowledge [57]. Their work does not address whether or not a summary statistic provides actionable information; their work, although similar, addresses another issue.
- Parker [68] compares seven binary classifier summary statistics, rating them on their degree of agreement. His hypothesis was that in the specific cases where all summary statistics but one agree on the best classifier, then the one disagreeing must be in error. The study focused on classifier selection quality, but was silent on a summary statistic's other actionable information.
- Jamain and Hand executed a meta-analysis of classification performance. After a rigorous selection process, six studies were considered qualified for a meta-analysis. Their three analytical methods, analysis of linearly scaled test results, logistic regression analysis and linear regression analysis were inconclusive. The study suggested that this might have been due to the confounding effects of unspecified test data characteristics. The authors suggested that significantly more test data characterization might have allowed compensation for test data effects. A secondary conclusion was that classifier evaluations may exhibit a tendency to selective reporting. The study had no hard recommendations. However, the authors did note that a wise selection strategy might be for performance which is "consistently reasonably good and never very bad", rather than seeking optimization. They further note that "the real question a user wants to answer is 'which classification method is best for *me* to use on *my* problem with *my* data' " [40]. Jamain and Hand's work affirms the value of this study and suggests two characteristics for efficacious summary statistics for end users. This work leverages Jamain and Hand's insights.
- Seliya, et al., compare twenty-two summary statistics for correlation. They ultimately identified three unspecified factors by which they binned the summary statistics. Their recommendation was to use one summary statistic for each factor, thereby giving a more complete view of CPD performance [75]. Their work does not address whether or not a summary statistic provides actionable information; their work, although similar, addresses another issue.
- Baldi, et al., evaluate nine summary statistics, ultimately concluding that the Matthews correlation coefficient and Mutual information summary statistics are most balanced, based on their observation that they both use all four JPT values [3]. Their work does not address whether or not a summary statistic provides actionable information; their work, although similar, addresses another issue.

- Sokolova, et al., evaluate eight summary statistics, proposing that summary statistics other than *TAR*, *ROC-AUC* and F_β*-score* are better for assessing CPD ability to distinguish classes [78]. Their work does not address whether or not a summary statistic provides actionable information. Their work, although similar, addresses another issue.

- Caruna and Niculesdu-Mizil compared nine summary statistics applied to test results from fourteen thousand classifier-data set pairings. Each summary statistic value generated represented a dimension, with each summary statistic described by a fourteen thousand dimension vector. summary statistics were then compared based on euclidean distances between the vectors. Multidimensional scaling indicated that the vectors could be partitioned into three to five dimension groups. However, the report did not hypothesize what the partitions represented. The study also considered pairwise summary statistic correlations. This study did not consider end user interests [10].

- Lavesson, et al., propose a multi-faceted machine learning evaluation protocol when evaluating machine learning protocols for application-specific suitability. One factor, "Accuracy", refers to the quality of the machine learning component's output. The work, however is silent on the actual summary statistic selected and the selection process [49].

- Oreški and Oreški investigate analysis with unbalanced data sets. They conclude that summary statistic selection is critical. Researchers can effectively skew their results toward different conclusions by their summary statistic selection. No means to mitigate this risk is provided. [65].

- Bifet, et al., [4] examine the challenges of stream classification. Their discussion addresses managing imbalanced data sets, but does not map summary statistic selection to end user needs.

None of the studies reviewed specifically focused on end user efficacy: this was a gap addressed by my doctoral research.

2.1 CPD summary statistics evaluated

Based on the literature review, eight summary statistics that appeared to be well, but not universally, accepted were selected for in-depth analysis. A brief description of each follows. Each summary statistic is more fully addressed in Chapter 5.

Diagnostic Odds Ratio (DOR) [32]
DOR is defined as

$$DOR = \frac{\frac{T_+}{F_-}}{\frac{F_+}{T_-}},$$

where $\frac{T_+}{F_-}$ =*True Positive Odds (TPO)* and $\frac{F_+}{T_-}$ =*False Positive Odds (FPO)*. After simplification,

$$DOR = \frac{T_+ T_-}{F_+ F_-}.$$

Discriminant Power (DP) [5]

Discriminant power is defined as

$$DP = \frac{\sqrt{3}}{\pi}(logX + logW),$$

where

$$X = \frac{sensitivity}{1 - sensitivity} = \frac{T_+}{F_-} \text{ and } Y = \frac{specificity}{1 - specificity} = \frac{T_-}{F_+}.$$

Recasting the equation, we get

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

DOR/DP Refers to both the diagnostic odds ratio and discriminant power. *DOR/DP* are related summary statistics:

$$DP = \frac{\sqrt{3}}{\pi} log(DOR).$$

F-score [88]

F_β-score is defined as:

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

where β is the relative weight of precision and recall:

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

IC [70]

Mutual Information Coefficient is the mutual information (I) contained in ground truth regarding the test set S ($Y \cup \overline{Y}$) and the CPD prediction of ground truth, as contained in $Z \cup \overline{Z}$. Normalized by the entropy in ground truth (H):

$$IC = \frac{I(Y \cup \overline{Y}, Z \cup \overline{Z})}{H(Y \cup \overline{Y})}.$$

Expressing I and H in terms of JPT categories,

$$I(Y \cup \overline{Y}, Z \cup \overline{Z}) = -H\left(\frac{T_+}{N}, \frac{F_+}{N}, \frac{F_-}{N}, \frac{T_-}{N}\right)$$
$$-\frac{T_+}{N}\log(|Y| * |Z|) - \frac{F_+}{N}\log(|\overline{Y}| * |Z|) - \frac{F_-}{N}\log(|Y| * |\overline{Z}|) - \frac{T_+}{N}\log(|\overline{Y}| * |\overline{Z}|), \tag{2.1}$$

where

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

J [97]

Youden Index has a number of expressions. 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

$$J = sensitivity + specificity - 1, \text{ where}$$

$$sensitivity = \frac{T_+}{Y} \text{ and } specificity = \frac{T_-}{\overline{Y}}.$$

Further, *sensitivity* is also known as the *true positive rate* (TPR) and *specificity* is the complement of the *false positive rate*, $specificity = 1 - FPR = 1 - \frac{F_+}{\overline{Y}}$. Hence an even simpler (thus better, according to the Minimum Description Length principal) definition would be

$$J = TPR - FPR.$$

MCC [59]

Mathews Correlation Coefficient, in the form commonly seen today:

$$MCC = \frac{(T_+ * T_-) - (F_+ * F_-)}{\sqrt{Y * \overline{Y} * Z * \overline{Z}}}$$

ROC-AUC [84, 85, 86]

Receiver Operating Characteristic Area Under the Curve. The title originates from the fact that it is the area under a "ROC curve', a curve defined by false positive ($FPR = \frac{F_+}{\overline{M}}$) and true positive ($TPR = \frac{T_+}{M}$) rates. These values are calculated from JPTs of classifier output for a number of thresholds across the observed range, then graphed as the ROC curve.

TAR

Total Accuracy Rate is an intuitive summary statistic. It has been in use so long, its origin is no longer cited:

$$TAR = \frac{T_+ + T_-}{T_+ + T_- + F_+ + F_-}$$

These summary statistics were selected because they were either de facto standards in particular fields of study, or based on anecdotal evidence, appeared to be gaining in popularity.

CHAPTER 3

RESEARCH PROTOCOL

As noted in Chapter 1, investigator sentiment has been that efficacious end user CPD evaluation is an intractable problem. This belief is given credence by the myriad summary statistics seen in the literature, often introduced with comments on frustrations with existing summary statistics. Clearly, an ad hoc approach has failed to successfully resolve the problem.

My study employed *The Theory of Inventive Problem Solving* (TRIZ), a process first codified in the 1940s by Genrich Altshuller. The field has evolved significantly since Altshuller's first work; a text on the process has been published recently by Gordon Cameron [7]. A thorough discussion of the field is out of scope, but the four basic requirements are:

- **Identify the problem's root cause.** Insights into the root cause primarily constituted invalidating commonly held beliefs. This emerged from the summary statistic characterization study reported in Chapter 5.
- **Identify the sufficient explanatory variables.** Use of the four JPT categories is well established. However, the role played by JPT category impact resulted from the summary statistic characterization study reported in Chapter 5 and an exercise working through use cases to precisely define what constituted actionable information for end users.
- **Identify constraints on the solution.** The use case analysis provide some insight, its results are reported in Chapter 1.2. A study of measurement theory was provided key insights [36, 69, 83].
- **Define success criteria for the solution.** The good summary statistic study reported in Chapter 7 and the use case analysis provided the major insights for this component.

Once these four requirements were met, a satisfactory solution to this dissertation's problem emerged. The balance of the study consisted of verifying that the proposed summary statistics satisfied the four axioms and validating their utility on real-world applications.

CHAPTER 4

SUMMARY STATISTIC CHARACTERIZATION

Characterizing the summary statistics required a protocol that was scientifically sound (applied appropriate statistics and tests were repeatable) and that tested attribute relevance to end users.

4.1 Scientific soundness

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

As a class, CPD problems can have any number of categories greater than one. In order to avoid confounding variables, this study is limited to binary CPD problems. Extension to n-airy problems is left for future work.

CPD evaluation studies can be partitioned into two groups: those that use "real-world" data and those that use simulated data. Characterizing CPD evaluation summary statistics requires observing how the summary statistics respond as CPD output varies. Real-world data, such as those available in repositories, e.g., the UCI Machine Learning Repository, provide the opportunity to test against a wide variety of complex data types [30]. However, observing the effect of incremental changes on real world data is difficult at best. For this purpose, I use simulated CPD output in Chapters 6 and 9. Results using the simulated data were validated in Chapter 10 using real world applications, with real world data.

4.2 Test framework

I use a supervised test to ascertain CPD quality. This test framework provides the ability to compare "ground truth" to CPD output. Abstractly described,

test set elements interact with the defined process (in this context, a CPD). 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 $\overline{Y}$ membership. The probability that a randomly selected output will be detected as a member of class $\overline{Y}$ is the subsequent probability ($P_{subsequent}$). $P_{subsequent}$ 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 [$P_{leading} = |Y|/|S|$] or an odds ratio [$ratio_{+} = |Y|/|\overline{Y}|$]) and the defined process. The defined process contributes its own uncertainty (P_{event}) to the observed output. Thus the test system can be described by the equation $P_{subsequent} = f(P_{leading}, P_{event})$. The CPD test set model is illustrated in Figure 4.1.

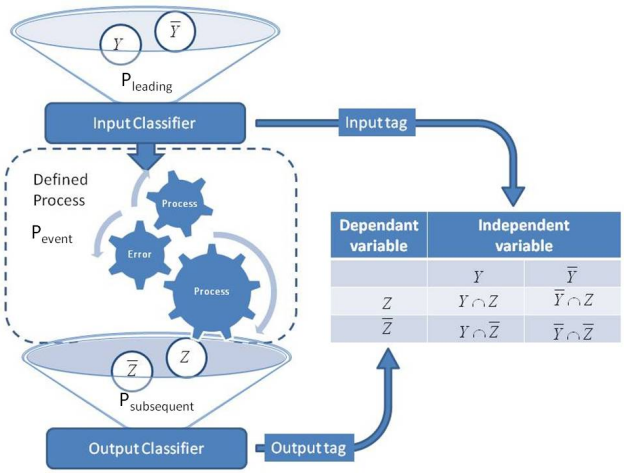


Figure 4.1: The test system "ground truth" inputs have a specific mix, representing the underlying probability for the system ($P_{leading}$). The test system outputs have a specific mix ($P_{subsequent}$), representing the interaction of the defined process and the inputs. The defined process contribution to the uncertainty observed in the output is represented by P_{event}. Often, the results are presented in JPTs.

Tying the test set model to the examples in Chapter 11, $P_{subsequent}$ consists of the patient's rheumatoid arthritis (RA) diagnosis, the bank's loan funding decision and the stream of intrusion detector classifications. P_{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_{event} for the bank's loan decision is subject to the appropriateness of the model, the variables used and the quality of the information received. P_{event} for the intrusion detection examples 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/not malicious determination.

In the CPD test system described, $\hat{P}_{leading}$ is a characteristic of the input

test data set, hence, it is always fixed[1,2]. It is, in fact, related to $ratio_+$:

$$\hat{P}_{leading} = 1 - \frac{ratio_+}{1 + ratio_+}. \tag{4.1}$$

Test system inputs are also a function of class probability distribution functions (pdf) in the source population. This study considers the effects of both $ratio_+$ and pdf on summary statistic output.

4.3 $ratio_+$ sensitivity

The $ratio_+$ sensitivity tests used normally distributed classes ($N(\overline{m}, \sigma^2)$; $\overline{m}$ is the distribution mean and σ is the standard deviation). For all but *IC*, the analyses were based on four hundred data sets consisting of two hundred thousand randomly drawn observations from two source populations; $Positive = N(1.0, 0.04)$ and $Negative = N(2.0, 0.04)$. Separate tests were run with data sets having $ratio_+$s of

$$2^0 : 1,\ \ 2^1 : 1,\ \ 2^2 : 1,\ \ \cdots,\ \ 2^{13} : 1.$$

A total of 5,600 independent data sets were used in this study. Due to *IC*'s computational complexity, its analysis was based on four hundred data sets consisting of twenty thousand randomly drawn observations from two source populations. For each summary statistic evaluated, I observed how the reported metric was affected by $ratio_+$ vs. boundary vs. metric output. The 3-D results are presented as contour plots. Because the summary statistic values are asymptotic to one (thus non-linear), I 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 3-D data, but on the 2-D graphs in 6.1.3, the ninety percent confidence interval (90% CI) is displayed for each $ratio_+$ vs. peak summary statistic value graph. To illustrate, the 90% CI is indicated by the vertical lines at each $ratio_+$ tested in Figure 4.2; the plotted line indicates the median.

4.4 pdf sensitivity

Pdf sensitivity was based on ten different source population probability distribution functions: uniform, normal, extreme value, Gamma, Cauchy and Beta. Taking advantage of Beta's configurability, I generate five pdf shapes. Figure 4.3 shows the pdfs used. pdf selection was intended to provide distributions

[1] $\hat{P}$ "P hat" indicates the actual data set. When not "hatted", P refers to the source population. The difference between the two is that the data set is always known and only represents the source population. Generally, I assume data sets accurately represent their source populations. Confidence intervals indicate the quality of that assumption.

[2] Such concrete knowledge is unlikely in a field deployment.

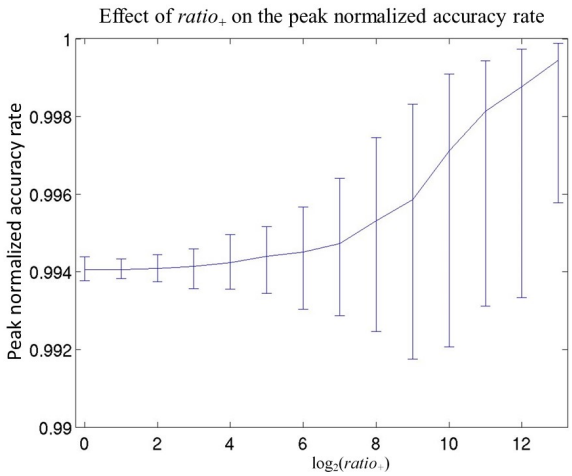


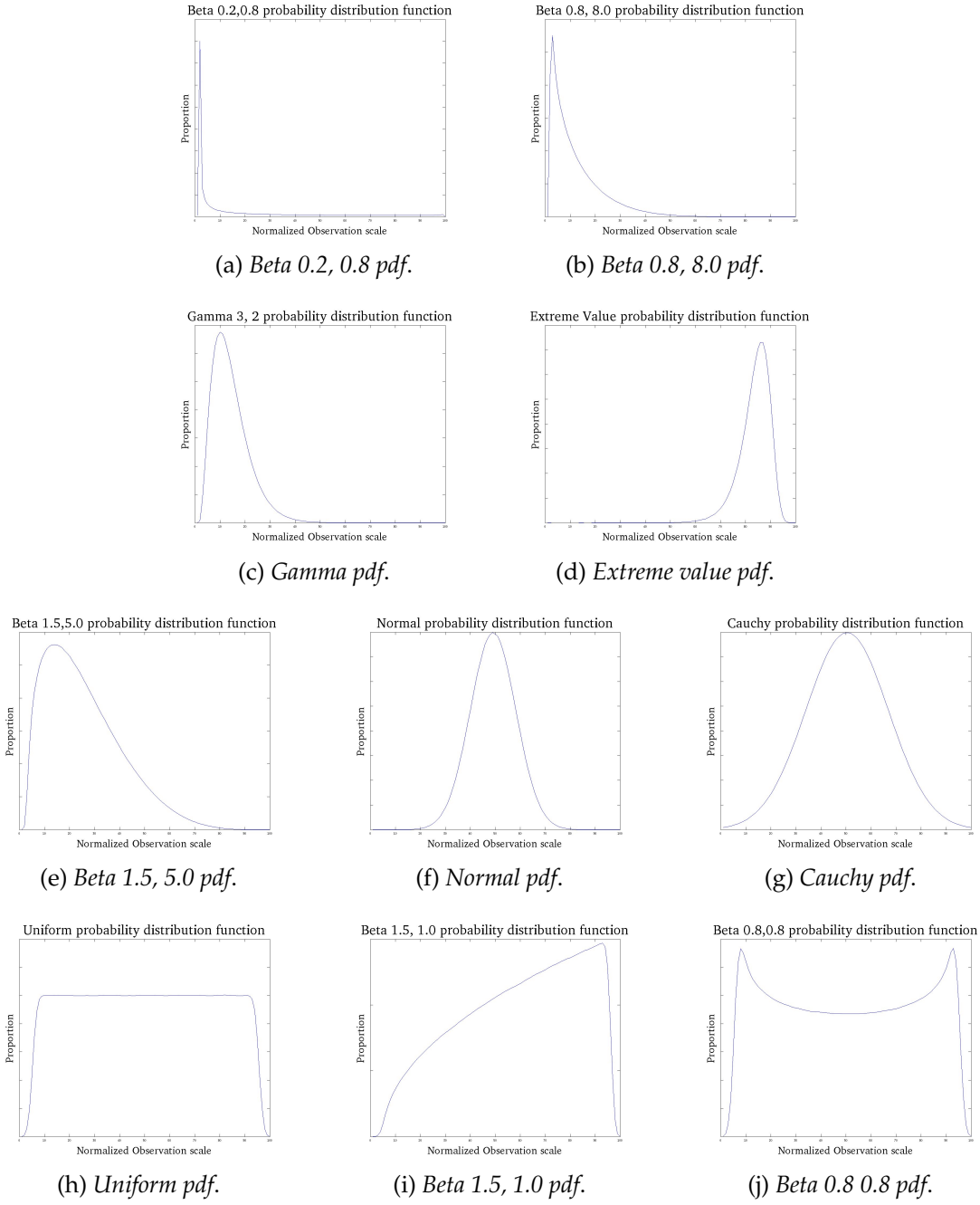
Figure 4.2: The vertical bars in this graph indicate the 90% confidence interval for measurements at each point observed. The plotted line is the median.

with a wide range of skewness and kurtosis. (Thereby, mitigating the risk of coming to conclusions that are not generally applicable.) By definition, the area under pdf curves equal one. The X and Y axis values are artifacts of the probability density estimator used (MATLAB ksdensity.m) and have no bearing on this research.

Pdf sensitivity testing used Monte Carlo evaluation of two identical source distributions, varying only by their means (for the Cauchy distribution, I varied the median). To accommodate *IC*'s computational complexity, data sets consisted of twenty thousand randomly drawn observations; repetitions were kept at four hundred. To avoid class imbalance effects, the test set $ratio_+$ was kept at one. Using each summary statistic, I calculated values for each boundary threshold. Summary statistic output versus boundary was then graphed. Boundary ranges varied between distributions tested, as well as by the extent of the test distributions overlaps. In order to avoid potential difficulties comparing summary statistics, the boundary ranges were normalized.

Basing conclusions on a test which only considers a test series with one amount of overlap between Y and $\overline{Y}$ could overlook important trends. Hence, tests series included class degrees of overlap ranging from full overlap to nearly full separation.

This protocol provides the flexibility and repeatability necessary for analysis, abides by the constraints necessary for analysis of less tractable problem domains with difficult problem environments (e.g., complex CPD input and output distributions) and considers the environmental factors to which end users must contend.



(a) *Beta 0.2, 0.8 pdf.* (b) *Beta 0.8, 8.0 pdf.*

(c) *Gamma pdf.* (d) *Extreme value pdf.*

(e) *Beta 1.5, 5.0 pdf.* (f) *Normal pdf.* (g) *Cauchy pdf.*

(h) *Uniform pdf.* (i) *Beta 1.5, 1.0 pdf.* (j) *Beta 0.8 0.8 pdf.*

Figure 4.3: Graphs of the ten pdfs used for sensitivity testing. In this context, only the distribution shapes are important: the tests included distributions with a wide range of skewness and kurtosis. By definition, the areas under pdf curves equal one; the X and Y axis values are artifacts of the probability density estimator used (MATLAB ksdensity.m) and have no bearing on this research.

4.5 Summary statistic output comparison

Test results from the pdf sensitivity analysis is rich, but the numerous graphs makes comparison between summary statistics difficult. Accordingly, I use summary graphs which show the summary statistic outputs on the y-axis and quantify the degree of class difference on the x-axis. It turns out that *TAR* is a useful measure of class difference:

$$\textit{degree of class overlap } = \frac{f_+ + f_-}{|S|}, \tag{4.2}$$

where 0.5 indicates complete overlap (the classes are totally indistinguishable) and 1.0 indicates the classes are completely separate. From the JPT definition,

$$f_+ + f_- = |S| - (t_+ + t_-). \tag{4.3}$$

Substituting equation 4.3 into 4.2, and a bit of algebra,

$$\textit{degree of class overlap } = 1 - \frac{t_+ + t_-}{|S|}. \tag{4.4}$$

Noting that

$$TAR = \frac{t_+ + t_-}{|S|}, \tag{4.5}$$

substituting equation 4.5 into Equation 4.4 and re-arranging terms,

$$TAR = 1 - \textit{degree of class overlap.} \tag{4.6}$$

Hence, *TAR* is the complement of degree of class overlap, therefore useful to quantify class difference. Each line on the pdf sensitivity graphs represent the summary statistic output for a test run with a specific degree of class overlap; the peak values in each graph in Figure 5.10 provides an indicator for the test set's degree of overlap. Result distortions caused by class imbalance were avoided by using class sizes with $ratio_+ = 1$. Class separability was simulated by varying values in one class to create a bias between classes. Each line on the graph represents tests with one bias.

CHAPTER 5

THE CURRENT STATE - OF - THE - ART

Dissertations must "contribute to the body of knowledge". Thus, it is essential to know the problem domain's current status. Once the status is known, gaps can be identified. The requirement to contribute to the body of knowledge can thus be satisfied with a research project designed to address a gap.

Summary statistics are intended to provide actionable information for specific stakeholder actions. As noted in Chapter 1, my focus is end user actions. While investigating the current state-of-the-art, three motivations emerged as the drivers resulting in the eight common summary statistics reviewed. This chapter organizes the summary statistics based on their root motivation.

Taking the criteria identified in Chapter 1.2 as providing actionable information for end users, each evaluation considers what actionable information a specific summary statistic provides for end users.

5.1 Summary statistic motivations

When making evaluations, stakeholders require actionable information. Summary statistics are a means of providing that information. Measurement theory, a topic not seen much outside the social sciences, defines two of these information types. Measurement theory distinguishes between two entity attribute (in this context, summary statistic) types, *intrinsic* summary statistics; those that are part of an entity's definition (for example, density or mass) and *extrinsic* summary statistics; those that are expressions of the entity's interaction with the environment (weight, for example) [36]. When reported in joint probability tables (JPT), classifier output is partitioned into four distinct categories, T_+, F_+, F_- and T_-. After any data set has been tested, the final object count in each category is influenced by the environmental factors relative class size (also known as class imbalance; quantified herein as $ratio_+ = \overline{Y}/Y$.), class probability distribution functions (pdf) and boundary used (B, the boundary that determines to which class each observation is allocated). Thus T_+, F_+, F_- and T_- are extrinsic summary statistics. Six of the eight summary statistics used herein address the perceived needs defined by measurement theory; controlling CPD's extrinsic factors; and quantifying CPD's intrinsic qualities. Two of the summary statistics address the third

perceived need, quantifying an end user's performance criteria. The summary statistic discussions in this chapter are organized by stakeholder context.

5.1.1 Mitigating class imbalance ($ratio_+$)

One challenge researchers face is comparing their CPD results to results reported by others. One of the major difficulties is not all test sets have the same characteristics. Test sets used may well have different relative class sizes, which can cloud results. As an example, I ran a classifier on two test sets drawn from the same class populations. Since both the classes and classifier were the same, one would expect statistically indistinguishable output. However, since JPT categories are extrinsic, the anticipated similarity may be masked. Table 5.1 illustrates the issue.

$ratio_+$	F_+	F_-
1	125	250
9	25	450

Table 5.1: This table shows the total F_+ and F_- for two tests with the same classifier on equally sized data sets (2250 observations), drawn from the same populations. The only difference is the samples have different $ratio_+$. Because the tests were run on data sets with different sized classes, the equivalence of the classifier's effectiveness is not apparent. Without prior knowledge that the same classifier was used, an observer could conclude the classifiers were significantly different.

The JPT values shown in Table 5.1 are output from one classifier on two equally sized, two-class test sets. Both test sets were drawn from the same two class populations. The only difference is one test set has a class divergence of $1:9$ and the other a class divergence of $1:1$. The CPD tool performs equally well in each test; however, the outcome's dependence on $ratio_+$ masks the CPD performance equality in the JPTs. When class sizes are equal ($ratio_+ = 1$, there are twice as many F_- observations than F_+. However, when $ratio_+ = 9$, here are eighteen times as many F_- observations than F_+! Without knowledge of the test sets used, an observer could well conclude that these were two significantly different CPD tools.

There are three approaches to mitigating class imbalance.

Balance class sizes by either decreasing the size of the dominant class (undersampling) or increasing the size of the lesser class (oversampling). The approach is appealing, but can be a source of error. Oversampling can distort a class pdf and undersampling can increase uncertainty. He and Ma present a nice discussion of these issues in "Imbalanced Learning"[39].

Use a common data set. Natural language studies often test on common corpora. For more general use, there are University of California, Irvine's Machine Learning Repository and the KDD repository [30, 63]. This approach is effective, but in dynamic problem domains, the data sets can cause results to lose relevance rapidly. CPD test results are test set sensitive. Regardless of the stakeholder being a researcher, developer or end user, this sensitivity must be taken into account.

Use $ratio_+$ invariant summary statistics. These summary statistics avoid the risks associated with creating balanced test sets and allow results to be based on relevant test sets. This dissertation is about summary statistic efficacy; the challenges of this approach is the focus of this chapter section (5.1.1).

Gary M. Weiss, author of Chapter Two in "Imbalanced Learning"[39], concludes that the underlying issue is lack of knowledge about the lesser class, particularly regarding rare events. This deficiency is endemic to the problem domain; the effectiveness of all class imbalance mitigation strategies are limited as a result.

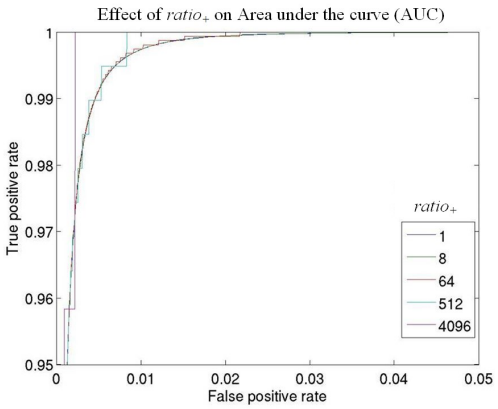
Receiver operating characteristic area under the curve (ROC-AUC) The receiver operating characteristic has a solid history. Swets campaigned diligently to establish it as the evaluation criterion of choice [84, 85, 86]. The $\{TPR, FPR\}$ summary statistic suite is the basis for the *ROC-AUC* summary statistic. The title originates from the fact that it is the area under a "ROC curve', a curve defined by false positive ($FPR = F_+/\overline{Y}$) and true positive ($TPR = T_+/Y$) 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 [27, 43][1]. *ROC-AUC* is measured in the interval $[0.5, 1.0]$.

In contrast to the other summary statistics reviewed, *ROC-AUC* is generally accompanied by the *ROC* curve. (Indeed, the *ROC* curve may be presented without providing *ROC-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 *ROC-AUC* summary statistic[2]. (This is, of course, true for any summary statistic suite, since consolidation of multiple values into a single summary statistic value means that information is lost.)

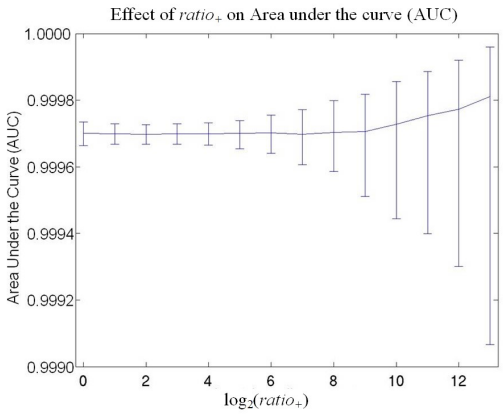
There are numerous ROC-AUC variants [54, 89]. Vanderlooy and Hüllermeier determined in their comparison, that despite intuitive appeal, none of the variants confer any CPD selection improvement. From the end user perspective,

[1] There is a similar summary statistic the "Detection Error Tradeoff" (DET) [58]. 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 summary statistics with a lower bound of zero.

[2] Since all of the inherently $ratio_+$ invariant summary statistics studies have $\{TPR, FPR\}$ as summary statistic suites, the *ROC* curve could be presented for each of them as well.



(a) As class sizes diverge, the *ROC* curve inflection point moves closer and closer to (0,1), thus driving the *ROC-AUC* toward 1. This is significant; a classifier with a *ROC* curve that runs linearly from (0,0) to (1,1) (*ROC-AUC* $= 0.5$) has an accuracy rate equivalent to random selection and a classifier represented by a *ROC* curve that passes through (0,1) (*ROC-AUC* $=$ 1.0) is perfect. Thus as the *ROC* curve inflection point approaches (0,1), classifiers appear more accurate. *ROC* curves exhibit an unexpected sensitivity to class divergence. Hence, the *ROC-AUC* may overstate a classifier's accuracy under certain conditions.



(b) Probability theory, specifically, the strong law of large numbers, predicts that normalization will increase uncertainty. This is certainly true for the *ROC-AUC*. The 90% confidence intervals in this graph (the vertical bars at each observation point) of the *ROC-AUC* vs $ratio_+$ shows how the *ROC* summary statistic becomes increasingly uncertain as $ratio_+$ increases. The graph also shows how the *ROC* starts to overstate classifier quality (the *ROC-AUC* increases).

Figure 5.1: These figures provide two views on how *ROC-AUC* is affected by class divergence. Historically, the *ROC* (and summary statistic *ROC-AUC*) have been a standard classifier assessment protocol. This figure shows that class divergence may cause *ROC-AUC* to overly optimistically report classification effectiveness, while simultaneously incurring a substantial increase in the confidence interval.

since the underlying summary statistic units remain the same, they all have the same limited efficacy.

The *ROC-AUC* has been criticized on more theoretical terms recently [37, 38, 52]. David J. Hand, the main proponent of the issue states

> "...when interpreted in terms of a balance of the relative costs of the two kinds of misclassification, the AUC is incoherent in the sense that it requires that the relative costs of the two kinds of misclassification differ from classifier to classifier [38]."

However, because it is boundary and $ratio_+$ invariant, the *ROC-AUC* is a valued tool in this research environment. Figure 5.1a illustrate how the summary statistic is $ratio_+$ invariant: it's values change very little. Figure 5.1b shows that *AUC*'s $ratio_+$ invariance does not defy the strong law of large numbers; below $ratio_+ = 2^6$,

the *ROC-AUC* exhibits invariance, however, above 2^6 ($|\overline{Y}| < 400$), the median *ROC-AUC* tended to overstate classifier quality in my tests[3]. I also observed that the 90% CI increased markedly as class divergence increased[4]. *ROC-AUC* is boundary invariant, so it cannot be used to identify the optimum boundary or boundary sensitivity.

The *ROC-AUC* has value for researchers. How does *ROC-AUC* fare for end user efficacy?

What question does the summary statistic quantify? The *ROC-AUC* has been defined as the probability that, given one each, randomly chosen elements of class Y and class $\overline{Y}$, a classifier will rank the class Y instance higher than class $\overline{Y}$ one (assuming Y ranks higher than $\overline{Y}$) [27]. From the end user's perspective, there are two difficulties: i) the probability has nothing to do with actual classification and ii) the observer needs a priori knowledge of the instance's ground truth. If the end user already knows ground truth, then the CPD is unnecessary.

Is the summary statistic measured on a ratio scale? *ROC-AUC* is bounded within the interval $[0.5, 1.0]$, so it does not have a meaningful zero, nor does it have a standard interval. *ROC-AUC* is measured on an ordinal scale; impact is out of scope.

Does the summary statistic exhibit boundary sensitivity? *ROC-AUC* is boundary invariant, so it can't be used to identify the optimum boundary.

For end users, *ROC-AUC* has vanishingly little value.

Youden index (J) The Youden index (traditionally represented by J) was proposed in 1950 and is seen in medical diagnostic studies [97]. There are a number of expressions of J. The original is

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

[3]A test series wherein I varied $|S|$ showed that $ratio_+$ the observed invariance failure was a function of $|\overline{X}|$, not $ratio_+$.

[4]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 under-represented. Since process accuracy is inversely related to class overlap, a reduction in estimated class overlap will result in process accuracy over-estimation. In my tests, the difference became statistically significant when sample sizes fell below four hundred members.

Perhaps a more common representation is

$$J = sensitivity + specificity - 1, \text{ where}$$

$$sensitivity = \frac{T_+}{Y} \text{ and } specificity = \frac{T_-}{\overline{Y}}.$$

Further, *sensitivity* is also known as the *true positive rate* (TPR) and *specificity* is the complement of the *false positive rate*, $specificity = 1 - FPR = 1 - F_+/\overline{Y}$. Hence an even simpler (thus better, according to the Minimum Description Length principal) definition would be

$$J = TPR - FPR. \tag{5.1}$$

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

J is special in that $J = 0$ indicates a CPD with an output equal to that of tossing a fair coin. $J = 1$ with a perfect CPD and $J = -1$ for a CPD that misclassifies everything. Practically, J is measured on the interval $[0, 1.0]$. As noted in their respective literature bases, J shares a characteristic with *ROC-AUC*, in that it is insensitive to $ratio_+$. This can be seen in Figure 5.2.

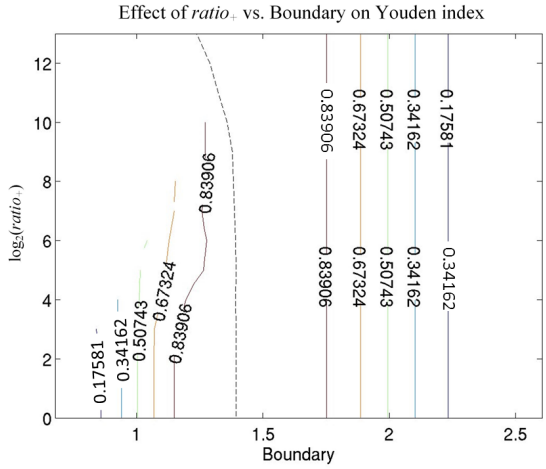
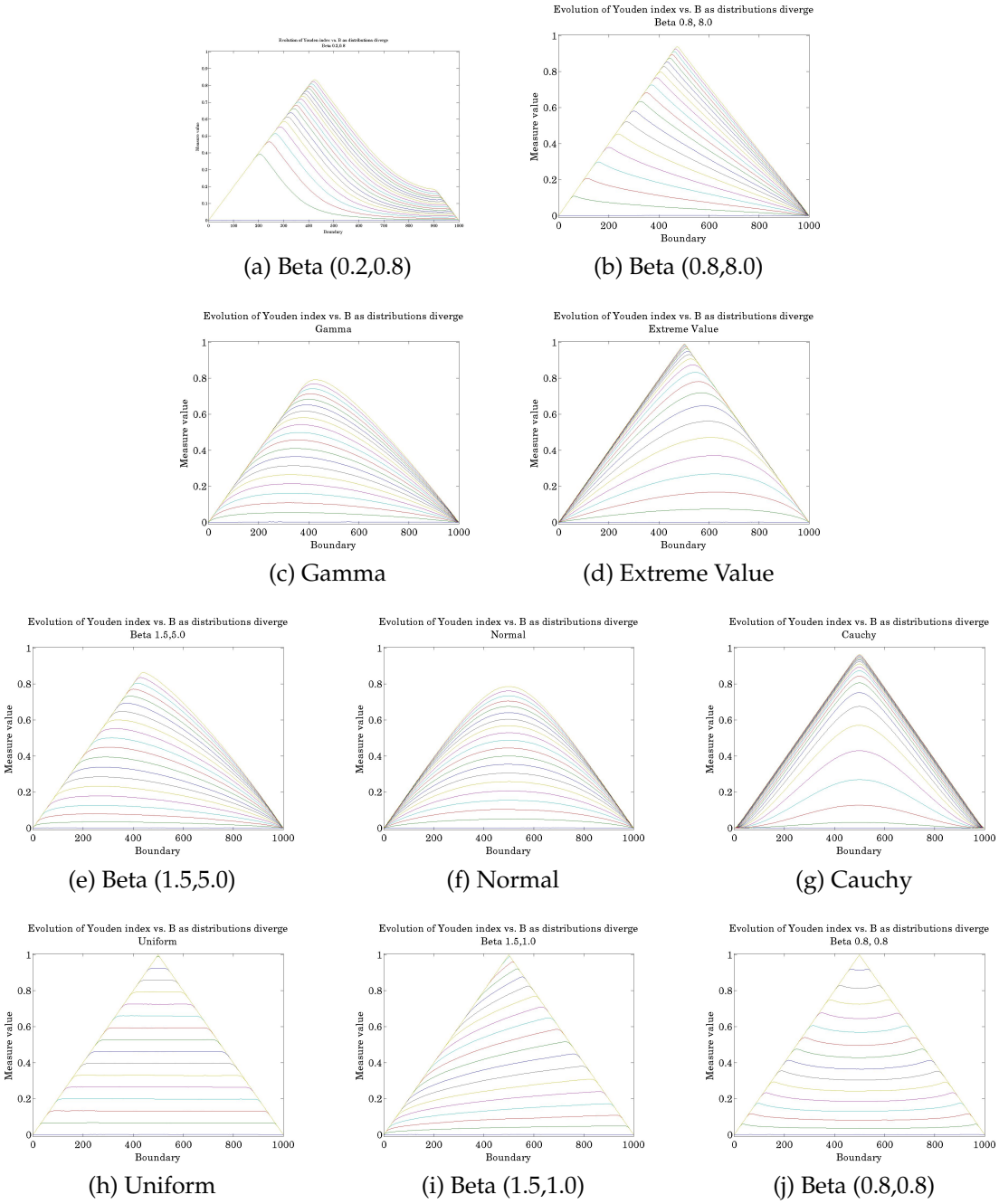


Figure 5.2: The Youden Index has a very uniform shape and the optimum boundary lies along the peak of the Youden Index ridge. This exhibits the expected $ratio_+$ invariance.

J's sensitivity to $ratio_+$ and pdf are presented in Figures 5.3 and 5.2. How does it fare for end user efficacy?

What question does the summary statistic quantify? J quantifies the spread between the TPR and FPR — this is not actionable information for end users.



(a) Beta (0.2,0.8)

(b) Beta (0.8,8.0)

(c) Gamma

(d) Extreme Value

(e) Beta (1.5,5.0)

(f) Normal

(g) Cauchy

(h) Uniform

(i) Beta (1.5,1.0)

(j) Beta (0.8,0.8)

Figure 5.3: The Youden index optimum boundary is sensitive to pdf, and its optimum boundary is consistent with other summary statistics so it provides actionable information to end users. J's optimum boundary sensitivity may not be actionable for end users.

Is the summary statistic measured on a ratio scale? J is measured on the interval $[-1, 1]$. It's zero may be meaningful in some cases, but not the general case, nor does it have a standard unit. The Youden index is measured on an ordinal scale; impact is out of scope.

Does the summary statistic exhibit boundary sensitivity? The Youden index exhibits boundary sensitivity. In fact, the sensitivity is equivalent to *TAR*, when $ratio_+ = 1$. This equivalence is discussed in section 5.1.3.

Diagnostic odds ratio and discriminant power (DOR/DP) Two related summary statistics are the diagnostic odds ratio (*DOR*) [32] and Discriminant Power(*DP*) [5]. Diagnostic odds ratio (*DOR*) is defined as

$$DOR = \frac{\frac{T_+}{F_-}}{\frac{F_+}{T_-}},$$

where $\frac{T_+}{F_-}$ =*True Positive Odds (TPO)* and $\frac{F_+}{T_-}$ =*False Positive Odds (FPO)*. After simplification,

$$DOR = \frac{T_+ T_-}{F_+ F_-}$$

Discriminant power(*DP*) is defined as

$$DP = \frac{\sqrt{3}}{\pi}(logX + logW),$$

where

$$X = \frac{sensitivity}{1 - sensitivity} \text{ and } Y = \frac{specificity}{1 - specificity}.$$

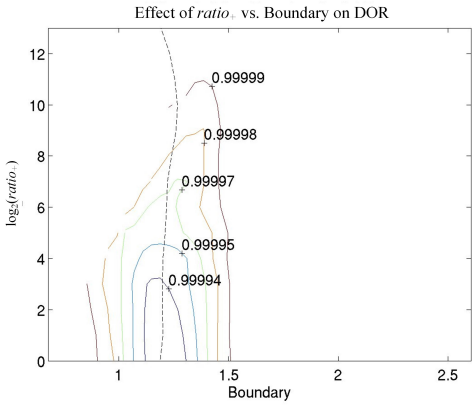
Recasting the equation yields

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

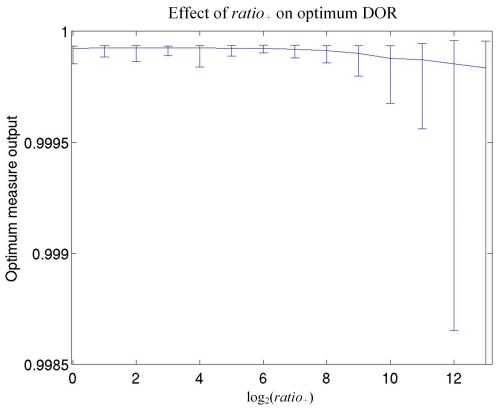
The derivation is provided in Appendix B. Comparing the two summary statistics,

$$DP = \frac{\sqrt{3}}{\pi} log(DOR).$$

DOR and *DP* are found in medical research. *DOR* is measured in the interval $[0.0, \infty]$, *DP* is measured in the interval $[-\infty, \infty]$. Interestingly, $DP = -\infty$ and $DOR = 0$ when *either* $T_+ = 0$ *or* $T_- = 0$, both need not equal zero. Similarly, $DP = \infty$ and $DOR = \infty$ when *either* $F_+ = 0$ *or* $F_- = 0$, both need not equal zero. Hence a CPD can classify some observations correctly (*Total Accuracy* > 0), yet

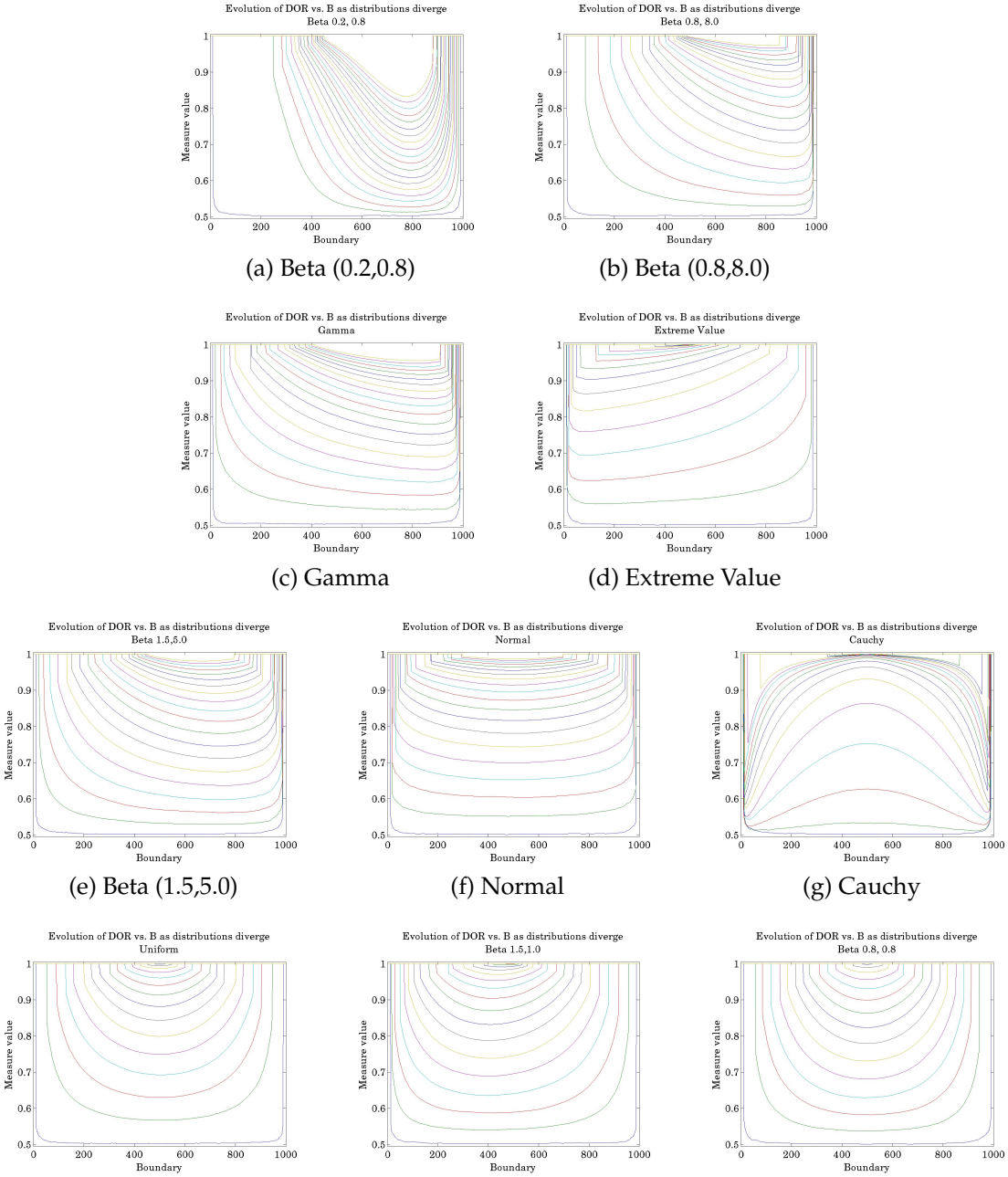


(a) Contour graph of $ratio_+$ vs. boundary vs. *DOR* value. Scaling makes the summary statistic seem somewhat $ratio_+$ sensitive. However, Graph 5.4b shows *DOR* is actually $ratio_+$ invariant.



(b) Graph of $ratio_+$ vs. *DOR*, with error bars.

Figure 5.4: Instead of the optimum value being maxima, like the other summary statistics evaluated, the optimum *DOR* value is a minimum. Hence, the contours show a valley instead of a ridge. Also contrary to the other summary statistics, *DOR* decreases when the absolute class size effect becomes noticeable. This means that the contours are closed, instead of open like the others. *DOR*'s vertical optimum boundary line and constant value (seen in Graph 5.4b) indicates *DOR*. (and hence, *DP*) is $ratio_+$ invariant. *DOR/DP* optimum boundaries (approx. 1.2) are offset from the optimum boundaries seen in the other $ratio_+$ invariant summary statistics (approx 1.4). *DP*, the log form of *DOR*, has the same characteristics as *DOR*.



(a) Beta (0.2,0.8) (b) Beta (0.8,8.0)

(c) Gamma (d) Extreme Value

(e) Beta (1.5,5.0) (f) Normal (g) Cauchy

(h) Uniform (i) Beta (1.5,1.0) (j) Beta (0.8,0.8)

Figure 5.5: Graphs showing *DOR*'s boundary sensitivity to pdf.

have $DP = -\infty$ and $DOR = 0$. This is counter intuitive, since one would expect $DP = -\infty$ and $DOR = 0$ to indicate a totally ineffective CPD and $DP = \infty$ and $DOR = \infty$ to indicate a perfect CPD, rather than something in-between. 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 X versus the rate it was observed in group $\overline{Y}$. 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 $\min(DOR)$ (or $\min(DP)$). Thus for any test run, the boundary with the *smallest* T_+T_- relative to F_+F_- gives the best accuracy. Not only is this counter intuitive, but also a potential error source. The problem originates from the fact that the greater the $\min(DOR)$ (or $\min(DP)$), the better the results. Thus, if the boundary used to partition the test output is not at the optimum boundary, the results may appear better than they really are[5]. Most observations regarding *DOR* apply to *DP* as well. For example, $DP = 0$ when $T_+T_- = F_+F_-$.

Another worrying characteristic is seen in Figure 5.5g. In the distributions tested, $B^*(DOR)$ were minima in all but the Cauchy. With the Cauchy distribution, *DOR* exhibited two minima. As an indicator of CPD quality, multiple optima are acceptable. However, the minima seen in Figure 5.5g do not seem to indicate CPD quality. Instead, the maximum near boundary point 500 seems to indicate CPD quality. This is a challenging situation. If a stakeholder is using *DOR* to compare two CPDs, care must be taken. Are the observed $B^*(DOR)$ maxima or minima? *DOR* and *DP*'s sensitivity to $ratio_+$ and pdf are presented in Figures 5.5 and 5.4.

One important characteristic of *DOR* and *DP* is that they are $ratio_+$ invariant. An important difference between *DOR/DP* and the other $ratio_+$ invariant summary statistics is that their optimum boundaries, although constant in my tests, are offset from the "minimum error boundary". These effects can be seen in Figure 5.4[6]. Since *DOR* and *DP* are minima, they follow a valley in the contour

[5] *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 vs. cholesterol levels; cholesterol level is a continuous variable. However, in a study of heart attacks vs. 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 re-appear.

[6] *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 summary statistic (bounded by $[0, \infty]$) to a "probability" type summary statistic (bounded by $[0, 1]$). The relation between the two forms is

$$\textit{probability measure} = 1 - \frac{1}{\textit{odds measure} + 1}.$$

graph, instead of a ridge, Also contrary to the other summary statistics, *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 summary statistics. *DOR*'s vertical optimum boundary line and constant value (seen in Graph 5.4b) indicates *DOR* (and hence, DP) are $ratio_+$ invariant.

DP's and *DOR*'s observed optimum boundary differs from the optimum boundary for the other $ratio_+$ invariant summary statistics tested. Because of this boundary bias, they may not be useful for selecting boundaries. For example, in my test environment, *TAR* at the common optimum boundary is 0.994, *TAR* at *DOR*'s optimum boundary is 0.958; the difference is significant at the 95% confidence level.

How do *DP* and *DOR* fare for end user efficacy?

What question does the summary statistic quantify? Since T_+ and T_- are (statistically) independent[7], (as are F_+ and F_-), *DP* and *DOR* could, in a probabilistic sense, be interpreted as quantifying the odds that, given two random observations, one will be classified T_+ and the other T_- rather than one being classified F_+ and the other F_-. As with *ROC-AUC* and *J*, this value does not seem actionable information. Thus *DOR* and *DP* seem to be relevant in niche CPD scenarios, but not to general CPD problem types.

Are the summary statistics measured on a ratio scale? Although *DOR* is measured in the interval $[0.0, \infty]$ and *DP* is measured in the interval $[-\infty, \infty]$, their zeros are not meaningful. *DOR* does have a standard unit; *DP*, being a log function, does not. *DP* and *DOR* are not measured on a rational scale, so they cannot quantify end user impact.

Do the summary statistics exhibit boundary sensitivity? These summary statistics are boundary sensitive. However, they are not internally consistent. Sometimes, min *DOR* and min *DP* are optima and (as seen in Figure 5.5g), maxima are optimum.

5.1.2 Measuring CPD's intrinsic characteristic

Ideally, a CPD's output matches ground truth perfectly. Rarely, however, are the two identical — the CPD model does not describe its input completely. Intuitively, the match between a CPD's output and ground truth has the ring of being a CPD's intrinsic characteristic. Two concepts, one from statistics (correlation), the other from information theory (mutual information), seem to quantify the agreement between ground truth and a CPD's model. Both have been applied to CPD problems.

---

[7] Independence is a highly overloaded term. In this context, it means that any change to T_+ will not affect T_-.

Matthews correlation coefficient (MCC) The Matthews Correlation Coefficient (MCC) was introduced by B. W. Matthews [59]. In a subsequent classifier summary statistic survey, Baldi, et al. restated the summary statistic in the form commonly seen today [3]:

$$MCC = \frac{(t_+ * t_-) - (f_+ * f_-)}{\sqrt{|Y| * |\overline{Y}| * |Z| * |\overline{Z}|}}. \tag{5.2}$$

MCC is the application of Pearson correlation coefficient to CPD evaluation and is a summary statistic of the similarity between ground truth and CPD output. *MCC* is measured across the interval $[-1, 1]$. When *MCC*$= 0$, the CPD is as effective as a fair coin; *MCC*$= 1$ when ground truth and the CPD output are in complete agreement. Consequently, *MCC*'s practical range is $[0, 1.0]$.

Actual target classification		Y	$\overline{Y}$	
Test Result	*Positive*	$\frac{T_+}{Y}$	$\frac{F_+}{\overline{Y}}$	
	Negative	$\frac{F_-}{Y}$	$\frac{T_-}{\overline{Y}}$	
Normalized totals		1	1	2

Table 5.2: The expressions in this JPT normalize the category values.

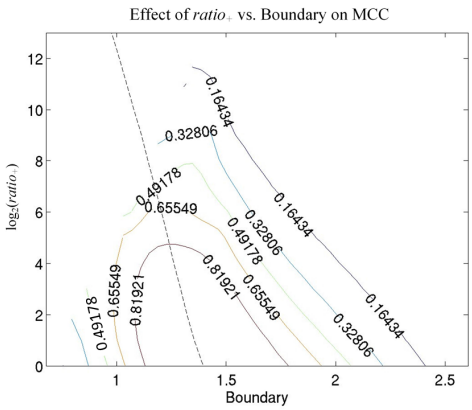
Although not mentioned explicitly in Matthews' work, Baldi et. al. notes that the equation requires normalized distributions:

$$\frac{a - \overline{a}}{\sigma_A}, \text{ where } a \in \{A\},$$

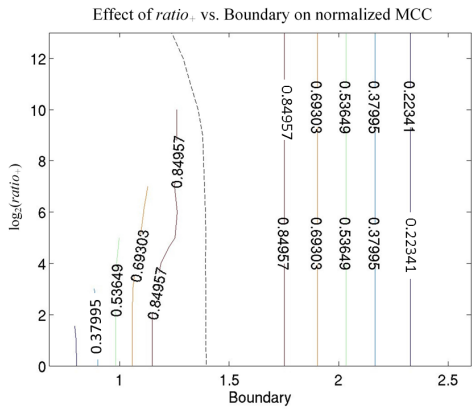
$A = \{a_1, a_2, a_3, \ldots, a_s\}$ is the class of the input data set (Y or $\overline{Y}$ in my lexicon) and $s = |A|$. $\overline{a}$ is the mean of A. $\overline{A}$ is treated in the same manner. One effect of distribution normalization is class size equalization: $ratio_+ = 1$ (This is discussed in Chapter 6). Table 5.2 shows the expressions used to generate normalized JPTs; class size equalization is indicated by the ones in the "normalized totals" row. To demonstrate Matthews initial intent to use normalized distributions, I 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 pre-requisite applied by Matthews, et al. and noted by Baldi, et al. seems to have been lost, although the belief that *MCC* is $ratio_+$ invariant persists [8, 9, 12, 20, 31, 44, 50][8]. As a consequence, Baldi, et al.'s equation

[8] A Google Scholar search for "Matthews correlation coefficient" turned up well over one thousand articles. The publications cited are but a small sample.



(a) MCC exhibits $ratio_+$ sensitivity on non-normalized JPTs.



(b) On normalized JPTs, *MCC* exhibits $ratio_+$ invariance. The increasing boundary curvature when $ratio_+ > 2^6$ is a JPT normalization artifact explained in Chapter 6.1.3.

Figure 5.6: If *MCC* inputs are not normalized, it is $ratio_+$ sensitive.

is sometimes applied without first normalizing the JPTs. Two of these reports, Cannon, et al. [8] and Mirceva, et al. [31] include JPTs. Upon recalculating their results, I determined that the values presented were based on non-normalized JPTs. In both cases, there were substantial differences between the results on normalized and non-normalized JPTs. In the Cannon, et al. results, the difference affected not only the values, but also the process rankings. Using normalized JPTs, the process ranked last by Cannon, et al., (*MOLPRINT*) 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 sub optimal process for use in real-world settings.

Figure 5.6 shows how normalization impacts the $ratio_+$ sensitivity for *MCC*. Graph 5.6a shows *MCC*'s response when the raw JPT values are used. The peak boundary (indicated on the graph by the dashed line) shifts as $ratio_+$ varies and the value decreases as $ratio_+$ increases (indicated by the sloping dashed line that intersects the contours). In contrast, Graph 5.6b 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 I surveyed did not identify whether or not the *MCC* values reported were on normalized JPTs. As seen in Figures 5.6a and 5.6b, comparing results across tests where $ratio_+$ is not normalized could lead to errors. Since $ratio_+$ 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 sub

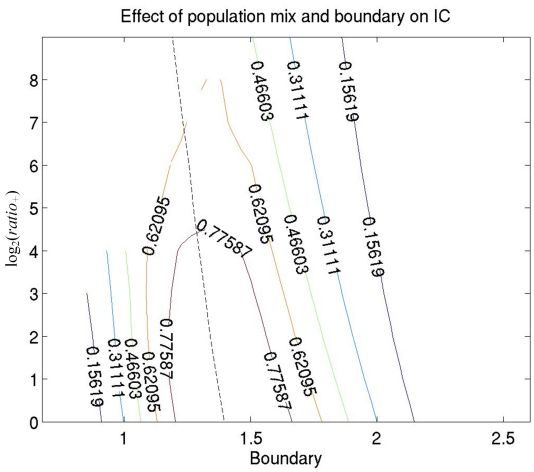


Figure 5.7: The sloped dotted line on the contour graph shows that *IC* is not $ratio_+$ invariant.

optimal boundary, thus the corrected *MCC* value will also be sub optimal[9]. In this section, I characterized *MCC* using non-normalized data. Appendix C derives a $ratio_+$ invariant *MCC* expression. *MCC*'s sensitivity to $ratio_+$ and pdf are presented in Figures 5.6 and 5.8.

Given that *MCC* does quantify CPD's intrinsic characteristic, when calculated on normalized JPTs, the summary statistic appears to be valuable for researchers. How does it fare for end user efficacy?

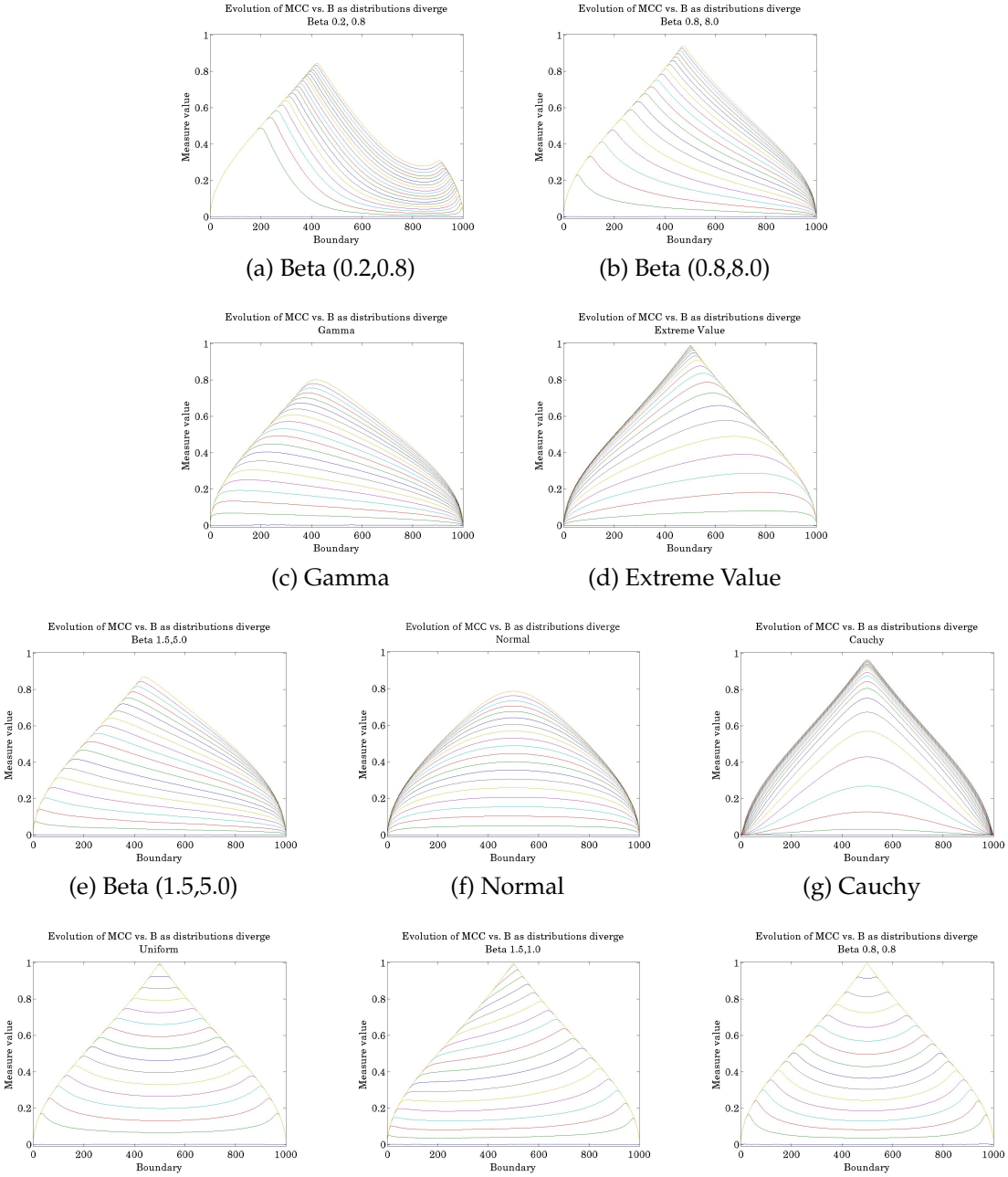
What question does the summary statistic quantify? *MCC* quantifies the similarity between ground truth and CPD output.

Is the summary statistic measured on a ratio scale? *MCC* does have a meaningful zero. However, since it is measured in the interval $[-1, 1]$, it is not a ratio scale summary statistic. *MCC* is measured on an ordinal scale; impact is out of scope.

Does the summary statistic exhibit boundary sensitivity? *MCC* is sensitive to both $ratio_+$ and pdf. Surprisingly, on the uniform distribution, it is bimodal. This characteristic is discussed in Chapter 9.4.

Mutual information coefficient (IC) Rost and Sander introduced an information theory-based summary statistic into the literature in 1993 [70]. It was subsequently included in a summary statistic comparison by Baldi [3]. Since then, it

[9]Using the results shown in Figures 5.6a and 5.6b as an example, if a test was run on a sample with $ratio_+ = 2^8$ on raw JPTs, the $MCC \simeq 0.33$ and $B^* \simeq 1.1$. Recalculating *MCC* for the normalized JPT observed at $ratio_+ = 2^8$ and $B^* \simeq 1.1$, results in $MCC < 0.69$. However, the actual peak is $MCC > 0.85$.



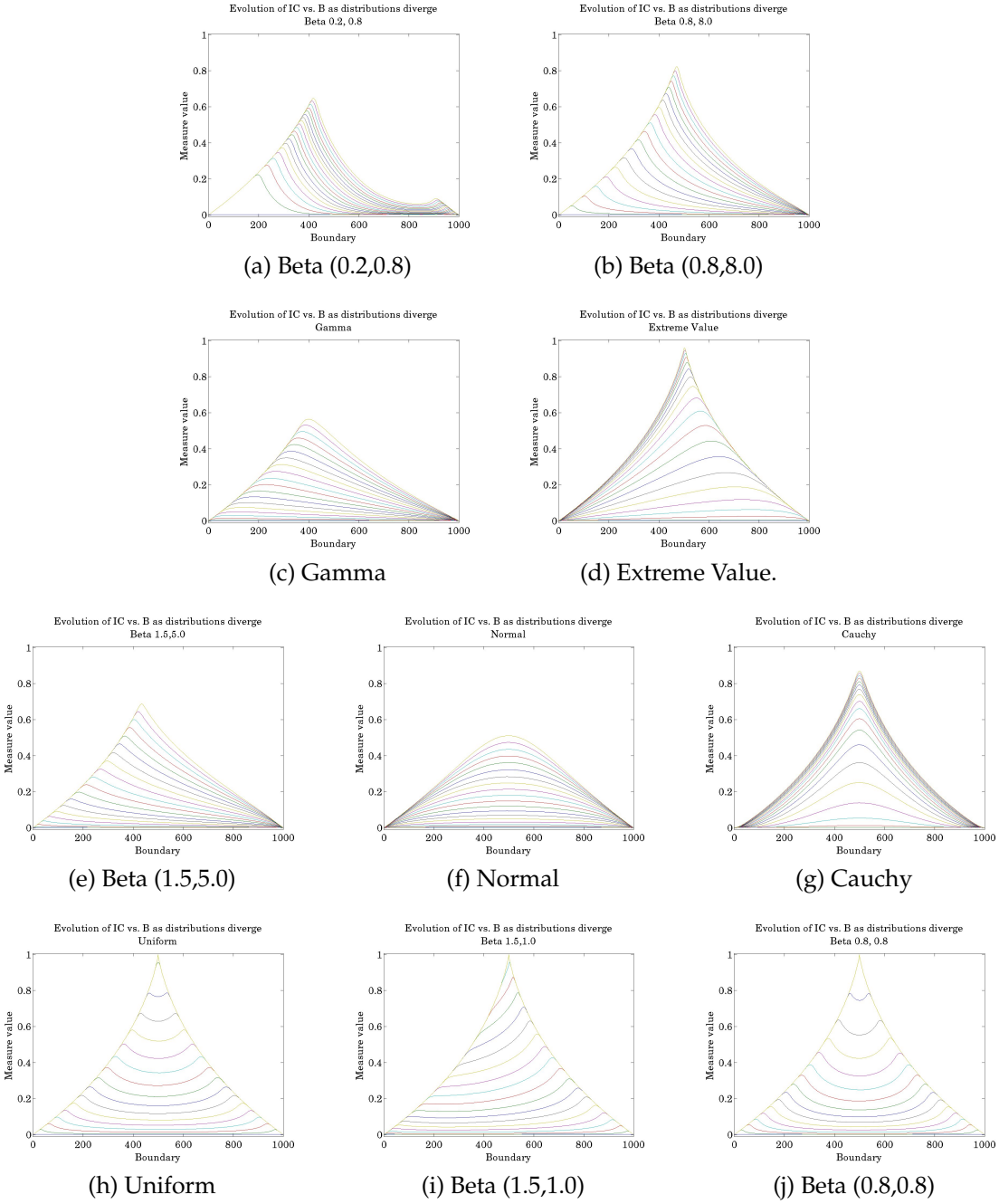
(a) Beta (0.2,0.8) (b) Beta (0.8,8.0)

(c) Gamma (d) Extreme Value

(e) Beta (1.5,5.0) (f) Normal (g) Cauchy

(h) Uniform (i) Beta (1.5,1.0) (j) Beta (0.8,0.8)

Figure 5.8: Graphs showing *MCC*'s boundary sensitivity to pdf.



(a) Beta (0.2,0.8) (b) Beta (0.8,8.0)

(c) Gamma (d) Extreme Value.

(e) Beta (1.5,5.0) (f) Normal (g) Cauchy

(h) Uniform (i) Beta (1.5,1.0) (j) Beta (0.8,0.8)

Figure 5.9: Graphs showing *IC*'s boundary sensitivity to pdf.

has gained some traction in biological literature [15, 45, 47, 53, 61, 66, 80, 81, 94] and been seen in network management literature [29]. The summary statistic is sometimes called the information coefficient or mutual information coefficient; I use the acronym *IC*. *IC* is measured on the interval $[0, 1.0]$. *IC* quantifies the proportion of full knowledge an observer has of ground truth, given the target CPD tool's output.

As explained by Baldi, et al., *IC* is the mutual information (I) contained in ground truth regarding the test set S ($Y \cup \overline{Y}$) and the CPD prediction of ground truth, as contained in $Z \cup \overline{Z}$. Normalized by the entropy in ground truth (H):

$$IC = \frac{I(Y \cup \overline{Y}, Z \cup \overline{Z})}{H(Y \cup \overline{Y})}.$$

Expressing I and H in terms of JPT categories,

$$I(Y \cup \overline{Y}, Z \cup \overline{Z}) = -H\left(\frac{T_+}{N}, \frac{F_+}{N}, \frac{F_-}{N}, \frac{T_-}{N}\right)$$
$$- \frac{T_+}{N}\log(|Y| * |Z|) - \frac{F_+}{N}\log(|\overline{Y}| * |Z|) - \frac{F_-}{N}\log(|Y| * |\overline{Z}|) - \frac{T_+}{N}\log(|\overline{Y}| * |\overline{Z}|), \tag{5.3}$$

where

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

Some reports indicate the belief that the summary statistics are $ratio_+$ invariant [53, 66, 94]. Solis and Rackovsky note that their particular information theoretic summary statistic may not be $ratio_+$ invariant [80]. The belief that information theoretic summary statistics are $ratio_+$ invariant comes from the fact that information theory applies to probability density functions, which are always normalized ($ratio_+ = 1$) [17, 96]. Unless JPTs are normalized prior to use, *IC* and related summary statistics cannot be guaranteed to be $ratio_+$ invariant.

Like other summary statistics, *IC* compares target CPD output to a 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 random selection, then $IC = 0$. A side effect of IC's use of logs is increased computational complexity. All of the other summary statistics 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 my analysis. Had I calculated *IC* on the two hundred thousand element test sets used for the other summary statistics, it would have taken approximately six months. Consequently, I tested *IC* on twenty thousand element test sets. In

Figure 5.7, I can see that the peak boundary shifts as $ratio_+$ increases, thus *IC* is not $ratio_+$ invariant. As with the other $ratio_+$ sensitive summary statistics, JPT normalization, can confer $ratio_+$ invariance.

For the most part, *IC*'s optimum boundary are consistent with the other summary statistics. However, *MCC* and *IC* are bimodal with a uniform distribution (Figure 5.8h). Potential implications of this behavior are discussed in Section 9.4. *IC*'s sensitivity to $ratio_+$ and pdf are presented in Figures 5.7 and 5.9. How does it fare for end user efficacy?

What question does the summary statistic quantify? *IC* quantifies the proportion of full knowledge an observer has of ground truth, given the target CPD tool's output.

Is the summary statistic measured on a ratio scale? *IC* is measured on the interval $[0, 1.0]$, so it is measured on an ordinal scale; impact is out of scope.

Does the summary statistic exhibit boundary sensitivity? *IC* is sensitive to both $ratio_+$ and pdf. Surprisingly, on the uniform distribution, it is bimodal. This characteristic is discussed in Chapter 9.4.

5.1.3 Tailoring to an end user's interest

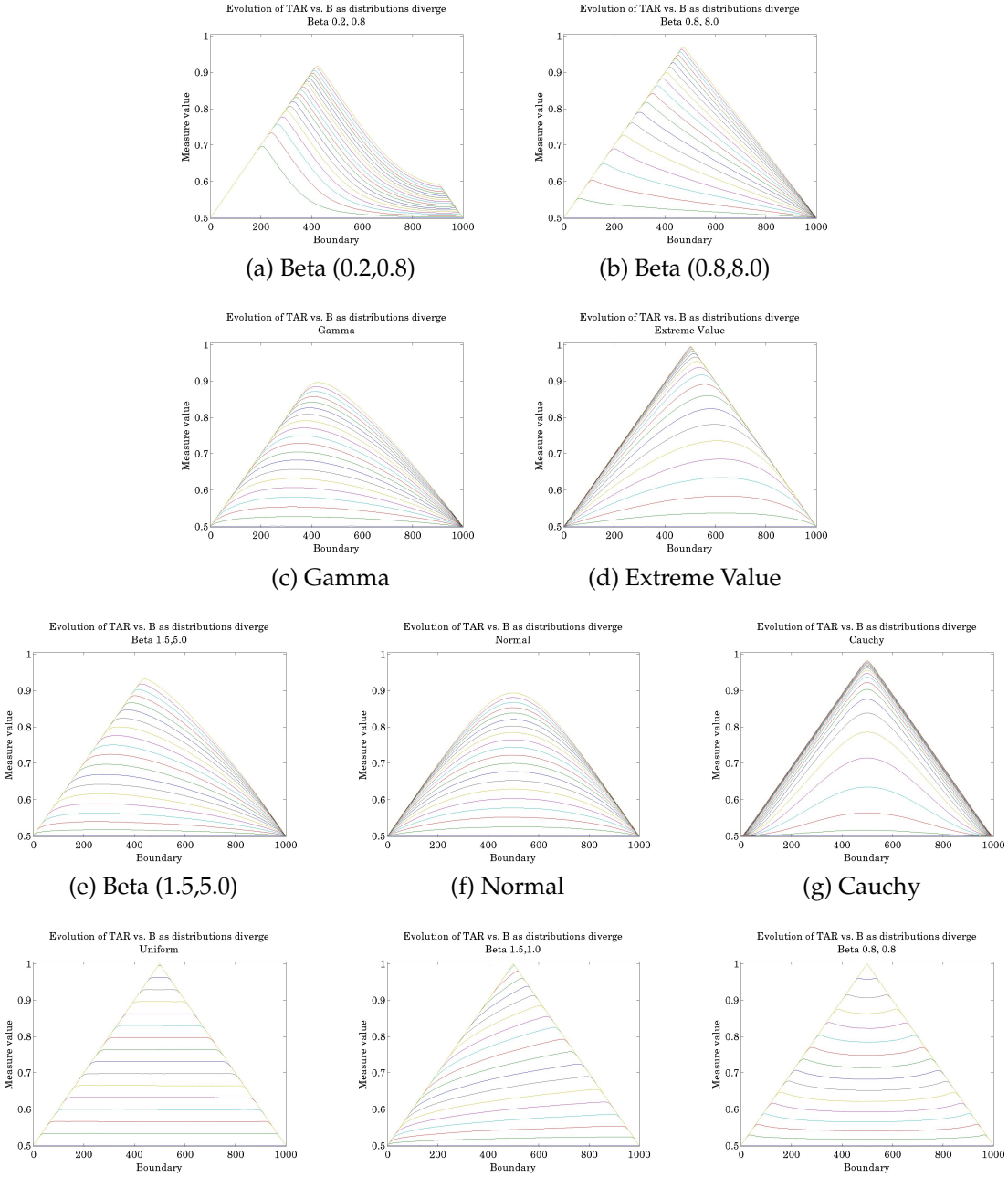
So far, I have considered summary statistics mitigating an extrinsic factor confounding CPD comparisons by researchers (Section 5.1.1) and summary statistics quantifying a CPD's "intrinsic" characteristic (Section 5.1.2). The third summary statistic creation strategy is addressing the end user's need. In a poorly posed form, end users want to know "how well with this CPD work for me?". This section considers two solutions to that question.

Total accuracy rate (TAR) In the general case, one would expect maximizing correct classifications to be the goal for end users. Based on that assumption, an intuitive summary statistic would simply be a matter of measuring correctly partitioned targets in the test data and reporting the percent-correct. (An equivalent strategy would measure the incorrectly partitioned targets and report the percent-error.):

$$\textit{Total Accuracy Rate (TAR)} = \frac{t_+ + t_-}{|S|}.$$

TAR's summary statistic suite consists of the accuracy rates for the two "True" categories:

$$T_+\textit{Accuracy} = \frac{t_+}{|S|},$$



(a) Beta (0.2,0.8) (b) Beta (0.8,8.0)

(c) Gamma (d) Extreme Value

(e) Beta (1.5,5.0) (f) Normal (g) Cauchy

(h) Uniform (i) Beta (1.5,1.0) (j) Beta (0.8,0.8)

Figure 5.10: Graphs showing *TAR*'s boundary sensitivity to pdf.

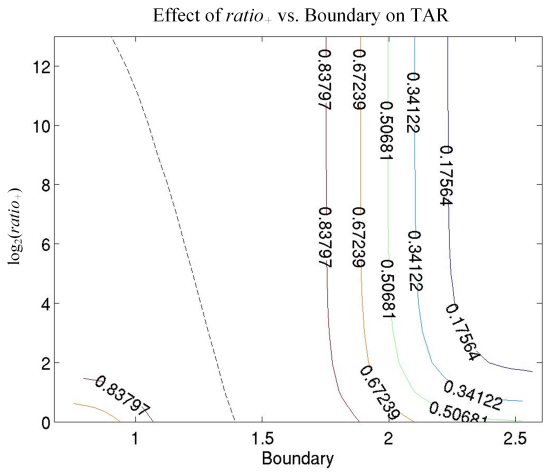


Figure 5.11: Being the sum of the observed correct classifications. It is significant that the dashed line, indicating the optimum boundary, is not vertical; this shows that *TAR* is $ratio_+$ sensitive.

$$T\_Accuracy = \frac{t_-}{|S|}.$$

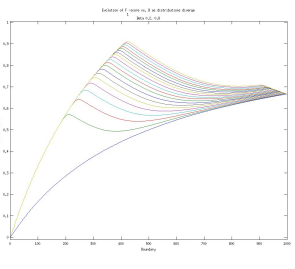
TAR's sensitivity to $ratio_+$ and pdf are presented in Figures 5.11 and 5.10.

Figure 5.11 shows *TAR* as boundary and $ratio_+$ 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 $ratio_+$ increases. The optimum boundary and the reported accuracy rate both change as $ratio_+$ varies. *TAR* is seen in the literature, but often used as an example of a poor summary statistic. *TAR* is $ratio_+$ sensitive, a characteristic which is confounding for researchers. This topic will arise in following chapters. How does it fare for end user efficacy?

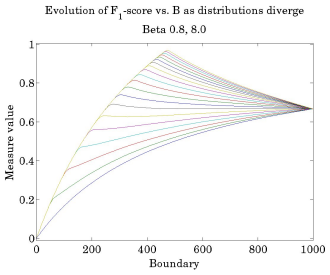
What question does the summary statistic quantify? *TAR* quantifies the proportion of correctly partitioned targets in the test data.

Is the summary statistic measured on a ratio scale? *TAR* is measured on the interval $[0, 1.0]$. However, the zero is not meaningful for end users and it does not have a standard unit, so is measured on an ordinal scale; impact is out of scope.

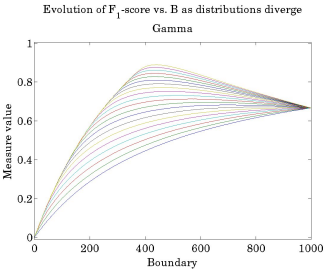
Does the summary statistic exhibit boundary sensitivity? *TAR* exhibits boundary sensitivity. In fact, the sensitivity is equivalent to J, when $ratio_+ = 1$. This equivalence is discussed in section 5.1.3.



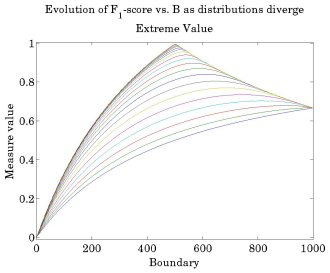
(a) Beta (0.2,0.8)



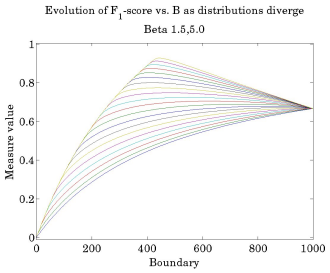
(b) Beta (0.8,8.0))



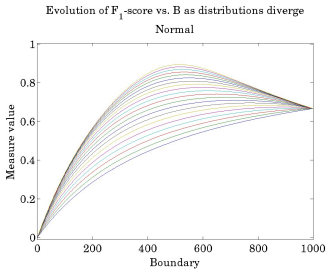
(c) Gamma



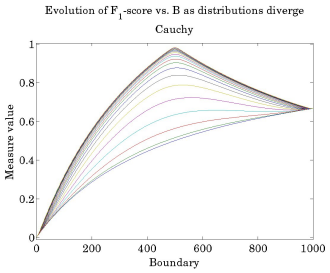
(d) Extreme value



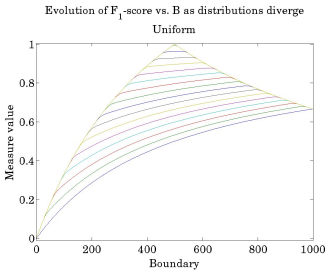
(e) Beta (1.5,5.0)



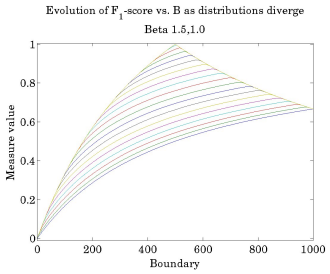
(f) Normal



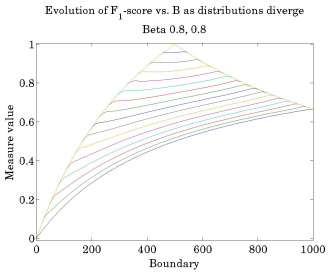
(g) Cauchy



(h) Uniform



(i) Beta (1.5,1.0)



(j) Beta (0.8,0.8)

Figure 5.12: Graphs showing F_β-*score*'s boundary sensitivity to pdf.

F_β-score The summary statistic suite for *F_β-score* is recall and precision, information retrieval performance criteria put forth by Cleverdon as actionable information in that problem domain [14]. *Recall* quantifies a CPD's completeness (what is the probability that the desired observations in the database are correctly identified). *Precision* quantifies what the probability is that undesired observations are mistakenly labeled as desired. Van Rijsbergen used precision and recall as the basis for an effectiveness summary statistic [88]. *F_β-score* is the complement to Van Rijsbergen's summary statistic and is now used in other problem domains.

Recall and precision correspond to the conditional probabilities $P(T_+|Y)$ and $P(T_+|Z)$ (also known as "True Positive Rate" (TPR) and "Positive Predictive Value" (PPV)). In the problem domain within which they were introduced (information retrieval), these summary statistics quantify how well a CPD relates an object to a concept, such as selecting a document based on keywords. *F_β-score* is defined as:

$$F_\beta\text{-}score = \frac{(1+\beta^2)(precision)(recall)}{(\beta^2)(precision + recall)},$$

where β is the relative importance of precision and recall:

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

Substituting for precision and recall and rearranging terms,

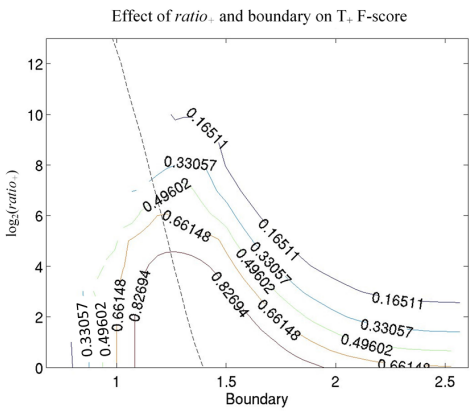
$$F_\beta\text{-}score = \frac{t_+}{t_+ + \frac{\beta^2}{1+\beta^2} f_+ + \frac{1}{1+\beta^2} f_-}$$

In information retrieval, irrelevant documents (T_-) are valueless. Because of this, *F_β-score* excludes T_-. What is important, is T_+ and the degrading effects of error (F_+ and F_-) on the results received by the end user. This is what *F_β-score* quantifies. One of the information retrieval system evaluation questions posed by Van Rijsbergen is "is it worth it?". Unfortunately, *F_β-score* is measured in the interval $[0, 1.0]$, so is an ordinal scale summary statistic. Systems can be ranked, but the intervals are not easily interpretable – "worthiness" cannot be determined.

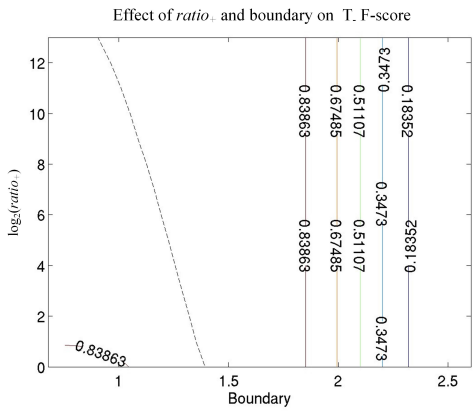
In contrast to *TAR*, *F_β-score*'s *$ratio_+$* sensitivity varies, depending upon the category ignored (T_+ or T_-). The effect can be seen in Figures 5.13a and 5.13b. *F_β-score* is based on factors which information retrieval end users and researchers seemingly agree are important. Even though it is *$ratio_+$* sensitive, it has become a de facto standard in some CPD evaluation problem domains. Considering *F_β-score*'s end user efficacy, it is not clear what *F_β-score* is quantifying. *F_β-score*'s sensitivity to *$ratio_+$* and pdf are presented in Figures 5.12 and 5.13.

How does it fare for end user efficacy?

What question does the summary statistic quantify? *F_β-score* measures the degrading effects of error (F_+ and F_-) on the results received by the end user.



(a) The relatively flat area around the optimum boundary (black dashed line in the graph) with low $ratio_+$ suggests a low boundary sensitivity. The ridge follows the same optimum boundary as that of the total accuracy rate.



(b) If, instead of selecting T_+, I select T_-, then F_β-*score* looks remarkably similar to *TAR* . Thus, F_β-*score* quantifies the categorical process's effect on a specific category.

Figure 5.13: These graphs show that F_β-*score*, a summary statistic commonly used to compare CPD effectiveness, is $ratio_+$ sensitive. This is a desirable characteristic for problem domains such as information retrieval. In addition to $ratio_+$ sensitivity, F_β-*score* is also sensitive to the target class.

This is the problem in information retrieval, the domain for which F_β-*score* was created. Unfortunately, this is not impact.

Is the summary statistic measured on a ratio scale? F_β-*score* falls within the interval $[0, 1.0]$. It has a meaningful zero, but it does not have a standard unit. F_β-*score* is an ordinal scale summary statistic; impact is out of scope.

Does the summary statistic exhibit boundary sensitivity? F_β-*score* is sensitive to both $ratio_+$ and pdf.

5.2 Gap analysis

	Summary statistic						
	TAR	F_β*-score*	Youden	*MCC*	*IC*	*DOR*	*AUC*
Type III error avoided	No	No	No	No	No	No	No
Ratio scale	No	No	No	No	No	No	No
Boundary sensitivity	Yes	Yes	Yes	Yes	Yes	Yes	No

Table 5.3: None of the summary statistics considered meet end user needs.

Statisticians have introduced a term useful in addressing summary statistic efficacy: *Type III error* — getting the right answer to the wrong problem. End user's attempting to glean actionable information for CPD assessment are presented with Type-III results. As an example of the problem end users face, consider $ratio_+$ sensitivity. Two commonly seen summary statistics, total accuracy rate (*TAR*), an intuitive summary statistic and the F_β*-score* [88], have opposite responses to $ratio_+$ [24]. The F_β*-score* monotonically decreases as $ratio_+$ increases. Under the same conditions, *TAR* monotonically increases. It would seem that in some problem domains, the summary statistic used (TAR versus F_β*-score*) would lead to contradictory conclusions. Table 5.3 summarizes the end user efficacy of the summary statistics tested; none are end user efficacious.

This may seem a surprising state of affairs. However, perhaps it should not be. From a purely academic perspective, the goal for many researchers is to characterize the CPD performance independent of environment. Hence, a plethora of environmentally insensitive summary statistics have emerged and summary statistics have been tested for environmental invariance. For example, Sokolova and Lapalme summary statistics used for CPD evaluation for invariance to various JPT perturbations [79].

The situation may be the result of a more basic deficiency. I was blinded by the absence of any framework for characterizing CPD problems. Without any points of reference, how can one navigate? One supporting observation was how similar problem domains, instead of converging on a common summary statistic, may use dissimilar summary statistics. For example, intrusion detection and information retrieval are in many ways similar, yet *ROC-AUC* prevails in intrusion detection and information retrieval generally uses F_β*-score*. These two summary statistics have very little similarity: *ROC-AUC* is class size imbalance and boundary invariant, F_β*-score* is neither. The two summary statistics also have different theoretical foundations. F_β*-score* is based on efficacy criteria for information retrieval proposed by Cleverdon; *ROC-AUC* is based on the receiver operating characteristic, a summary statistic originating in signal detection.

The gap, then, is the lack of a framework for characterizing CPD problems. The lack of end user efficacious CPD evaluation summary statistics is a symptom of that underlying deficiency.

5.3 Bridging the gap

Having identified the root cause for the dearth of end user efficacious CPD evaluation summary statistics, I now have the two end points for this dissertation. The process has five milestones:

1. **Identify key differentiators for CPD problems.** It might be fair to state that each end user's need is, in some way, unique. However, uniqueness does not imply a lack of useful common problem characteristics. Chapter 6 develops the deliverable for this milestone.

2. **Define Axioms for efficacious CPD evaluation summary statistics.** Once key differentiators have been identified, they can be applied to establish specific performance indicators for efficacious CPD evaluation summary statistics. Chapter 8 develops the deliverable for this milestone.

3. **Develop Axiom-compliant CPD evaluation summary statistics.** Given specific needs for each problem type, two summary statistics that provide actionable information and comply with the axioms are developed. Chapter 9 develops the deliverable for this milestone.

4. **Test the efficacious CPD evaluation summary statistics.** Re-analyzing published results shows the actionable information previously unavailable to end users. The increased summary statistic efficacy is shown in Chapter 10.

5. **Determine the CPD evaluation summary statistics bounds.** Creating a CPD tool starts with the original idea, moves through a development process and ultimately results in a mature, marketable product. This process has been quantized into technical readiness levels. Chapter section 12.2 develops the deliverable for this milestone.

The end result will be an improved understanding of the CPD domain. This knowledge leads to better summary statistics and summary statistic selection and application.

CHAPTER 6

END USER EXPLANATORY VARIABLES

Basic research is application agnostic. As a consequence, application-specific factors are confounding. In contrast, end users are concerned with how CPD tools function for their specific application: application-specific factors are explanatory / independent. This chapter identifies two end user explanatory variables and summarizes the means by which they are addressed in the summary statistics reviewed. For each explanatory variable, practical considerations and implications are presented, relative to two published CPD problems published in the literature, Egyptian Bank loan decisions [1] and a medical diagnostic test [64].

6.1 The role of $ratio_+$

What is $ratio_+$ to the end user? In Chapter 5, some of the summary statistics characterized were intentionally created to be $ratio_+$ invariant, while others were created without a concern for invariance. Does this imply that for some end user problems, $ratio_+$ is an explanatory factor and for others it is confounding? If so, then how can the two problem types be differentiated? (Note that in the test model presented in Figure 4.1, $ratio_+$ and $P_{leading}$ quantify the same environmental characteristic.)

When could $ratio_+$ be explanatory? One such situation arises where individual results are significant only to the extent to which they contribute to a cumulative result. Consider a bank lending decision. The impact is cumulative, with each evaluation activity contributing to the bank's overall profitability. In this case, relative class size (expressed as $P_{leading}$ or $ratio_+$ in this work) is important. If the end user were to base a deployed classifier's boundary on P_{event} by using a $ratio_+$ invariant summary statistic (a summary statistic that could not reflect the bank's estimate of their lending environment), there would likely be either excessive losses due to lending to unqualified borrowers, or unrealized gain due to rejecting qualified borrowers. Cases of this type, where each individual outcome contributes to a cumulative result, require knowledge of both $P_{leading}$ and P_{event}.

- Of the eight summary statistics reviewed, four are $ratio_+$ sensitive: *MCC*, *IC*, *TAR* and F_β*-score*.

- The opposite responses of *TAR* and F_β*-score* to $ratio_+$ suggests the existence of other explanatory variables in the CPD problem structure. This is addressed in Section 6.2.

Actual target classification		Y	$\overline{Y}$	
Test	*Positive*	$\frac{t_+}{\|Y\|}$	$\frac{f_+}{\|\overline{Y}\|}$	
Result	*Negative*	$\frac{f_-}{\|Y\|}$	$\frac{t_-}{\|\overline{Y}\|}$	
Normalized totals		1	1	2

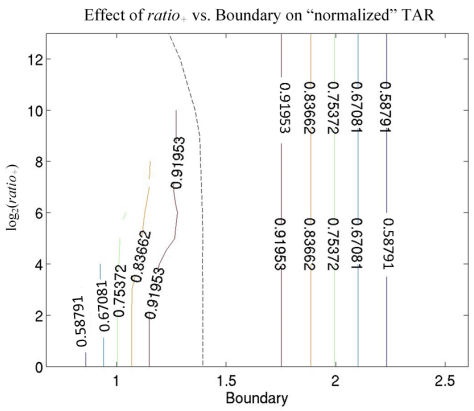
Table 6.1: The values in this JPT have been normalized. Normalization results in equal class sizes (the totals both equal one).

When could $ratio_+$ be confounding? 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 patients positive for rheumatoid arthritis (RA_+) tested could vary considerably. The example in Section 10 presents a case with two physician's offices. One was a general practitioner, where patients tested for RA had a $ratio_+$ of 0.01 ($|RA_+|/|RA_-|$). The other was a rheumatologist. In that office, patients tested for RA had a $ratio_+$ of 100 ($|RA_+|/|RA_-|$). 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 $ratio_+$ based $P_{leading}$ does not minimize the error for the patient: $ratio_+$ is confounding.

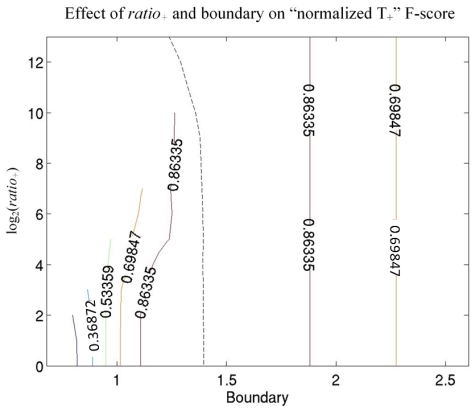
Four of the summary statistics reviewed in this dissertation are $ratio_+$ invariant: *ROC-AUC*, *J*, *DOR* and *DP*. As shown in Chapter 5, none are end user efficacious. Fortunately, $ratio_+$ invariance is *NOT* strictly a summary statistic characteristic. It can be conferred by using an intermediate JPT representation: a normalized JPT.

6.1.1 JPT tuning

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 curve can be interpreted as the probability of an event occurring within the bounds defined by that segment [11]. Similarly, distribution standardization facilitates pdf comparisons. In CPD analysis, distribution standardization takes the form of tuning JPTs so that $ratio_+ = 1$: JPT normalization. Table 1 shows a



(a) Contour graph of normalized Accuracy rate. The reader may note that, other than the contour values, the graph is almost exactly the same as the Youden Index graph.



(b) Contour graph of normalized F-score. As with Youden Index and normalized Accuracy rate, the optimum boundary follows the "minimum error boundary".

Figure 6.1: The normalized Accuracy rate and F_β-*score* seem to be relatively invariant to $ratio_+$. Not only is the value relatively constant, but the boundary stays constant as well.

JPT displaying "raw" data – actual category cardinality. After normalization, the class totals (bottom row in the Table 6.1) are one. Thus, JPT normalization seems to be a cause for $ratio_+$ invariance in summary statistics. As such, it provides a benefit to end users with problems where $ratio_+$ is confounding: any JPT-based CPD evaluation summary statistic will have output for $ratio_+ = 1$ test data input, if the input JPTs are normalized. (Illustrated in Figure 6.1). Although any summary statistic can be $ratio_+$ invariant when the JPTs are normalized, some summary statistics have emerged which have intrinsic $ratio_+$ invariance. These inherently $ratio_+$ invariant summary statistics all have $\{TPR, FPR\}$ (ratios that normalize the JPTs) as summary statistic suites, thus rather than being counter examples, they provide empirical evidence that JPT normalization is the root cause for $ratio_+$ invariance in summary statistics; a proof is beyond the scope of this dissertation.

There is also a benefit for end users with $ratio_+$ sensitive problems. Statisticians use distribution standardization to mitigate $ratio_+$ effects, however, the process is reversible. JPTs with $ratio_+ = 1$ can be "tuned" to any desired $ratio_+$ simply by multiplying one class by a constant c so that $\frac{cY}{\overline{Y}}$ equals the desired value[1]. Thus, an end user with a $ratio_+$ 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 $ratio_+$ range. These insights are applied to real-world problems in Chapter 10.

[1]This expression does not require that $ratio_+ = 1$ initially. With the exception of Y or $\overline{Y}$ equaling zero, any JPT can be transformed (tuned) from one $ratio_+$ to another.

Source population				
		Y	$\overline{Y}$	
Test Result	*Positive*	TPR	FPR	
	Negative	$FPR = 1 - TPR$	$TNR = 1 - FPR$	
	Totals	1	1	2

Table 6.2: A normalized JPT has class sizes adjusted to one. The four classification categories are expressed as proportions of the test set class of which they actually are members.

However, the optimum boundary is $ratio_+$ dependent, thus the tool is not complete. To apply to all end users, results for all possible optimum boundaries would need to be provided[2]. This is impractical, if not impossible for CPD test reports to include. As illustrated in chapter 5, footnote 9, the tuned JPTs will indicate trends, but cannot be considered definitive. Nonetheless, JPT tuning extends JPT normalization in a way I have not previously seen in the literature and provides end users with a valuable capability.

6.1.2 Practical considerations and implications

JPT tuning has value, regardless of whether an end user's problem domain is $ratio_+$ sensitive or not. If the domain is $ratio_+$ sensitive, such as banks making loan decisions, then JPT tuning allows end users to adjust reported results to their specific $ratio_+$. If $ratio_+$ confounding for the domain, such as a medical diagnosis, then JPT tuning can normalize the test results ($ratio_+ = 1$).

An important implication is that CPD evaluations are no longer tied to inherently $ratio_+$ invariant summary statistics. Hence, end users are free to use any summary statistic that quantifies the characteristic of interest. This possibility is further developed in this study.

6.1.3 Summary statistics with intrinsic $ratio_+$ invariance

Although the *ROC-AUC*, Youden Index and *DOR/DP* are distinctly different summary statistics, they all have one key similarity: normalized input. The *ROC-AUC* and Youden Index both are (TPR, FPR) and since TPR and FPR are conditional probabilities $P(T_+|Y)$ and $P(F_+|\overline{Y})$. Likewise, TNR and FNR are conditional probabilities $P(T_-|Y)$ and $P(F_-|\overline{Y})$. If, in the JPT, I replace T_+ by

[2]There may be a solution to this deficiency; I will investigate this in future work.

Source population				
		Y	$\overline{Y}$	$Totals \downarrow$
Test Result	$Positive$	$c_Y * TPR$	$c_{\overline{Y}} * FPR$	Z
	$Negative$	$c_Y * (1 - TPR)$	$c_{\overline{Y}} * (1 - FPR)$	$\overline{Z}$
	$Totals$	c_Y	$c_{\overline{Y}}$	N

Table 6.3: JPTs can be defined in terms of the TPR and FPR. c_Y and $c_{\overline{Y}}$ are the class sizes in the test set.

TPR, F_+ by FPR, T_- by TNR, F_- by FNR, then the marginal totals Y and $\overline{Y}$ are replaced by 1s and N becomes 2. This is shown in Table 6.2. Since the two marginal totals representing class size are equal, this process compensates for *ratio*$_+$: the CPD output JPTs have been normalized. In this dissertation, calculations and discussion using the *ratio*$_+$ invariant JPT form shown in Table 6.3 will refer to "normalized" versions. Any discussions not referring to "normalization" are of summary statistics using the "raw" JPT form as presented in the Lexicon, Table 1.

Regardless of the actual test set *ratio*$_+$s, the input values for the *ROC-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 6.3. Using Table 6.3 definitions,

$$DOR = \frac{(c_Y * TPR)(c_{\overline{Y}} * (1 - FPR))}{(c_Y * (1 - TPR)(c_{\overline{Y}} * FPR)}$$

after simplification,

$$DOR = \frac{TPR(1 - FPR)}{FPR(1 - TPR)}.$$

Thus I find that *DOR* and *DP* are based on normalized JPTs as well.

DOR and *DP* do have one important difference from the other *ratio*$_+$ invariant summary statistics (this includes *TAR*, *MCC* and F_β*-score* on normalized JPTs). Their optimum boundaries is not the same.

Figure 6.2 shows the optimum boundary vs. *ratio*$_+$ for the normalized *TAR*, normalized F_1*-score*, normalized *MCC*, Youden index and *DOR*. As can be seen in the Figure, *DOR* peaks at a different boundary than the other *ratio*$_+$ invariant summary statistics tested and (excepting DOR) the optimum boundary is relatively stable until *ratio*$_+ > 2^6$, after which the detected optimum boundary starts dropping rapidly[3]. Figure 6.2 supports my observation in Chapter 5 *DOR* and *DP* should not be used to identify the optimum boundary.

[3]We noticed a similar effect on the summary statistic's values. The values started becoming overly optimistic (once again, excepting DOR, the values of which dropped). The cause turned out to be a result of the Strong Law of Large Numbers. The effect became significant when class Y's size fell below 400 elements.

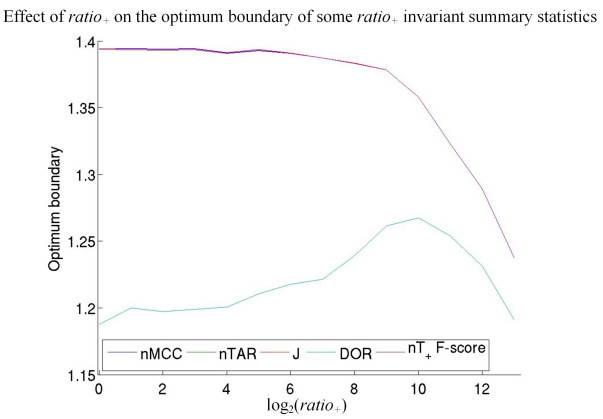


Figure 6.2: This figure plots the optimum boundary for five summary statistics, but seems to only have two lines. This is because all but *DOR/DP* identified essentially the same optimum boundary. Hence, the upper line is an overlay of plots for four separate summary statistics.

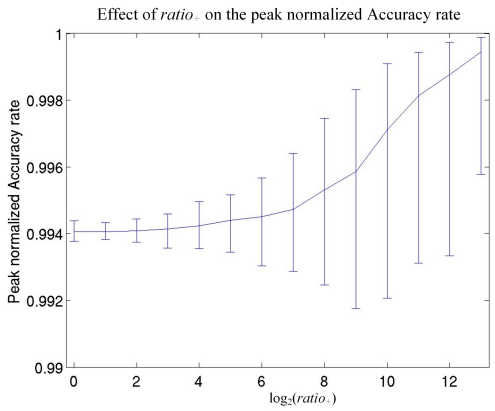
From the literature, I see that *MCC* is $ratio_+$ invariant when calculated on normalized JPTs. Presumably, other $ratio_+$ sensitive summary statistics will be $ratio_+$ invariant when calculated on normalized JPTs as well. I tested this hypothesis by calculating *TAR*, F_β*-score* and *MCC* values on normalized versions. Figure 6.3 displays the peak Accuracy rate and F_β*-score* on normalized JPTs and compares them to the output of the established $ratio_+$ invariant summary statistics, *ROC-AUC*, *DOR*[4] and Youden Index. (*DP*, being just a log expression of *DOR*. was left out.) The graphs are provided solely to compare their response to $ratio_+$. Any conclusions from Figure 6.3 beyond that must be made with care.

Figure 6.3 brings out some interesting insights:

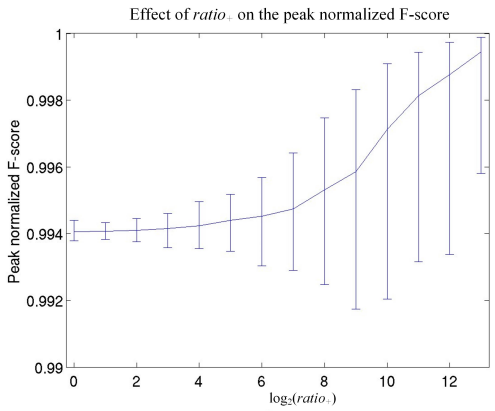
- *Confidence interval* response to $ratio_+$ seems to fall into two categories. All of the normalized summary statistics (including *ROC-AUC*, Youden index, *DOR* and *DP*) have relatively stable CIs below $ratio_+ S = 2^6$. Above $ratio_+ = 2^6$, there is an observable trend away from the stable value. This is due to an issue with absolute sample size resulting from the strong law of large numbers. In my tests, the problem became statistically significant when the smaller sample had less than four hundred members.

  For normalized *TAR*, Youden Index, normalized F_β*-score* and normalized *MCC*, the 90% confidence interval generally increases as $ratio_+$ 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

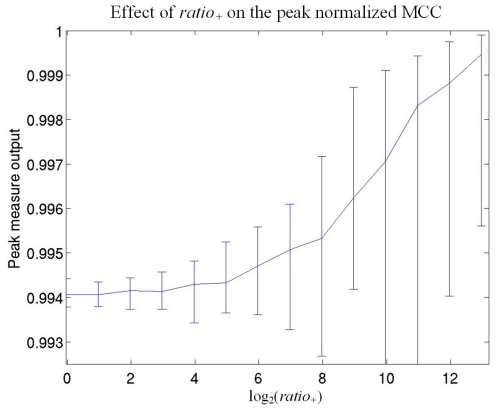
[4]All of the other summary statistics are bound. In order to facilitate comparison, *DOR*. was transformed from an "odds" format to the equivalent "probability" format.



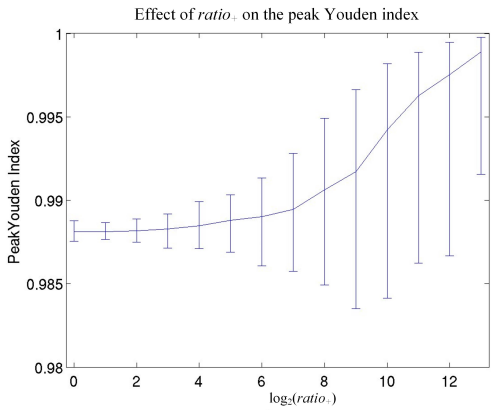
(a) Peak normalized Accuracy rate (90% CI). It strongly resembles the Youden Index.



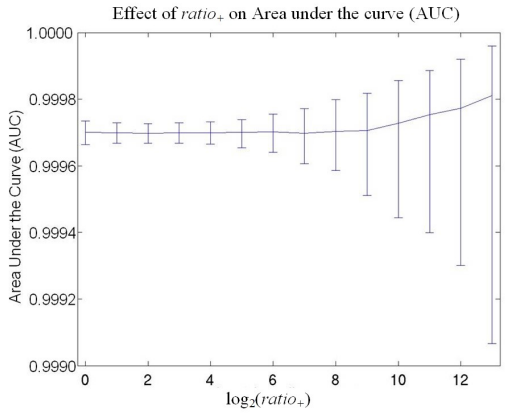
(b) Peak normalized F_β-*score* (90% CI). It exhibits the most $ratio_+$ instability.



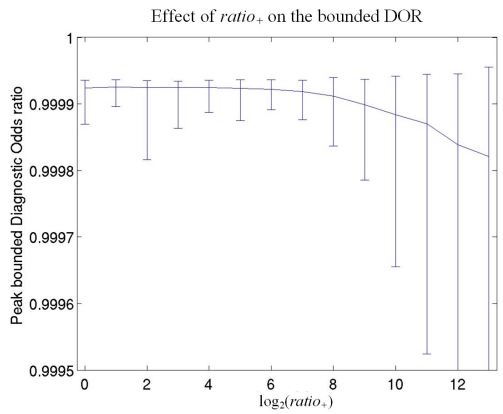
(c) Peak normalized *MCC* (90% CI). The earlier similarity noted between *MCC* and F_β-*score* does not as strong on normalized JPTs.



(d) Peak Youden Index (90% CI). This summary statistic turns out to be related to the normalized Accuracy rate.



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



(f) Best *DOR*. (90% CI). As $ratio_+$ invariance weakens, the *DOR* value drops.

Figure 6.3: The normalized summary statistics are $ratio_+$ invariant, but a well-known absolute sample size effect shows when $ratio_+ > 2^6$. The effect was statistically significant when $|Y| < 400$.

size decreases, normalization magnifies any changes in T_+ and F_- 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.)

- *Summary statistic 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:

$$\textit{Youden Index} = TPR - FPR$$

$$\textit{norm}\ \text{TAR} = \frac{TPR + 1 - FPR}{2} \text{ so that}$$

$$\textit{norm}\ \text{TAR} = \frac{\textit{Youden Index} + 1}{2}. \tag{6.1}$$

  Thus I see that normalized accuracy rate and Youden Index are related.

- *JPT normalization can inflate reported process accuracy*. Each graph in Figure 6.3 exhibits $ratio_+$ stability when $ratio_+ < 2^6$. However, when $ratio_+ > 2^6$, $ratio_+$ 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 under-represented. Since process accuracy is inversely related to class overlap, a reduction in estimated class overlap will result in process accuracy over-estimation. In my 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 $ratio_+$, my protocol decreases $|Y|$. Y has the lower mean, thus as $ratio_+$ increases, the calculated optimum boundary starts shifting toward μ_Y. In my tests, when $ratio_+ > 2^6$, the shift becomes statistically significant.

6.1.4 Practical considerations and implications

JPT tuning is valuable to end users regardless of whether $ratio_+$ is confounding or explanatory in their problem domain. If the domain is $ratio_+$ sensitive, such as banks making loan decisions, then JPT tuning allows end users

to adjust reported results to their specific $ratio_+$. If $ratio_+$ is confounding for the domain, such as a medical diagnosis, then JPT tuning can normalize the test results ($ratio_+ = 1$). This mitigates $ratio_+$'s effect on CPD problems where $ratio_+$ is confounding (as constrained by the strong law of large numbers).

An important implication is that CPD evaluations are no longer tied to inherently $ratio_+$ invariant summary statistics. Hence, end users are free to use any summary statistic that quantifies the characteristic of interest.

Another valuable possibility is that JPT tuning enables sensitivity analysis for CPD problems where $ratio_+$ is explanatory. These possibilities are further developed in this study, with examples in Chapter 10.

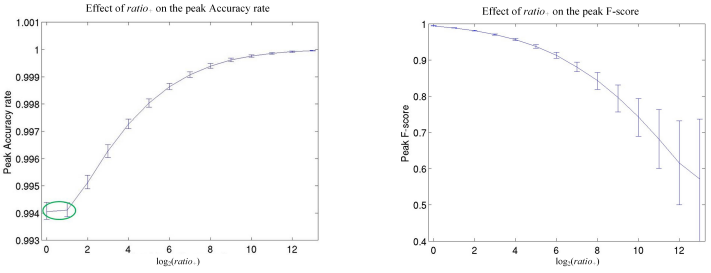
6.2 JPT category impact

While characterizing summary statistics in Section 5, I noted that two summary statistics which are based on end user interests, *TAR* and F_β*-score*, differ in their $ratio_+$ sensitivity. As can be seen in Figure 6.4, when *TAR* monotonically increases, F_β*-score* monotonically decreases. Clearly, the two summary statistics are answering different questions.

What is the factor that causes the differences observed between *TAR* and F_β*-score*? Comparing the *TAR* and F_β*-score* equations, I see different treatments of the four JPT categories. The most notable difference is with T_-. *TAR* includes T_- and F_β*-score* does not:

$$TAR = \frac{t_+ + t_-}{|S|} \text{ while}$$

$$F_\beta\text{-}score = \frac{t_+}{t_+ + \frac{\beta^2 f_+}{1+\beta^2} + \frac{f_-}{1+\beta^2}}.$$



(a) When $ratio_+$ increases, *TAR* increases as well.

(b) When $ratio_+$ increases, F_β*-score* decreases.

Figure 6.4: Although both *TAR* and F_β*-score* reputedly address end user interest, they exhibit different $ratio_+$ sensitivities. This brings into question for what problems are these summary statistics appropriate?

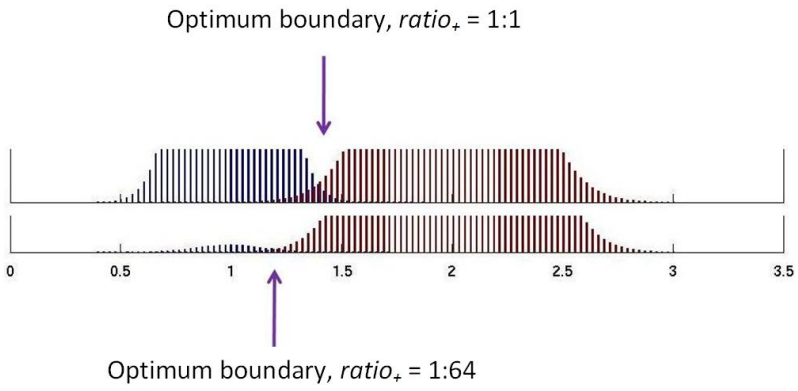


Figure 6.5: As $ratio_+$ increases, the optimum boundary shifts out the dominant class's tail. Classification accuracy of the dominant class improves, to the detriment of the other class's classification accuracy.

Also, the errors (F_+ and F_-) are discounted in F_β-*score*'s denominator. These differences have a significant effect. As $\overline{Y}$ increasingly dominates the test set, the optimum boundary shifts out the tail of $\overline{Y}$. The shift moves the optimum boundary closer to the expected value of $\overline{Y}$, increasing the number of Y elements in F_-, adversely affecting Y's classification. The effect can be seen in Figure 6.5. The *TAR* indicates the $ratio_+$'s effect on overall accuracy, the F_β-*score* only indicates the change in Y's accuracy. The F_β-*score* originated as a summary statistic for information retrieval algorithm evaluation. In that problem domain, the value is in the correct information retrieved (T_+), there is no value in incorrect information not retrieved (T_-). Hence, considering category importance, it is reasonable that T_- be excluded. β allows end users to further tailor the F_β-*score* to meet their particular category importance combinations. Consequently, as $ratio_+$ increases, *TAR* monotonically increases and the F_β-*score* monotonically decreases. The difference between *TAR* and F_β-*score*, then, results from weighing the JPT categories differently. *TAR* assumes all JPT categories have an equal impact on end users, F_β-*score* sets JPT category impacts as appropriate for information retrieval.

Based on these insights, I can now generalize impact. Considering the individual test set elements, each element will have an *impact* (ι) on the end user. Since all elements are placed into one of the four possible categories, a statistical expectation (E) can be calculated for each category.

For each category, there is a gain or loss (positive or negative ι) associated with each element output. ι will be tied to the JPT category, with each category tied to a specific (not necessarily unique) quantity $I = (\iota_{T_+}, \iota_{F_+}, \iota_{F_-}, \iota_{T_-})$. Thus, the expected effect on the end user of each and every element output will be reflected by the element of I applicable to the category to which the element is binned. (Typically, gains are viewed as positive values and losses are negative values, although there are exceptions.) If the bin counts in a JPT are expressed

as proportions as described in the Lexicon and there is a sufficiently large (i.e., statistically significant) $|S|$, then ι can be viewed in a number of contexts. From the raw data, *individual element impacts* can be viewed: ι_{s_n}, ι_{y_n}, $\iota_{\overline{y}_n}$, ι_{z_n} or $\iota_{\overline{z}_n}$ (n denotes the specific element of the source set, S, Y, $\overline{Y}$, Z or $\overline{Z}$). Impact can also be expressed as statistical expectations (*expected individual element impact*) by category or class, $I = (\iota_{T_+}, \iota_{F_+}, \iota_{F_-}, \iota_{T_-})$, where
$\iota_{T_+} = \sum_{s \in T_+} \iota_s / t_+$,
$\iota_{F_+} = \sum_{s \in F_+} \iota_s / f_+$,
$\iota_{F_-} = \sum_{s \in F_-} \iota_s / f_-$,
$\iota_{T_-} = \sum_{s \in T_-} \iota_s / t_-$,
$\iota_Z = \sum_{s \in Z} \iota_s / |Z| = \iota_{T_+} t_+ / |Z| + \iota_{F_+} f_+ / |Z|$,
$\iota_{\overline{Z}} = \sum_{s \in \overline{Z}} \iota_s / |\overline{Z}| = \iota_{F_-} f_- / |\overline{Z}| + \iota_{T_-} t_- / |\overline{Z}|$,
$\iota_Y = \sum_{s \in Y} \iota_s / |Y| = \iota_{T_+} t_+ / |Y| + \iota_{F_-} f_- / |Y|$ and
$\iota_{\overline{Y}} = \sum_{s \in \overline{Y}} \iota_s / |\overline{Y}| = \iota_{F_+} f_+ / |\overline{Y}| + \iota_{T_-} t_- / |\overline{Y}|$.

TAR and *F_β-score* are members of the same summary statistic (SS) family:

$$SS = \frac{\iota_{T_+} t_+ + \iota_{T_-} t_-}{\iota_{T_+} t_+ + \iota_{F_-} f_- + \iota_{F_+} + f_+ + \iota_{T_-} t_-}. \text{ By applying different } I \text{ vectors,} \quad (6.2)$$

$$TAR = SS|I(TAR) = (1, 1, 1, 1) \text{ and} \quad (6.3)$$

$$F_\beta\text{-}score = SS|I(F_\beta\text{-}score) = (1, \frac{\beta^2}{1+\beta^2}, \frac{1}{1+\beta^2}, 0). \quad (6.4)$$

In essence, *F_β-score* is a weighted *TAR*. Although heretofore unrecognized, impact is already an element of some commonly seen summary statistics.

6.2.1 Practical considerations and implications

The impact vector I meets an essential end user need: tailoring CPD evaluations to their specific problem. In the bank loan decision problem,
$I = \{1.0, -0.01, -0.05, -0.01\}$, based on background information provided in the study leveraged. The rheumatoid arthritis diagnostic test comparison set $I = \{\$0, \$-7,900, \$-13,000, \$0\}$, using reported misdiagnosis costs. In both cases, factoring I into the analysis provides insights not previously available.

This Chapter reports the first actionable insights for end users:

- End user problems can be partitioned into two types, those in which $ratio_+$ is important and those in which it is confounding.
- JPT normalization can compensate for test set $ratio_+$ on problems where the characteristic is confounding.

- JPT tuning can allow end users to adjust test outputs for different $ratio_+$ conditions.
- Summary statistic values are influenced by their (often implicit) JPT category impacts. Hence, end users need to assure that summary statistics used have relevant JPT category impacts.

These insights will contribute to the other research outcomes.

CHAPTER 7

CHARACTERISTICS OF A GOOD SUMMARY STATISTIC

Although measurement theory provides a foundation (such as defining numbering systems and "necessary and sufficient conditions" for summary statistics) [36, 69], there seems to be a dearth of literature on "what constitutes a good summary statistic". As a concept, quality has tended to be ill-defined; it would seem that "goodness" is a somewhat elusive concept. This work's goal is to identify efficacious summary statistics for end users. To that end, I looked at how other problem domains defined "goodness".

In engineering (which I am classing as a "hard" discipline), the criteria for assessing the relative quality of tools used to quantify some material characteristic seem to be be clearly defined. Soft disciplines, such as psychometry, have struggled with understanding measurement; psychological characteristics (emotions, for instance) are not so readily quantified. Nonetheless, the "soft sciences" have some general understanding of "goodness". In my review of the CPD assessment discipline, any such discussions were very narrowly scoped; a generally applicable framework was conspicuously absent.

Often, understanding principles used in different problem domains can bring to light commonalities and suggest fruitful strategies elsewhere. Hence, I start by identifying "good summary statistic" insights from "hard" and "soft" disciplines as well as measurement theory.

7.1 Observations from soft disciplines

Soft sciences have addressed the good summary statistic issue. Their effort has produced a framework, measure validity, to rank measurement tools [93]. The framework's underlying ideas are useful:

- *Utility* Users are expecting to receive information relevant to their needs. Hence defining stakeholders and their issues establishes a foundation for quantifying a summary statistic's usefulness (Utility is equivalent to efficacy; I use both terms). This reinforces Cleverdon's points.
- *Quality* Users need to know how much confidence can be placed on the information. Measure validity defines three ranked categories. The least reliable are based on conjecture — the summary statistics are "self-evident",

but have not had their effectiveness tested. Somewhat better summary statistics are pragmatic — empirical experience has demonstrated that the values can be mapped to the characteristic of interest, but the function relating them is unknown. The most reliable summary statistics both *i*) work in practice and *ii*) apply underlying principles — they have a theoretical basis which supports their use.

Utility and quality are not summary statistics. Rather, they are "good summary statistic" attributes identified by the soft sciences. Efficacy and quality are sufficiently vague terms that they can be perceived as being virtually synonymous. However, quality in this context relates to the degree to which the underlying model or process(es) affecting a summary statistic's values are understood — and there is a close association between the two. This study's topic is improving CPD evaluation efficacy by providing a means of selecting context relevant summary statistics. Applying the three quality tiers given above, this study results in the highest quality summary statistics: *i*) it provides a theoretical basis for proposed summary statistics in Section 9.1, *ii*) the summary statistic's effectiveness in practice are demonstrated in Section 10 [11, 42, 67].

7.2 Observations from hard disciplines

Although the soft sciences have struggled with understanding measurement, "hard disciplines" such as engineering, seem to have settled on two specific summary statistics for process output quality: precision and accuracy. These summary statistics have been widely and successfully applied to a wide range of engineering measurement problems. CPD problems do not use measurement in the engineering sense; the two problem domains are disjoint. Nonetheless, an understanding of the characteristics that make precision and accuracy useful may be mapped onto the CPD solution space.

David Hand defines precision and accuracy as:

- *Precision*: The degree to which multiple measurements are the same or similar. (result uncertainty) This is often quantified with uncertainty measures such as standard deviation and confidence interval.
- *Accuracy*: The degree of conformity of a measured quantity to its true value. (result bias) This is often quantified by calculating the difference between ground truth (the known true value) and a result centrality measure such as mean or median [36].

I see four summary statistic attributes in precision and accuracy:

- They are useful and they provide information by which an end user can make comparisons relevant to their problem.

- They describe the process's output set distribution. Engineering system characteristics are generally assumed to be Gaussian (or normally) distributed ($\mathcal{N}(\mu, \sigma^2)$), so the precision can be mapped to the output's standard deviation (σ) and accuracy can be mapped to the bias between the sample population's expected value ($\overline{x}$) and the domain's true value (μ): $accuracy = |\mu - \overline{x}|$. Using a non-parametric mapping, accuracy reflects the measure's bias while precision reflects the measure's uncertainty. Given the precision and accuracy describing an engineering measurement system's output, someone could create data sets consistent with actual output from that system.

- They are unrelated. Knowing an engineering system's standard deviation (precision) provides no insight into that system's bias (accuracy); $precision \neq f(accuracy)$ and $accuracy \neq f(precision)$. An engineering system can be completely described by orthogonal summary statistics.

- They are quantified using ratio scales (a term coined by Stevens) [83]. It is desirable for summary statistics to respond equally to system changes they detect (a "standard sequence"[1]). Also, a ratio scale measure has a "meaningful" zero[2] [36].

7.3 Building the "good summary statistic" framework

Considering the user benefits derived from the soft and hard science good summary statistic attribute lists, this section identifies good summary statistic criteria for CPD evaluation. Comparing the two lists, we find there is only one common attribute; efficacy. It is reassuring that such an intuitive criterion is common in both hard and soft summary statistics. It is, however, the only common criterion. The balance of this section considers the disjoint criteria; Table 7.1 recaps the key points.

As mentioned at the beginning of this Chapter, CPD and engineering-type summary statistics are disjoint problem domains. One characteristic where their lack of similarity shows is metric independence. As previously noted, using the engineering summary statistics precision and accuracy, it is possible to characterize an engineering system with independent summary statistics. However,

[1] A detailed treatment of standard sequence is beyond the scope of this paper; those interested are referred to measurement theory texts such as Hand [36]. In the context used here, the summary statistic must satisfy concatenation: $(a+1)x = ax + x$, where x is some standard unit of measure and $a \in [0, \infty)$.

[2] This is important because ratio scales are capable of expressing magnitude as well as rank. Temperature measured in degrees Kelvin is an example of a ratio scale. There can be no temperature less than zero degrees Kelvin, and temperature effects are generally linear: $f(n)$ will yield a result that is one half of the same calculation using the temperature $f(2n)$; temperature measurements with the Kelvin scale exhibit linearity.

with categorical problems, characterizing the system may require using dependent variables. Of the four variables necessary to completely describe a system state, two are fixed by the data set. Thus a CPD evaluation only controls two variables. As an example, let a JPT be defined by the four marginal totals, Y, $\overline{Y}$, Z and $\overline{Z}$. Test result classifier output is expressed in the quantities Z and $\overline{Z}$. No matter what that test's output is, it can never alter Y or $\overline{Y}$ [3]. Since test result classifier output consists of only Z and $\overline{Z}$, the only other JPT value that can be derived is S; the others require knowledge of the sample's ground truth. Thus, it is not possible to completely define a system state using only classifier output (Z and $\overline{Z}$). Limiting usable variables to those that are unrelated, results in an incomplete system state description; requiring completeness forces the use of dependent variables. Hence in the CPD problem domain, we can have either a complete system description, or all unrelated summary statistics, but not both. How do we choose? What we learn from the engineering domain, however, is that there are two important summary statistic characteristics, bias and uncertainty[4]. So long as we can quantify these two factors relevant to their problem domain, a stakeholder has actionable information.

How is accuracy and precision expressed in CPD problems? Engineering problems of the type discussed here have some sense of distance, a continuous variable. This is not true of CPD problems. Instead, results are binned and the discrete bin counts analyzed. The concept of precision still exists, however. It is the rate at which observations are mis-classified; perfect precision (a zero bias; the CPD model matches reality exactly) is equivalent to error-free test results. As noted earlier, we are limiting this discussion to 2x2 JPTs. Problems so expressed have two error types, F_+ and F_-. Quantifying precision requires that both error types be accounted for, I consider both.

Precision's definition stated earlier is related to the statistical concept of uncertainty. In both engineering quantification requires repeated tests; the concept maps directly to CPD problems. Uncertainty is quantified as the degree of variation between test outcomes. An important point is that uncertainty is a test characteristic, not a summary statistic characteristic, thus we can disregard it as a good summary statistic criterion.

One feature inherent to CPD problems and absent in the soft and hard discipline's discussion is classifier output sensitivity to boundary. JPTs are actually snapshots of system performance at specific boundaries. The output classifier's tags shown in Figure 4.1 could be more fully labeled as Z and $\overline{Z}$. Y and $\overline{Y}$ are ground truth[5], thus not influenced by boundary. The other JPT elements are influenced by boundary, however, so could be more fully labeled $Z(B)$, $\overline{Z}(B)$,

[3] The same case can be made for any other set of four JPT values sufficient to re-create the JPT. Value sets can be transformed by solving linear equations.

[4] Quantifying bias and uncertainty are, to a great extent the motivation behind statistics.

[5] Ground truth is the presumed objective reality or "known good" values, derived from sources or by means believed to be highly reliable.

$T_-(B)$, $T_+(B)$, $F_-(B)$ and $F_+(B)$. The summary statistics we are considering can all be expressed as $f\{Z(B), \overline{Z}(B), T_-(B), T_+(B), F_-(B), F_+(B)\}$, thus we can see that the summary statistics can also be expressed as $f(B)$. For a practitioner, optimizing output from a system such as diagrammed in Figure 4.1 requires using the optimum boundary; the boundary that maximizes impact for the end user.

The soft discipline's perspective on summary statistic quality differs from the hard discipline's by focusing on the strength of a summary statistic's tie to the underlying model. In essence, the stronger the tie, the more confidence users have in the results: quality is an expression of user confidence. Whereas precision and accuracy are quantifiable, summary statistic validity's "quality" characteristic is qualitative. Despite the differences between the three summary statistic characteristics, each is addressing the same user concern: "How confident am I that my results are actionable?". This study directly addresses summary statistic validity. My hypothesis is that a summary statistic with strong validity will also be demonstrably more efficacious than summary statistics with weak validity.

Sifting through the candidate good summary statistic criteria, we arrive at three characteristics necessary for a "good CPD evaluation summary statistic": ratio scale behavior, efficacy and boundary sensitivity:

- *Ratio scale behavior*: Since JPT values consist of counts, it is impossible for any category to have a negative value. Also, an empty bin (category count equals zero) is meaningful, thus there is a meaningful zero. Also counts, by definition, are standard sequences. Hence, ratio scale measures should be possible.
- *Efficacy*: The information provided by the summary statistics must clearly address the stakeholder's problem. For each CPD problem type, we now add the requirement that the summary statistic must quantify bias.
- *Boundary sensitivity*: If a summary statistic is boundary invariant, then it is useless for labeling output from a CPD. Thus for the practitioner, boundary sensitivity is essential.

Table 7.1 lists the specific criteria mined from the problem domains considered and the final criteria determined to be relevant for CPD evaluation. The purpose of this exercise was to identify candidate criteria. The designations "hard disciplines" and "soft disciplines" were not intended to uniquely or completely partition the universe of problem domains. Indeed, the partitioning has little, if any, effect on CPD problem density; CPD evaluation problems can be found in virtually all disciplines.

The four "good summary statistic criteria" gleaned from other problem domains inform the efficacious summary statistic axioms defined for CPD evaluation.

	Source			
Candidate Criteria	Soft Disciplines	Hard Disciplines	Final Criteria	Relevant Discussion
Efficacy	X	X	X	Four axia Chapter 8
Quality	X		X	Compliant Summary statistics (Chapter 9)
Ratio scale	X	X	X	Measurement theory Chapter 8.5
Uncertainty		X	—	—
Bias		X	X	Compliant Summary statistics (Chapter 9)
Boundary sensitivity			X	Axiom 2 Chapter 8

Table 7.1: Of the candidate good summary statistic criteria considered, all but one (uncertainty) are relevant to end users evaluating CPDs. Since these criteria influence further work, the relevant discussions are mapped to the criteria.

CHAPTER 8

FOUR AXIOMS FOR END USER EFFICACIOUS SUMMARY STATISTICS

My intent is to improve the ability of end users to use CPD test results; efficacious end user summary statistics of CPDs must reflect the CPD's performance in the end users environment and their problem's context. There are many types of performance. For instance, classifier speed, memory usage or CPU usage are factors that might affect CPD efficacy. I focus on CPD output utility, the CPD's ability to approach the optimal response for an end user. In the previous sections, I have identified some problem characteristics that can partition end user problems and three questions end users have about how a CPD tool will work in their environment. I now express the insights gained as Axioms. After presenting each Axiom, I evaluate seven commonly seen summary statistics' compliance with the Axiom.

8.1 Axiom 1, Category importance

An efficacious end user summary statistic must be sensitive to the same factors and to the same degree as the end users are to their respective problems. With regard to utility, the end user's context is defined by the importance, or impact, of elements from each JPT category on the end user. The axiom follows directly from the previous discussion. A small change to any element of I will generate corresponding changes in a compliant summary statistic. *TAR* and the *F_β-score* are examples discussed in Chapter 6.2. The response of these two summary statistics to the same input JPTs are distinctly different. The commonality between *TAR* and *F_β-score* leads to my first axiom.

Axiom 1 (Category importance) *An end user efficacious summary statistic must be a function of problem specific impact vector $I = (\iota_{T_+}, \iota_{F_+}, \iota_{F_-}, \iota_{T_-})$, where each element of $I \in \mathbb{Q}$.*

A summary statistic that complies with Axiom 1 is sensitive to I. Thus end users can tune the summary statistic output to match their particular problem.

None of the summary statistics I reviewed in Chapter 5 satisfy Axiom 1. The *F_β-score*, conditioned by β, provides some ability to incorporate impact.

However, since $\iota_{T_-} = 0$ regardless of β, it fails. The other summary statistics considered, *TAR*, Youden index, *ROC-AUC*, *DOR/DP*, *MCC* and *IC* do not have a provision for setting impacts. Instead, all have fixed impacts: implicitly, $I = (1, 1, 1, 1)$.

8.1.1 Practical considerations and implications

The observations made in Chapter Section 6.2.1 are directly applicable to this Axiom.

8.2 Axiom 2, Environmental sensitivity

In order to receive the greatest benefit from a deployed CPD, end users need to identify and use the optimum boundary, as well as be able to estimate the optimum impact and boundary sensitivity in their particular situation. In Chapter 5, I evaluated summary statistic sensitivity to pdf; in Chapter 6.1.3, I explored how the optimum boundary is sensitive to the $ratio_+$. For the type of CPD problems under consideration, optimum boundary is sensitive to both $ratio_+$ and class distribution shapes (pdf). Usually, pdfs are considered to have equal areas under their curves, however, since end users are sensitive to $ratio_+$, I use the term "pdf" to indicate the problem appropriate relationship, either observed, or normalized. It may be possible to identify a CPD problem where the optimum boundary is not sensitive to input class pdfs. However, I contend that one end user need is to identify an appropriate classifier boundary for their problem. Consider an end user environment with two classes, defined by $f(\mu, \sigma)$ and $g(\mu', \sigma')$, where μ and μ' are the distribution means and σ and σ' are the distribution standard deviations. Let the end user have optimum boundary B^*. If I change both distributions by adding $\Delta\mu$ to μ and μ', then the optimum boundary becomes $B^* + \Delta\mu$ and ι remains constant. There are myriad permutations that can be made to this simple end user environment. Most will affect the optimum boundary and/or ι; some will not. The point is that class distribution invariance is not an end-user-efficacious summary statistic characteristic; end users are better served by summary statistics that are class distribution sensitive.

As defined by Schaeffer, a pdf can be interpreted as the long-run relative frequency of occurrence of outcomes resulting from a random experiment (in my case, an experiment using test sample S drawn from $pdf(S)$) [73]. From such a pdf, the *expectation* (E), or most common outcome can be calculated. From the category definitions and Table 1 in Chapter 1, it can be seen that T_+ and F_- are both unique subsets of Y; $\overline{Y}$ is similarly partitioned into unique subsets T_- and F_+. Table 1 also shows that $S = Y \cup \overline{Y}$. The classification of each element of S into Z and $\overline{Z}$ is determined by the boundary (B) used by the target *CPD*. Thus for any given *CPD* processing multiple Ss, there will exist a long-run expected object

count for each JPT category ($E(JPT)$). $E(JPT)$ is a function of boundary and $pdf(S)$ (alternatively expressed as B, $pdf(Y)$ and $pdf(\overline{Y})$). I will present an example where $pdf(S)$ is changed by varying $pdf(Y)$. The example applies to $pdf(S)$, regardless of whether $pdf(Y)$, $pdf(\overline{Y})$, or both vary.

1. For every element in a test set, regardless of ground truth, a CPD boundary can be configured such that the element under evaluation can be classified as Z. Similarly, a boundary can be configured such that the same element can be classified as $\overline{Z}$.

2. Y is an ordered set, such that a boundary change will cause the CPD to reclassify elements of Y in order.

3. For each element of Y in T_+, a boundary, B_k, exists such that $\{y_1, y_2, \cdots, y_k\} \in T_+$, but $\{y_{k+1}, y_{k+2}, \cdots, y_{|Y|}\} \in F_-$. For simplicity, I discuss a single boundary. In reality, a family of boundaries could exist that satisfy the criterion.

4. Therefore, an ordered set of boundaries also exists where there is a one-to-one mapping between the elements of Y and $\{B_1, B_2, \cdots, B_{|Y|}\}$. Thus, for all $i <= |Y|$, when the subject CPD is using boundary B_i ($CPD(B_i)$), $t_+ = i$ and $f_- = |Y| - i$.

5. Now let there be two test set elements, $(y_m, y_{m+n}) \in Y$, where m and n are arbitrary. Each element is defined by vectors found on ordered boundaries B_m and B_{m+n}, with $B_m < B_{m+n}$. The boundary used (B_k) effects y_m's and y_{m+n}'s classification:
   When $k \geq m + n$, $CPD(B_k \in \{B_{m+n}, B_{m+n+1}, \cdots, B_{|Y|}\})$ will classify both
   y_m and y_{m+n} as Z: $\{y_m, y_{m+n}\} \subset T_+(B_k)$.
   When $m + n > k \geq m$, $CPD(B_k \in \{B_m, \cdots, B_{m+n-1}\}$ will label y_m as Z (JPT category T_+) and y_{m+n} as $\overline{Z}$ (JPT category F_-):
   $y_m \in T_+(B_k)$ and $y_{m+n} \in F_-(B_k)$.
   When $k < m$, $CPD(B_k \in \{B_1, B_2, \cdots, B_{m-1}\})$ will classify both
   y_m and y_{m+n} as $\overline{Z}$ (JPT category F_-): $\{y_m, y_{m+n}\} \subset F_-(B_k)$.

6. After executing a sufficiently large number of experiments, I can determine expected cardinality for each JPT category within each boundary interval; $E(t_+|B_k)$ and $E(f_-|B_k)$.

7. Using the expectations for each JPT category for a specific boundary, I can also determine the expected summary statistic value: $E(SS|B_k)$.

A summary statistic valuable to end users will be sensitive to changes in the end user's environment. $Pdf(S)$ represents the end user's environment. Thus, I show how such a change will be reflected in a suitably sensitive summary statistic.

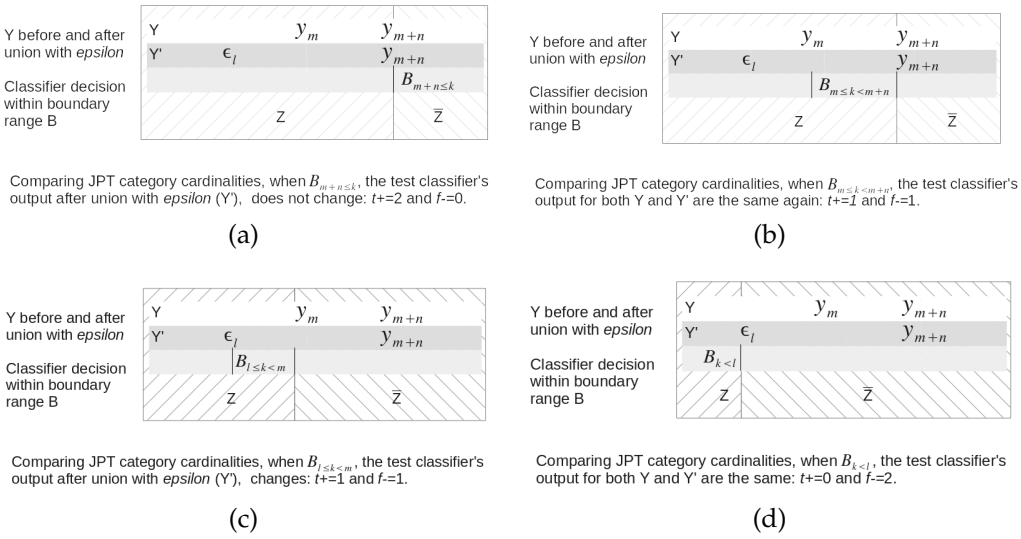


Figure 8.1: As the boundary shifts from left to right, the difference between Y and Y' are reflected in the JPT categories. An end user efficacious summary statistic will be sensitive to these changes.

If Y is replaced by Y', a set drawn from a source population with a probability density function different from $pdf(Y)$, the difference between the two can be described by $\Delta = pdf(Y') - pdf(Y)$. Because the difference can be negative across some interval where the pdfs are defined, Δ is not a pdf. I can, however, draw samples from Δ and create a set, *epsilon*. Set elements drawn from negative regions of Δ are indicated by ϵ^-, set elements drawn from positive regions of Δ are indicated by ϵ. The difference between *epsilon* and a conventional set such as S or Y, is that the interaction between elements in *epsilon* and S is destructive for ϵ^-: $\{s_l\} \bigcup \{\epsilon_l^-\} = \emptyset$; s_l indicates ϵ_l^-'s nearest neighbor. The interaction between elements in *epsilon* and S is constructive for *epsilon*'s elements within a boundary interval where Δ is positive: $\{s_k\} \bigcup \{\epsilon_k\} = \{s_k, \epsilon_k\}$. One consequence of this behavior is that individual elements of Δ will cause $|S|$ to change. I am considering a finite sample of size $|S|$, so Δ must have an equal number of ϵ and ϵ^-.

1. Given Y and its subset $\{y_m, y_{m+n}\} \in Y$ from the previous discussion, let there be a Δ from which a small representative sample, *epsilon*$= \{\epsilon_l, \epsilon_{m+n}^-\}$ is drawn. ϵ_{m+n}^- is from a range in Δ that is negative, hence $\{y_{m+n}\} \cup \{\epsilon_{m+n}^-\} = \emptyset$. The resulting union of Y and *epsilon*, generates $\{\epsilon_l, y_m\} \in Y'$ where $|Y'| = |Y|$. For this example, I stipulate that $\epsilon_l < y_m$, thus, the boundaries for this example have the same relationship: $B_l < B_m < B_{m+n}$.

2. I now consider the effect of the change on the JPT generated from processing Y', as B_k varies across the interval $[B_{Y'(1)}, B_{Y'(|Y'|)}]$. The four cases are illustrated on Figure 8.1.

a) When $k \geq m+n$, both elements are correctly classified as Z and the JPTs for S and S' are the same: $(\epsilon_l, y_m) \in T'_+(B_k)$; $t_+(B_k|Y') = t_+(B_k|Y)$.[1]
b) When $m+n > k \geq m$, both elements are correctly classified as Z but the JPTs for S and S' differ: $(\epsilon_l) \in T'_+(B_k)$; $t_+(B_k|Y') = t_+(B_k|Y)$.
c) When $m > k \geq l$, only ϵ_l is correctly classified as Z; the JPTs for S and S' differ:$\epsilon_l \in T'_+(B_k)$ and $y_{m+n} \in F'_-(B_k)$; $t_+(B_k|Y') = t_+(B_k|Y) + 1$.
d) When $l > k$, both elements are misclassified as $\overline{Z}$ and the JPTs for S and S' are the same: $(\epsilon_l, y_{m+n}) \in F'_-(B_k)$; $t_+(B_k|Y') = t_+(B_k|Y)$.

After executing a sufficiently large number of experiments, I can determine $E(t_+(B_k|Y'))$ and $E(f_-(B_k|Y'))$ within each boundary interval. I see that within the boundary range $[B_l, B_{m+n}]$ defined by the pdf perturbation $\Delta = pdf(Y') - pdf(Y)$, $Y' \cap Y \neq \emptyset$; the expectation for some JPT categories will differ from the unperturbed JPT expectations. For clarity, I have limited the $pdf(Y)$'s perturbation to substituting one element; ϵ_l replaces y_{m+n}. The perturbation could be of any size; however, the discussion would be substantially more involved.

An efficacious summary statistic will reflect those differences. Tables 8.1 and 8.2 summarize the result of the described pdf change.

Thus, my second axiom for end user efficacious summary statistics (SS) is:

Axiom 2 (Environmental sensitivity) *With a change in* pdf*(Y),*
(e.g., pdf*(Y′)*= Δ+pdf*(Y),* pdf*(S))*=pdf*(Y)*+ pdf*($\overline{Y}$) and* pdf*(S′)*=pdf*(S)*+Δ*), where* Δ *describes a perturbation in Y's and S's source population, for all boundaries within* Δ*, there exists a* $E(\text{SS}(B)|S')) - E(\text{SS}(B)|S)) \neq 0$*. The same is true for a change in* $\overline{Y}$ *and for any* ratio_+.

For any given boundary within the interval affected by a probability distribution change, an effective summary statistic's expected output will reflect that probability distribution change. The input class pdfs of any environment can vary by user. Thus, an optimum boundary for one end user may not be optimum for another. A summary statistic compliant with Axiom 2 will reflect how the target CPD impacts the end user when provided with different inputs. Each of these points are effects of boundary sensitivity.

Of the summary statistics I reviewed in Chapter 5, only *TAR*, *MCC* and *IC* comply fully. *AUC* fails to satisfy Axiom 2: it is boundary invariant. The Youden index and *DOR/DP*, being $ratio_+$ invariant, fail for $ratio_+$ sensitive problems. The F_β*-score* fails because of its T_- invariance.

[1] At this point, I have extended my notation. In the previous paragraph, the discussion regards only one input distribution, so there was need to only distinguish between JPTs created using different boundaries. Hence, the notation presents category membership when the CPD uses a specific boundary: $X(B_k)$, where X indicates the JPT category and B_k represents the boundary used. However, I now start comparing JPTs with inputs drawn from different source populations. To do so, I compare the JPT category cardinalities. $t_+(B_k|Y')$ refers to the cardinality of the T_+ category, when B_k was the boundary and the source population was $pdf(Y')$.

Effect of a change in pdf (Y versus Y') on $E(t_+(B_k))$			
B_k	$(y_m, y_{m+n}) \in Y$?	$(\epsilon_l, y_m) \in Y'$?	$E(t_+(B_k\|Y')) - E(t_+(B_k\|Y))$
$k \geq m+n$	$(y_m, y_{m+n}) \in Y$	$(\epsilon_l, y_m) \in Y'$	0
$m+n > k \geq m$	$(y_m) \in Y$	$(\epsilon_l) \in Y'$	0
$m > k \geq l$	$() \in Y$	$(\epsilon_l) \in Y'$	1
$k < l$	$() \in Y$	$() \in Y'$	0

Table 8.1: Changes in class pdfs effect JPT categories. In the Axiom 2 example, $E(t_+(B_k))$ is affected by a pdf change within the interval $[B_l, B_m)$. Outside that interval, $E(t_+(B_k))$ remains unaffected.

Effect of a change in pdf (Y versus Y') on $E(f_-(B_k))$			
B_k	$(y_m, y_{m+n}) \in Y$?	$(\epsilon_l, y_m) \in Y'$?	$E(f_-(B_k\|Y')) - E(f_-(B_k\|Y))$
$k \geq m+n$	$(y_m, y_{m+n}) \in Y$	$(\epsilon_l, y_m) \in Y'$	0
$m+n > k \geq m$	$(y_m) \in Y$	$(\epsilon_l) \in Y'$	0
$m > k \geq l$	$() \in Y$	$(\epsilon_l) \in Y'$	-1
$k < l$	$() \in Y$	$() \in Y'$	0

Table 8.2: The t_+ increase noted in Table 8.1 triggers the corresponding f_- decrease shown here. An end user efficacious summary statistic should be sensitive to such changes.

8.2.1 Practical considerations and implications

Summary statistics that are environmentally sensitive, confer three benefits to end users. i) They provide end users with optimum boundary estimates. ii) They provide end users with estimates of actual impact under their field conditions. iii) They exhibit boundary sensitivity. This is important because CPD inputs are stochastic in nature, so over any particular observation window, end user's actual experience will vary. Boundary sensitive measures allow end users to estimate the range over which to expect their actual outcomes to occur. Sensitivity results are shown in both the banking and medical diagnostic test examples.

$ratio_+$ sensitive summary statistics have broader applicability than measures that are $ratio_+$ invariant. The observations made in Chapter Section 6.1.4 are directly applicable to this Axiom.

8.3 Axiom 3: CPD output basis

End users have limited visibility into a deployed CPD's process. For instance, they have knowledge of the inputs and outputs, but not their ground truth. Thus, end users may find CPD evaluation summary statistics that can be calculated from, or decomposed into Z and $\overline{Z}$ more useful than others.

Consider an end user desiring to prevent intrusions (unauthorized access and activity) into an information system. Ideally, these end users would be able to detect and prevent all attempted system intrusions. Unfortunately, the only CPD information directly observable by end users is the fallible output of their intrusion detection and prevention systems (Z and $\overline{Z}$). The end user needs to assess the impact of the CPD output on their problem. Typical end user questions might be:

a) "Given that the test result is positive, what is the expected impact (ι_Z)?"

$\overline{a}$) "Given that the test result is negative, what is the expected impact ($\iota_{\overline{Z}}$)?"

(End users also have knowledge regarding their problem environment to assist in calculating impact. This, however, is independent of the CPD output.) In the field, end users will have access to only the test results (Z and $\overline{Z}$). Quantitative answers to a and $\overline{a}$ provide readily actionable information for deployed systems.

In contrast, answering the questions:

b) "Given that ground truth is Y, how probable is it that the test result is positive ($P(T_+|Y)$)?"

$\overline{b}$) "Given that ground truth is $\overline{Y}$, how probable is it that the test result is negative ($P(T_-|\overline{Y})$)?"

may be less readily usable, since it assumes knowledge of the ground truth – often, the information end users don't know and are trying to obtain. This is the situation in intrusion detection and prevention. f_- and t_- cannot be known with certainty, thus it is impossible to determine $|Y|$ and $|\overline{Y}|$. Questions b and $\overline{b}$ are conditioned by $|Y|$ and $|\overline{Y}|$. If these two variables are unknown, then answers to questions b and $\overline{b}$, although estimable in a supervised test, may have little value to end users. Because ground truth is unknown in the field, applying answers to these two questions may be difficult to calculate for deployed systems. Answers to questions a and $\overline{a}$, however, are estimable from field observations. Because they can be directly mapped to field results, end users may find a and $\overline{a}$ easier to use, thus more valuable than b and $\overline{b}$.

One advantage Axiom 3 confers to end users is the ability to compare predicted outcomes to field observations. Assume that intrusion detection and prevention is a "$ratio_+$ is important" problem. An end user can gather information on t_+ and f_+ and verify whether or not the expected impact of Z is satisfactorily close to the observed impact. However, if the observed and actual Z impacts are not acceptably close, then either the end user's expected $ratio_+$ is incorrect, or the intrusion detection system is not performing as expected. An end user may be able to use the discrepancy to justify an in-depth investigation.

If the actual and expected Z impacts *are* acceptably close, an end user can then impute the expected f_-. The expected f_- quantifies the missed attack rate,

thus the end user can project an expected impact for the missed attacks. This information can then be used for decisions regarding risk reduction options such as adding detection or planning for remediation.

In the case of a "$ratio_+$ is confounding" problem such as a medical diagnosis, efficacious end user information is based on normalized test results; $|Z_n|$ and $|\overline{Z_n}|$ are not directly observable. However, $E(|Z|)$ and $E(|\overline{Z}|)$ can be calculated: $E(|Z|/|S|) = t_{+n}/|S| + ratio_+ * f_{+n}/|S|$ and $E(|\overline{Z}|/|S|) = f_{-n}/|S| + ratio_+ * t_{-n}/|S|$. Thus an end user can, using their observed results, calculate the apparent $ratio_+$, if a CPD application vendor publishes $(t_{+n}/|S|, f_{+n}/|S|, f_{-n}/|S|, t_{-n}/|S|)$. The end user can then compare the apparent $ratio_+$ against their expectations and decide whether or not the apparent $ratio_+$ is reasonable. An end user may want to further investigate an apparently anomalous $ratio_+$.

Given the proper summary statistic, end users can better assess their CPD options and monitor CPD effectiveness. These end user visibility observations lead to the third Axiom.

Axiom 3 (CPD output basis) *An end user efficacious summary statistic must be quantifiable in terms relative to information known and visible to the end user (Z and $\overline{Z}$).*

As noted in Chapter 6.1, I find that the $ratio_+$ invariant summary statistics, *ROC-AUC* and the Youden index have as their basis, the summary statistic suite $\{t_+/|Y|, f_+/|\overline{Y}|\}$. Both of these summary statistics are conditioned by ground truth (Y and $\overline{Y}$), not the CPD outputs visible to the end user (Z and $\overline{Z}$), hence they do not satisfy Axiom 3. However, the $ratio_+$ invariant summary statistic pair *DOR/DP* do satisfy Axiom 3. I can demonstrate Axiom 3 compliance by substituting the four conditional ratios

$$\frac{t_+}{|Z|}, \qquad \frac{f_+}{|Z|}, \qquad \frac{f_-}{|\overline{Z}|}, \qquad \frac{t_-}{|\overline{Z}|} \tag{8.1}$$

for t_+, f_+, f_- and t_- in the summary statistics. In the *DOR*. equation, I have

$$DOR = \frac{\frac{t_+}{|Z|} * \frac{t_-}{|\overline{Z}|}}{\frac{f_+}{|\overline{Z}|} * \frac{f_-}{|Z|}}.$$

Multiplying the numerator and denominator by $|Z| * |\overline{Z}|$, results in

$$DOR = \frac{t_+ * t_-}{f_+ * f_-},$$

the original *DOR*. equation. As noted in Chapter 5, $DP = \frac{\sqrt{3}}{\pi} log(DOR)$, thus *DP* also satisfies Axiom 3.

If I use the same substitution applied in *DOR*. into the equations for *TAR*, *F_β-score*, *ROC-AUC*, *MCC* and the *IC*, I find that none are equivalent to the original equations. Of the commonly seen summary statistics considered, only the *DOR/DP* satisfies Axiom 3.

8.3.1 Practical considerations and implications

This capability is indeed new and the information generated is potentially valuable to end users. For instance, in the banking example, prudent institutions will be tracking performance indicators for loans made (perhaps average income per loan) and loan applications rejected (perhaps total application processing cost) and application acceptance ratio. Summary statistics that comply with this Axiom will provide bankers information on the expected performance ratio values. If the expected and actual values differ substantially, the lending institution may want to more closely examine their processes and CPD tool.

A similar situation exists with medical diagnoses CPDs. The RA example includes expected patient impacts for both positive and negative test results. As noted above, these values can be used to calculate the apparent $ratio_+$ value. If these differ from the observed value significantly, then medical practitioners can more closely examine their processes and diagnostic tool.

8.4 Axiom 4: summary statistic value appropriateness

In order for a summary statistic to be end user efficacious, the end user must be able to map the summary statistic output to their problem. The end user concern is to avoid the inverse of "type III" errors (receiving the right answer to the wrong question [46]); having to make a decision, given information on an unrelated question. An informed end user may know what the values presented quantify. However, if the values are not relevant to their problem and cannot be transformed into values that are applicable, the end user must rely on "soft" transformations, such as expert opinion. Opinions can vary from expert to expert, so soft transformations incur considerable uncertainty. As the saying goes "your results may vary".

Consider the intrusion detection and prevention problem presented in Section 8.3. In selecting an intrusion detection and prevention program, an information system administrator may want to compare the cost of operation of each program. F_β*-score* provides some ability to adjust for cost. However, it fully discounts the T_- category, this bias skews the result. The F_β*-score* and *TAR* are ratios, so they cannot provide net costs. Instead, they provide relative costs. In the general case, the information system will not operate in an environment where the volume of malicious activity is approximately the same as the volume of legitimate activity, so the $ratio_+$ invariant summary statistics, *ROC-AUC*, *DOR/DP* and Youden Index are not suitable. *MCC* and *IC* summary statistic basic CPD characteristics, but their output cannot be mapped to the system administrator's need; these summary statistic values are not appropriate either. In order to use these summary statistic values, the end users may attempt a "seat of the pants" transformation, based on experience. Such results are subjective and may vary from expert to expert. End users are better served when subjective transformations can be avoided; when the information received answers the correct question.

Axiom 4 (summary statistic value appropriateness) *An end user efficacious summary statistic output must quantify the CPD's impact on the end user's characteristic of interest.*

While Axiom 4 might seem self-evident, not all commonly seen summary statistics satisfy it. For example, the *ROC-AUC* quantifies the probability that a randomly selected member of class Y will have a lower test value than a randomly selected member of class $\overline{Y}$. Thus the *ROC-AUC* value would be difficult for an end user to use directly, since it assumes prior knowledge of ground truth, a situation not likely to exist in the field.

DOR/DP quantify the odds of two randomly selected elements of the test set being one each T_+ and T_-, rather than one each F_+ and F_-. *DOR/DP* do not require prior knowledge of ground truth, but it is a very specific scenario. One limitation is that it requires output pairs, rather than considering individual outputs. A second limitation is that there are ten possible pairings (e.g., two T_+s) and ninety unique and potentially useful ratios. (When counting, I do not consider ratio inverses to be unique. A useless ratio would be one where the numerator and denominator are the same, for example t_+t_+/t_+t_+.) Thus, the *DOR/DP* output is not broadly applicable.

From the end user perspective, the *ROC-AUC*, *DOR/DP*, *TAR* and *F_β-score* all share another failing; all have lower bounds of zero, thus cannot quantify a negative impact.

8.4.1 Practical considerations and implications

The bottom line for Axiom 4 is that the units in which evaluation output is quantified must be relevant to the end user. For example, the banking example uses a "standard loan income" unit. This value can be mapped to local currency. The original work was in Egyptian pounds. Such a value would be readily understood by persons in the Egyptian banking industry.

The rheumatoid arthritis example quantifies the mis-diagnosis impact on quality of life in U.S. dollars. Once again, results explained in these terms will be broadly understood. Patients, for example, could readily understand such values, whereas probabilities and likelihood ratios would be less familiar.

8.5 Preconditions from measurement theory

At a more abstract level, measurement theory has addressed end user summary statistic efficacy [36]. One relevant insight is that numbers are used in different ways and that these uses constrain their information content and hence, their utility. I use the scale-type definitions proposed by Stevens [83]. Stevens

defined four scale types, nominal, ordinal, interval and ratio. Ratio scales are the most information rich and have the least functional constraints, so summary statistics using ratio scales are preferred.

8.5.1 Practical considerations and implications

Re-analysis of the bank example and medical diagnoses examples both output currency values. Currency scales have both a meaningful zero and are quantified with a standard sequence. In both examples, the information generated with summary statistics results on ratio scales demonstrate the benefit of using ratio scales.

Of the summary statistics reviewed, none are measured on a ratio scale. Hence, I find that none of the commonly seen summary statistics satisfy all of the criteria. Table 8.3 recaps how each summary statistic tested conforms to the axioms and two ratio scale properties, having a meaningful zero an being standard sequences. Sokolova, et al. tested invariance to JPT perturbations [79]. Where their tests are relevant to my axioms, their results corroborate ours.

8.6 Relevance for research

Over time, useful technical concepts evolve into useful tools and project underwriters shift from basic researchers to developers and end users. To manage technology development and product acquisition risk, NASA developed an assessment framework, based on a nine level "technical readiness" (TRL) scale. TRL has been subsequently adopted by the US Department of Defense (DoD) [22, 56]. Early technology development fleshes out the technical concepts. The lack of insight means that assessments are generally qualitative. By TRL 3, however, the technology is well enough defined that quantitative assessments are

	Summary statistic						
	TAR	F_β*-score*	Youden	*MCC*	*IC*	*DOR*	AUC
Axiom 1: category impacts	No	Partial	No	No	No	No	No
Axiom 2: pdf sensitivity	Yes	Yes	Yes	Yes	Yes	Yes	No
Axiom 3: perspective	No	No	No	No	No	Yes	No
Axiom 4: relevance	No	No	No	No	No	No	No
Ratio: meaningful zero	No	No	Yes	Yes	Yes	No	No
Ratio: standard sequence	No	No	No	No	No	Yes	No

Table 8.3: None of the summary statistics considered satisfy all four axioms, nor do any exhibit both of the characteristics necessary to be ratio scale measures; having a meaningful zero and being standard sequences.

possible. At this stage, experimental critical function and/or characteristic proof of concept can be executed. While some specific problem domains may have been proposed, initial critical function and/or characteristic assessments should be application agnostic.

Challenges for basic research CPD assessment are that i) the source population ($Y \bigcup \overline{Y}$) for any test set (and hence the test set itself) will always have a $ratio_+$ and pdf. ii) all summary statistics weigh their variables. In the example presented in Chapter 6.2 (Equation 6.2), *TAR* has implicit weights (Equation 6.3) and F_β*-score* has explicit weights (Equation 6.4). Taking the Bayesian perspective, the choices should reflect the lack of prior information: $ratio_+ = 1$, $I = (1, 1, 1, 1)$ and source populations with uniform pdfs.

This dissertation addresses end user interests (TRL$\geq$ 4), so I treat $ratio_+$ as appropriate, use the commonly seen normal distribution and impacts as appropriate ($I = (1, 1, 1, 1)$ for non-specific evaluation; for the examples, where available, published I were used).

CHAPTER 9

TWO AXIOM COMPLIANT SUMMARY STATISTICS

Applying TRIZ, I have identified the root cause of the end user's CPD evaluation problem (the starting point), identified summary statistic characteristics valuable to end users (the desired end point) and the solutions' dependent variables. The final task is to devise a solution.

9.1 Measuring impact

As noted in Chapter 5 and supported in Chapter 8, the commonly seen CPD evaluation summary statistics do not well quantify end user impact. In this chapter, I propose suitable summary statistics. I have identified two problem types, so I consider each separately.

The efficacious CPD evaluation summary statistic discussion makes some assumptions regarding the end user's need:

- The end user's problem either treats each input set element individually (as in the case of a medical diagnosis or intrusion detection) or the problem is based on the cumulative effect of the elements in the input stream (as in the case of bank loan application decisions or information retrieval).
- The problem does not involve a multiplicative or exponential effect. Such problem domains do exist, but must be treated as additive problems. Consider an information retrieval task gathering all of the information known about topic x. As information sources are identified, there will most likely start to be information overlaps, thus individual resource contributions to the body of knowledge gradually diminish. After some point, additional resources add no new information, thus have zero value. Ideally, it would be useful to be able to rate the contribution of each resource. However, there are difficulties:
  - The valuation of each resource is sequence dependent. Assuming resources are randomly selected for evaluation, the same resource, if an early selection, could have no duplicate information, thus retain its full value. If it is a late selection, it could be valueless.

– The end user cannot precisely know the full extent of the information. Thus, even though it appears that the end point is being approached, the possibility exists that the next resource evaluated will be totally unique. In such a case, the resource's value would not be discounted.

In such a scenario, it may not be practical for an end user to identify the uniqueness of each individual resource. Relevance, the basis of information retrieval, however, is based on key word analysis and has been demonstrated to be practical. It is an additive function and thus avoids the two difficulties presented.

- All CPD events are independent. Sequence dependence is discussed in the previous bullet.
- The problem is restricted to a 2x2 matrix. (The matrix size constraint is for discussion clarity. Extension to MxM matrices is deferred for future work.).

When the impact is cumulative, there will be either a gain or loss (impact, ι) associated with each element output. ι can be expressed as a statistical, not necessarily unique, expectation for each JPT category: $I = (\iota_{T_+}, \iota_{F_+}, \iota_{F_-}, \iota_{T_-})$. Thus, each and every element output will affect the end user by the element of I applicable to the category to which the element is binned[1]. An end user can expect the net gain or loss to be the sum of the individual element gains and losses. If the elements of I are proportions as described in the Lexicon and there is a sufficiently large (i.e., statistically significant) test sample ($|S|$), then ι can be viewed in the contexts introduced in Section 6.2. As noted in that section, impacts can also be expressed as a statistical expectation for each element of S, regardless of category. For problems where impact is cumulative,

$$\iota_I = \iota_{T_+}\frac{t_+}{|S|} + \iota_{F_+}\frac{f_+}{|S|} + \iota_{F_-}\frac{f_-}{|S|} + \iota_{T_-}\frac{t_-}{|S|}. \tag{9.1}$$

ι_I can also be expressed on Z and $\overline{Z}$, the outputs actually observed by the end user:

$$\iota_I = \iota_Z\frac{|Z|}{|S|} + \iota_{\overline{Z}}\frac{|\overline{Z}|}{|S|} \tag{9.2}$$

The estimated total impact for the test set S, then is:

estimated total impact : $\iota_{tot} = |S|\iota_I$.

[1]The elements of I can be defined in different ways, depending upon the information known about each element of S. For instance, if elements were bank loan applications, then the impact could be measured per dollar requested. Alternatively, impact could be based on the statistical expectation for the category. Once again using bank loan applications, the impact could be per loan, where an average loan amount is known. I assume the latter situation.

A similar summary statistic, "Profit", has been introduced for customer churn prediction models by Verbraken, et al. [92]. It differs from ι_I in two ways: *i*) Profit's costs and benefits must consist of all positive values and *ii*) misclassification costs are deducted, correct classification gains are added. Intuitively, Verbraken et al.'s constraints seem correct: gains are positive values and losses are negative values. As will be seen in Section 10, intuition does not hold in every case. Recasting a problem to fit Profit's requirements could cause the measurement scale to no longer have a meaningful zero. Should this occur, the analysis is now using an interval scale rather than a ratio scale. Consequently, analyses such as Verbraken, et al.'s proposed cost-benefit ratio would no longer be valid. ι_I, as seen in Equation 9.1, does not have Profit's constraints, so is not susceptible to summary statistic degradation.

There are occasions when the impact is not cumulative, but each output is important individually, e.g., a medical diagnosis. The individual result is important. In these situations, $ratio_+$ is confounding, so normalized JPTs are used; normalization mathematically balances the relative class sizes, thus it mitigates any skew resulting from $ratio_+$. In order to facilitate comparison with non-normalized JPTs, the sum of all categories is kept at one ($|S| = 1$). Another commonly seen normalized JPT would have $|S| = 2$. In this table, the JPT category proportions are real numbers that sum up to one and the individual input class totals add up to 0.5.

		Actual classification		
		Y	$\overline{Y}$	*Totals* $\downarrow$
Test	$+ : s_i \in \{Z\}$	$t_{+n} = \frac{t_+}{2\|Y\|}$	$f_{+n} = \frac{f_+}{2\|\overline{Y}\|}$	$\|Z_n\|$
Result	$- : s_i \notin \{Z\}$	$f_{-n} = \frac{f_-}{2\|Y\|}$	$t_{-n} = \frac{t_-}{2\|\overline{Y}\|}$	$\|\overline{Z_n}\|$
Normalized totals		0.5	0.5	1

Table 9.1: The values in this JPT have been normalized.

The end user's concern, a) "given that a result is rendered, how am I affected?", however, can be partitioned into two questions;
$a1$) "given that the result is positive, how am I affected?" and
$a2$) "given that the result is negative, how am I affected?"[2]. These questions indi-

[2]Partitioning based on CPD output is not unique to situations where outputs are important individually. Problems where CPD results are cumulative can be partitioned in the same way. However, when results are cumulative, question a is the most useful for an end user; questions $a1$ and $a2$ may be of secondary importance. When individual outputs are important, questions $a1$ and $a2$ are primary. They are directly applicable by an end user to individual results; question a is less useful.

cate that the values of interest are weighted conditional expectations:

$$\iota_Z = \frac{\iota_{T_+} t_{+n} + \iota_{F_+} f_{+n}}{|Z_n|} \quad \text{and} \quad \iota_{\overline{Z}} = \frac{\iota_{T_-} t_{-n} + \iota_{F_-} f_{-n}}{|\overline{Z_n}|}. \tag{9.3}$$

Because each output is independent, the average of these two values provides the expected impact (ι) per outcome:

$$\iota_\sigma = \frac{\iota_Z + \iota_{\overline{Z}}}{2}. \tag{9.4}$$

Substituting the normalized expressions from Table 9.1, the expected impact becomes:

$$\iota_\sigma = \frac{1}{2}\left(\frac{\iota_{T_+} t_{+n}}{|Z_n|} + \frac{\iota_{F_+} f_{+n}}{|Z_n|} + \frac{\iota_{T_-} t_{-n}}{|\overline{Z_n}|} + \frac{\iota_{F_-} f_{-n}}{|\overline{Z_n}|}\right). \tag{9.5}$$

The summary statistic and associated suite are valuable to end users. Those directly affected (e.g., medical patients) may find the normalized JPT values and I more informative.

The two problem types (individual impact and cumulative impact) have their unique characteristics, resulting in different sets of relevant summary statistics. For problems where impact is cumulative, the summary statistic,

$$\iota_I = \frac{1}{|S|}\left(\iota_{T_+} t_+ + \iota_{F_+} f_+ + \iota_{F_-} f_- + \iota_{T_-} t_-\right), \tag{9.6}$$

provides actionable information to the end user. The monotonic summary statistics upon which it is based and which provide more insight into CPD impact are the two end user visible values:

$$\iota_Z = \frac{1}{|S|}\left(\iota_{T_+} t_+ + \iota_{F_+} f_+\right) \text{ and } \iota_{\overline{Z}} = \frac{1}{|S|}\left(\iota_{F_-} f_- + \iota_{T_-} t_-\right). \tag{9.7}$$

The factors addressed in developing ι_I and ι_σ satisfy the underlying model requirement for the highest quality summary statistics, as stated in Chapter 7.3.

Impact vectors are problem specific, but there may be occasions when some problems have Is that differ by a multiplicative constant. Such problem sets can be considered problem families, represented by a single I with different scaling. Appendix D presents a means of scaling.

Both ι_I and ι_σ have optima, so they are summary statistics. In the case of ι_I, the associated suite (summary statistics that provide insight into a specific aspect of the CPD output) consists of ι_{Z_I} and $\iota_{\overline{Z}_I}$, the two outputs observable by end users. ι_I differs from the usual summary statistic in that it directly quantifies the characteristic of interest to end users. ι_σ's summary statistic suite consists of the conditional expectations of the two outputs observable by an end user, ι_{Z_σ} and $\iota_{\overline{Z}_\sigma}$. For problem domains where ι_σ is appropriate, the summary statistic does

contain less information. Consider, for instance, the example where a person receives a medical diagnosis. If the test result is positive, then that person's impact will be *either* ι_{T+} *or* ι_{F+}. Likewise, if the test result is negative, then that person's impact will be *either* ι_{T-} *or* ι_{F-}. The three composite summary statistics, ι_σ, ι_{Z_σ} and $\iota_{\overline{Z}_\sigma}$ may have little utility for the patient. The values are, however useful to diagnosticians in assessing diagnostic and treatment strategies.

Interestingly, ι_I is, in a sense, already in use. It is implicit in *TAR*: *TAR*= $\iota_I|I = (1, 0, 0, 1)$. Equation 6.1 shows that J is a rescaled *TAR* on normalized JPTs. Hence, J is also a rescaled $\iota_I|I = (1, 0, 0, 1)$ on normalized JPTs.

9.1.1 Practical considerations and implications

An abstract analysis of ι_I and ι_σ suggest that they will generate more information rich output than other measures currently in use. The four examples in Chapter 10 illustrate the differences. In the bank loan example, not only does the conclusion regarding the best selection algorithm change, but actual income expectations and sensitivity to $ratio_+$ are revealed. The rheumatoid arthritis diagnostic test comparison conclusions on the better test stay the same with both the original and impact-based measures. However, the outputs, being quantified in currency, are much easier to understand.

9.2 Summary statistic usage

There are two end user motivations for measuring CPD impact. One is to directly estimate performance on their problem and in their environment. The summary statistics proposed in this chapter directly provide that utility. CPD output bias, as introduced in Chapter 7.3, is measurable with ι_I and ι_σ.

The second motivation is to compare relative CPD impacts on their problem and in their environment. Intuitively, generating a ratio of two CPD impacts,

$$\iota \; ratio = \frac{\iota(1)}{\iota(2)},$$

would quantify the desired relationship. However, since the values can be either positive or negative, interpretation is problematic. If $\iota(1) > 0$, $\iota(2) > 0$ and $\iota(1) > \iota(2)$, then $\iota \; ratio > 1$ and an end user can infer that CPD 1 is better than CPD 2 by a certain magnitude. If $\iota \; ratio < 1$, then the reverse is true; CPD 2 is better than CPD 1 by a certain magnitude. However, if $\iota(1) < 0$, $\iota(2) < 0$ and $\iota(1) > \iota(2)$, then $\iota \; ratio < 1$, the opposite of when $\iota(1) > 0$ and $\iota(2) > 0$. summary statistic interpretation is not $\iota(n)$ sign invariant. Depending upon conditions, either the larger value or the smaller value can indicate greater effect. Attention must be paid to this detail.

Special care must be taken if the two impacts have different signs. As an example, consider two CPD evaluation scenarios: *i*) $\iota(A) = -1000$ and $\iota(B) = 1$ and $\iota(C) = 1000$ and $\iota(D) = -1$,. In both cases, the ratio is -1000. One might reasonably assume equivalent impact. However, the opposite is true: selecting CPD A instead of CPD B will result in a small gain, rather than a large loss. In the second case, a small loss can be converted into a large gain. The end result on an end user is nowhere near the same; simple ratios may be difficult for an end user to interpret.

The difficulty is a result of potentially having both positive and negative estimated impacts. There is a solution, shifting the output range such that the minimum is zero:

$$\max(\iota_I) = \iota_{T_+} + \iota_{T_-} - \iota_{F_+} - \iota_{F-} \quad \text{and}$$

$$\min(\iota_I) = \iota_{F_+} + \iota_{F_-} - \iota_{F_+} - \iota_{F-} = 0.$$

After the translation,

$$\textit{biased } \iota_I = \iota_I - \iota_{F_+} - \iota_{F-} \text{ and} \tag{9.8}$$

$$\textit{biased } \iota_\sigma = \iota_\sigma - \frac{\iota_{F_+} + \iota_{F-}}{2}. \tag{9.9}$$

Now the simple ratio is

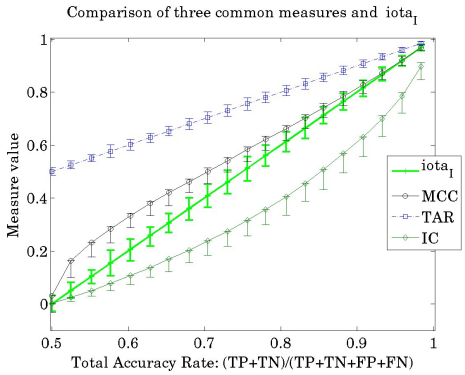
$$\iota \textit{ ratio} = \frac{\textit{biased } \iota(1)}{\textit{biased } \iota(2)}.$$

Equations 9.8 and 9.9 both still satisfy the four Axioms, but for these equations, zero is now defined as "the worst possible output", rather than "no effect". Consequently, the biased impacts do not have meaningful zeros. The solution defeats the purpose: the transformed scales are now interval, rather than ratio scales, so ratios are no longer valid.

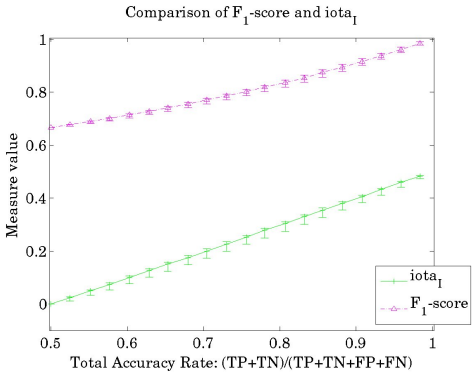
The problem caused by impacts with different signs, however, may be moot. If the impacts being compared have different signs, then one must provide a desired result and the other will provide an undesirable result. In such a case, magnitude is irrelevant: an end user will always choose the CPD which provides a desirable result. This practicality aside, such a ratio quantifies the extent to which one CPD benefits the end user to the extent to which the end user is hindered by the other: the comparison is irrelevant.

9.3 Addition and ratio based summary statistic comparison

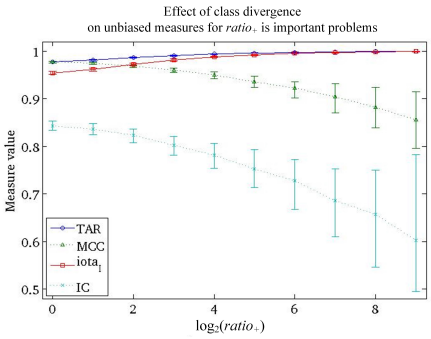
ι_I and ι_σ are addition-based summary statistics that represent the impact of a CPD on an end user; a new perspective on CPD evaluation. This requires incorporating relevant environmental variables that quantify the effect individual CPD outcomes have on the end user. Ratio-based summary statistics (e.g., *TAR*, *F_β-score*, *MCC*, *ROC-AUC*, *DOR*, *Youden index* and *IC*) do not incorporate relevant



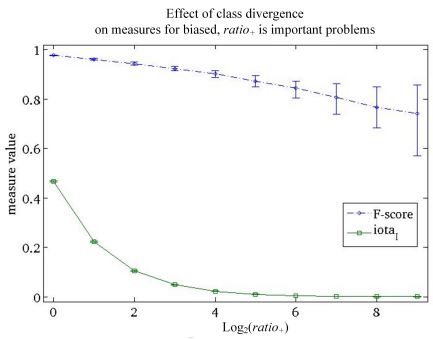
(a) When I is unbiased ($\iota_{T-} = 1$) and $ratio_+ = 1$, ι_I responds linearly to changes in class distinguishability, *MCC* and *IC* respond non-linearly. As expected, they converge when $TAR = 1$.



(b) Only two summary statistics are valid for biased I (e.g., $\iota_{T-} = 0$). Compared to ι_I, F_β*-score*'s response to class distinguishability is non-linear and the summary statistics do not converge when $TAR = 1$.



(c) When $ratio_+$ varies and I is unbiased ($\iota_{T-} = 1$), impact (ι_I) is still most similar to *TAR* . *MCC* and *IC* both decrease as $ratio_+$ increases; ι_I and *TAR* increase.



(d) As $ratio_+$ increases and I is biased ($\iota_{T-} = 0$), impact (ι_I) exhibits exponential decay; F_β*-score* does not.

Figure 9.1: ι_I reflects the target CPD's effect on end users. These figures show that, under the conditions tested, ratio summary statistics do not. Thus, the benefit end users might expect based on ratio summary statistics, may not exist.

environmental variables, hence fail in this regard and are not directly applicable to the end user [3]. ι_I and ι_σ address this deficiency. However, this comparison is not tied to any particular problem or domain, so I use unbiased Is; for the additive summary statistics, $I = (1, -1, -1, 1)$; the ratio summary statistics (excepting F_1-score) implicitly use $I = (1, 1, 1, 1)$. The F_1-score is inherently biased, so I also ran tests with $\iota_{T_-} = 0$.

Figure 9.1 shows the results of comparing ι_I versus summary statistics valid for $ratio_+$ sensitive problems. Two test series were run:

- One series held $ratio_+ = 1$. I randomly drew two equally sized samples from the same probability density distribution, then adjusted one to be offset (exhibit bias) from the other ($pdf(\overline{Z}) = pdf(Z) + offset$. The offsets were selected so that the class overlap ranged from full overlap ($\mu_{\overline{Z}} - \mu_Z = 0$) to full separation ($\mu_{\overline{Z}} - \mu_Z \geq 4\sigma$)).

  TAR is my baseline summary statistic. It is a simple, intuitive, distribution invariant CPD summary statistic. $TAR = 0.5$ means that one-half of the CPD output is correctly classified (the case when two class distributions overlap completely). $TAR = 1.0$ means that all of the CPD output is correctly classified. The top two graphs (Figures 9.1a and 9.1b) show the effect of varying class overlap.

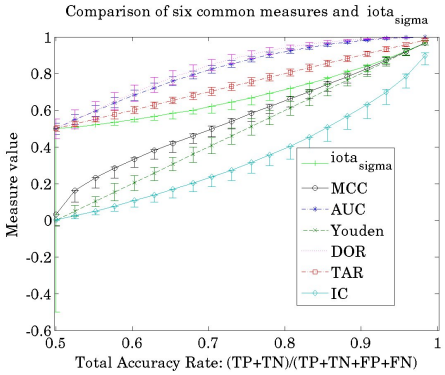
- The other series held a fixed offset ($\mu_{\overline{Z}} \neq \mu_Z$) between classes (the source distribution for class $\overline{Z}$ not the same as the source distribution for class Z) and varying $ratio_+ \in [1, \infty]$. $\log_2(ratio_+) = 0$ indicates that the test sample had equally sized CPD input classes ($|Y| = |\overline{Y}|$). The bottom two graphs (Figures 9.1c and 9.1d) show the effect of varying $ratio_+$.

The left column (Figures 9.1a and 9.1c) are tests on summary statistics when all four JPT categories are unbiased; the right column (Figures 9.1b and 9.1d) shows the effect of bias on summary statistic value; I test when T_- is unimportant ($\iota_{T_-} = 0$).

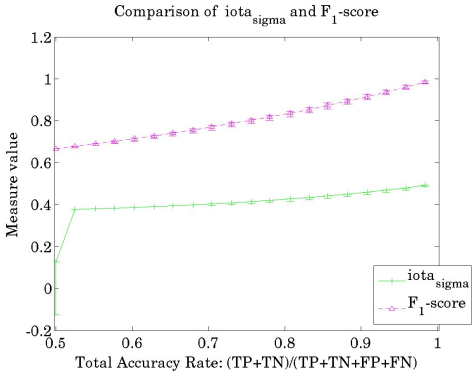
Comparing the effect of varying ι_{T_-} (Figures 9.1a versus 9.1b), I see that when $\iota_{T_-} = 1$ (and $ratio_+ = 1$), ι_I converges with the other applicable summary statistics as *TAR* approaches 1. When $\iota_{T_-} = 0$, ι_I does not intersect with the acceptable ratio-type summary statistic, F_β*-score*. Comparing ι_I and *TAR*, I see in Figure 9.1a, that they share the same endpoint, varying only by slope. In Figure 9.1c, both summary statistics monotonically increase as $ratio_+$ increases. These similarities result from a relationship between the two summary statistics: $|S| = t_+ + f_- + f_+ + t_-$, thus $TAR = (t_+ + t_-)/|S|$. *TAR* is a member of the ι_I family, with $I = (1, 0, 0, 1)$. In this context, *TAR* is a biased form of ι_I, so it is not applicable when I must be unbiased.

[3] F_β*-score* has the β variable, which expresses the relative importance of precision to recall. However, when used, end users have trouble mapping the F_β*-score* values to their situation.

Figures 9.1c versus 9.1d show how $ratio_+$ sensitivity is also markedly different. When $\iota_{T_-} = 0$, (Figure 9.1d), ι_I and F_β-*score* have dissimilar responses to changes in $ratio_+$. Comparing ι_I in Figures 9.1c and 9.1d, I can see a dramatic change in ι_I's character. In Figure 9.1c, ι_I asymptotically approaches 1; in Figure 9.1d, ι_I asymptotically approaches 0. An end user basing a decision on a ratio summary statistic might be led to expect performance different than will actually be experienced.



(a) When all of the category impacts are equally weighted ($\iota_{T_-} = 1$) and class overlap varies from 100% to 0%, ι_σ is functionally equivalent to the Youden index. The equivalence does not hold, however, over all possible category weights.



(b) When $\iota_{T_-} = 0$, ι_σ responds to the change from a balanced I by reducing the impact. The weighted *TAR* increases.

Figure 9.2: ι_σ is the expectation of the impact on the end user, conditioned on the CPD output. Regardless of the value of ι_{T_-} the ratio summary statistics converge on one when the classes are fully separable. When $\iota_{T_-} = 0$, the loss of that positive contribution reduces ι_σ.

Comparing ι_I, *MCC* and *IC* in Figure 9.1c, I see that *MCC* and *IC* monotonically decrease as class sizes diverge, whereas end user impact monotonically increases. Hence neither *MCC* nor *IC* characterize end user impact. Of the four potential summary statistics, only ι_I is end user efficacious.

ι_σ is for problems requiring $ratio_+$ invariance, so I execute the test series varying class median and forgo the $ratio_+$ sensitivity test. When $\iota_{T_-} = 1$, I compare ι_σ against *ROC-AUC*, the Youden Index and *DOR*. I also compare ι_σ against *TAR* [4], *MCC* and IC, commonly seen $ratio_+$ sensitive summary statistics, calculated on normalized JPTs. The results of the comparison is shown in Figure 9.2a. All seven summary statistics follow unique tracks, thus ι_σ uniquely quantifies impact.

[4]TAR is my baseline (x-axis) summary statistic, against which all summary statistics are compared. As a biased version of ι_I, it is not valid for this problem type. However, it is commonly seen; I include it here so all summary statistics have the same visual representation.

When $\iota_{T_-} = 0$, I compare ι_σ against the *F_β-score*. For the reasons noted above, the other summary statistics are excluded. The results of the $\iota_{T_-} = 0$ comparison is shown in Figure 9.2b. Both summary statistics follow unique tracks, thus ι_σ uniquely quantifies impact.

Comparing ι_σ in Figures 9.2a and 9.2b, I see that when $\iota_{T_-} = 0$, ι_σ is about half of *F_β-score*. The change is due to T_- being discounted. As with ι_I in my biased I test, the summary statistic does not converge with *F_β-score*, when the classes are perfectly separable. The key point is that from the end user's perspective, ι_σ accurately reflects the effect of $\iota_{T_-} = 0$; the increase observed in *F_β-score* does not.

9.4 Completing impact summary statistic comparison

Chapter 5 considered some summary statistics seen in CPD evaluation literature. In this chapter, two axiom compliant summary statistics were identified. For completeness, the three questions posed for each of the pre-existing summary statistics are answered for ι_I and ι_σ. Since ι_I and ι_σ are not impact invariant, the pdf sensitivity analysis includes tests with $I = (1, -1, -1, 1)$ for comparison with *TAR*, *J*, *MCC*, *IC* and *DOR/DP*; $I = (1, -1/2, -1/2, 0)$ is used for comparison with F_1*-score*.

ι_I **What question does the summary statistic quantify?** ι_I quantifies the expected per object impact (measured in a standard interval unit relevant to the end user) of a CPD's output, in the end user's environment; $ratio_+$ is important.

Is the summary statistic measured on a ratio scale? ι_I has a meaningful zero. By definition, a standard interval unit is used, so ι_I is a ratio scale summary statistic.

Does the summary statistic exhibit boundary sensitivity? ι_I is sensitive to both $ratio_+$ and pdf. Figure 9.3 shows the pdf sensitivity with a balanced I; Figure 9.4 shows the pdf sensitivity when I is unbalanced.

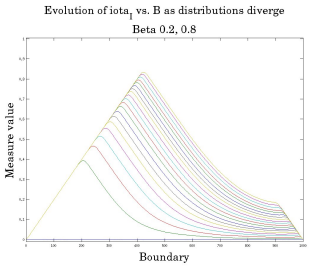
ι_σ **What question does the summary statistic quantify?** ι_σ quantifies the expected per object impact (measured in a standard interval unit relevant to the end user) of a CPD's output when the end user's environment is confounding; $ratio_+$ is confounding.

Is the summary statistic measured on a ratio scale? ι_σ has a meaningful zero. By definition, a standard interval unit is used, so ι_I is a ratio scale summary statistic.

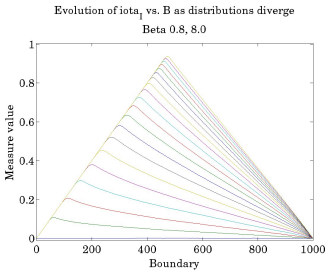
Does the summary statistic exhibit boundary sensitivity? ι_σ is sensitive to both $ratio_+$ and pdf. Figure 9.5 shows the pdf sensitivity with a balanced I; Figure 9.6 shows the pdf sensitivity when I is unbalanced. Surprisingly, on some distributions, (e.g., the uniform distribution as seen in Figure 9.5h), it is bimodal. Some of the distributions tested resulted in bi-modal curves. This is not an issue for CPD comparison, particularly if one peak is dominant. However, in situations where peaks are statistically equivalent, it does pose a problem for optimal boundary selection. If the two maxima are statistically equivalent, how can one be selected over the other? What is the significance of two maxima? Is this actionable information? I suggest that one maximum provides the best ι_Z while minimizing $\iota_{\overline{Z}}$, the other provides the best $\iota_{\overline{Z}}$ while minimizing ι_Z. Considering a medical diagnosis, perhaps the optimal diagnostic strategy would be to have three possible test results, "Positive", "Negative" and "Unknown". A test result in the inter-modal space would be inconclusive; a test result above the upper optimal boundary would be conclusive for one condition and a

test result below the lower optimal boundary would be conclusive for the alternate condition. A similar approach may be appropriate when there are multiple, statistically different maxima. This issue will be dealt with in the future.

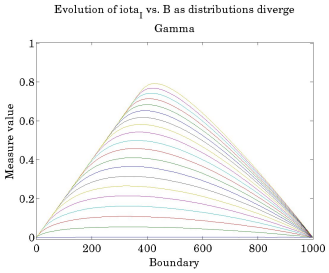
This Chapter reports the final actionable insights for end users: two summary statistics which are tailored for end user efficacy. The following chapters evaluate the measurement problem from different stakeholder perspectives.



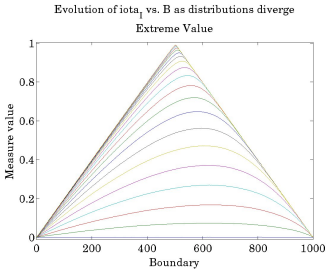
(a) Beta (0.2,0.8)



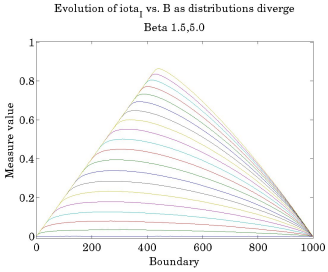
(b) Beta (0.8,8.0))



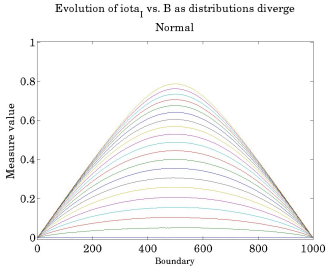
(c) Gamma



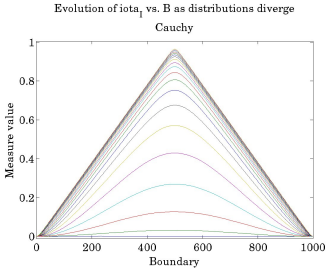
(d) Extreme value



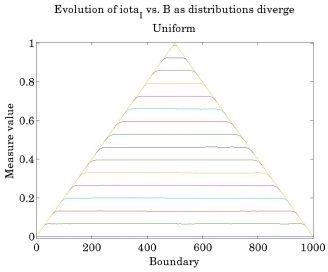
(e) Beta (1.5,5.0)



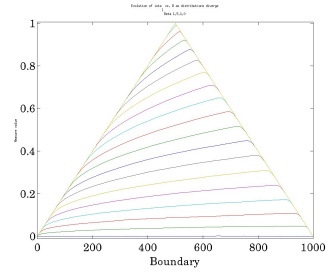
(f) Normal



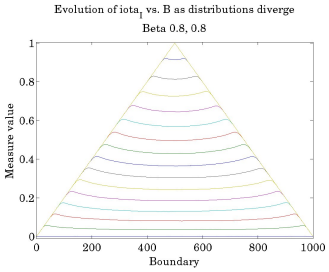
(g) Cauchy



(h) Uniform

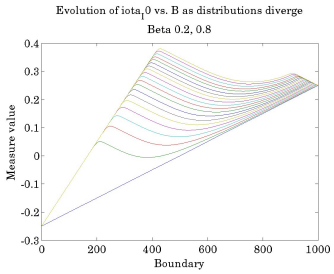


(i) Beta (1.5,1.0)

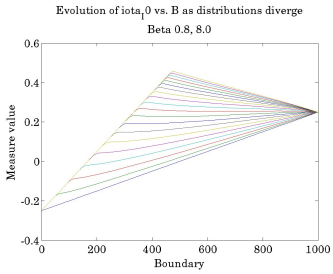


(j) Beta (0.8,0.8)

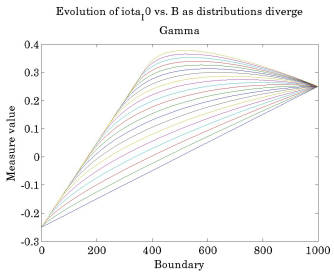
Figure 9.3: Graphs of the ten pdfs used for ι_I sensitivity testing when $I = (1, -1, -1, 1)$.



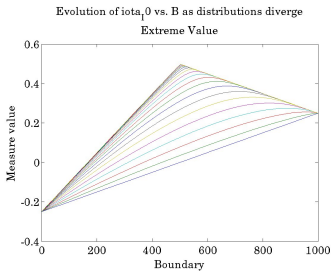
(a) Beta (0.2,0.8)



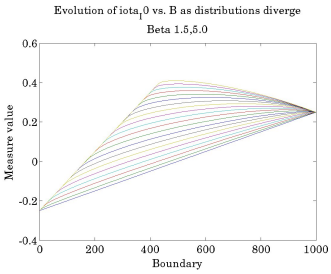
(b) Beta (0.8,8.0))



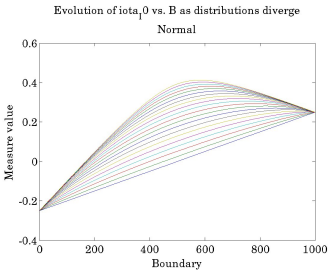
(c) Gamma



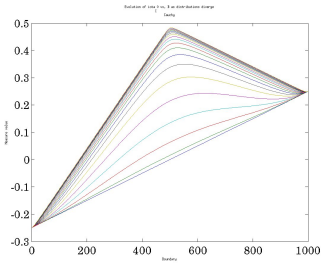
(d) Extreme value



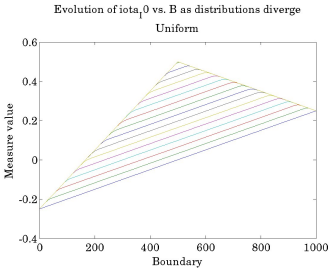
(e) Beta (1.5,5.0)



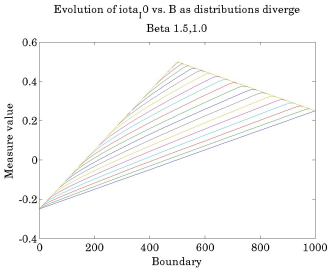
(f) Normal



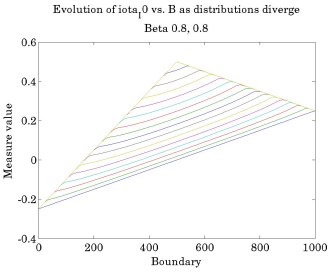
(g) Cauchy



(h) Uniform

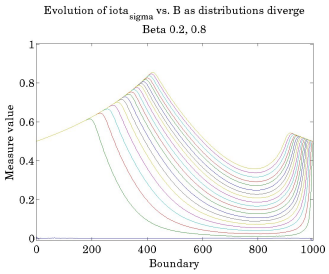


(i) Beta (1.5,1.0)

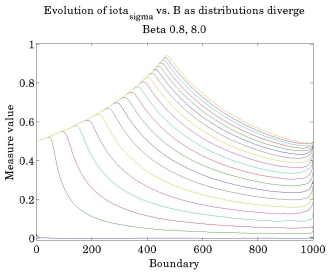


(j) Beta (0.8,0.8)

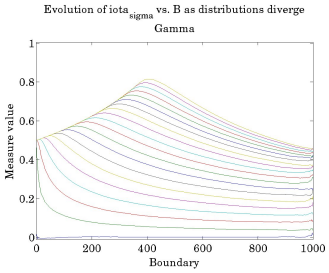
Figure 9.4: Graphs of the ten pdfs used for ι_I sensitivity testing when $I = (1, -1/2, -1/2, 0)$.



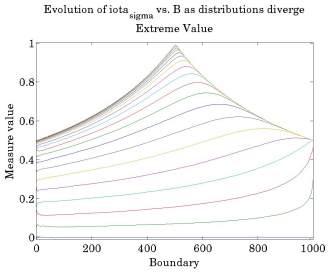
(a) Beta (0.2,0.8)



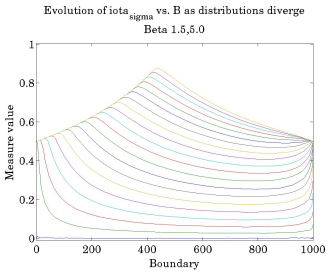
(b) Beta (0.8,8.0))



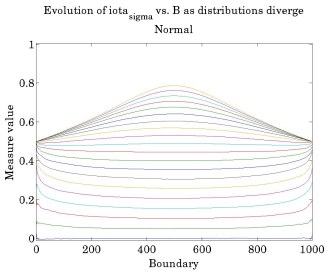
(c) Gamma



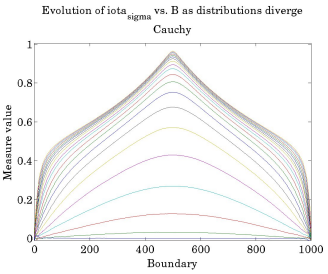
(d) Extreme value



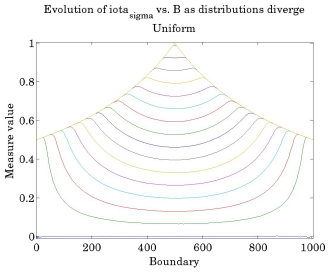
(e) Beta (1.5,5.0)



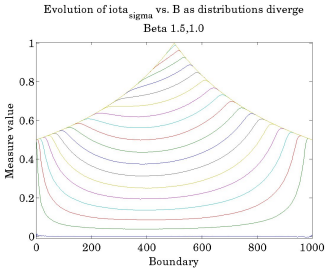
(f) Normal



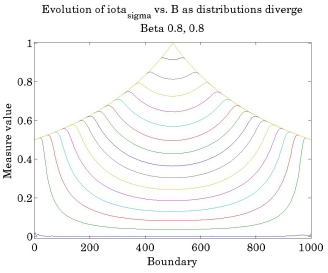
(g) Cauchy



(h) Uniform

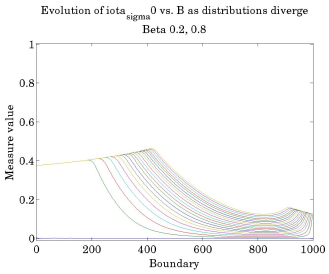


(i) Beta (1.5,1.0)

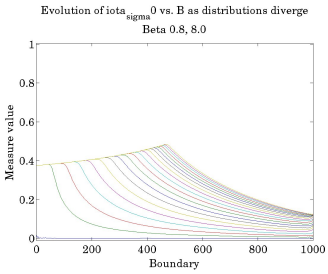


(j) Beta (0.8,0.8)

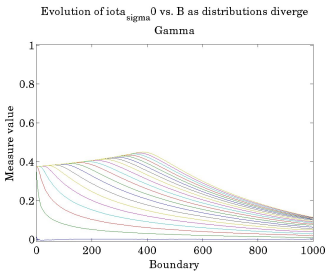
Figure 9.5: Graphs of the ten pdfs used for ι_σ sensitivity testing when $I = (1, -1, -1, 1)$.



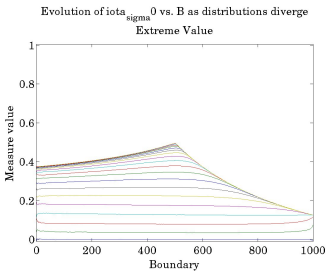
(a) Beta (0.2,0.8)



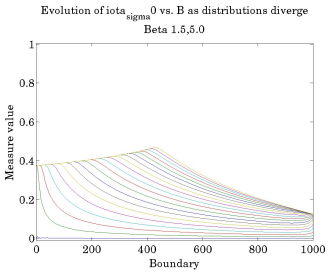
(b) Beta (0.8,8.0))



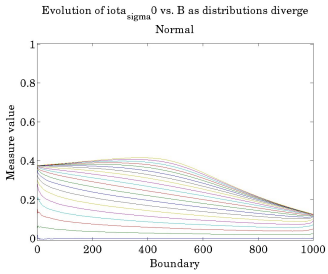
(c) Gamma



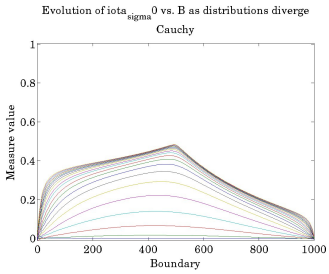
(d) Extreme value



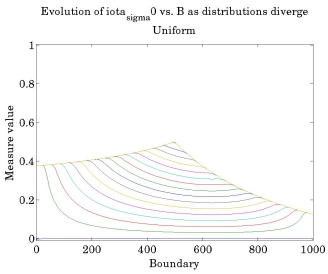
(e) Beta (1.5,5.0)



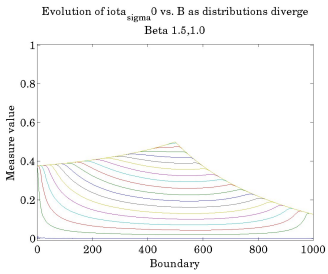
(f) Normal



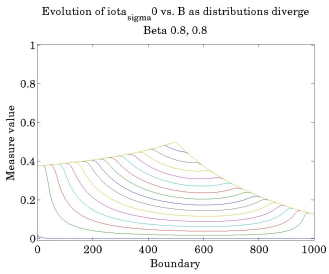
(g) Cauchy



(h) Uniform



(i) Beta (1.5,1.0)



(j) Beta (0.8,0.8)

Figure 9.6: Graphs of the ten pdfs used for ι_σ sensitivity testing when $I = (1, -1/2, -1/2, 0)$.

CHAPTER 10

EXAMPLES

How much better do the impact summary statistics address end user concerns on real problems? To demonstrate, I re-analyze four published CPD problems, a bank loan decision problem where $ratio_+$ is important (ι_I is relevant), a rheumatoid arthritis meta-analysis where $ratio_+$ is confounding (ι_σ is relevant), detecting a masquerade-type cyber attack and an intrusion detection problem where summary statistic suitability is context sensitive. The fourth problem is the first known to apply the impact summary statistics.

10.1 Bank loan decisions

Optimizing bank loan decisions is a "cumulative output" type of CPD problem. End user impact is best quantified by ι_I. Although there is a body of credit scoring algorithms tests, I found none with sufficient data available for a full re-analysis. The H. A. Abdou work selected [1], included sufficient detail to compare peak outputs identified by the algorithms tested.

10.1.1 Test protocol

Abdou provides the normalized JPT proportions and reports a misclassification cost ratio of 5 : 1; *MCR*=*cost of Type II errors* (f_+)/*cost of Type 1 errors* (f_-). (*MCR* considers direct costs only. It does not include the opportunity cost, the lost income attributable to qualified applicants not being funded.) Abdou does not provide loan amount information, so I define a "standard loan unit" of some arbitrary number of Egyptian pounds (EGP) and calculate the impact per loan unit. The JPT categories are defined as:

Good applicants ($\iota_{T_+} = 1.0$) These are loans that are made and pay off as expected.

Known defaulters ($\iota_{T_-} = -.01$) These are applicants rejected where ground truth is a known default. Abdou is not clear on how this JPT category is quantified. In a strict sense, these loans are not made, so how can ground truth be

Original and reanalysis results					
	EMC	ι_I at $ratio_+$=			
Test	5 : 1	$ratio_+$=1	$ratio_+$=0.67	$ratio_+$=0.48	$ratio_+$=0.40
WOE_{T2}	**0.4627**	0.3484	0.4223	0.4764	0.5068
PA	0.6232	0.4377	0.5299	0.5975	0.6354
GP_p	0.5679	0.4315	0.5555	0.5896	0.6259
GP_t	0.6964	**0.4590**	**0.5561**	**0.6271**	**0.6670**

Table 10.1: This table compares Abdou's original credit scoring algorithm results [1] using estimated misclassification cost (EMC) and ι_I as summary statistics. The highlighted values indicate the best results. EMC is a cost, thus the lower the value, the better; WOE is best. ι_I estimates the net impact; the higher the number the better; GP_t is best. The ι_I results are consistent with other studies; generally, AI-derived algorithms outperform manual algorithms.

a known default? All applicants, including those rejected, incur an application processing cost, hence, ι_{T_-} is negative.

Unknown defaulters ($\iota_{F_+} = -.05$) These are loans made that defaulted. The value is based on $MCR = 5 : 1$.

Unknown good applicants ($\iota_{F_-} = -.01$) These are rejected applicants that later proved to be good.

In order to limit complexity, I assume that the standard loan unit has a defined annual profit expectation. Intuitively, I might expect $\iota_{T_-} > 0$. Contrary to intuition, in Abdou's scenario, ι_{T_-} has a slight negative impact. This is due to application processing costs incurred, regardless of the loan decision made.

Abdou also normalizes his data $ratio_+ = 1.0$. However, the problem is $ratio_+$ sensitive, so I use JPT tuning to adjust to the reported value, $ratio_+ = 0.48$. Abdou ran a sensitivity analysis on EMC; I will use JPT tuning to illustrate how an end user can run a $ratio_+$ sensitivity analysis. (As I noted Chapter 6.1, such a sensitivity analysis can test results at the identified boundary, however, $ratio_+$ causes the optimum boundary to shift. So without the actual data, JPT tuning cannot be used to estimate the peak impact.) I tune the JPT to two other relative class sizes, $ratio_+ = 0.67$ and $ratio_+ = 0.40$. The four weight of the evidence (WOE) JPTs and four genetic programming (team) (GP_t) JPTs are shown in Appendix E; Table 10.1 compares my ι_I results with the estimated misclassification cost (EMC) reported by Abdou.

10.1.2 Results

Abdou concludes that the WOE model is the best performer, based on EMC. However, ι_I shows that genetic programming performs the best. Abdou

does not report confidence intervals, but because the results for probit analysis (PA) and GP_p are so similar, there is reason to suspect that the the difference is statistically insignificant. Based on ι_I, WOE seems to be the worst performer. WOE has a substantial negative impact on the lender compared to either PA and GP. My re-analysis indicates that GP is at least equivalent to the best non-artificial intelligence method tested – this is consistent with other tests comparing Artificial Intelligence (AI) and non-AI methods.

Using the Egyptian banking environment assumptions presented here, my sensitivity analysis of GP shows that for $ratio_+ = [0.40, 0.67]$, the annual profit per loan unit would range from fifty-six to sixty-seven percent of the amount that would be received if loan decisions were perfect. Thus by using ι_I, the bank decision makers receive valuable information that can be used to define loan application scoring policy and procedures. I want to emphasize that the banking environment assumptions used will probably not be extensible to a wide bank pool. Thus, ι_I will be most useful when each institution tunes the values to their specific environment.

10.2 Rheumatoid arthritis testing

The Hippocratic oath is commonly held to characterize the medical practitioner's mantra: do no harm. This dedication to their patient's welfare would seem to make impact a suitable measure. Indeed, in recent years, there has been a discussion thread in the medical community regarding difficulties medical practitioners have interpreting and properly applying medical tests. Steurer, et al., sound the alarm by observing that medical practitioners do not have a clear understanding of commonly used summary statistics such as sensitivity and positive predictive value. The authors found that adding non-technical language improved clinicians ability to correctly interpret the information [82]. Whiting, et al., performed a meta-analysis on the issue. They found that practitioners consistently have difficulty interpreting Bayesian values [95]. A report by Zhelev, et al., is a case in point. This study focused on Cochrane diagnostic test accuracy reviews. Cochrane reviews are a well-trusted evidence-based health care resource, yet Zhelev, et al., discovered that regardless of experience with the reviews, practitioner's understanding was poor [98]. Of the papers reviewed, the one most aligned with my work was Gopalakrishna, et al. The authors confirmed the difficulty observed by others and identified three contributing factors "methodological issues, resource limitations and a lack of awareness on the need for evidence that links testing to patient outcomes". Gopalakrishna, et al. recommend education as a quick mitigation, but go on to note that "a shift in the way we view the value of a test is required: to move away from solely considering how accurate a test may be in diagnosing a condition to including the value it may bring to the patient receiving the test" [34]. The impact summary statistic used in this example satisfy Gopalakrishna, et al.'s suggested paradigm shift.

Nishimura, et al. [64] published a meta-analysis [18] of two rheumatoid arthritis (RA) diagnostic tests comparing two medical diagnostic tests for RA. The meta-analysis is quite thorough and accounts for many potential variations between studies. The team concludes that one test is better than the other, however, does so without using a summary statistic. My re-analysis adds ι_σ, the appropriate impact summary statistic identified in Chapter 9.

Nishimura, et al.'s study uses two summary statistics, positive likelihood ratio (LR_+) and negative likelihood ratio (LR_-). Both of these ratios are conditioned on ground truth (Y and $\overline{Y}$) rather than (Z and $\overline{Z}$), the outputs observable by end users. Given a typical test using supervised inputs (where ground truth is known), these two values are efficacious for researchers. They are less so in the field, where end users have only the CPD output; ground truth is unknown.

	Normalized odds ratio summary statistics	
Test	LR_+	LR_-
Anti-CCP	12.46 (9.72–15.98)	0.36 (0.31–0.42)
RF	4.86 (3.95–5.97)	0.38 (0.33–0.44)

	Expected (annual economic) impact ($\$10^3$)		
Test	ι_Z	$\iota_{\overline{Z}}$	ι_σ
Anti-CCP	-0.55 (-0.56– -0.44)	-9.6 (-9.7– -9.5)	-5.1 (-5.1– -5.0)
RF	-1.4 (-1.6– -1.3)	-9.5 (-9.6– -9.4)	-5.5 (-5.5– -5.5)

Table 10.2: These tables compare the summary likelihood ratios originally reported [64] (top table) and the corresponding ι_Z, $\iota_{\overline{Z}}$ and ι_σ (bottom table). Both summary statistic suites show that the anti-CCP test is better, as does the summary statistic ι_σ. The additional insight gained by assessing impact may lead an end user to want substantial corroboration of a negative test result.

10.2.1 Test protocol

The authors observe that RA treatment is harmful to and costly for persons with false positive results. Regardless of the diagnosis, a correct diagnosis maximizes the subject's quality of life. Accordingly, I define the meaningful zero as the costs associated with a correct diagnosis: $\{\iota_{T_+} = 0, \iota_{T_-} = 0\}$. An incorrect diagnosis results in reduced quality of life. Rounding Lajas, et al.'s reported costs [48] to Nishimura et al.'s degree of precision(two significant digits), the mis-diagnosis costs are: $\iota_{F_+} = -\$7,900$ and $\iota_{F_-} = -\$13,000$.

Since each individual diagnosis is important, ι_σ is the appropriate summary statistic. In my extension to Nishimura et al.'s report, I calculate the $\iota_{Z(\sigma)}$, $\iota_{\overline{Z}(\sigma)}$ and ι_σ on the pooled test data.

		Actual RA condition	
		Diseased	Not diseased
Anti-CCP test Result	*Positive*	0.67 (0.65–0.68)	0.05 (0.04–0.06)
	Negative	0.33 (0.32–0.35)	0.95 (0.94–0.95)
	Totals	1	1

		Actual RA condition	
		Diseased	Not diseased
RF test Result	*Positive*	0.69 (0.68–0.7)	0.15 (0.14–0.16)
	Negative	0.31 (0.3–0.32)	0.85 (0.84–0.86)
	Totals	1	1

Table 10.3: These *normalized JPTs* of Nishimura et al.'s pooled anti-CCP and RF test data were calculated from Nishimura, et al.'s reported sensitivities and specificities [64]. A person without RA is far less likely to be mis-diagnosed than one with the disease when the anti-CCP test is used.

Anti-CCP test Result	Actual RA condition		
	Diseased	Not diseased	ι
Positive	0	-.55 (-.56– -.44)	$\iota_{Z(\sigma)}$=-.55 (-.56– -.44)
Negative	-9.6 (-9.7– -9.5)	0	$\iota_{\overline{Z}(\sigma)}$=-9.6 (-9.7– -9.5)
			ι_{σ}=-5.1 (-5.1– -5.0)

RF test Result	Actual RA condition		
	Diseased	Not diseased	ι
Positive	0	-1.4 (-1.5– -1.3)	$\iota_{Z(\sigma)}$=-1.4 (-1.6– -1.3)
Negative	-9.5 (-9.6– -9.4)	0	$\iota_{\overline{Z}(\sigma)}$=-9.5 (-9.6– -9.4)
			ι_{σ}=-5.5 (-5.5– -5.5)

Table 10.4: These tables show the $\iota_{Z(\sigma)}$, $\iota_{\overline{Z}(\sigma)}$ and ι_{σ} as well as the proportional contribution of each JPT category. All values are in thousands of dollars. The values indicate the unnecessary annual economic cost resulting from an incorrect diagnosis.

10.2.2 Results

Table 10.2 shows the original likelihood ratios reported by Nishimura et al. and the proposed new summary statistics. (The parenthesized range is the 95% confidence interval. On a single tailed test as used here, only one bound is relevant, thus the bound indicates a 97.5% confidence.) Reassuringly, the results of the proposed summary statistics and the original summary statistics used are similar and the same conclusion reached (the anti-CCP test is better than the RF test). Comparing ι_σ for each test and keeping in mind the end user context requires $ratio_+$ invariance, the anti-CCP test estimated annual economic impact on patients is four hundred dollars less than the RF test's estimated annual economic impact.

End users inspecting the raw JPT values in Table 10.3 may note that anti-CCP has a substantially lower F_+ rate than the RF test. They might therefore be inclined to place substantially more trust in the RF test's negative result, rather than in a negative result from the anti-CCP test. However, that trust does not result in a better outcome for the patient. Actually both tests have statistically equivalent negative impacts. In fact, the end user can see that in contrast to a positive test result, a negative result can have a substantial negative annual cost: an end user may not want to conclude a patient with a negative test result is RA free without strong corroboration. Regarding positive test results, the RF test has approximately three times worse (negative) impact than anti-CCP.

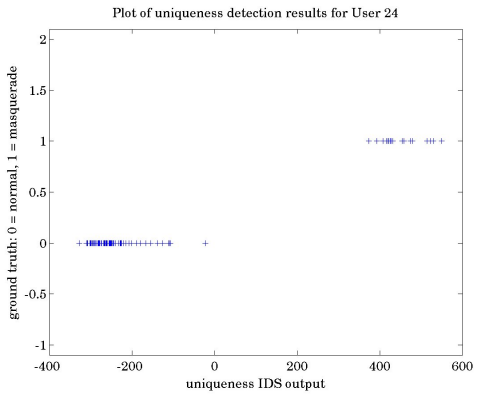
Table 10.4 shows the JPTs for both raw proportions and impacts. Assessing the RA test's impacts provides information to the researcher, however, the real contribution may be that it its a shift in the way medical practitioners view test value. Gopalakrishna, et al. recommend including the value it may bring to the patient receiving the test; ι_σ is such a summary statistic. This example shows its value to the end user.

recommend education as a quick mitigation, but go on to note that "a shift in the way we view the value of a test is required: to move away from solely considering how accurate a test may be in diagnosing a condition to including the value it may bring to the patient receiving the test

10.3 A Cyber security masquerade study

This re-analysis and the next address cyber security issues. Early cyber security strategy focused on "keeping the bad guys out". Over time, this strategy has evolved into a risk-based strategy that considers the impact of specific malicious activities. The cost of low impact events may be dominated by processing false alarms (F_+) in which case cumulative effects are important: ι_I is the appropriate summary statistic. The cost of high impact events may be dominated by processing missed attacks (F_-) and for some impact events, a single missed

(a) Ground truth for the compression algorithm



(b) Ground truth for the uniqueness algorithm

Figure 10.1: These plots show that both classifiers can perfectly distinguish between normal and masquerade traffic. Uniqueness has a wider gap between classes, so is a more robust classifier.

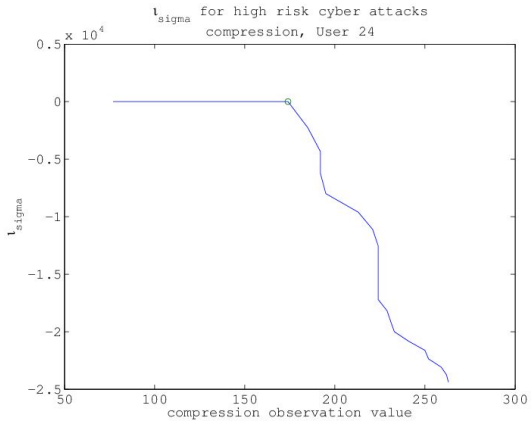
attack may be disruptive, perhaps catastrophic. In such cases, ι_σ is the appropriate summary statistic. Hence, for both cyber security examples, both impact summary statistics are used.

10.3.1 Test protocol

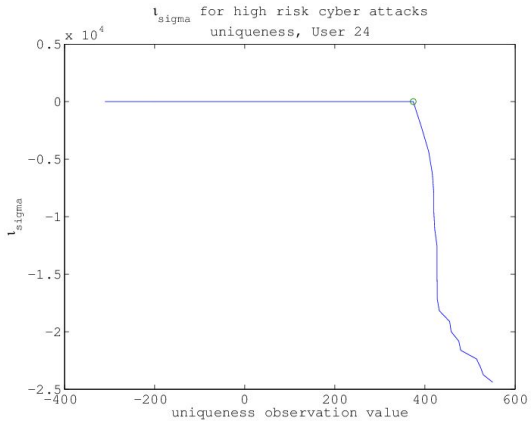
Schonlau, et al. simulate a masquerade attack by capturing UNIX commands resulting from specific user activity, then inserting UNIX commands generated by another user into the original command stream [74]. In the other two examples, the data available only allowed re-analyses of the original work. However, Schonlau's test data and raw test results are available online (click the Masquerading User Data tab on http://www.schonlau.net), so it is possible to illustrate a hypothetical end user analysis.

Using Schonlau, et al.'s data, I simulated two end users assessing the performance of two anomaly-based intrusion detectors. Schonlau simulated fifty users and tested six detection algorithms; I selected two each. Schonlau's study used test sets that were moderately sized, one hundred blocks of test data, of which a small, random portion were masquerade activity. As noted in Chapter 5, my experience with small class sizes indicate that the evaluation results tend to yield optimistic results. To minimize this effect, I selected the two test sets (users) with the most balanced class sizes. Users 9 and 24 had $ratio_+ = 3.8$ and 3.3, respectively.

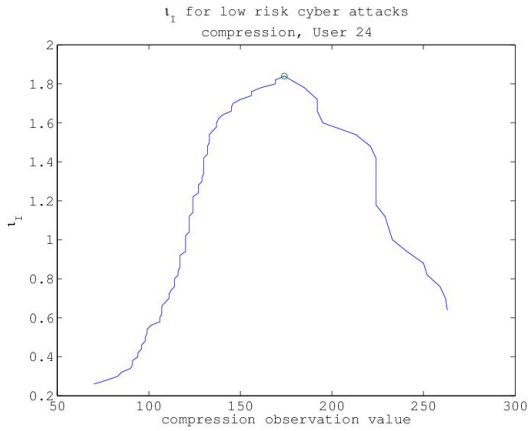
Schonlau tested six different CPD algorithms. Overall, Schonlau, et al. reported that "uniqueness" provided the best detection rate, "compression" the worst. I chose the two extremes.



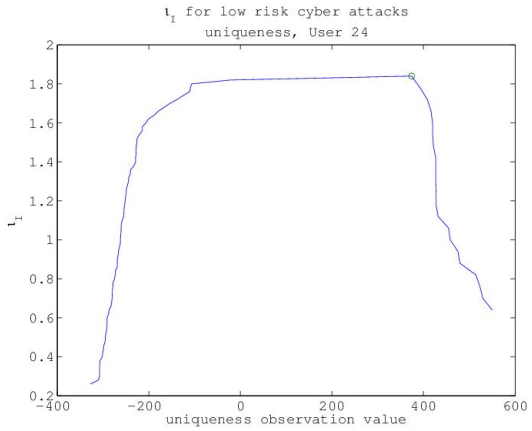
(a) high risk curve for the compression algorithm



(b) high risk curve for the uniqueness algorithm

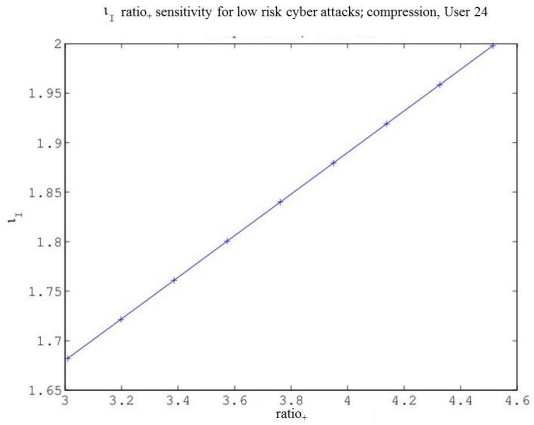


(c) low risk curve for the compression algorithm

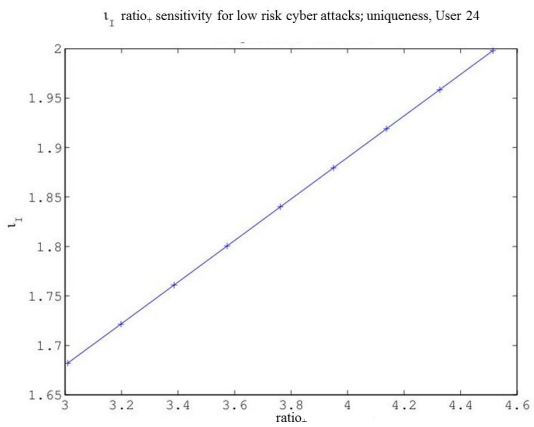


(d) low risk curve for the uniqueness algorithm

Figure 10.2: Impact graphs for User 24. For both high and low risk events, the uniqueness algorithm exhibits less optimum boundary sensitivity. The optimum boundary is indicated on the graphs as a green circle.



(a) low risk curve for the compression algorithm



(b) low risk curve for the uniqueness algorithm

Figure 10.3: ι_I sensitivity to $ratio_+$ graphs for User 24. Both the compression and uniqueness algorithms exhibit the same $ratio_+$ sensitivity over a 20% variation, $\iota_I \approx [1.68, 1.99]$.

For a low impact scenario, $I = (\$5, -\$1, -\$1, \$1)$ is used;
$I = (\$10, -\$1, -\$100000, \$1)$ is used for a high impact scenario. Impact is measured in currency (USD), so the greater the value, the better the expected result for the user. In this analysis, I calculate ι_I and ι_σ, test $ratio_+$ sensitivity for ι_I and assess the optimum boundary sensitivities for both ι_I and ι_σ. All four data sets have approximately equal $ratio_+$. In keeping with the user product selection theme, both user's evaluations will end with a downselect matrix and recommendations.

10.3.2 Results

Users 9 and 24 represent two independent stakeholders doing their own independent analyses for their separate environments. Table 10.5 shows the ι_I and ι_σ results for user 24. "+" indicates the better algorithm; "-" indicates the worse algorithm; and "=" indicates that the algorithms are equivalent. In the table, "@x" indicates the normalized boundary at which the value was observed (the boundaries are normalized to a 0 to 100 interval). In this table, both classifiers have the same peak impacts at the same optimum boundaries. This is because, as can be seen in Figure 10.1, both classifiers correctly identify the entire test set. ι_I reflects the effect of the test set's $ratio_+$. ι_σ is $ratio_+$ invariant; for a perfect classifier, its value is the average of ι_{T_+} and ι_{T_-}.

How do the two classifier optimum boundary sensitivities compare for high risk events? Figures 10.2a and 10.2b show that the uniqueness algorithm

Scenario	Classifier type	
	Compression	Uniqueness
Low risk (ι_I)	\$1.84 @ 21	\$1.84 @ 21
High risk (ι_σ)	\$5.50 @ 21	\$5.50 @ 21
Low risk B^* sensitivity	-	+
High risk B^* sensitivity	-	+
Low risk $ratio_+$ sensitivity	=	=

Table 10.5: User 24's downselect matrix shows that although the impacts and optimum boundaries are the same for both tests, the uniqueness algorithm is less sensitive, thus performs better than compression. User 24 would choose the uniqueness-based detector. "+" indicates the better algorithm; "-" indicates the worse algorithm; and "=" indicates that the algorithms are equivalent. "@x" indicates the normalized boundary at which the value was observed.

drops off much faster to the right of B^* the optimum boundary than does compression. However, when the observation value ranges are normalized, uniqueness has a substantially larger area under the curve, so provides the better outcome overall. This is the result of the difference in the gaps between the two classes. Whereas compression detection could be fooled by edge cases (Schonlau's test sets are not large, so are likely to not include less frequent events for either group), uniqueness has a large gap, so will be less sensitive to edge cases. Considering the severe impact of a missed attack, the gap provides a buffer against missed attacks.

How do the two classifier optimum boundary sensitivities compare for low risk events? Figure 10.2d shows that the uniqueness algorithm exhibits an impact plateau near the optimum boundary, whereas Figure 10.2c shows that the impact for the compression algorithm rapidly degrades. This shows that whereas the uniqueness impact is constant across the interval $[-100, 400]$ (the gap between the two classes), compression's impact shows no stability around B^*. Also, when the observation value ranges are normalized, uniqueness has a substantially larger area under the curve, so provides the better outcome overall.

ι_σ is $ratio_+$ invariant, but ι_I is not. How sensitive to $ratio_+$ is ι_I? Both classifiers exhibit the same $ratio_+$ sensitivity. This can be seen in Figure 10.3.

These findings are summarized in the downselect matrix for User 24 (Table 10.5). The two classifiers only vary in their optimum boundary sensitivity: uniqueness is better. Hence, User 24 would deploy the uniqueness-based classifier. Figure 10.1 shows the data plots for User 24's ground truth. Both classifiers correctly label the entire test set. However, uniqueness has a greater gap between classes, so is a more robust tool for User 24.

Table 10.6 shows the peak expected impact test results for User 9. For the low risk events and the I selected for this analysis, the compression algorithm expected impact is approximately 2.4 times better than the uniqueness al-

gorithm. For low risk events, the uniqueness algorithm performs optimally when all events are treated as normal ($B^* = 1$). This boundary suggests that uniqueness cannot differentiate between normal and masquerade traffic.

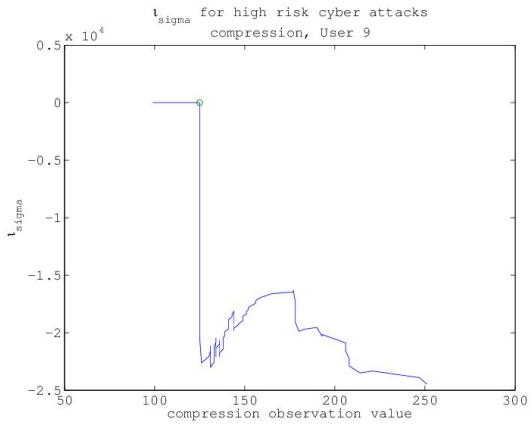
For high risk events, uniqueness performs optimally when all events are treated as masquerades ($B^* = 100$). Once again, this boundary suggests that uniqueness cannot differentiate between normal and masquerade traffic, however, now the end user is better off assuming all traffic is masquerade. Table 10.6 shows that compression provides a positive benefit for end users, uniqueness has a large loss.

How do the two classifier optimum boundary sensitivities compare? For both high and low impact scenarios, the compression algorithm exhibits relative impact stability while B is slightly less than B^*. When $B > B^*$, the compression algorithm's impact degrades precipitously. In contrast, *any* change in uniqueness's B away from B^* causes a precipitous drop in impact. This is shown in Figure 10.4. Since uniqueness provides its best result when all values are classified the same, it is really an ineffective algorithm for User 9.

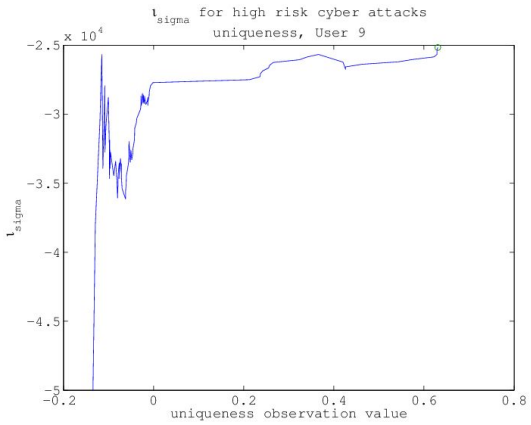
How sensitive to $ratio_+$ is ι_I? This is shown in Figure 10.5. Over a 20% variation, the compression algorithm's impact range is $\iota_I \approx [1.08, 1.37]$. The uniqueness algorithm's sensitivity differs markedly from that of compression and from that seen with User 24. The "V" shape was unexpected. An investigation of the figure's underlying data revealed that the slope change is caused by a change from $B^* = 1$ to $B^* = 100$ around $ratio_+ = 3$. The figure shows the peak impact value, regardless of $B*$, thus, the cause is not apparent in the figure. User 9, however, is unlikely to know when to shift the optimum boundary, so this "V" performance is unlikely to be experienced in the field. (The discussion is academic, however, since uniqueness is ineffective for User 9.)

Table 10.6 is the downselect matrix for User 9. In every category, the compression-based classifier is better than the uniqueness-based classifier. Hence, User 9 would deploy the compression-based classifier. Figure 10.6 shows the data plots for User 9's ground truth. Both classifiers have large overlaps between normal and masquerade traffic. Compression's overlap is slightly less, so is a more robust tool for User 9. User 9, however, may want to use two thresholds. Comparing the high impact optimum boundary's result in the low impact scenario in Figure 10.4 and vice versa, the optimum threshold for high risk events only has about seventy percent of the benefit received with the low risk impact. Similarly, the low impact optimum threshold results in a large negative impact from high risk events.

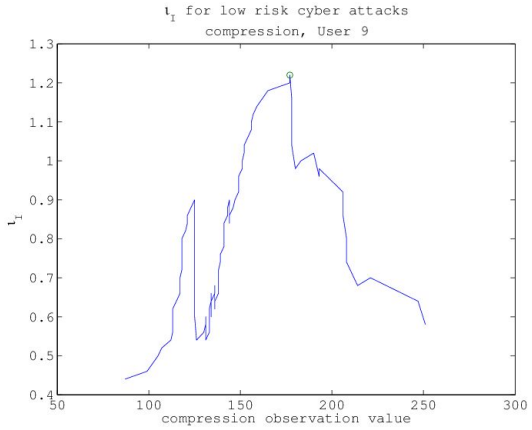
Schonlau. et al. use a modified *ROC* curve to analyze their results. Since they use a *ROC*-like summary statistic, they are unable to assess risk-based impact, nor can they assess B^* or $ratio_+$ sensitivities. Impact-based analyses can. This example also shows how sensitive CPD performance (and thus, end user tool selection) is to the end user's environment. In User 24's environment, both algorithms work well, but uniqueness exhibits less boundary sensitivity, it is the better performer. In User 9's environment, neither algorithm works well, but uniqueness is essentially useless, with compression being the better performer.



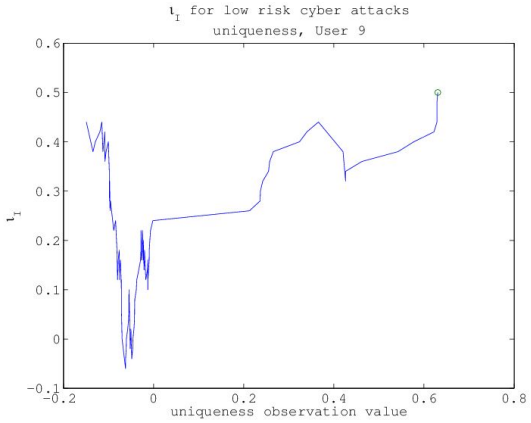
(a) high risk curve for the compression algorithm



(b) high risk curve for the uniqueness algorithm



(c) low risk curve for the compression algorithm

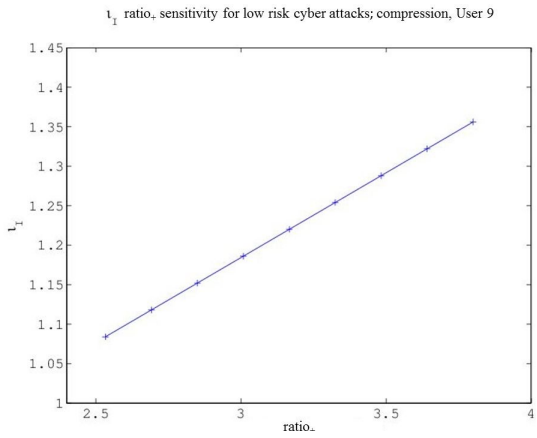


(d) low risk curve for the uniqueness algorithm

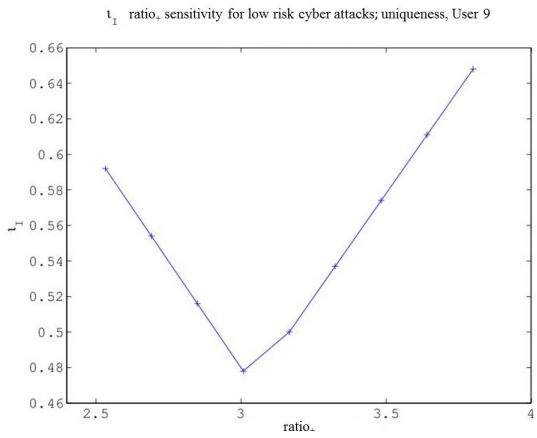
Figure 10.4: Impact graphs for User 9. For both high and low risk events, the compression algorithm exhibits less optimum boundary sensitivity. Optimum boundary is indicated on the graphs as a green circle.

	Classifier type	
Scenario	Compression	Uniqueness
Low risk (ι_I)	\$1.22 @ 17	\$0.50 @ 1
High risk (ι_σ)	\$3.24 @ 77	-\$25,200 @ 100
Low risk B^* sensitivity	+	-
High risk B^* sensitivity	+	-
Low risk $ratio_+$ sensitivity	+	-

Table 10.6: User 9's downselect matrix shows that the compression algorithm performs better for User 9 than uniqueness.

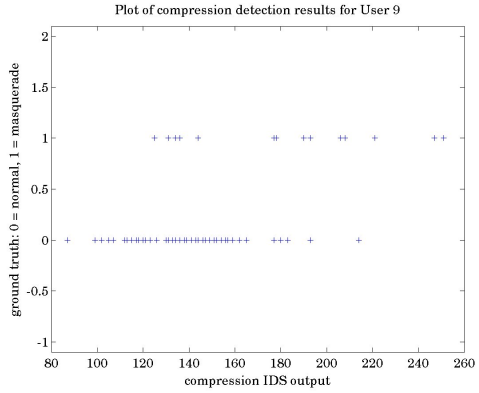


(a) low risk curve for the compression algorithm

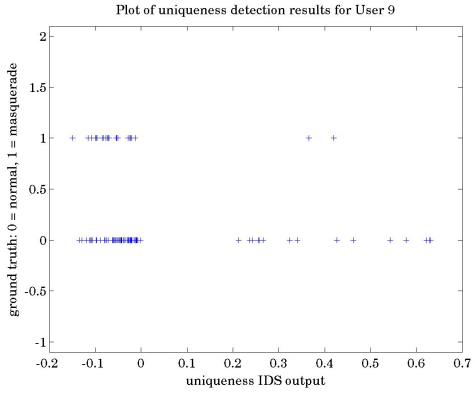


(b) low risk curve for the uniqueness algorithm

Figure 10.5: ι_I sensitivity to $ratio_+$ graphs for User 9. Over a 20% variation, the compression algorithm's impact range is $\iota_I \approx [1.08, 1.37]$. The uniqueness algorithms sensitivity differs markedly from that of compression and from that seen with User 24. The "V" shape was unexpected. Investigation into the result determined that it was caused by a change from $B^* = 1$ to $B^* = 100$. The test tracked the peak impact, regardless of B^*. User 9, however, is unlikely to know when to shift the optimum boundary, so this "V" performance is unlikely to be experienced in the field.



(a) Ground truth for the compression algorithm



(b) Ground truth for the uniqueness algorithm

Figure 10.6: These plots show that both classifiers have a large overlap in normal and masquerade traffic. Compression has somewhat less overlap, so is the more robust classifier.

End users are ill-advised to rely on "generic" test results for tool selection and configuration decisions.

10.4 Selecting an intrusion detection system for an industrial control system

Section 10.3, illustrates comparing two masquerade detection algorithms. This section illustrates how ι_I and ι_σ can aid in a "GO / NO GO" intrusion detector deployment decision. Similar to the Rheumatoid Arthritis example and the cyber security masquerade example, there are decision options. In this example, the options are i) keeping the current system with only security incident response and recovery capabilities (equivalent to classifying all events as normal [*test=negative*]) or ii) adding an intrusion detection system. The impact assessment uses actual test results for a novel intrusion detection system (IDS) tailored for industrial control systems (ICS) (Kratos) under development by root9B, LLC [25]. The assessment assumes scenario-specific I vectors for two incident types, irrecoverable (for which I has an extreme negative impact for missed attacks) and recoverable (for which there are six I vectors with detection error impacts representing different impact levels, ranging from very high to very low).

Irrecoverable incidents would have devastating effects on the enterprise and its mission: they must be completely avoided. Even a single instance is too many, thus ι_σ is the applicable measure. In contrast, recoverable incidents negatively affect the end user or enterprise, but are not catastrophic. In this example, we assume the effect is cumulative. As in the bank loan decision example, ι_I is the appropriate measure.

10.4.1 Test protocol

Kratos, root9B's novel anomaly detector monitors eight indicators of compromise (IoC). One important Kratos feature is that it eases deployment and maintenance by automatically selecting classification boundaries for each IoC. This affects Kratos' assessment, since impact's dependence upon boundary cannot be tested; test results consist solely of a JPT table reflecting results on the test set, with $ratio_+ = 0.00065$. Also, input complexity doesn't allow a graphical presentation as in Section 10.3.

Kratos was trained on actual data from a manufacturing plant in which the Purdue Level two SCADA traffic was over MODBUS/TCP. Without changing any Kratos algorithms, the IDS was tested on previously unseen DNP3 traffic. Kratos's anomaly detection algorithm exhibited exceptional robustness. In tests of eight indicators of compromise, in a body of over three million events, Kratos missed very few anomalies and generated no false alarms. Table 10.7 shows the results Kratos's poorest showing, a test where only two bytes were changed in each target payload.

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

Table 10.7: Observed test results, organized into a joint probability table.

10.4.2 Analysis protocol

Conceptually, the four Axioms introduced in Chapter 8 confer specific insights to end users. This Kratos ICS IDS test analysis is organized to illustrate each Axiom's contribution. Axioms One and Four and the ratio scale requirement from measurement theory (recapped below), allow end users to map CPD tool test output to expected JPT category impact. This impact can be quantified using their preferred unit of measure (assuming it is ratio scale).

Axiom 1 (Category importance) *An end user efficacious summary statistic must be a function of problem specific impact vector* $I = (\iota_{T_+}, \iota_{F_+}, \iota_{F_-}, \iota_{T_-})$*, where each element of* $I \in \mathbb{Q}$*.*

Axiom 4 (Summary statistic value appropriateness) *An end user efficacious summary statistic's output must quantify the CPD's impact on the end user's characteristic of interest.*

Measurement theory. *Ratio scales allow the most extensive analysis, thus end users are best served by ratio scale summary statistics.*

I use currency (United States Dollars [USD]) to quantify the CPD's end user impact. Although the end user in this discussion is hypothetical, often currency is used for comparison. One benefit of currency is that it is almost universally comprehensible by decision makers, regardless of their technical expertise: hence it satisfies Axiom Four. Currency is also measured on a ratio scale, so satisfies the measurement theory requirement. Currency can also quantify the impact of each CPD outcome relative to JPT category, so satisfies Axiom One.

Table 10.8 shows the I vectors for this use case. The values for each category are based on the following scenario. These values are for illustration purposes only:
Legitimate events processed as legitimate (T_-): The enterprise realizes a one hundred dollar benefit from every legitimate event.
Legitimate events processed as malicious (F_+): Although there is loss incurred, the enterprise still realizes a ninety-nine dollar 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; one tenth of one cent.
Malicious events processed as legitimate (F_-): The enterprise realizes a cost from every missed attack. For this use case, the specific event's impact class is mapped directly to the expected $\iota_{F\_}$. High impact events generate a major loss; low impact events generate a small reduction in per-event income. This evaluation considers six $\iota_{F\_}$ values.

Table 10.8 summarizes the scenario impacts, with each row representing an I vector. All JPT category impacts are the same, except F_-. This would probably not be true in the field. However, this simplifies analysis with no loss of generality.

As noted in Section 6.2 regarding the different behaviors of *TAR* and F_β-*score*, JPT category impact can have a dramatic effect on test results: each end user must determine values relevant for their specific situation.

In this scenario, we assume there are a number of potential attacks, with a range of impacts. However, the I vector presents statistical expectations, so in the general case, there will be some degree of uncertainty associated with each value. To illustrate end user utility, the low and medium F_- impacts are upper and lower bounds. I execute an $\iota_{F\_}$ sensitivity analysis by evaluating the bounds plus three intermediate values.

	Impact value			
Impact Class	ι_{T_+}	ι_{F_+}	ι_{F_-}	ι_{T_-}
Low	\$−0.001	\$99	\$90.00	\$100
	\$−0.001	\$99	\$0.0010	\$100
	\$−0.001	\$99	\$−1.00	\$100
	\$−0.001	\$99	\$−20.00	\$100
Medium	\$−0.001	\$99	\$−100.00	\$100
High	\$−0.001	\$99	\$−10,000	\$100

Table 10.8: Impact values used in this use case. For clarity, Impact classes are tied to missed attacks ($\iota_{F\_}$). All other vector values are kept constant.

The previous test demonstrates axiom compliant summary statistic incorporation of system characteristics into a CPD tool assessment. This quantifies its effect on the target performance characteristic. Similarly, Axiom Two (recapped below) allows end users to incorporate environmental characteristics into their CPD tool assessment.

Axiom 2 (Environmental sensitivity) *With a change in* pdf(Y), *e.g.,* pdf(Y')= Δ+pdf(Y), pdf(S)=pdf(Y)+ pdf$(\overline{Y})$ *and* pdf(S')=pdf(S)+Δ, *where* Δ

describes a perturbation in Y's and S's source population. For all boundaries within Δ, *there exists a* $E(\mathrm{SS}(B)|S')) - E(\mathrm{SS}(B)|S)) \neq 0$. *The same is true for a change in* $\overline{Y}$ *and for any* ratio_+.

In this analysis and with no loss in generality, I assume that the test normal and malicious ICS traffic samples are representative of their respective classes, but that attack frequency (and hence, $ratio_+$) will vary. JPT normalization compensates for $ratio_+$s for problems where it is confounding. As noted in Section 10.1, any test set is an environmental expectation, so JPT tuning allows the end user to get a sense of the possible impact range. In my test data, $ratio_+ = 0.00065$; I use JPT tuning to estimate the CPD tool's impact at five points across $ratio_+ \in [0.0001, 0.007]$.

In research conditions, investigators executing supervised tests will know ground truth. However, in the field, end users will not. This greatly limits the practitioner's ability to compare actual with expected field results. An Axiom Three (recapped below) compliant summary statistic enables this comparison.

Axiom 3 (CPD output basis) *An end user efficacious summary statistic must be quantifiable in terms relative to information known and visible to the end user: CPD output (*Z *and* $\overline{Z}$*).*

Although both ι_σ and ι_I are Axiom Three compliant, only ι_I's values enable actual vs. expected impact comparisons[1], the motivation for Axiom Three. Therefore, only $\iota_{I(Z)}$ and $\iota_{I(\overline{Z})}$ will be presented and discussed.

[1] Any irrecoverable event can only occur once, but ι_{F_-} is only observable after the event. For the same reason, ι_σ and $\iota_{\sigma(Z)}$ are only calculable after the event. Since by definition, ι_σ is appropriate for irrecoverable events, if such an event has occurred, then the diagnostic value of ι_σ and $\iota_{\sigma(Z)}$ in the field to validate model assumptions and possibly optimize detector impact is vanishingly low. The summary statistics are useful, however, as risk indicators.

summary statistic	summary statistic value	
	NO GO	GO
Accuracy (TAR)	0.9993	**0.9999**
F_1-score	undefined	0.9389
Youden Index	undefined	0.88500
Low impact ι_I ($\iota_{F\_} = 90.00$)	**$99.99**	$99.94
ι_I ($\iota_{F\_} = 0.001$)	$99.94	$99.94
ι_I ($\iota_{F\_} = -1.00$)	$99.93	$99.93
ι_I ($\iota_{F\_} = -20.00$)	$99.92	$**99.93**
Medium impact ι_I ($\iota_{F\_} = -100.00$)	$99.87	$**99.93**
High impact ι_σ ($\iota_{F\_} = -10,000.00$)	–$4937	**—$941.7**

Table 10.9: The bolded values for each summary statistic indicate the decision supported. The F_1-*score* and Youden Index are undefined for the "NO GO" case. However, *TAR* supports a "GO" decision. In contrast, the impact summary statistics show that the decision is not so clear-cut. A "GO" decision is justifiable when $\iota_{F\_} \leq -20.00$, but is only strongly supported for high impact events. However deploying the IDS has a significant stabilizing effect on impact. This may be important to decision makers.

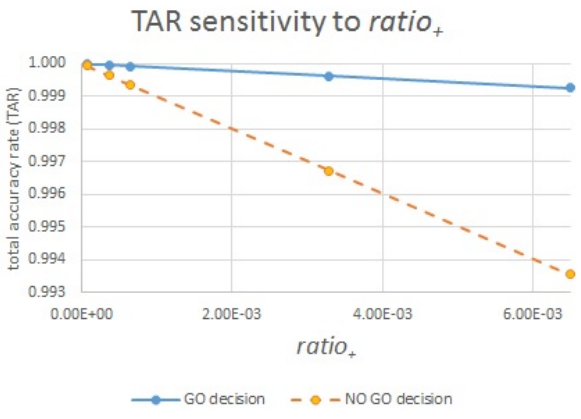


Figure 10.7: TAR suggests that in all cases, the target system is more effective with the classifier than without.

10.4.3 Results

Table 10.9 shows the results on the raw data. To illustrate the difference in actionable information decision makers receive when using ι_I and ι_σ, *TAR*, F_1-*score* and Youden Index, values are calculated. Since the "NO GO" scenario has

no *test=negative* output, the F_1*-score* and Youden index are undefined[2]. However, *TAR* supports a "GO" decision. Figure 10.7 shows *TAR*'s GO/NO GO results for the five $ratio_+$ scenarios evaluated; at all points, the "NO GO" line is below the "GO" line, implying that deploying Kratos will reduce the incidence of malicious events, thereby reducing their overall impact on the enterprise.

In contrast, Table 10.9 shows that the impact summary statistics only strongly support a "GO" decision for high impact events. Figure 10.8 shows that for events with modest impact, a "GO" decision is supported when $\iota_{F\_} \leq -20.00$. The detector is contraindicated for lower impact events. However, the IDS does have a strong stabilizing effect on the expected impact — which decision makers may find appealing. These insights are not available from the commonly seen summary statistics.

Comparing the ι_σ values for high impact (irrecoverable) events, both decisions have negative values, but the "NO GO" decision is about five times more negative than the "GO" decision. A "GO" decision is strongly supported for high impact (irrecoverable) events. However, the large negative value indicates that the enterprise still has substantial residual risk; decision makers may want to implement additional risk mitigations.

ι_I also enables field practitioners to compare actual and expected results. ι_Z and $\iota_{\overline{Z}}$ show the expected per event impact of CPD output. Figures 10.9a and 10.9b show the expected impacts observable in the field. ι_Z is roughly a factor of 10^7 greater than $\iota_{\overline{Z}}$. Hence, CPD *test=true* output is the controlling observable. Figure 10.9b shows that CPD *test=false* output is $\iota_{F\_}$ invariant, but this is an artifact of the values selected (see Table 10.8); the invariance might not exist in a real-world evaluation. Once the observed $\iota_{F\_}$ and ι_I are calculated, a specific $ratio_+$ line will pass through that ($\iota_{F\_}$, ι_I) point. This is the implicit $ratio_+$. If this value varies unacceptably from the expected value, then the underlying CPD tool assessment assumptions may not represent reality; further investigation could be justified.

[2]This is an artifact of the "NO GO" scenario. If the alternate had even one *test=positive* event, then the F_1*-score* and Youden Index could be used for comparison.

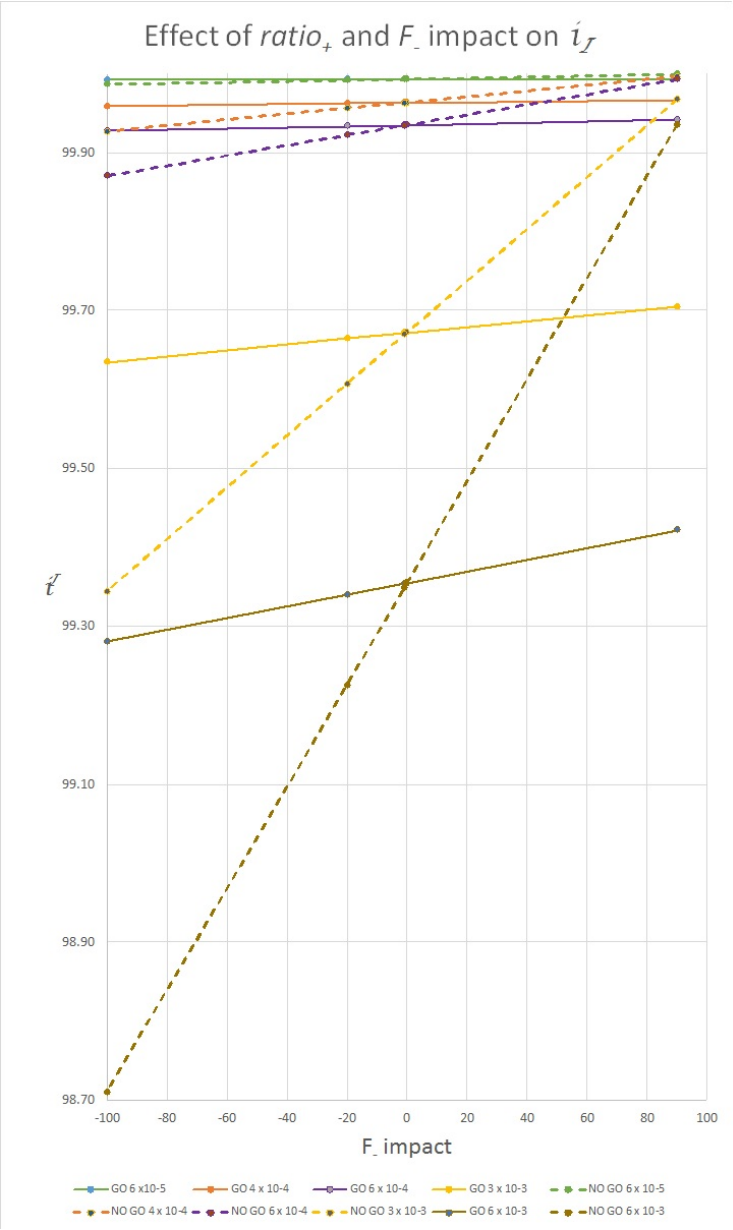
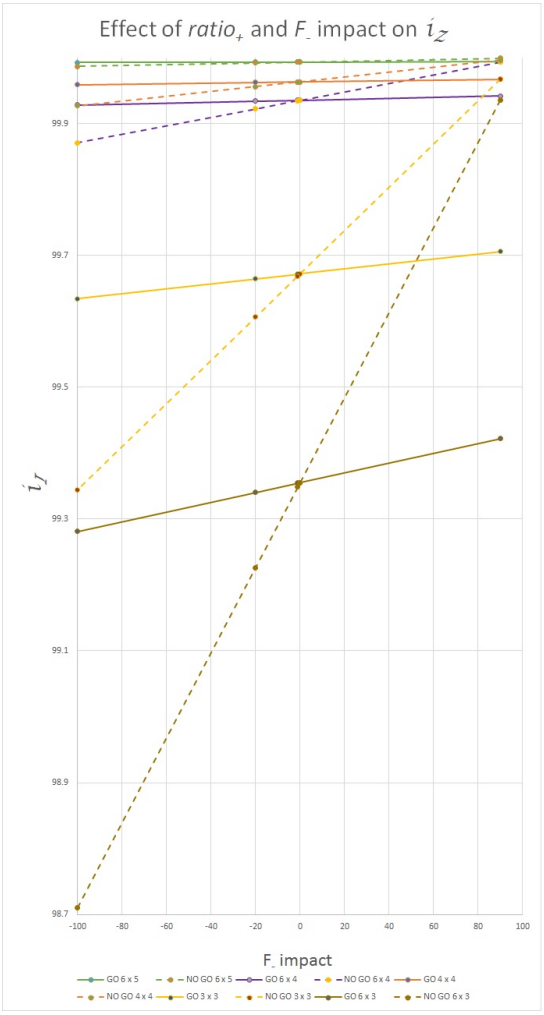


Figure 10.8: This figure shows ι_I versus $\iota_{F\_}$ for GO decisions (solid lines) and NO GO decisions (dotted lines) under five different $ratio_+$ conditions. In every case, the two lines intersect. This indicates that the "GO" decision (deploying Kratos ICS IDS) is supported for events where $\iota_{F\_}$ is negative (left of the intersection), but not when $\iota_{F\_}$ is positive (right of the intersection). The figure also shows that with Kratos installed, the system's output (as quantified by ι_I) is more stable than without Kratos. This reduction in volatility may be important to decision makers.

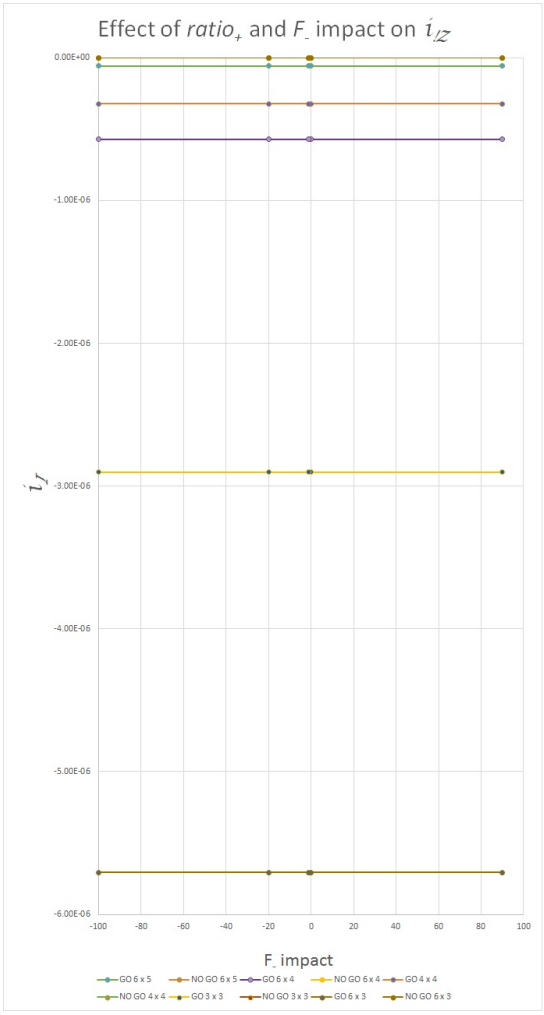
10.5 Example findings

Each of the examples in this chapter apply the impact summary statistics. In the re-evaluation cases, impact-based test results were compared with the original findings. The original research included both conventional and impact summary statistics. In every case, end users would have benefited from the additional insights resulting from impact-based evaluation.

In the Egyptian bank loan analysis, JPT tuning showed that the original conclusion was incorrect. A $ratio_+$ sensitivity analysis illustrated that an end user



(a) Expected *test=true* CPD output impact



(b) Expected *test=false* CPD output impact

Figure 10.9: Expected CPD output impact (that is, the expected impact of the CPD's "event is normal" and "event is anomalous" outputs) can be calculated. End users can use the test results shown in these figures to determine if the CPD is performing as expected. Discrepancies may be grounds for further investigation.

could estimate the expected loan decision impact. These additional insights could improve bank loan policies.

The rheumatoid arthritis diagnostic test re-analysis puts the researcher's findings into an easily understood context: currency. ι_σ has a meaningful zero: a correct diagnosis results in the appropriate treatment and the best quality of life (zero negative benefits). The greater the negative value, the worse the expected result is for patients. On the other hand, likelihood ratios do not have a meaningful zero. Instead, the distance from one is important. However, the two summary

statistics have different bounds. LR_- has a upper bound of one and LR_+ has an upper bound of infinity. This makes interpretation less intuitive.

The cyber security masquerade study re-analyzed Schonlau's data for two users. My analysis culminated with a downselect matrix, which considered the expected impact for both low and high risk events as well as boundary and $ratio_+$ sensitivity. The insights gained from the impact-based analysis showed that the two users would realize the best results from different CPD algorithms.

The industrial control system intrusion detection study showed that each of the axioms (and the measure theory derived insight) have value to end users. The proposed summary statistics provide end users with a richer understanding of CPD tool effectiveness, enabling a more informed deployment decision.

In each of these examples, applying the appropriate impact summary statistic generated actionable information not revealed by other summary statistics. Hence, the theoretically derived solution holds up in real-world application.

CHAPTER 11

USER STUDY

The initial work was inspired by anecdotal stories and informal discussions with end users in cyber security. While this may be sufficient to initiate a work such as this, the claimed utility (and significance of the contribution to the body of knowledge) remains hypothetical until validated. Hence, when the opportunity arose, a group of cyber security end users were surveyed. Their response provided a qualitative means of assessing whether actual end users believed this dissertation's outcomes would improve their CPD tool evaluations. In conjunction with the body of research on medical practitioners mentioned in Chapter 10.2, this study illustrates that end users see value in quantifying impact.

11.1 Study protocol

The study consisted of an initial questionnaire to evaluate the respondent's experience with classifier assessment and selection. This was followed by working through the ICS IDS use case in Section 10.4, then discussing the respondent's opinion of the proposed summary statistic's utility.

11.2 Results

There were five participants, hence the group size was too small for a rigorous statistical analysis. However, three conclusions were possible:

- the more experience end users had with classifier evaluation, the less satisfied they were with the existing summary statistics.
- after exposure to ι_I and ι_σ and viewing a use case applying them as well as Accuracy, F_1-score and Youden index, the respondents overwhelmingly felt that ι_I and ι_σ provided substantial additional actionable information.
- the extra effort needed to define I was justified.

Respondents observed some potential impact summary statistic limitations:

- *There are classifier selection considerations other than impact.* Issues such as ease-of-use and response timeliness also can sway a decision. Impact provides better insight into only one performance factor.

- *The impact summary statistics are too new.* The impact summary statistics may be used in conjunction with others. That will provide different views of performance and give the user community time to decide the impact summary statistic's value.

- *Determining I will consume time.* There may be occasions when, for whatever reason, I cannot be determined. Over time, default I values may be determined for certain problem domains. However, just knowing that one factor is missing may benefit decision makers. They will have a better sense of the information quality upon which they are deciding.

- *I quantifies statistical expectations.* The entire classifier evaluation is based on inputs that are presumed representative. The impact summary statistics do not change that challenge. They do, however, provide the means for better quantifying how well a classifier will satisfy their need.

- *By definition, F_-s are not observed, so ι_{F_-} must be imputed.* Estimating ι_{F_-} is more prone to uncertainty than quantifying ι_{T_+}, ι_{F_+} and ι_{T_-}, which can be based on direct observation. However, uncertainty can be processed. The option of ignoring the value puts the decision maker in a weaker position.

These end user comments show that the impact summary statistics are not a panacea. When compared to the utility of other summary statistics, however, in no case does using an impact summary statistic degrade an end user's decision making ability.

CHAPTER 12

CONCLUSIONS

The purpose of this dissertation was to identify end user efficacious summary statistics for CPD evaluation. Such summary statistics, then, constitute the target deliverable.

12.1 Key findings

A gap analysis indicated the root cause for the need was the lack of a framework characterizing CPD problems. The first deliverable was the missing framework. The framework consists of two factors:

Relative class size ($ratio_+$) End user CPD problems can be partitioned into two types. Either $ratio_+$ is confounding or $ratio_+$ is explanatory.

Expected impact per JPT category element Every CPD event impacts end users. Since events can be binned into one of four types, an expected impact can be calculated for each bin.

One obstacle to finding end user efficacious summary statistics was the commonly held belief that useful summary statistics must be $ratio_+$ invariant. I determined that this belief was incorrect; JPT normalization confers $ratio_+$ invariance. This breakthrough opened up the set of possible summary statistics, making the target deliverable achievable. Hence, one key outcome was breaking the dependence between CPD summary statistics and $ratio_+$ invariance.

A secondary use of JPT normalization that emerged was JPT tuning. It is possible to adjust a JPT to reflect any desired $ratio_+$. This makes it possible for end users with problems where $ratio_+$ is explanatory to execute $ratio_+$ sensitivity analyses. Thus the second outcome was introducing JPT normalization and JPT tuning to the CPD problem domain. (JPT normalization is an established procedure in probability, but not universally known in CPD research and development.)

Before a summary statistic can be tested for end user efficacy, success criteria must be defined. The third key outcome consists of four Axioms which an end user efficacious summary statistic must satisfy and tying efficacy to measurement theory, particularly scale type:

Axiom 1 (Category importance) *An end user efficacious summary statistic must be a function of problem specific impact vector* $I = (\iota_{T_+}, \iota_{F_+}, \iota_{F_-}, \iota_{T_-})$*, where each element of* $I \in \mathbb{Q}$*.*

Axiom 2 (Environmental sensitivity) *With a change in* pdf*(Y),*
(e.g., pdf*(Y′)*= Δ+pdf*(Y),* pdf*(S))*=pdf*(Y)*+ pdf*($\overline{Y}$) and* pdf*(S′)*=pdf*(S)*+Δ*), where* Δ *describes a perturbation in Y's and S's source population, for all boundaries within* Δ*, there exists a* $E(\text{SS}(B)|S')) - E(\text{SS}(B)|S)) \neq 0$*. The same is true for a change in* $\overline{Y}$ *and for any* ratio_+*.*

Axiom 3 (CPD output basis) *An end user efficacious summary statistic must be quantifiable in terms relative to information known and visible to the end user: CPD output (Z and* $\overline{Z}$*).*

Axiom 4 (Summary statistic value appropriateness) *An end user efficacious summary statistic's output must quantify the CPD's impact on the end user's characteristic of interest.*

Measurement theory. *Ratio scales allow the most extensive analysis, thus end users are best served by ratio scale summary statistics.*

The fourth key outcome is defining two end user efficacious summary statistics and their respective summary statistic suites:

$$\iota_I = \iota_{T_+}\frac{t_+}{|S|} + \iota_{F_+}\frac{f_+}{|S|} + \iota_{F_-}\frac{f_-}{|S|} + \iota_{T_-}\frac{t_-}{|S|}. \tag{12.1}$$

ι_I can also be expressed on Z and $\overline{Z}$, the outputs actually observed by the end user:

$$\iota_I = \iota_Z\frac{|Z|}{|S|} + \iota_{\overline{Z}}\frac{|\overline{Z}|}{|S|}. \tag{12.2}$$

ι_I is appropriate for CPD problems where $ratio_+$ is explanatory.

$$\iota_\sigma = \frac{1}{2}\left(\frac{\iota_{T_+}t_{+n}}{|Z_n|} + \frac{\iota_{F_+}f_{+n}}{|Z_n|} + \frac{\iota_{T_-}t_{-n}}{|\overline{Z_n}|} + \frac{\iota_{F_-}f_{-n}}{|\overline{Z_n}|}\right). \tag{12.3}$$

$$\iota_Z = \frac{\iota_{T_+}t_{+n} + \iota_{F_+}f_{+n}}{|Z_n|} \quad \text{and} \quad \iota_{\overline{Z}} = \frac{\iota_{T_-}t_{-n} + \iota_{F_-}f_{-n}}{|\overline{Z_n}|}. \tag{12.4}$$

ι_σ is appropriate for CPD problems where $ratio_+$ is confounding.

	Unbiased I		Biased I	
$ratio_+$ is	Summary Statistic	$I =$	Summary Statistic	$I =$
confounding	$n\iota_\sigma$	$(1,-1,-1,1)$	$n\iota_\sigma$	$(\iota_{T_+}, \iota_{F_+}, \iota_{F_-}, \iota_{T_-})$
	ROC-AUC	$(1,1,1,1)$	nF_β-*score*	$(1, \frac{\beta^2}{1+\beta^2}, \frac{1}{1+\beta^2}, 0)$
	DOR\DP	$(1,1,1,1)$		
	n*MCC*	$(1,1,1,1)$		
	n*IC*	$(1,1,1,1)$		
explanatory	ι_I	$(1,-1,-1,1)$	ι_I	$(\iota_{T_+}, \iota_{F_+}, \iota_{F_-}, \iota_{T_-})$
			F_β-*score*	$(1, \frac{\beta^2}{1+\beta^2}, \frac{1}{1+\beta^2}, 0)$
			TAR	$\iota_I \vert (1,0,0,1)$

Table 12.1: This matrix maps the summary statistics discussed to the CPD problem factors, $ratio_+$ and impact. The summary statistic names with an "n" prefix indicate these summary statistics must have JPT values normalized. J is not listed. As a scale transformed $n\iota_I$, it is not appropriate for $ratio_+$ is confounding CPD problems and since J is $ratio_+$ invariant, it is not suitable for $ratio_+$ is explanatory problems.

12.2 Summary statistic selection recommendations

Given the CPD problem space factors $ratio_+$ and impact, for which combinations are the summary statistics tested valid? Table 12.1 maps summary statistics to the problem space. The relationship between ι_I, J and *TAR* means that J and *TAR* are not suitable for CPD problems where $ratio_+$ is confounding. Since J is $ratio_+$ invariant, it is also not suitable for CPD problems where $ratio_+$ is explanatory.

Unbiased I is actually only a specific I, so the unbiased I column in Table 12.1 can be eliminated. *TAR* is only suitable for one specific I and F_β-*score* has a limited I range. Taking into consideration end user efficacy, all that remain are ι_I and ι_σ.

Although this dissertation's focus is end users, an end user efficacious summary statistic, $n\iota_I \vert (1,-1,-1,1)$, is suitable for application agnostic (TRL 3) evaluations.

12.3 Results reporting Recommendations

Basic research is application agnostic. As such, environmental effects are confounding. In contrast, environmental effects are explanatory for applied research and field deployments. In addition to using efficacious summary statistics,

published CPD test reports need to provide sufficient information for end users to estimate results in their own environments. This information goes beyond that needed by a basic researcher, so aiding end users may require intentionally including JPTs for a range of boundaries and $ratio_+$s.

12.4 Impact measuring process for practitioners

The value of this analysis is greatly diminished if end users cannot readily use the results. Fortunately, using the impact summary statistics only entails four steps:

1. Identify the expected impacts and define I. If the problem is complex, like the intrusion detection example, then multiple Is may be needed.
2. Identify the appropriate summary statistic, ι_I or ι_σ. Various aspects of a complex problem may need to be addressed separately and the same summary statistic may not be appropriate for all aspects.
3. If ι_σ 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 summary statistic. The boundary with the JPT which generates the best impact value is the optimum boundary. The optimum boundary will provide the best results for the target classifier. In the example given, the best impact is the the maximum. This, however, is problem dependent. In some cases, the best result may be a minimum.

With these values, end users can also determine the classifier output's boundary sensitivity.

12.5 Future work

It seems that one research project invariably leads to other opportunities. Potential extensions to this work include:

- In addition to CPD evaluation, other problem types can be expressed using the model illustrated in Figure 4.1. This suggests that other problems domains may benefit from this work.
- There is a relationship between the *ROC* curve and impact. Does a similar relationship exist for a CPD output conditioned "*ROC*-like" curve?
- The report recommendations in Section 12.3 identify necessary information. However, researchers and end users alike may benefit from a uniform format and means of identifying useful $ratio_+$ and boundary ranges.

- My results show that sometimes summary statistics, including ι_I and ι_σ can be multi-modal. How can end users take advantage of this information?
- There are an infinite number of possible I. However, some I ranges may be more common than others, particularly when considering single problem domains. This information could facilitate end user efficacy.
- This study focuses on classifier evaluation. However, any decisioning system or subsystem's output can be expressed in a JPT, hence, the summary statistics developed herein are potentially applicable. One field of note is Cyber Security Econometrics. Since risk-based cyber security is gaining traction, the impact summary statistics may be valuable.
- In this work, class samples needed to be greater than four hundred observations to benefit from the strong law of large numbers. Are there factors that influence this sample size and can the negative effect of the law on small sample sizes be mitigated?
- Given I and normalized JPT data, end users can calculate the expected impact for both Z and $\overline{Z}$ CPD outputs for any $ratio_+$. the reverse should also be true: given I, normalized JPT data and the observed, expected impact for both Z and $\overline{Z}$ CPD outputs, $ratio_+$ can be estimated. This is a new capability. What are its practical implications?

APPENDIX A

RESTATING F_1 IN TERMS OF JPT VALUES

As defined,

$$F_1 = 2\frac{(precision)(recall)}{precision + recall}.$$

where

$$precision = \frac{T_+}{T_+ + F_-}$$

and

$$recall = \frac{T_+}{T_+ + F_+}.$$

Substituting yields

$$F_1 = 2\frac{\frac{T_+}{T_+ + F_-}\frac{T_+}{T_+ + F_+}}{\frac{T_+}{T_+ + F_-} + \frac{T_+}{T_+ + F_+}}.$$

Multiplying and creating common denominators,

$$F_1 = \frac{\frac{2T_+^2}{(T_+ + F_-)(T_+ + F_+)}}{\frac{T_+(T_+ + F_+) + T_+(T_+ + F_-)}{(T_+ + F_-)(T_+ + F_+)}}.$$

Multiplying numerator and denominator by $\frac{(T_+ + F_-)(T_+ + F_+)}{T_+}$ leaves

$$F_1 = \frac{2T_+}{2T_+ + F_+ + F_-} = \frac{T_+}{T_+ + \frac{F_+}{2} + \frac{F_-}{2}}.$$

APPENDIX B

RESTATING *DP* IN TERMS OF JPT VALUES

$$
\begin{aligned}
DP &= \frac{\sqrt{3}}{\pi}(logU + logW) \text{ ,where} \\
U &= \frac{sensitivity}{1 - sensitivity}, \\
W &= \frac{specificity}{1 - specificity}, \\
sensitivity &= \frac{T_+}{Y}, \\
1 - sensitivity &= \frac{F_-}{Y}, \\
specificity &= \frac{T_-}{\overline{Y}} \text{ and} \\
1 - specificity &= \frac{F_+}{\overline{Y}}.
\end{aligned}
$$

combining the logs yields

$$DP = \frac{\sqrt{3}}{\pi}(log(UW)).$$

Then substituting for U and W,

$$DP = \frac{\sqrt{3}}{\pi}\left(log\left(\frac{sensitivity}{1 - sensitivity}\frac{specificity}{1 - specificity}\right)\right).$$

Substituting for sensitivity and specificity,

$$DP = \frac{\sqrt{3}}{\pi}log\left(\frac{\frac{T_+}{Y}\frac{T_-}{\overline{Y}}}{\frac{F_-}{Y}\frac{F_+}{\overline{Y}}}\right).$$

Multiplying top and bottom by $Y\overline{Y}$ yields

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

APPENDIX C

DERIVATION OF NORMALIZED *MCC* EQUATION

MCC, as commonly calculated,

$$MCC = \frac{(T_+ * T_-) - (F_+ * F_-)}{\sqrt{Y * \overline{Y} * Z * \overline{Z}}}, \tag{C.1}$$

is *not ratio*$_+$ invariant as is sometimes reported [8, 9, 12, 20, 44, 50, 31]; it must use normalized JPT values (as in Table C.1).

Actual target classification		Y	$\overline{Y}$	
Test	*Positive*	$\frac{T_+}{Y}$	$\frac{F_+}{\overline{Y}}$	
Result	*Negative*	$\frac{F_-}{Y}$	$\frac{T_-}{\overline{Y}}$	
Normalized totals		1	1	2

Table C.1: The values in this JPT have been normalized.

Substituting the normalized JPT values in Equation C.1 and collecting terms, the *ratio*$_+$ invariant *MCC* is:

$$\textit{normalized MCC} = \frac{(T_+ * T_-) - (F_+ * F_-)}{\sqrt{Y * \overline{Y} * (\overline{Y} * T_+ + Y * F_+) * (Y * T_- + \overline{Y} * F_+)}}. \tag{C.2}$$

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

APPENDIX D

CREATING A SCALABLE IMPACT SUMMARY STATISTIC

Impacts are problem specific, but it is possible for I sets to differ by a multiplicative constant (e.g., $I_1 = (1, -2, -3, 4)$ and $I_2 = (2, -4, -6, 8)$). It may be possible for an analysis to address multiple end users by reporting results on normalized Is (hereafter represented by $\Im_x$, where x represents a specific normalized I), from which end users can readily extract their relevant impact values.

Considering I_1 and I_2, $\Im_a = (0.25, -0.50, -0.75, 1.0)$. Hence, $I_1 = 4 * \Im_a$ and $I_2 = 8 * \Im_a$, where 4 and 8 are weights (ω) specific for end users 1 and 2. Conveniently, if ι_I and ι_σ are calculated using $\Im_a$ (i.e. $\iota_I(\Im_a)$ and $\iota_\sigma(\Im_a)$) instead of user-specific I, then the impacts for suitable end users are ω multiples of $\iota_I(\Im_a)$ and $\iota_\sigma(\Im_a)$. Considering my two end users, $\iota_I(1) = \omega_1 * \iota_I(\Im_a)$ and $\iota_\sigma(1) = \omega_1 * \iota_\sigma(\Im_a))$; $\iota_I(2) = \omega_2 * \iota_I(\Im_a)$ and $\iota_\sigma(2) = \omega_2 * \iota_\sigma(\Im_a))$.

It is possible that end users may not need exact $\Im$s, approximations may provide acceptable accuracy. Considering the bounds on using approximations is reserved for future work.

APPENDIX E

ABDOU CREDIT SCORING JPTS

My re-analysis of Abdou's results used JPT tuning to estimate ι_I for the actual test set. Two other $ratio_+$s ($ratio_+ = 0.67$ and $ratio_+ = 0.40$) were estimated to illustrate how JPT tuning can be used for sensitivity analysis. The JPTs for WOE and GP_t are shown in Tables E.1 and E.2 below. Sometimes normalized JPTs are set such that the total test set size is two ($|S| = 2$). Since ι_I is additive and I want to calculate the estimated impact per loan unit, I scale the the proportions so that the total test set size is one $|S| = 1$. The proportional totals line shows the scaling factors used for each JPT. The weight of evidence (WOE)and GP_t JPTs are shown; Abdou's other JPTs are scaled in the same manner.

JPTs used for Abdou GP$_t$ credit scoring re-analysis

$ratio_+ = 1$		Actual classification		
		Good risk	bad risk	
Loan decision	make loan	0.4724	0.2024	
	deny loan	0.0276	0.2976	
Proportional totals		0.5000	0.5000	1

$ratio_+ = 0.67$		Actual classification		
		Good risk	bad risk	
Loan decision	make loan	0.5669	0.1619	
	deny loan	0.0331	0.2381	
Proportional totals		0.6000	0.4000	1

$ratio_+ = 0.48$		Actual classification		
		Good risk	bad risk	
Loan decision	make loan	0.6361	0.1323	
	deny loan	0.0371	0.1945	
Proportional totals		0.6732	0.3268	1

$ratio_+ = 0.40$		Actual classification		
		Good risk	bad risk	
Loan decision	make loan	0.6749	0.1156	
	deny loan	0.0394	0.1701	
Proportional totals		0.7143	0.2857	1

Table E.1: Four GP$_t$ JPTs tuned for $ratio_+$ used to illustrate how the proposed summary statistic (ι_I) enables sensitivity testing.

JPTs used for Abdou WOE_{T2} credit scoring re-analysis

$ratio_+ = 1$		Actual classification		
		Good risk	bad risk	
Loan decision	make loan	0.3603	0.1389	
	deny loan	0.1397	0.3611	
Proportional totals		0.5000	0.5000	1

$ratio_+ = 0.67$		Actual classification		
		Good risk	bad risk	
Loan decision	make loan	0.4324	0.1111	
	deny loan	0.1676	0.2889	
Proportional totals		0.6000	0.4000	1

$ratio_+ = 0.48$		Actual classification		
		Good risk	bad risk	
Loan decision	make loan	0.4869	0.0901	
	deny loan	0.1888	0.2342	
Proportional totals		0.6757	0.3243	1

$ratio_+ = 0.40$		Actual classification		
		Good risk	bad risk	
Loan decision	make loan	0.6749	0.1156	
	deny loan	0.0394	0.1701	
Proportional totals		0.7143	0.2857	1

Table E.2: Four WOE JPTs tuned for $ratio_+$ used to illustrate how the proposed summary statistic (ι_I) enables sensitivity testing.

APPENDIX F

MAPPING STAKEHOLDER NEEDS TO RELEVANT SUMMARY STATISTIC VARIABLES

Over time, useful technical concepts evolve into useful tools and project underwriters shift from basic researchers to developers and end users. To manage technology development and product acquisition risk, NASA developed an assessment framework, based on a nine level "technical readiness" (TRL) scale. TRL has been subsequently adopted by the US Department of Defense (DoD) [22, 56]. As technology development programs progress, project underwriters will want to assess technical maturity. DoD's protocol specifies integrated project review (IPR) teams for these assessments. This work has shown that summary statistics relevant for basic research may not be relevant for end users. This Chapter maps CPD technology maturity to relevant summary statistics, ultimately making recommendations regarding TRL level and appropriate summary statistics.

F.1 Related figures of merit

Sequentially staged maturity modeling, the basis upon which technology readiness levels may have been based, was originally published by Crosby in 1979 [19]. TRL's use by NASA was first reported in 1989 [72]. TRL brought structure to a process that, hitherto, had been ad hoc, based on intuition and experience. TRL provided a problem domain independent frame of reference for assessing technology maturity. Although a major improvement in technical innovation management, TRL is not a complete solution and has spawned a number of enhancements.

One difficulty, whenever a complex analysis is condensed into a single value, is that information is lost. To more closely match TRL to program management needs, readiness level development continues [28, 55, 56, 76, 87].

Marketing research considers a seemingly similar concept, technology acceptance. In contrast to technology maturity's utility emphasis, technology acceptance heavily emphasizes human response. Each of the technology acceptance assessment schemes include technology utility (or perceived utility), but human factors dominate the model [33, 51, 13, 91].

Applicability of the two assessment types may be tied to the related problem's risk. If a sub-optimal solution can result in a critically unacceptable stakeholder state (lead to a crisis), then a TRL-based assessment may be appropriate. If a sub-optimal solution cannot lead to a crisis, or a substantial negative impact on stakeholders, then technology acceptance may be the appropriate approach. I focus on the first scenario, not the second.

Although each of the TRL-based assessment scales mentioned is in some way unique, I focus on one similarity; the effectiveness of the technology's core capability must be tracked. My interest is CPD tools, hence, I discuss the need using software TRL [87].

F.2 CPD stakeholders, TRL and summary statistic variables

The DoD [22], defines nine TRL levels. The initial TRL (TRL 1) is achieved when basic principals have been identified, but methods and uses have not; a fully mature technology (TRL 9) has been deployed and has proven operational capabilities. Along the path from TRL 1 to TRL 9, stakeholders and stakeholder interests (and thus IPR perspectives) change. This dissertation considers summary statistic efficacy for the various stakeholders, mapping summary statistic efficacy to TRL level.

In this section, I consider the characteristics of the seven CPD tool evaluation summary statistics mentioned in Chapter 5 and the two summary statistics I introduced in Chapter 9 and map them to the software TRL scale [87].

Qualitative core concept development At inception, a novel idea may not be well defined, or perhaps even clearly understood. Without a clear understanding, quantitative testing has little value; evaluations are qualitative. TRL 1 and TRL 2 recognize these early stages:

TRL 1 Basic principles observed and reported

TRL 2 Technology concept and/or expression formulated

TRL 1 and TRL 2 are qualitative, hence, no summary statistics are relevant.

Quantitative core concept development In initial CPD concept testing, the primary audience is basic researchers. As a concept CPD's capabilities are refined, applied researchers become interested in testing CPD tools against specific problems. An effective concept CPD will ultimately gain the attention of developers and end users, who are interested in estimating results in actual deployments. TRL 3, 4 and 5 can be mapped to these three groups.

TRL 3 — Environment agnostic: Analytical and experimental critical function and/or characteristic proof of concept.

At this level, focus is on CPD algorithm function. Because of the nature of the process, environmental effects cannot be avoided, but environmental bias can, by specifying $ratio_+$ invariance and unbiased I. Of the summary statistics reviewed, ROC-AUC, J and *DOR/DP* are $ratio_+$ invariant, so meet the $ratio_+$ requirement. After JPT normalization, *MCC* and *IC* are also $ratio_+$ invariant and have an unbiased I.

$TAR = \iota_I$, but with $I = (1, 0, 0, 1)$, thus is inherently biased. The same is true of F_β-score. Regardless of β chosen, I is biased ($I = (1, 1/(1+\beta^2), \beta^2/(1+\beta^2), 0)$). F_β-score and *TAR* are unsuitable for TRL 3 evaluations. Unfortunately, *ROC-AUC*, *J*, *DOR/DP*, *TAR*, *MCC* and *IC* are measured on ordinal, rather than ratio scales, so valid analyses are limited; ratio scale summary statistics are better [36].

The summary statistic ι_I introduced in Chapter 9, when applied to normalized JPTs and with $I = (1, -1, -1, 1)$ is unbiased and is measured on a ratio scale, so does not suffer the analysis limitations of the other summary statistics reviewed. A second summary statistic also introduced in Chapter 9, ι_σ, is conditioned on CPD algorithm output (Z and $\overline{Z}$), this conditioning is useful for end users, but TRL 3 studies are end user agnostic, so the additional operation makes ι_σ unsuitable for TRL 3 evaluation.

In summary,

Invalid summary statistics: *F_β-score*, *TAR* (inherent I bias) and ι_σ (unnecessary JPT conditioning)

Limited summary statistics: *ROC-AUC*, *J*, *DOR/DP*, *MCC* and *IC* (measured on ordinal scales)

Unlimited summary statistic: ι_I (with $I = (1, -1, -1, 1)$ on normalized JPT; measured on a ratio scale)

TRL 4 — Environment perceived: Module and/or subsystem validation in a laboratory environment.

At this level, focus is on defined problem domain(s). Regardless of problem domain, JPT category impact (I) is important. However, I cannot be as precisely defined as for a specific deployment ($I = ([P, P'], [Q, Q'], [R, R'], [S, S'])$, where the capitalized range boundaries represent domain specific bounds). *ROC-AUC*, *J*, *DOR/DP*, *MCC* and *IC* have implicitly fixed $I = (1, 1, 1, 1)$, *TAR* has implicitly fixed $I = (1, 0, 0, 1)$, so these have little value for TRL 4 evaluations. Only F_β-score, ι_σ and ι_I have a variable I. Of these three I sensitive summary statistics, F_β-score's I only has one degree of freedom (β): there are some domains where it is not useful. Also, F_β-score is measured on ordinal, rather than a ratio scale, so valid analyses are limited. Of the summary statistics reviewed, only ι_σ and ι_I are valid for all I.

Problem domains are partitioned into two types; those where environmental $ratio_+$ is confounding ($ratio_+ = 1$) and those where environmental $ratio_+$ is important ($ratio_+ = [N, N']$). ι_σ is tailored for $ratio_+$

is important problem domains, ι_I is tailored for $ratio_+$ is confounding problem domains.

In summary,

Not I-tunable, ordinal scale summary statistics: *ROC-AUC*, *J*, *TAR*, *DOR/DP*, *MCC* and *IC*

Limited I-tunable, ordinal scale summary statistic: F_β-score

Unlimited I-tunable, ratio scale summary statistics: ι_σ ($ratio_+$ is important problem domains), ι_I ($ratio_+$ is confounding problem domains)

TRL 5 — Environment defined: Module and/or subsystem validation in a relevant environment

At this level, focus is on a specific deployment, rather than a problem domain. As such, I and $ratio_+$ (if important) can be more narrowly defined: $I = ([p, p'], [q, q'], [r, r'], [s, s'])$ and $ratio_+ = [n, n']$. Lower case range boundaries represent deployment specific bounds. Problem knowledge is greater than for TRL 4 evaluations, otherwise the evaluation criteria are the same. Hence, the TRL 4 summary statistic conclusion applies for TRL 5 evaluations:

Not I-tunable, ordinal scale summary statistics: *ROC-AUC*, *J*, *TAR*, *DOR/DP*, *MCC* and *IC*

Limited I-tunable, ordinal scale summary statistic: F_β-score

Unlimited I-tunable, ratio scale summary statistics: ι_σ ($ratio_+$ is important problem domains), ι_I ($ratio_+$ is confounding problem domains)

Peripheral feature development Rarely can a deployable, effective CPD tool exhibit solely the core capability. End users will need additional features for effective field operation. Such features include a user interface, error management and cyber security. These features are independent of the environment within which the tool will be deployed. Thus, the discussion and evaluation for TRL 4 applies for TRL 6 through TRL 9.

TRL 6 Module and/or subsystem validation in a relevant end-to-end environment

TRL 7 System prototype demonstration in an operational high fidelity environment

TRL 8 Actual system completed and mission qualified through test and demonstration in an operational environment

TRL 9 Actual system proven through successful, mission-proven operational capabilities

F.3 Summary statistic recommendations

The software TRL, as defined by the DoD, has nine milestones. Of these, TRL 1 and TRL 2 are CPD algorithm evaluation precursors. TRL 3, TRL 4 and TRL 5 vary in their CPD algorithm's core capability evaluation; TRL 6 through TRL 9 focus on peripheral deployment issues. My interest is an IPR efficacious evaluation of a CPD algorithm's core capability, the focus of TRL 3, TRL 4 and TRL 5. Of these, TRL 4 and TRL 5 vary by the extent of environmental knowledge — the accuracy and precision with which $ratio_+$ and I can be specified. Regarding summary statistic selection, the appropriate summary statistic characteristics change between TRL 3 and TRL 4. A TRL 3 evaluation is intended to quantify results for researchers requiring application and environmental invariance. TRL 4 and TRL 5 evaluations are intended to quantify results for applied researchers and end users, both of whom are affected by environmental effects. Table F.1 summarizes this dissertation's findings.

The key requirement for TRL 3 evaluation is that the summary statistic be unbiased, both with regard to I and $ratio_+$. Two summary statistics are invalid because they are inherently biased, *TAR* and F_β-score. Although otherwise suitable, ι_σ incorporates JPT conditioning unnecessary for TRL 3 studies, so is excluded; the other six summary statistics reviewed are either inherently unbiased, or unbiased on normalized JPTs. *ROC-AUC*, *J*, *DOR/DP*, *MCC* and *IC* are all measured on ordinal scales, so are limited. ι_I includes JPT category impacts; potential bias is avoided by normalizing JPTs and using $I = (1, -1, -1, 1)$ — impacts for correct and incorrect CPD algorithm output are equally scaled and oppositely signed. It also has the advantage of being measured on a ratio scale. Ratio scale summary statistics do not have any limitations on valid statistical analyses. Hence, ι_I is the best TRL 3 summary statistic considered.

TRL 4 and TRL 5 CPD tool studies factor in problem distinctives and environmental factors. F_β-score is suitable for "$ratio_+$ is important" domains where $I = (1, 1/(1+\beta^2), \beta^2/(1+\beta^2), 0))$ is valid, but is still limited by not being measured on a ratio scale. ι_I is suitable for all "$ratio_+$ is important" end user problem

	Summary statistic								
Criterion	ι_I	ι_σ	*TAR*	F_β*-score*	J	*MCC*	*IC*	*DOR*	AUC
TRL 3	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes
TRL $\geq$4	Yes	Yes	Limited	Limited	No	No	No	No	No
Ratio scale	Yes	Yes	Yes	No	No	No	No	No	No

Table F.1: Summary statistic applicability: In contrast to the other summary statistics considered, ι_I addresses the user's needs for all three relevant TRL (TRL 3 for basic research, TRL 4 and TRL 5 for $ratio_+$ sensitive problems). For $TRL \geq 4$, ι_I and ι_σ (for $ratio_+$ invariant problems) address end user's needs. The ι summary statistics are also the only ratio scale summary statistics.

domains and basic research, ι_σ is suitable for "$ratio_+$ is confounding" end user problem domains. Further, both ι_σ and ι_I are measured on ratio scales, so all statistical analyses are valid. Hence, ι_σ and ι_I are the best TRL 4 and TRL 5 summary statistics considered.

Because TRL 6 through TRL 9 do not develop a CPD tool's core capabilities, ι_σ and ι_I are the best summary statistics for any TRL ≥ 4.

APPENDIX G

PUBLISHER TERMS OF SERVICE

G.1 Crosstalk Magazine

G.2 Hindawi Publishing

Terms of Service
Updated: March 02, 2017

G.2.1 Introduction

These terms of service (Terms) together with our Privacy Policy (https://www.hindawi.com/privacy/) set out the basis on which you may browse and use our website, available at https://www.hindawi.com and its subdomains and any services offered through it, including the Manuscript Tracking System (MTS) (collectively, the Site).

The Site is owned and operated by Hindawi Limited, a company registered in England and Wales with registered no. 08671628 and registered address at Hindawi Ltd., Adam House, Third Floor, 1 Fitzroy Square, London W1T 5HF, UK (we, us or our).

Your use of the Site will be governed by these Terms which serve as a legal contract between us and you. By browsing or using the Site and/or any services, content, or materials made available through the Site, you are agreeing

to be legally bound by the Terms and our Privacy Policy (https://www.hindawi.com/privacy/).

You may access the Site as an unregistered user but will need to register with us for an account (MTS Account) in order to submit, edit, or review any articles intended for publication in one of our journals. Any submissions, edits, information, or data (in any form) you make via your MTS Account including without limitation uploading of any articles or edits to articles are described in these Terms as your Contribution. In addition to these Terms, you acknowledge and agree that you will comply with any Author Guidelines, Editorial Workflows, Publication Ethics, or other policies applicable to the journal(s) you contribute to via your MTS Account (Policies).

In these terms Intellectual Property Rights means any patents, registered and unregistered trade-marks and service marks, domain names, registered designs and design rights, copyright (including such rights in computer software and databases), database rights and moral rights (in each case for the full period thereof and extensions, revivals, and renewals thereof), applications for the foregoing and the right to apply for any of the foregoing anywhere in the world, and all similar rights anywhere in the world including those subsisting in inventions, designs, drawings, and computer programs.

We may revise the Terms at any time by amending this page. Your use of the Site will be subject to the most recent version of the Terms available on the Site. We recommend that you read through the Terms available on the Site regularly so that you can be sure that you are aware of any changes that may apply to you. General Availability of the Site

We do not guarantee that the Site, or any content on it, including the MTS, will always be available, uninterrupted, up to date, or otherwise free from errors, omissions, bugs, or viruses.

Access to the Site is permitted on a temporary and as-is basis. We may suspend, withdraw, discontinue, or change all or any part of the Site without notice. We will not be liable to you if for any reason our site is unavailable at any time or for any period.

You are responsible for regularly creating backup copies of your Contribution.

You are responsible for your access to the Site, including:

- ensuring that all persons who access the Site through your internet connection are aware of these Terms and other applicable terms and conditions, and that they comply with them; and
- that your internet enabled device and telecommunications systems carry the appropriate anti-virus software necessary to minimise the risk of any harmful viruses infecting your internet enabled device.

G.2.2 General Conduct

At all times when using or accessing the Site you represent, warrant and undertake that you will:

- not breach any applicable laws, statutes, rules, regulations, guidelines, directives, and codes;
- not use the Site to seek or offer any goods or services;
- if you are a MTS Account holder, respond promptly to communications from us or any other MTS Account holder;
- not publish, post, upload, store, distribute, or disseminate anything which solicits or harvests another users password or other account information;
- not reverse engineer, decompile, disassemble, or otherwise attempt to obtain the Sites source code;
- not misuse the Site by knowingly introducing viruses, Trojans, worms, logic bombs, third-party or external links, or other malicious or harmful material;
- not attempt to gain unauthorised access to the Site, the server on which the Site is stored or any server, computer, or database connected to the Site;
- not attack the Site via a denial of service attack or a distributed denial-of-service attack; and
- comply with the Policies and all reasonable instructions given by us from time to time.

A failure to comply with our conduct expectations set out above, or any other part of these Terms, all of which we reserve the right to determine in our sole discretion, may result in the:

- suspension or termination of your MTS Account;
- deletion of your Contribution; and/or
- suspension or termination of your right to use the Site.

The actions above shall be without prejudice to any other rights or remedies which may be available to us. Your MTS Account

You must ensure that the information you provide during your registration is complete, accurate, and kept up to date. Your MTS Account is personal to you.

We reserve the right to validate the account information supplied at any time.

Once registered, it is your responsibility to keep your password and any other security information confidential. We will be entitled to assume that any person that logs into or uses your MTS Account is you. You must notify us immediately at hindawi@hindawi.com if you know or suspect that anyone other than you has had access to or knows your password. Intellectual Property Rights and Confidentiality

With the exception of articles published on our Site marked Open Access, which may be used in accordance with the terms of the licenses described in the Publication of Contributions section below, for all other pages of the site:

- all Intellectual Property Rights in the Site are owned, and will remain owned, by us or our licensors at all times. You may not copy or use any part of the Site other than as expressly permitted by the Terms; and
- you may print off one copy, and may download extracts, of any page(s) from the Site for your personal use and link to the Site as permitted under the Linking to Hindawi section below.

If you are a MTS Account holder you may deal with any Contribution in accordance with these Terms, our Policies, and our instructions to you.

You acknowledge and agree that articles (including all drafts, notes, and preparatory material and/or other Contributions related to the article) prior to first publication are highly confidential (Confidential Information). You warrant and agree that you will (a) hold all Confidential Information at all times under conditions of secrecy and will take all reasonable steps to preserve its confidentiality; (b) use Confidential Information solely for the purpose of performing any services authorised by us and for no other purpose whatsoever; and (c) not disclose Confidential Information to any third party or use it for the benefit of any third party other than as permitted by us (which may only be given on obtaining undertakings similar to those in this paragraph from such parties).

Prior to publication of your Contribution on the licence terms set out in the Publication of Contributions section below, you grant, and you represent and warrant, that you have the right to grant to us a non-exclusive, irrevocable, perpetual, transferable, sub-licensable, worldwide, royalty-free licence to publicly perform, copy, reproduce, display, make available, communicate to the public, modify, edit, manage, distribute, store, and publish any and all of your Contribution in print and electronic form, including on the MTS in order to prepare an article for publication in our journal(s). Publication of Contributions (Articles)

You acknowledge and agree that, if and once editorially accepted for publication, your Contribution shall be published by us under the Creative Commons Attribution 4.0 License (http://creativecommons.org/licenses/by/4.0/) and any data relating to the article (including without limitation any reference lists) is distributed pursuant to the Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/).

For the avoidance of doubt, you shall retain the non-exclusive right to publish or otherwise use your Contribution and shall be the Licensor of the Licensed Material for the purposes of the Creative Commons Attribution 4.0 Licence and Creative Commons Public Domain Dedication waiver.

If we consider that the copyright in your Contribution has been or is likely to be infringed, we shall be entitled (on giving notice where possible to you) to take such steps to deal with the matter as we may consider appropriate. You undertake not to do anything which may compromise or prejudice our conduct of such claim and we shall be entitled on giving notice to you to use your name and the name of any other co-author as a party to such proceedings and to control, settle, or compromise such proceedings as we see fit. You shall (and shall procure that any other co-authors shall) co-operate with and give all reasonable assistance to us in the conduct of such claim. After deduction of our costs in connection with such claim, half of any profits or damages which may be received in respect of any infringement of the copyright shall be retained by us with the remainder shared equally among you and any co-authors.

We reserve the right to refuse to publish any article or other Contribution for any reason.

By submitting or making a Contribution (including without limitation all tables, abstracts, illustrations, data, diagrams, and other materials accompanying the text of the Contribution) for the purposes of publication by us you represent and warrant that:

- you are the sole author of the Contribution, or, if not, you have been authorised by and have the full permission, licence, and consent of all co-authors to submit the Contribution;
- you are legally able and entitled to submit the Contribution and authorise us to publish the Contribution in accordance with the terms this section;
- the Contribution is original, has not already been published in any other form or is not currently under consideration for publication by another journal;
- you have obtained all necessary permissions and licences for any third-party material, data, or code contained in the
-
- Contribution (including without limitation all illustrations, charts, photographs, maps, or other material) and the
-
- Contribution does not infringe any copyright, trade mark right, or any other rights of any third party;
- you have disclosed to us in accordance with any Policies all potential conflicts of interest, including without limitation,

- personal or political interests, commercial associations, or financial interests held by you and any co-authors;
- nothing in the Contribution has been obtained in contempt of court or any violation of the Interception of Communications Act 1985, the Data Protection Act 1998 or the Official Secrets Act 1989 or any other analogous foreign legislation;
- nothing in the Contribution contains any information that may violate the State Secrets Laws of the Peoples Republic of China (PRC), including (a) any information that has a vital bearing on Chinese state security and national interests; (b) any information defined as a State Secret under the PRC law, or (c) any classified information belonging to governmental authorities of the Peoples Republic of China, including government agencies, quasi-government agencies, public institutions, or state-owned enterprises (together the State Secret Laws). For the purposes of PRC law and these Terms State Secrets shall include, but not be limited to: (i) unpublished information concerning major policy decisions on Chinese State affairs; (ii) confidential information concerning national defense and the activities of the armed forces; (iii) confidential information concerning national diplomatic policies and activities; (iv) confidential information concerning national economic and social development; (v) matters concerning classified science and technology; and (vi) unpublished Chinese State security matters and non-public information about the on-going investigation of criminal offenses.
- the Contribution contains nothing that is unlawful, defamatory, legally privileged, or which would, if published, constitute a breach of contract, or of confidentiality, or breach of privacy;
- no advice, recipe, formula, or instruction in the Contribution will if followed or implemented by any person cause loss, damage, or injury to them or any other person;
- any research conducted for the Contribution has been conducted in accordance with applicable laws, regulations and codes of practice;
- you have adhered to the Policies in your preparation and submission of the Contribution; and due care, diligence, and all other requisite investigations were carried out in the preparation of the Contribution to ensure its
- accuracy and all statements contained in it purporting to be factual are true and correct.

G.2.3 Article Processing Charges

Publication of an article with us requires payment of a non-refundable article processing charge in accordance with the applicable Article Processing

Charges Policy
(for example, https://www.hindawi.com/journals/aaa/apc/) (APC).

In the event an article is accepted for publication, the author originally submitting the article intended for publication (whether for himself/herself or on behalf of a group of authors or an institution) will be invoiced for the applicable APC. Payments of APCs must be made in the period stipulated on our invoice. Payments must be made in full via our third-party payment service provider and in accordance with such providers terms of service. Payment by any other means is permitted only in our sole discretion.

You acknowledge and agree that payment of an APC does not guarantee publication of an article. Reporting

You acknowledge that, other than for the purposes of verifying MTS Account registrations and article submissions in accordance with the Policies, we do not actively monitor Contributions and/or any of the other content that is made available on the Site.

If you believe that any part of the Site:

- infringes your Intellectual Property Rights or other proprietary rights;
- is defamatory to you; and/or
- is otherwise in breach of these Terms,

please notify us by writing to us at hindawi@hindawi.com.

G.2.4 Third-Party Links and Resources

The Site may contain links to other sites and resources. These links are provided for your information only and we make no warranties or representations whatsoever about any third-party websites which you may access through the Site. We assume no responsibility for the content of sites linked to on the Site. We will not be liable for any loss or damage that may arise from your use of any such third-party sites.

Third-party websites are in no way approved, vetted, checked or endorsed by us and you agree that we shall not be responsible or in any way liable for the content, accuracy, compliance with relevant laws, or accessibility or any information, data, advice, or statements, or for the quality of any products or services available on such sites. Links do not necessarily imply that we are or that the Site is affiliated to or associated with such third-party sites. If you decide to visit any other site, you do so at your own risk. In addition, use of any other site may be subject to your acceptance of additional terms and conditions which we suggest you read carefully before proceeding. Linking to Hindawi

You may link to the Site provided you do so in a way that is fair and legal and does not damage our reputation or take advantage of it. However, you must not suggest any form of association, approval, or endorsement on our part where none exists and you must not establish a link to the Site in any site that is not owned by you.

The Site must not be framed on any other site.

You may link to any part of the Site but we reserve the right to withdraw linking permission without notice. Data Protection

Please refer to our Privacy Policy (https://www.hindawi.com/privacy/), for details of how personally identifiable information is collected by us and may be processed or shared with others. Our Liability

Nothing in these Terms excludes or limits our liability for anything that cannot be excluded or limited by law.

To the extent permitted by law, we exclude all conditions, warranties, representations, or other terms which may apply to the Site or any content made available on it, whether express or implied.

Subject to this section, you agree that we will not be liable for any loss or damage, (whether direct or indirect or arising under contract, tort (including negligence), breach of statutory duty, or otherwise) even if foreseeable, arising under or in connection with any:

- use of, availability of, or inability to use the Site; or
- use of or reliance on any content displayed on the Site.

Without limiting the effect of the paragraph above, due to the inherent risks of using the Internet, we cannot be liable for any damage to, or viruses that may infect, your internet enabled device or any other property when you are using the Site. The uploading, posting, downloading, or accessing of any content (including Contributions), material, and/or other information made available by the Site is done at your own discretion and risk and with your agreement that you will be solely responsible for any damage to your device or loss of data that results from the downloading or acquisition of any such content, material, and/or information.

You agree to indemnify us against any claims, liabilities, losses, damages, expenses, or legal proceedings arising out of:

- your use of the Site;
- your Contribution (including any articles);
- any breach or alleged breach of your warranties describe in the section Publication of Contributions above; or
- your failure to comply with these Terms.

G.2.5 General

Any failure or delay by us to enforce any of our rights under these Terms will not be taken as or deemed to be a waiver of that or any other right unless we acknowledge and agree to such a waiver in writing.

These Terms are not intended to be for the benefit of, nor exercisable by, any person who is not a party to these Terms.

If a court deems any part of the terms set out in these Terms to be invalid, illegal or unenforceable, the remainder of the terms will remain unaffected.

These Terms and our Privacy Policy (https://www.hindawi.com/privacy/) set out the full extent of our obligations and liabilities concerning the Site and replace any previous agreements and understandings between us and you.

We may assign our rights under this agreement and transfer our obligations under this agreement in our sole discretion by providing notice to you via the email address that you provide to us or by notifying you on our site.

These Terms, their subject matter and formation (and any non-contractual disputes or claims arising out of or in connection with these Terms), are governed and construed in accordance with English law. You and we both agree that the courts of England and Wales will have exclusive jurisdiction.

REFERENCES

[1] H. A. Abdou. Genetic programming for credit scoring: the case of Egyptian public sector banks. *Expert systems with applications*, 36:11402–11417, 2009.

[2] B. A. Almquist. PhD thesis, Univrsity of Iowa, 2011.

[3] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16:412–424, 2000.

[4] A. Bifet, G. De Francisci Morales, J. Read, G. Holmes, and B. Pfahringer. Efficient online evaluation of big data stream classifiers. In *Proceedings of the SIGKDD Conference on Knowledge discovery and Data Mining*, pages 59–68, Sydney, NSW, Australia, August 10-13 2015.

[5] D. D. Blakely, E. Z. Oddone, V. Hasselblad, D.L. Simel, and D. B. Matchar. Noninvasive carotid artery testing: a meta-analytic view. *Annals of Internal Medicine*, 122:360–367, 1995.

[6] D. Bohning, W. Bohning, and H. Holling. Revisiting Youden's Index as a useful measure of the misclassification of diagnostic studies. *Statistical Methods in Medical Research*, pages 1–12, 2008.

[7] G. Cameron. *Engineer at Large: The Essential Guide to Structured Problem Solving and Creative Innovation*. CreateSpace, Scotts Valley, Cal., first edition, 2015.

[8] E. O. Cannon, A. Bender, D. S. Palmer, and J. B. O. Mitchell. Chemoinformatics-based classification of prohibited substances employed for doping in sport. *J. Chem. Inf. Model.*, 46:2369–2380, 2006.

[9] O. Carugo. Detailed estimation of bioinformatics prediction reliability through the fragmented prediction performance plots. *BMC Bioinformatics*, 8:380, 2007.

[10] R. Caruna and A. Niculescu-Mizil. Data mining in metric space: An emperical analysis of supervised learning performance criteria. In *Proceedings of Knowledge Discovery and Data Mining 2004*, pages 69–78, 2004.

[11] G. Casela and R. L. Berger. *Statistical Inference*. Duxbury Press, Pacific Grove, Cal., second edition, 2001.

[12] P. Chatterjee, S. Basu, M. Kundu, M. Nasipuri, and D. Plewczynski. PSP MCSVM: brainstorming consensus prediction of protein secondary structures using two-stage multiclass support vector machines. *Journal of Molecular Modeling*, 17:2191–2201, 2011.

[13] M. Y. Chuttur. Overview of the technology acceptance model: Origins, developments and future directions. *Sprouts: Working Papers*, pages 9–37, 2009.

[14] C. W. Cleverdon. The critical appraisal of information retrieval systems. http://hdl.handle.net/1826/1366, 1968.

[15] M. S. Cline, K. Karplus, R. H. Lathrop, T. F. Smith, R. G. Rogers Jr., and D. Haussler. Information-theoretic dissection of pairwise contact potentials. *Protiens*, 49:4–14, 2002.

[16] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[17] T. M. Cover and J. A. Thomas. *Elements of Infor. Theory*. Wiley Series in Telecomm. John Wiley and Sons, Inc., New York, 1991.

[18] I. K. Crombie and H. T. O. Davies. What is meta-analysis? "What is ... ?" series NPR09/1112, Hayward Medical Communications, 2009.

[19] P.B. Crosby. *Quality is Free: The Art of Making Quality Certain*. McGraw-Hill, New York, NY, USA, 1979.

[20] P. Dao, K. Wang, C. Collins, M. Esterand, A Lapuk, and S. C. Sahinalp. Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics*, 27:i205–i213, 2011.

[21] J. Davis and M. Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the* 23^{rd} *International Conference on Machine Learning*, 2006.

[22] DoD. Mandatory procedures for major defense acquistion programs (MDAPS) and major automated information system (MAIS) acquision programs (DoD 5000.2-R). http://www.acq.osd.mil/ie/bei/pm/reflibrary/dodi/p50002r.pdf, April 5, 2002.

[23] R. P. W. Duin. A note on comparing classifiers. *Pattern Recoginition Letters*, 17:529–536, 1996.

[24] E. E. Eiland and L. M. Liebrock. Efficacious end user measures, Part 1: Relative class size and end user problem domains. *Adv. in Artif. Intell.*, 2013:2:2–2:23, 2013.

[25] E. E. Eiland and E. Stride. root9B, LLC unpublished data set, June 2016.

[26] J. M. Fardy. Evaluation of diagnostic tests. *Methods of Molecular Biology, Clinical Epidemiology*, 473:127–136, 2009.

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

[28] J. A. Fernandez. Contextual role of TRLs and MRLs in technology management. Technical Report SAND2010-7595, Sandia National Laboratory, 2010.

[29] J. Francois, H. Abdelnur, R. State, and O. Festor. PTF: Passive temporal fingerprinting. In *Proceedings of 12th IFIP/IEEE International Symposium on Integrated Network Management*, pages 1–8, 2011.

[30] A. Frank and A. Asuncion. UCI machine learning repository. http://archive.ics.uci.edu/ml, 2010.

[31] A. Naumoski G. Mirceva and D. Davcev. A novel fuzzy decision tree based method for detecting protein active sites. *Advances in Intelligent and Soft Computing*, 150:51–60, 2012.

[32] A. S. Glas, J. G. Lijmer, M. H. Prins, G. J. Bonsel, and P. M. M. Bossuyt. The Diagnostic Odds Ratio: a single indicator of test performance. *Journal of Clinical Epidemiology*, 56:1129–1135, 2003.

[33] P. Godoe and T. Johansen. Understanding adoption of new technologies: Technology readiness and technology acceptance as an integrated concept. *Journal of European Psychology Students*, 3(0), 2012.

[34] G. Gopalakrishna, M.M. Leeflang, C. Davenport, A.J. Sanabria, P. Alonso-Coello, K. McCaffery, P. Bossuyt, and M.W. Langendam. Barriers to making recommendations about medical tests: a qualitative study of European guideline developers. *The British Medical Journal Open*, 6:e010549, 2015.

[35] Q. Gu, L. Zhu, and Z. Cai. Evaluation measures of classification performance of imbalanced data sets. In *Proceedings of 4th International Symposium on Intelligence Computation and Applications*, volume 51, pages 461–471, 2009.

[36] D. J. Hand. *Measurement Theory and Practice: the World Through Quantification*. Oxford University Press, New York, 2004.

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

[38] D.J. Hand and C. Anagnostopoulos. When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? *Pattern Recognition Letterss*, 34:492–495, 2013.

[39] H. He and Y. MA. *Imbalanced Lerning: Foundations, Algorithms and Applications*. IEEE Press, 2013.

[40] A. Jamain and D. J. Hand. Mining supervised classification performance studies: A meta-analytic investigation. *Journal of Classification*, 25:87–112, 2008.

[41] N. Japkowicz. Why question machine learning evaluation methods? *Proceedings of AAAI-06; Evaluation Methods for Machine Learning Workshop 6-11*, pages 6–11, 2006.

[42] N. Japkowicz and M. Shah. *Evaluating Learning Algorithms A Classification Perspective*. Cambridge University Press, New York, New York, 2014.

[43] D. Johnson. Performance evaluation. cnx.rice.edu/ content/m11274/latest, August 11, 2003.

[44] K. K. Kandaswamy, K. Chou, T. Martinetz, S. Moller, P. N. Suganthan, S. Sridharan, and G. Pugalenthi. Detailed estimation of bioinformatics prediction reliability through the fragmented prediction performance plots. *Journal of Theoretical Biology*, 270:56–62, 2011.

[45] C. Kauffman and G. Karypis. An analysis of information content present in protein-DNA interactions. *Proceedings of the Pacific Symposium of Biocomputing*, pages 477–488, 2008.

[46] A. W. Kimball. Errors of the third kind in statistical consulting. *Journal of the American Statistician Associations*, 57:133–1427, 1957.

[47] M. Kulharia, R. S. Goody, and R. M. Jackson. Information theory based scoring function for the structure-based prediction of protein-ligand binding affinity. *Journal of Chemical Information and Modeling*, 48:1990–1998, 2008.

[48] C. Lajas, L. Abasolo, B. Bellajdel, C. Hernandez-Garcia, L. Carmona, E. Vargas, P. Lazaro, and J. A. Jover. Costs and predictors of costs in rheumatoid arthritis: A prevalence-based study. *Arthritis & Rheumatism*, 49(1):64–70, 2003.

[49] N. Lavesson, V. Boeva, E. Tsiporkova, and P. Davidssony. A method for evaluation of learning components. *Automated Software Engineering*, 21:41–63, 2014.

[50] T. Lee, C. Lu, S. Chen, N. A. Bretaa, T. Cheng, M. Su, and K. Huang. Investigation and identification of protein γ-glutamyl carboxylation sites. In *PROCEEDINGS OF Tenth International Conference on Bioinformatics. First ISCB Asia Joint Conference 2011: Bioinformatics*, 2011.

[51] P. Legris, J. Ingham, and P. Collerette. Why do people use information technology? a critical review of the technology acceptance model. *Information and Management*, 40(3):191–204, January 2009.

[52] J. M. Lobo, A. Jimnez-Valverde, and R. Real. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17:145–151, 2008.

[53] T. J. Magliery and L. Regan. Sequence variation in ligand binding sites in proteins. *BMC Bioinformatics*, 6:240, 2005.

[54] M. Majnik and Z. Bosnic. ROC analysis of classifers in machine learning: A survey. Technical Report, Faculty of Computer and Information Science MM-1/2011, University of Ljubljana, 2011.

[55] J. C. Mankins. Approaches to strategic research and technology (R&T) analysis and road maps. *Acta Astronautica*, 51:3–21, 2002.

[56] J. C. Mankins. Technology readiness assessments: A retrospective. *Acta Astronautica*, 65:1216–1223, 2009.

[57] D. Martens, J. Vanthienen, W. Verbake, and B. Baesens. Performance of classification models from a user perspective. *Decision Support Systems*, 51:782–793, 2011.

[58] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M Przybocki. The DET curve in assessment of detection task performance. In *Proceedings of the Fifth European Conference on Speech Communication and Technology, 1997*, pages 1895–1898, Rhodes, Greece, September 22-25 1997.

[59] B. W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta*, 405:442–451, 1975.

[60] W. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[61] C. S. Miller and D. Eisenberg. Using inferred residue contacts to distinguish between correct and incorrect protein models. *Bioinformatics*, 24:1575–1582, 2008.

[62] National Science Board. Globalization of science and engineering research science and engineering indicators 2010 (nsb 10-3). http://www.nsf.gov/statistics/nsb1003/definitions.htm, February 28, 2015.

[63] D. Newman. UCI KDD archive. kdd.ics.uci.edu, March 5, 2006.

[64] K. Nishimura, D. Sugiyama, Y. Kogata, G. Tsuji, T. Nakazawa, S. Kawano, K. Saigo, A. Morinobu, M. Koshiba, K. M. Kuntz, I. Kamae, and S. Kumagai. Meta-analysis: Diagnostic accuracy of anti-cyclic citrullinated peptide antibody and rheumatiod factor for rheumatoid arthritis. *Annals of Internal Medicine*, 146(11):797–808, 2007.

[65] G. Oreški and S. Oreški. A method for evaluation of learning components. *Journal of Information and Organizational Sciences*, 39:2:209–222, 2015.

[66] O. G. Othersen, A. G. Stefani, J. B. Huber, and H. Sticht. Application of information theory to feature selection in protein docking. *Journal of Molecular Modeling*, 18:1285–1297, 2012.

[67] R. L. Ott and M. Longnecker. *An Introduction to Statistical Methods and Data Analysis*. Duxbury Press, Pacific Grove, Cal., fifth edition, 2001.

[68] C. Parker. An analysis of performance measures for binary classifiers. pages 517–526, December 11 2011.

[69] R. W. Potter. *The Art of Measurement Theory and Practice*. Hewlett-Packard Professional Books. Prentice Hall PTR, Upper Saddle River, New Jersey, 2000.

[70] B. Rost and C. Sander. Prediction of protien secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, 232:584–599, 1993.

[71] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, third edition, 2009.

[72] S. R. Sadin, F. P. Povinelli, and R. Rosen. The NASA technology push towards future space mission systems. *Acta Astronautica*, 20:73–77, 1989.

[73] R. L. Schaeffer. *Introduction to Probability and its Applications*. Wadsworth Group/Thompson Learning, Belmont, California, USA, second edition, 1995.

[74] M. Schonlau, W. DuMouchel, W. Ju, A. Karr, M. Theus, and Y. Vardi. Computer intrusion: Detecting masquerades. *Statistical Science*, 16(1):58–74, 2001.

[75] N. Seliya, T. M. Khoshgoftaar, and J. Van Hulse. A study on the relationships of classifier performance metrics. In *Proceedings of the 21st IEEE International Conference on Tools with Artificial Intelligence*, pages 859–864, Newark, New Jersey, USA, November 2-5 2009.

[76] J. D. Smith II. An alternative to technology readiness levels for non-developmental item (NDI) software. *Proc. of the 38th Hawaii International Conference on System Sciences*, 2005, 2005.

[77] M. Sokolova, K. El Emam, S. Chowdhury, E. Neri, S. Rose, and E. Jonker. Evaluation of rare event detection. *Lecture Notes on Artificial Intelligence*, 6085:379–383, 2010.

[78] M. Sokolova, N. Japkowicz, and S. Szpakowicz. Beyond Accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Proceedings of the ACS Australian joint conference on artificial intelligence*, pages 1015–1021, 2006.

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

[80] A. D. Solis and S. Rackovsky. Contact prediction using mutual information and neural nets. *Protiens*, 71:1071–1087, 2008.

[81] B. Sterner, R. Singh, and B. Berger. Predicting and annotating catalytic residues: an information theoretic approach. *Journal of Computational Biology*, 14:1058–1073, 2007.

[82] J. Steurer, J.E. Fischer, L.M. Bachmann, M. Koller, and G. ter Riet. Communicating accuracy of tests to general practitioners: a controlled study. *The British Medical Journal*, 324:824–826, 2002.

[83] S. S. Stevens. On the theory of scales of measurement. *Science*, 103:677–680, 1946.

[84] J. A. Swets. Form of emperical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement performance. *Psychological Bulletin*, 99(2):181–198, 1986.

[85] J. A. Swets. Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin*, 99(1):100–117, 1986.

[86] J. A. Swets. Measuring accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, 1988.

[87] J. Taylor. Department of defense (DoD) technology readiness assessment (TRA) process. https://acc.dau.mil/adl/en-US/495629/file/62502/9

[88] C. J. Van Rijsbergen. *Information Retrieval*. 1979. http://www.dcs.gla.ac.uk/Keith/Preface.html.

[89] S. Vanderlooy and E. Hüllermeier. A critical analysis of variants of the AUC. *Machine Learning*, 72(3):247–262, 2008.

[90] V. N. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774780, 1963.

[91] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis. User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3):425–478, January 2009.

[92] T. Verbraken, Verbeke W., and B. Baesens. A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering*, 25(5):961–973, 2013.

[93] Washington State University. Validity of measurement. cbdd.wsu.edu/edev/NetTOM_ToT/Resources/ Other/TOM614/page69.htm, July 12,2007.

[94] A. M. Wassermann, B. Nisius, M. Vogt, and J. Bajorath. Identification of descriptors capturing compound class-specific features by mutual information analysis. *Journal of Chemical Information and Modeling*, 50:1935–1940, 2010.

[95] P.F. Whiting, C. Davenport, C. Jameson, M. Burke, J.A. Sterne, C. Hyde, and Y. Ben-Shlomo. How well do health professionals interpret diagnostic information? a systematic review. *The British Medical Journal Open*, 5:1–9, 2015.

[96] R. W. Yeung. *A First Course in Infor. Theory*. Information Technology: Transmission, Processing and Storage. Kluwer Academic, New York, 2002.

[97] W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3:32–35, 1950.

[98] Z. Zhelev, R. Garside, and C. Hyde. A qualitative study into the difficulties experienced by healthcare decision makers when reading a Cochrane diagnostic test accuracy review. *Systematic Reviews*, 2:DOI: 10.1186/2046–4053–2–32, 2013.

A COHERENT CLASSIFIER/PREDICTION/DIAGNOSTIC PROBLEM FRAMEWORK AND RELEVANT SUMMARY STATISTICS[1]

by

E. Earl Eiland