



Sensitivity and Specificity versus Precision and Recall, and Related Dilemmas

William Cullerne Bown¹ 

Accepted: 30 May 2024 / Published online: 26 June 2024
© The Author(s) under exclusive licence to The Classification Society 2024

Abstract

Many evaluations of binary classifiers begin by adopting a pair of indicators, most often sensitivity and specificity or precision and recall. Despite this, we lack a general, pan-disciplinary basis for choosing one pair over the other, or over one of four other sibling pairs. Related obscurity afflicts the choice between the receiver operating characteristic and the precision-recall curve. Here, I return to first principles to separate concerns and distinguish more than 50 foundational concepts. This allows me to establish six rules that allow one to identify which pair is correct. The choice depends on the context in which the classifier is to operate, the intended use of the classifications, their intended user(s), and the measurability of the underlying classes, but not skew. The rules can be applied by those who develop, operate, or regulate them to classifiers composed of technology, people, or combinations of the two.

Keywords Foundations of statistics · Diagnostics · Information retrieval · Signal processing · Machine learning · Evaluation · Skew

1 Introduction

Classification is a ubiquitous activity that we have for centuries attempted to master through quantification (Franklin, 2002). This effort has yielded more and more ways of evaluating the accuracy of the output of a classifier. It has not however led to anything that could be called a theory of evaluation, a lack noted by authors in recent decades (Flach, 2003; Hand, 2012; Van Rijsbergen, 1979). The central problem is that we have no accepted procedure for deciding between the many forms of evaluation. Why this way and not that?

This difficulty is clear in the wide-ranging review of evaluation methods by Hand (2012). On the one hand, he says, ‘failure to adopt an appropriate measure of performance means one is answering the wrong question ... with the consequence that one may well draw incorrect conclusions.’ On the other hand, he provides little guidance as to how one can go about working out which form of evaluation is appropriate.

✉ William Cullerne Bown
wockbah@gmail.com

¹ London, UK

Reviewing the literature, I have been able to find nine journal articles (eight in English, one in French) offering general surveys of classification evaluation methods published since Hand's, 2012 review, to which I have added Powers' influential 2011 article, already cited over 3000 times according to Google Scholar. None of these provides a solution to the appropriateness problem highlighted in Hand (2012). In some cases, the question of appropriateness is not addressed (Jiao & Du, 2016; Lever et al., 2016; Powers, 2011; Škrabánek & Doležel, 2017; Tharwat, 2018). Alternatively, while the question is addressed, an answer may be proffered without explicit recognition that it is inadequate (Dinga et al., 2019; Lever et al., 2016; Liu et al., 2014). In one case, the attack suggested is explicitly recognized as inadequate (Ducharme, 2018, p. 20). In another case, the authors' approach is opaque (Hossin & Sulaiman, 2015).

It is easy to lose sight of the pervasive nature of the appropriateness problem since it remains possible to make some progress even without an answer to it. If a binary classifier returns fewer false positives and fewer false negatives, it will always be better. If, like an online store looking to sales, the classifier can be assessed according to compelling consequences, that may be enough. If the work lacks a specific application, a convention can satisfy the demands of the discipline. However, these exemptions, significant as they are, leave a wide range of important real-world applications of classification bedeviled by a basic methodological problem.

Further, fields of classification often note that they are highly empirical. In the absence of adequate theory, this is understandable: both reliance on flawed theory and simply waiting for theory to arrive suffer from a lack of pragmatism. However, this does not mean the lack of a foundational theory is harmless. One consequence is that it is unclear under what circumstances the lessons learnt in one area are applicable in another, and this is the source of two problems. First, there is a temptation to adopt without an explicit rationale a form of evaluation from elsewhere that may be inappropriate. An example of this from history is that of information retrieval which in its early years struggled to rid itself of the assumption — eventually discarded — that the pair of sensitivity and specificity was the appropriate starting point. Second, there is an understandable hesitation in taking advantage of lessons learnt in other fields, the upshot of which is fragmentation in which each field develops in its own silo despite the evident commonality.

The difficulties can be traced back to statistical foundations. In *The design of experiments*, Fisher explains his inductive, or inferential, purpose through the example of the tea lady (Fisher, 1935). The question he is interested in is whether we should believe her when she says she can distinguish between cups poured tea-first and cups poured milk-first. He does not concern himself with how to make a better tea lady. So, if we had two tea boys who both claimed to be able to distinguish between tea-first and milk-first cups, but one mistook more tea-firsts for milk-firsts and the other vice-versa, Fisher's methods, and inferential statistics in general, would not allow us to decide which is better. The bit that is missing is a rigorous conception of what we mean by 'better', one that will allow us to compute a 'goodness' score for each tea boy and see which scores higher.

When we tackle the appropriateness problem, we do not have an inferential goal. Rather, the goal is meta-inferential. The classifier will make inferences (it is through a process of inference that the classifier arrives at a classification in any particular case). The question is, as we go about developing or operating the system that makes the inferences, what principles do we want to shape the way the inferences are made?

It is clear that some kind of preference is involved. For example, with the tea boys, you might be more concerned with missed tea-firsts and I with missed milk-firsts. When we look at the methods of evaluation of binary classifiers in use today, there are three

schools of thought regarding how to deal with this question of preference. In one school are those methods that axiomatically treat missed tea-firsts and missed milk-firsts as equally problematic, an approach that can be understood as aspiring to a mathematical degree of neutrality. This school can be traced back to the development by Fisher (1936) of quadratic discriminant analysis and hence, with simplifying assumptions, of linear discriminant analysis. A simpler example from machine learning is the Matthews Correlation Coefficient (MCC). The neutrality of this approach is popular when the work involved is purely methodological, where nothing can be assumed about the real-world circumstances in which the ultimate method may be applied. It can also be attractive when it seems the stakes are purely taxonomic; the original example in Fisher (1936) involved distinguishing plants from two species of iris.

However, the stakes in classification are often less abstract. If *Iris setosa* was a food and *Iris versicolor* a poison, and the classification being performed before we ate some irises, we would feel very differently about the potential errors. Equally, I will feel differently about the two kinds of error if I am about to be classified as having cancer (or not), eligible for a loan (or not), permitted to participate in a social media platform (or not), or guilty of a crime (or not). Such circumstances lead to a second school of thought that revolves around methods that provide for a judgement about the relative importance of the outcomes of classification. The prime examples here are those methods that start with a pair of indicators such as sensitivity and specificity or precision and recall. The exact origin of this approach is lost to us but has been traced to the 1920s (Hammond, 1996).

Both schools are popular. A search on Google Scholar in April 2024 for the strings ‘discriminant analysis’ and ‘Matthews Correlation Coefficient’ in the year 2021 finds about 48,100 and 4,000 articles; for the strings ‘sensitivity and specificity’ and ‘precision and recall’, it yields about 60,400 and 37,000 articles.

The third school is the method of adopting a currency and then allocating values in it to the successes and failures on the basis of empirical measurements.

The three schools are distinguished by their attitude to the question of preference; in their methods, they may overlap. For example, in the critical decision to choose one classifier over another, the cost-benefit approach is equivalent to adopting a linear combination of sensitivity and specificity (Cullerne Bown, 2023).

This article is concerned with the second school and specifically with pairs of indicators such as sensitivity and specificity. While such an approach both enables and potentially defers the element of judgement, it also shapes the way in which the output of successes and errors is understood, and hence creates the context for the judgement. Here, at the outset, the discipline of medicine typically relies on sensitivity and specificity for diagnostic tests. Information retrieval, by contrast, typically relies on precision and recall for search engines. So, this first step of choosing between the available pairs of indicators seems to be a matter of circumstance rather than preference.

Different fields have their own reasons for choosing their pair of indicators and this is a source of the fragmentation of the disciplinary landscape. We lack a general statistical basis, accepted across all disciplines, for choosing one pair over the other. Thus, a first step in reducing the appropriateness problem and establishing a general foundational theory of classification is to provide a general and non-preferential answer to the question of which pair of indicators should be used when. That is the primary purpose of this article.

To establish which pair is correct, the approach herein is to consider the context, or situation, in which the classifier is to be used. If that context is missing, for example because what is being evaluated is a statistical technique that could be used in any computer-based classifier, then it will not be possible to use the method to identify any pair as right or wrong.

To pursue this kind of analysis, I have found it necessary to return to first principles and work my way up. A basic commitment is that any assessment of accuracy should be an explicit procedure rooted in empirical observations and so my starting point is the theory of measurement. Then, I have encountered four kinds of concept. First, ideas that are familiar and named, such as a false positive. Second, ideas that are familiar but formally unnamed, such as the distinction between classification as a form of decision (or not). Third, ideas that are familiar but seem to lack clarity, such as that of a case class. Fourth, new ideas. By clarifying and systematizing all these ideas, I establish a conceptual architecture that also aims to advance the development of a foundational theory. A glossary of about 50 new and revised terms is included in the Supplementary Materials.

The article is in two parts. In Sects. 2 and 3, I define my terms. In Sects. 4, 5, 6, and 7, I use these building blocks to derive the method for choosing between pairs of indicators.

In Sect. 2, I provide definitions of classification, classifiers, and evaluation. I claim that my conception of the interior of a classifier is more basic and more general than the common score-threshold conception. By avoiding relying on the score-threshold pattern, I avoid making a canonical distinction between classifiers that are composed of technology, such as computers or diagnostic tests, and people, such as courts or a bureaucracy. For examples, I draw freely from both camps. This is not to say that there is no important difference between the two, only that the differences should not interfere with the methodology that is appropriate to assessing their accuracy. An advantage of this is that the rules developed herein provide a common basis for evaluation of both human and technological systems of classification and, as is increasingly common, systems that combine the two. In Sect. 3, I distinguish six pairs of indicators, such as precision and recall, based on tallies of outcomes in binary classification. To help erode fragmentation, I provide generic names for each indicator and pair that can be used across all disciplines.

In Sect. 4, I consider the role that classification plays, being sometimes a form of decision, sometimes not. In Sect. 5, I consider different potential users of the classifications and the implications of their different needs. This leads me to a common but tricky scenario, where a classifier makes decisions on behalf of someone who has a particular kind of action in mind. In Sect. 6, I argue that this is where the idea of a case class becomes useful, and provide a new definition of it. In these circumstances, I argue that the choice of pair depends on which classes are capable of being reliably counted. The upshot of this work is a set of six rules, forming a flow chart, that determines which pair should be used when. The guidance provided in this way is consistent with the disciplinary conventions in diagnostics, information retrieval and signal processing. An overview of the complete method is set out in Sect. 7.

In the Supplementary Materials, I illustrate how the method developed herein and the underlying principles may be applied in four ways:

1. Demonstration that my tallies-based method is consistent with the probabilities-based approach developed in signal processing
2. Application of the method to the choice between the receiver operating characteristic and the precision-recall curve
3. Application of the underlying principles to establish restrictions on the applicability of forms of evaluation not based on pairs of indicators (with potential implications for complex, resource-intensive applications such as large language models)
4. The case for retiring the term ‘unsupervised classification’.

2 Classification, Classifiers, and Evaluation

2.1 Classification

Consider a lottery. It puts labels on things, for example winner and loser, drafted for military service and not drafted. Thus, in an important respect, it looks like a classifier. But a lottery is not called a classifier because we do not have the idea that the labels can be right or wrong in reflecting some underlying reality; we understand that they are arbitrary. This absence can be articulated precisely: a lottery is not a form of measurement. And this in turn reveals the essential character of classification.

Definition *Classification* is measurement against a nominal scale. The elements of the scale are called *classes*.

When we are classifying, we always restrict the task to some particular kind, for example faces, words, court cases, patients. There is no limitation on what form the kind may take. It is not, for example, restricted to objects or events. If we can conceive of a kind, we can begin to classify its members. Drawing on the abstraction of set theory, we might call the members of the kind ‘elements’, but English provides a word that is both concrete and unlimited.

Definition A *thing* is member of a *population* that we are interested in classifying.

The nominal scale implies our deeming that each thing will fall into exactly one of the classes. The classes may be considered natural (for example is red, or not) or artificial (for example is entitled to enter the country, or not) and may impose an arbitrary distinction on a continuously varying quantity (for example is infected with COVID-19, or not).

When classifying, as with all measuring, we always are conscious of the possibility of error and so distinguish between the class that a thing truly belongs to and the class the classifier allocates to the thing.

Definition The class allocated to a thing by a classifier is called a *label*. Each label is either a *success* that truly represents the correct class or an *error* in which the label represents a wrong class. The varieties of success and failure are kinds of *outcome*.

Since a classifier may allocate a thing to any of the classes in the scale, when there are n classes there are n^2 kinds of outcome, of which n are kinds of success and $n^2 - n$ are kinds of error.

In the simplest form of classification, the population of things is conceptually divided between just two classes. Call the classes 1 and 0 and the corresponding labels *positive* and *negative*. The successes and errors found among the positives and negatives give us four kinds of outcome. Common names for these four are adopted here (Table 1): the outcomes from class 1 are divided between true positives and false negatives and from class 0 between true negatives and false positives (often abbreviated herein to TP, FN, TN, and FP).

Table 1 Four outcomes are possible in binary classification

	Positive	Negative
Class 1	True positive	False negative
Class 0	False positive	True negative

2.2 Classifiers

The physical form taken by a classifier varies widely and includes algorithms, signal processors, chemical tests, physical mechanisms such as sieves and bureaucracies (which is to say, any system in which the classifications are generated by people). It is also widely believed that classification is central to cognition in organisms including people — see for example *To Cognize is to Categorize: Cognition Is Categorization* (Harnad, 2005).

The sheer variety of forms makes it challenging to say anything universal about the interior of a classifier. The definition of classification in the preceding subsection treats the classifier itself as a black box and this has been enough for many authors (for example Lever et al., 2016). By contrast, a common idea of the interior has been formalized by Hand (2012), but I find it too narrow, twice over. First, Hand (2012) assumes that the classifier operates by allocating numerical values to attributes of the things that are being classified, an approach that seems to restrict us to technological systems of classification; we cannot say with confidence that this is, for example, how an expert witness comes to classify a bullet as having been fired by a particular gun. Second, he breaks a classifier into two components:

- (i) a score, s , which aims to capture the essence of the difference between classes, in the sense that the larger a score, the more likely the thing is to be from class 1;
- (ii) a threshold, t , such that labels are made by assigning things with $s > t$ to class 1, and otherwise to class 0.

This pattern is very common but there are two reasons for hesitating over its adoption. First, it is easy to think of classifiers that do not obviously have this pattern, for example an algorithm that seeks to identify bottles of red wine and simply looks for the word ‘red’ on the label. Second, there are plenty of classifiers — those that rely on people for example — where the internal process is opaque so that we do not know whether the score-threshold pattern is in use or not.

If we want a truly general theory of classification that, like the rest of statistics, is not intrinsically limited to certain disciplines or certain forms of physicality, then we should start with foundations that are encompassing rather than excluding, which are truly basic and irreducible. An important practical advantage of such an approach is that it will allow us to treat on a single, unified basis classifiers composed of technology, people, or as is increasingly common, combinations of the two.

Definition A *policy* is a rule that governs the operation of a system for a period of time.

The concept of a policy here should be interpreted generously, including the laws of nature that determine the movement of the planets, the physical arrangement of components that governs the functioning of a simple thermostat, a high-level definition of a subroutine in an algorithm and a written directive in a bureaucracy. The operation of a system therefore is determined by the policies that govern it.

Definition A *settled* system is one that is consistent in its policies over time.

Definition A *classifier* is a settled system tasked with classification.

This definition constrains a classifier to a single form of operation. For example, if the classifier relies on a test threshold, then a specific threshold must be adopted as any alteration in the threshold is an alteration in policy.

Definition The *classificatory mechanism* is the underlying and potentially alterable arrangement through which a classifier is instantiated.

Thus, if we construct a chart showing the receiver operating characteristic of a piece of signal processing equipment as the test threshold varies, then in the terms defined herein what is being assessed with the chart is not a classifier per se but the classificatory mechanism. Similarly, a ‘dynamic classifier’ that updates its own method of classification over time is understood here as a sequence of distinct classifiers.

This conception includes any classifier that follows the score-threshold pattern of Hand (2012) since: (i) the process through which a score is generated is defined by one or more policies; and (ii) the choice of threshold is another policy. Equally, the allocation of scores to attributes is again a matter of policy.

This article is exclusively concerned with the accuracy of classifiers as opposed to classificatory mechanisms (though see the Supplementary Materials for a discussion of the receiver operating characteristic). One must start somewhere, and the classifier seems to me the more basic unit, the method of evaluation applied to mechanisms often being built out of components developed to evaluate classifiers.

2.3 Evaluation

The meta-inferential aspect of classifier development implies a realm of managerial decision making that is meta to the classifier itself. Evaluation is central to this.

Definition A *discarding* is a decision that excludes from consideration one or more potential systems.

Discardings are the critical decisions that routinely provide the basis for future development, or deployment, and hence shape the form of the ultimate system in the real world. Examples include the adoption of a technological as opposed to bureaucratic approach to the mechanism, of a particular technological approach such as the reliance on a particular hormone in a pregnancy test, or of a classifier supplied by a specific company. It is always possible to discard all mechanisms and simply classify all things as positive or negative, and in some cases this may yield more accurate results. It is possible both that many classifiers are compared (as with evolutionary algorithms) and that more than one is retained (as when classificatory mechanisms are compared).

Definition A *sharp discarding* is one in which all but one of the systems are discarded.

A common example is an A/B comparison of two classifiers in which one is discarded and the other retained.

Definition *Evaluation* is the procedural assessment of the accuracy of a classifier on an empirical basis to enable sharp discarding.

This is not supposed to be a definition that captures the many ways in which the term ‘evaluation’ is used today. Rather, it is intended, for the purposes of this article, to restrict the idea of evaluation to one central concern. Evaluation of this kind has a paradigmatic starting point.

Definition A *ground truth* data set is a collection of things that have been given labels that we take to be always true. *Calibration* is the act of comparing the ground truth data

set with the labels allocated by the classifier, thus revealing the successes and errors of the classifier.

Sharp discardings can be stripped of human intuition by adopting as the form of evaluation some kind of measurement. The paradigmatic way to do this is with a control function, as found in engineering, that takes the empirical output of calibration (or indicators derived therefrom) as inputs and outputs a score in which higher is better. A utility function or fitness function plays the same role in other fields. In this way, we establish control over the meta-inferential process of discardings and can establish that classifier A has higher accuracy than classifier B. The kind of measurement involved here is pragmatic in the sense of Hand (2017) in that we choose what to measure and how to measure it at the same time.

Let us return to the score-threshold pattern of Hand (2012) and consider a classifier that has no component with a score-like role. It must nonetheless have an internal mechanism operating according to policies. Let us suppose that we can acquire measurements of the proportion of things of, say, class 1, correctly classified, and let this be p_1 . Then suppose a policy changes and this affects the labels so that the new proportion of things of class 1 classified as positive is p_2 . And then suppose a third variant yielding p_3 . The numbers p_1 , p_2 and p_3 can be placed in order and one way to interpret what is going on is that by making the changes in policy we are adjusting the probability of correctly classifying things of class 1. Thus, we can now construct an expanded classifier that has an additional attribute that we call ‘score’, and which has three settings, yielding the policies that result in p_1 , p_2 and p_3 , and which are ordered in the same sequence. Thus, even if the classifier has not been constructed with any intent to follow the score-threshold pattern, we can create a new iteration of the classifier: (i) that has such a score; and (ii) where higher scores do in fact reliably indicate that a thing is more likely to be from class 1.

To do this, we do not need any knowledge of, or interference in, the internal workings of the classifier. Thus — provided we can measure the proportion of things of a class classified correctly — the generality of the score-threshold pattern is re-established, not as something that is intrinsic to any classifier but as something we can always choose to impose on it. One benefit of this result concerns bureaucratic classifiers; we do not need to hypothesize that people work like machines to encompass such classifiers within a general conception familiar to those working with technological systems of classification.

Although I think it is generally taken for granted that the score and threshold will be on a ratio or interval scale, this is not in fact necessary. In our example, so long as p_1 , p_2 and p_3 can be ordered (so long as we have at least ordinal measurements of them), the same form of control can be established.

3 Four Ratios, Six Pairs

3.1 Four Ratios

Definition A *classificatory ratio* (*ratio* hereafter) has as its numerator a tally of one kind of success and as its denominator the tally of successes added to a tally of one kind of error.

This arrangement ensures the ratios vary between 0 (all errors) and 1 (no errors). Thus, an important aspect of these ratios that we will rely on in Sect. 6 is that we consider higher scores to be better. As there are two errors and two successes, we can form ratios in four

ways. To call such a ratio a success rate or an error rate is mildly misleading because it is always both of these things.

Definition A *supplementary ratio* has as its numerator a tally of one kind of error and as its denominator the tally of errors added to a tally of one kind of success.

In this case, we consider higher scores worse. If we take the ratio and the supplementary ratio constructed from the same pair of tallies, they will always add to 1. For this reason, although the supplementary ratios are widely used in some fields, we will have no use for them.

A wide variety of names is used for the four ratios in different fields. It seems to me that all the sets of names are problematic, albeit for different reasons. Some names, such as ‘true positive rate’ or ‘negative predictive value’ skate over the fact that the ratios always involve a combination of two outcomes and suggest an absoluteness that is misleading; only if one has been taught the convention can one decode the meaning of the words in such names. Others, such as ‘sensitivity’ spuriously invoke everyday words — all the ratios are sensitive to something. Others, such as ‘precision’ and ‘recall’, are nicely intuitive for their context but do not always translate well to other fields.

The different naming conventions may seem a minor hindrance within a field. But the convention in each case reflects a focus on a certain kind of problem and if the field begins to investigate other kinds of problem the language will intrude further. This is the case in medicine where language developed for diagnostic tests now coexists with language developed for information retrieval, so that two terms are in use for the same ratio in the same field:

$$\text{Sensitivity} = \text{Recall} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}|}$$

Thus, the terminology we have is a cause of fragmentation, duplication, and confusion. In an article such as this one, which aims to bring a general coherence to the issues involved, one finds oneself having to either constantly duplicate (or triplicate or quadruplicate) names to make oneself readily understood to all readers, or to accept that many readers will be obliged to keep referring to definitions.

To do better, we need new names. In this, we should recognize that: (i) there are two tallies in every ratio; and (ii) everyday language cannot nicely convey the differences between the ratios. A crude attempt would be ‘true-positive-false-negative-ness’, but this is a description rather than a name. We can get something more useful by creating new words according to a naming convention that relies on the first letter of the labels in the ratio, P or N.

Definition A *generic name* for a ratio is constructed as follows. The first letter is P or N, the letter that is characteristic of the success. The last letter is P or N, the letter that is characteristic of the error. The middle letter is Y. For a supplementary ratio, add S to the end of the name for its supplement.

This gives us words for the four ratios: pyn, nyp, pyp and nyn. It also gives us words for the supplementary ratios: pyns, nyps, pyps and nyns (Table 2).

These are technical terms like all the existing names and, like all those names, baffling to anyone who has not been given their formal definition. But once that hurdle is overcome, they have strengths. They are easy to say and to distinguish when heard. They are new words that when written down are hard to confuse with the many related ideas. If you have not yet memorized which is which, they are easy to decode; for example, pyn combines the true positives (P at the

front) with the false negatives (N at the back) and is hence another name for sensitivity (or recall). Equally, they are easy to construct when you start to use them for the first time.

The names remove an entire cognitive step so that one cannot make an error by misremembering a name, only by being substantively confused about which ratio is intended. When we combine them into pairs, we can immediately see what is going on just by looking at the names. For example, instead of ‘precision and recall’ we have ‘pyp and pyn’ and can see that the two both rely on the same yardstick of success (the true positives) but have an interest in different errors. For this article, the advantage of relying on generic names is that, even though they are unfamiliar to everyone, a single term can be used everywhere, and no reader will need to either refer back to definitions or be unsure of which ratio is being referred to.

In each ratio, the tallies in the denominator account for the entire set of either a class or a label. In pyn, the true positives and the false negatives decompose class 1, but in pyp it is instead the positives that are decomposed. The ratios are therefore telling us about two distinct concerns.

Definition *Achievement* concerns the proportion of things from a class that are identified. *Correctness* concerns the proportion of labels that are correct. The *focus* of a ratio is either achievement (pyn and nyp) or correctness (pyp and nyn), whichever concern is present in the denominator.

3.2 Six Pairs

The path we are following is to pick two ratios as the basis for evaluation. We can choose two ratios from the available four in six ways. Below, these six pairs of ratios are characterized and given intuitively suggestive names (Table 3).

In Pair 1 in Table 3 (nyp and pyn), both ratios are concerned with achievement. Call this the *achieving* pair; each ratio answers a question in the form, ‘How much have we achieved in terms of identifying the things in a class?’.

In Pair 2 (pyp and nyn), both ratios are concerned with correctness. Call this the *correctness* pair; each ratio answers a question in the form, ‘How likely is it that the label we have obtained is correct?’.

In these two pairs, all four tallies are used exactly once. In the remaining pairs however, one tally is used twice, and one tally is not used at all. The tally that is used twice is therefore a common concern that unites the ratios; and equally, the tally that is omitted represents a concern that is entirely neglected by the pair.

In Pair 3, one ratio (pyn) is concerned with achievement in class 0 and the other (pyp) with correctness in positive labels. This is unlike the previous two pairs in that focus does not unite the two ratios but divides them. The unifying feature here is that both ratios have as their measure of success the true positives. That is, both concern errors that can be made in the task of acquiring true positives. Thus, a quest for true positives, a kind of prospecting, is the unifying feature. Each ratio answers a question in the form, ‘To what extent is our prospecting failing?’ Call this the *prospecting* pair.

Pair 4 (nyp & nyn) is like Pair 3, but with the difference that the measure of success is now the true negatives. That is, both ratios concern errors made in the task of acquiring true negatives. Each ratio answers a question in the form, ‘To what extent is our prospecting (for true negatives) failing?’ Call this the *anti-prospecting* pair.

Table 2 Four ratios can be obtained by combining one tally of a success with one tally of an error in the standard pattern

Generic name	Medical name	Information Retrieval name	Success	Error	Denominator	Focus	Generic name for supplement
Pyn	Sensitivity	Recall	TP	FN	Class 1	Achievement	Pyns
Nyp	Specificity		TN	FP	Class 0	Achievement	Nyps
Pyp	Positive predictive value	Precision	TP	FP	Positives	Correctness	Pyys
Nyn	Negative predictive value		TN	FN	Negatives	Correctness	Nyns

Table 3 Six pairs of ratios can be obtained from the four ratios available

	Name	First ratio		Second ratio		Both include	Both exclude	Focus
		Success	Error	Success	Error			
1	Achieving <i>Nyp & Pyn</i>	TN	FP	TP	FN	-	-	Achievement
2	Correctness <i>Pyp & Nyn</i>	TP	FP	TN	FN	-	-	Correctness
3	Prospecting <i>Pyp & Pyn</i>	TP	FP	TP	FN	TP	TN	True positives
4	Anti-prospecting <i>Nyp & Nyn</i>	TN	FP	TN	FN	TN	TP	True negatives
5	Avoidant <i>Pyp & Nyp</i>	TP	FP	TN	FP	FP	FN	False positives
6	Anti-avoidant <i>Pyn & Nyn</i>	TP	FN	TN	FN	FN	FP	False negatives

Pair 5 (pyp and nyp) is distinguished by the appearance in both ratios of the false positives. Each ratio answers a question in the form, ‘To what extent are we managing to avoid false positives?’ Call this the *avoidant* pair.

Pair 6 (pyn and nyn) is distinguished by the appearance in both ratios of the false negatives. Each ratio answers a question in the form, ‘To what extent are we managing to avoid false negatives?’ Call this the *anti-avoidant* pair.

If we invert our labelling, so that what was class 1 becomes class 0 and vice-versa, Pairs 1 and 2 are unaffected. However, Pair 3 becomes Pair 4 (and vice-versa) and Pair 5 becomes Pair 6 (and vice-versa). This is the rationale for the names for Pairs 4 and 6. In addition, when the choice of labels is itself arbitrary, Pairs 4 and 6 (the antis) are superfluous. However, note that I will argue in Sect. 6 that the choice of labels is not arbitrary in common, well-defined circumstances.

Definition If both the ratios in a pair have the same focus, the *focus* of the pair is the same as the focus of the ratios; if the focus of the ratios is different, the focus of the pair is the tally that appears in both ratios.

None of the concepts introduced in this section can really claim to be new. On the other hand, the fact that so many of them lack commonly agreed names across disciplinary boundaries: (i) suggests they have not always come into focus; and (ii) inhibits their use as building blocks in the development of methods. And the work does leave us with one puzzle. The achieving, correctness and prospecting pairs are all in use. Why is it that the others, so far as I am aware, are completely unused?

4 Decisiveness

Now we begin to develop the method for identifying the correct pair of ratios. This enterprise is framed by a dominating concern for classification in the real world. As such, it stands in contrast to methodological work that may have little idea of the circumstances in which the methods will ultimately be used.

Sometimes the act of classification is more than mere labelling and takes the form of a decision, for example when a social media company bans a post (or not), a motor manufacturer's software stops a car (or not), a court finds a defendant guilty (or not), a doctor decides to operate (or not), or a credit scoring algorithm approves a loan (or not). In such cases the labelling does not disappear, it is just that there is no gap between the labelling and the deciding; the two occur in the same act.

Alternatively, the labels may be made and stored for any number of different uses, many of which may be opaque at the time of classification. This makes the label a kind of evidence-in-waiting. Or the classification may be explicitly conceived as a way of constructing evidence; examples include the tests relied on by a court or doctor, for example the question of whether fingerprints match or the patient's blood has 'abnormally' high or low levels of some component.

Definition A label that is also a form of decision is *decisive*; a label that is not decisive is *non-decisive*.

When a classifier is generating non-decisive labels, the ultimate decisionmaker, usually considering a wider range of evidence, needs some means of assessing how much weight should be placed on the label. This starts with the probability that it is correct and so there are two concerns, for the correctness of the positives and the correctness of the negatives. Hence, we have our first rule.

Rule 1 Non-decisiveness implies the correctness pair.

5 Perspective

Definition A *perspective* is a body whose interest in the labels is paramount to the evaluation.

In general, the bodies that may be considered for this role relate to a classifier in a way that is either *for*, *by* or *on*; *for* if the classification is carried out on behalf of the body; *by* if the body conducts the classification; *on* if the classification is of the body.

Definition If classification is carried out for a body, call the body the *principal*; if by a body, call the body an *agent*; if on a body, call the body a *subject*.

For example, if a company hires an agency to conduct drug tests on its staff, the classification is conducted for the company by the agency on the staff. The company is the principal, the agency is the agent, and the members of staff are the subjects.

There seem to be no grounds for conducting evaluation from the perspective of an agent. For example, in the medical sphere one can justify evaluating a classifier from the perspective of the patient (for example a pregnancy test) or the state, a public health attitude (for example a cancer screening program), but not from that of a doctor.

There is little scope for confusion or disagreement over the identity of agents or subjects. However, there can be disagreement over who the principal is, particularly in the political context. For example, in the legal sphere, different authors have developed different quantitative methodologies for evaluating the system of arriving at criminal verdicts by taking the principal to be a jury, the courts or the state (Barbin & Marec, 1987; Cullerne

Bown, 2018). Each of these implies a different population that is to provide the basis for evaluation: a jury — a single trial (and hence a reliance on subjective probability); the courts — multiple trials; the state — all actions by people in society. One cannot say in any absolute sense that one of these positions is right and the others wrong. However, one can say that: (i) no evaluation from the principal perspective is properly grounded until a principal is specified and the consequent population defined; (ii) this choice will shape the appropriate form of evaluation; and (iii) the choice will determine what kinds of questions the evaluation can address.

For example, if in considering criminal classification one adopts the courts as principal, then one restricts oneself to the population of cases that come to court, and the evaluation will be unable to tell us anything directly about the success or failure of the state in distinguishing criminal acts from innocent ones in society. Similarly, if in developing a radar system to detect incursions by enemy aircraft, one adopts as principal the manufacturer of an individual signal processing component in the radar system, then the evaluation will be unable to tell us anything directly about the success or failure of the system as a whole in identifying enemy aircraft.

In both cases, if we wish to improve the accuracy of the system as a whole, the most obvious approach will be to calibrate the system as a whole. Alterations to the policies governing the operation of the sub-system (the courts or signal processing component) can then be evaluated by observing their effect on the accuracy of the system as a whole. There is, so far as I am aware, no general basis for believing alterations to policies adopted on the basis of sub-system rather than whole-system evaluation will lead to improvements in whole-system accuracy.

If the perspective is that of a subject, then the only thing that matters is the correctness of the single label an individual is given. This may be positive or negative, and so there are again the two concerns addressed by the two ratios in the correctness pair. Hence, we have a second rule.

Rule 2 The subject perspective implies the correctness pair.

Classification may sometimes be performed in situations in which no one actor's concerns are considered paramount. For example, COVID-19 tests in the UK could be said to concern both their subjects and the state; both have a keen interest in their results and their effect on behavior. Such a situation calls for two parallel forms of evaluation, which may be based on different pairs.

6 Principal Decisions

6.1 The Case Class

Having dealt with non-decisiveness and the subject perspective, we are left with decisive classifiers evaluated from the principal perspective, a type of requirement that can be referred to as *principal decisions*. If we assume that the principal is purposeful rather than whimsical, then it follows that the labels are intended to result in actions of some kind with each distinct action implying a distinct underlying class and requiring a distinct label.

Definition The *null action* is what the principal intends when no classification takes place, if for example the classifier is broken. The *null class* is the class of things that should lead to the null action.

In a simple screening program for breast cancer, the null action is to not offer to treat the patient and the null class is mammogram scans that do not show cancer. In the criminal justice system, the null action is to not convict someone of a crime and the null class is acts that are not criminal. With a bank considering loans, the null action is to not grant a loan and the null class is loan applications that are unlikely to be profitable for the bank.

Definition Call any action intended by the principal other than the null action a *case action*. Call any class other than the null class a *case class*.

Note that the names now given to the classes are, unlike 1 and 0, not arbitrary but reflect a material distinction. In the binary case, this implies that any of the six pairs of ratios could be relevant. For this reason, and because of the precise clarity it encodes about the circumstances of the classification, it is preferable always to use the terms case class and null class when dealing with principal decisions. As a convention for determining how the ratios we have already defined are to be referred to, and in formulae generally, identify the case class with class 1 and the null class with class 0.

It may be argued that it is not always possible to distinguish a null class. Hand (2012) is of this opinion and briefly illustrates his view with a speech recognition system tasked with distinguishing between ‘yes’ and ‘no’. At first glance, we may think that we have a binary classifier in which both labels can be expected to result in actions that are not null. However, to treat this example properly, we must pay attention to a third possible response that must be present in the real world, which is that the speaker says neither ‘yes’ nor ‘no’, perhaps because they are busy eating popcorn.

Given the principal perspective, and assuming that a ‘yes’ is intended to result in an action, there are two distinct scenarios that we may face, depending on what happens when a ‘no’ is detected. If this leads to the same action as no response at all, then this is the null action and when a ‘no’ is detected the response will be placed in the null class. If a ‘no’ does not lead to the same action as no response at all, then we have a threefold classifier with: (i) a null class of sounds that should generate no action, call it 0; (ii) a case class of sounds that should generate the ‘yes’ action, call it Y; (iii) a case class of sounds that should generate the ‘no’ action, call it N. Either way, a null class emerges. The lesson is that once we have established the classification is decisive and from the principal perspective, we are able to acquire a clarity of analysis that is unavailable to Hand (2012).

The accuracy of the threefold classifier can be measured by re-conceiving of it as a sequence of two binary classifiers, each with their own null class. The first classifier classifies the sounds as Y or 0. The second classifier takes those sounds classified as 0 and divides them between N and 0. We can then measure the accuracy of the threefold classifier as a whole by acquiring measurements of the accuracy of each component classifier and combining these into a single score.

This example can be generalized. When dealing with principal decisions, we can always reduce matters to a branching path of binary classifiers in each of which there is a null class and a case class. An n -fold classifier will result in a branching path of $n-1$ binary classifiers. By measuring the accuracy of each component classifier and combining the measurements into a single score we can obtain a measurement of the accuracy of the classifier as a whole. The branching paths, however, are not unique and there is no particular reason why two different arrangements should yield identical discardings.

6.2 Skew, Subpopulations, and Attributes

Skew sometimes is defined in terms of probabilities, sometimes in terms of tallies, and this will often be a matter of convenience or habit. In this article, because our interest lies in the measurability of the underlying classes, it is simplest to define it in terms of tallies.

Definition $\text{Skew} = \frac{|\text{Class } 1|}{|\text{Class } 0|}$.

With principal decisions, the labelling of the classes is not arbitrary and hence high skew does not simply mean a highly unbalanced population; it implies a preponderance of the case class. It has direction. Bear this in mind when you see below terms such as ‘high skew’.

It is well established that as the skew of the population it works on varies, so the accuracy of a classifier, however measured, can be expected to vary. Problematically, one classifier may be more accurate than another at one skew and less accurate at another skew (Lampert & Gançarski, 2014) (this being one example of the general problem that a classifier trained on one kind of data may perform badly when it is applied to a different kind of data). Thus, one reason for measuring skew is to inform discardings.

We may be able to identify subpopulations that have different skew, some higher, some lower.

Definition Let skew_s be the skew of the subpopulation S and skew_{-s} the skew of the remainder of the population, then the *loadedness* of S is $\frac{\text{skew}_s}{\text{skew}_{-s}}$. S is *loaded towards class 1* when loadedness is greater than 1; *towards class 0* when loadedness is less than 1; and *towards neither class* when loadedness is equal to 1.

A subpopulation loaded towards the case class has, compared to the rest of the population, a higher proportion of things in the case class and a lower proportion in the null class. Everything else being equal, this means it will yield more true positives and false negatives (the outcomes of the case class) and fewer true negatives and false positives (the outcomes of the null class). Thus, pyn and nyp will be unchanged in the loaded subpopulation but pyp will be higher and nyn will be lower. Thus, for example, any control function based on the prospecting pair of pyn and pyp will yield a higher score.

The very final conclusion here depends on two assumptions about the form of the ultimate control function that I consider reasonable. First, if we adopt a pair of ratios, it is because we genuinely care about the concern that each one measures. The control function must therefore give some weight to both and cannot exclude one altogether from its computations. Second, as mentioned in Sect. 3.1, in any ratio, a higher score is better. This constraint implies that, for example, a control function can not allocate a higher score where both input ratios have lower values.

We can follow the same reasoning through for the other ratios and pairs, and for subpopulations loaded towards the null class. In some cases, the result is to improve the evaluative control score, in some cases to worsen it, in some cases to leave it unchanged and in some cases the upshot is unclear and depends on the form the later evaluation takes (Table 4).

Skew is also a helpful concept when considering the character of different attributes that we may ask a classifier to measure. For example, suppose we have a heap of rocks and are looking for nuggets containing gold, and that some of the nuggets containing gold have gold on the surface and some do not. Then the glitter of gold to our eyes will reliably indicate the case class but the absence of glitter will not reliably indicate the null class. We can make this difference precise as follows.

By imposing a threshold, any scalar attribute can be reconceived as a binary. Any binary attribute, B divides each class in two, between those that have the attribute and those that do not. We can thus regard each class as a distinct population that has two classes defined by the presence (by convention class 1) or absence (class 0) of the attribute.

Definition Let $skew_{1B}$ stand for the skew of class 1 when it is divided by the attribute B . Then:

$$\text{Indicativeness of } B = \frac{skew_{1B}}{skew_{0B}}$$

An attribute is said to be *indicative of class 1* when indicativeness is greater than 1; *indicative of class 0* when indicativeness is less than 1; and *indicative of neither class* when equal to 1.

In the above example, class 1 is nuggets that contain gold, class 0 is nuggets that don't and the attribute in question is surface glitter. We have $|gold\ nuggets\ that\ glitter| > 0$ and $|gold\ nuggets\ that\ don't\ glitter| > 0$ so that $skew_{1glitter} > 0$. And we have $|goldless\ nuggets\ that\ glitter| = 0$ and $|goldless\ nuggets\ that\ don't\ glitter| > 0$ so that $skew_{0glitter} = 0$. Hence the indicativeness of glitter is greater than 1 and glitter is indicative of gold.

Any binary attribute that can be measured can be used as a pre-processor. That is, we can establish a pathway of two classifiers in which: the first uses an attribute, say B , to allocate things to its case class; the second takes the case class from the pre-processor as its input and uses all other methods to allocate some of these things to its case class. In this, if B is indicative of the case class, and the preprocessor is reliable, it will have the effect

Table 4 The effect of adopting a loaded subpopulation or indicative binary attribute as a pre-processor on indicators is shown as follows: + increase; – decrease; = no change; $\pm$ unclear

	Loaded towards or indicative of class 1	Loaded towards or indicative of class 0	Not loaded towards or indicative of either class
Tallies			
TP	+	–	=
FN	+	–	=
TN	–	+	=
FP	–	+	=
Ratios			
Pyn	=	=	=
Pyp	+	–	=
Nyp	=	=	=
Nyn	–	+	=
Control functions based on pairs			
Achieving	=	=	=
Correctness	$\pm$	$\pm$	=
Prospecting	+	–	=
Anti-prospecting	–	+	=
Avoidant	+	–	=
Anti-avoidant	–	+	=

of increasing the skew of the population encountered by the second classifier. A move to deploy a new binary attribute that is indicative of the case class as a preprocessor therefore has the same effect on tallies, ratios, pairs, control functions and ultimately evaluation as a move to a subpopulation loaded towards the case class.

This is not to say that a scalar attribute should be turned into a binary and deployed in this way, only that this can always be done and so the effects detailed in Table 4 can always be achieved.

Now we can draw a conclusion that is specific to principal decisions.

Definition The *direct path* is the strategy of attempting to identify things of the case class; each true positive obtained is a *step* on the path.

The direct path is an unbalanced strategy in that it is uninterested in the null class and is an obvious response to the equally lopsided interest of the principal. The most obvious way of pursuing it is to identify: (i) subpopulations that are loaded towards the case class; and (ii) attributes that are indicative of the case class.

The direct path implies we are interested in only those ratios that tell us about success in identifying the case class. These are the two that have |TP| as their focus: pyp and pyn, the prospecting pair. Pyn tells us about our achievement in identifying things and pyp tells us about the correctness of the labels we get. Happily, as shown in Table 4, pursuing the direct path increases accuracy when measured with the prospecting pair. The question is, is the direct path the only way forward?

6.3 Indeterminacy

It is easy enough to set out how a ratio should be computed, but in what circumstances can it in fact be computed, and how should we react if it cannot be computed? To answer the second question first, if an indicator cannot be computed, it is of no use to us and must be abandoned. We cannot achieve a quantitative understanding of a classifier by relying on a quantity that cannot be established. To answer the first question, we need an additional concept.

Definition A quantity is *determinate* when it is: (i) finite; and (ii) has been reliably estimated, or we are confident that we can make a reliable estimate of it. Otherwise, it is *indeterminate*.

‘Reliable’ here refers to an idea that, so far as I am aware, has been discussed most fully in Lampert and Gañarski (2014) in their discussion of problems associated with skew. The two mention several kinds of scenario where a lack of stability in the skew may make any evaluation problematic, including:

- A. Scenarios, such as a petri dish covered with bacteria, where the skew varies over time but in a predictable way;
- B. Scenarios with a fixed skew that is unknown; for example, the authors do not know what proportion of all eye fundus images comprise blood vessels;
- C. Scenarios with skew that varies unpredictably in space, such as in satellite imagery the number of trees, buildings or fissures in a hectare of surface;

- D. Scenarios with skew that varies unpredictably over time, such the number of buildings in a hectare of surface image or the number of red cars in an hour of road surveillance video (described as chaotic).

Let us consider in turn whether the skew in these four scenarios is determinate or not.

- A. With the bacteria, the skew is not stable but can usually be reliably estimated for any point in time on the basis of a well-established equation — the skew is thus determinate;
- B. Until we have a way of acquiring a representative sample of the population of eye fundus images, the skew will remain indeterminate;
- C. Presumably, given some set of satellite images, it is possible to determine the skew in the images. The problem is rather that, after the algorithm is selected, it will be presented with new sets of images where the skew may be quite different and cannot be predicted. Given enough effort, it presumably would be possible to determine the skew even of these new sets but the practical reality, say Lampert & Gançarski, is that we lack the resources to do so. The skew of such new sets of images is thus indeterminate.
- D. Since the image is snapped at a point in time, it immediately becomes out of date. A year later, there may be significantly more or fewer trees. In 10 years, a hectare that was empty of trees may be full of them, or vice-versa. One can imagine models that might make estimates of future skew reliable, but in the absence of that, the skew is indeterminate.

Lampert & Gançarski consider only whether skew is indeterminate and do not define indeterminacy itself. In this article, the concept is made more general and more basic, and is defined. Its most important application for us is to the counting of the members in a set.

The cardinality of a set is the number of elements in the set or, in the case of a set with infinite elements, an infinity. It may also be called the size of the set. Thus, it is possible, for example, that $|\text{Null Class}|$ is indeterminate while $|\text{Case Class}|$ is determinate.

If a derived quantity is computed from a number of more basic components, then if one or more of the components is indeterminate, it will be impossible to reliably compute the derived quantity. Thus, we can say in such a case that the derived quantity is also indeterminate. If, say, $|\text{Null Class}|$ is indeterminate while $|\text{Case Class}|$ is determinate, then the skew will be indeterminate.

6.4 Determinate Skew

Now, let us consider evaluation of a classifier making principal decisions in the scenario in which the skew of the population to be classified is determinate. From this we can expect to be able to extract a representative sample with the same skew, and the accuracy of the classifier will be estimated on the basis of calibration on this. (Such an approach will not eliminate potential problems such as overfitting; however, these problems essentially arise from having a sample that is not truly representative and which contains a misleading distribution of attributes.)

In this case, there are in principle two ways to identify all the things of the case class, by labelling all the case class as positive or by labelling all the null class as negative. Furthermore, each thing in the null class that is correctly classified as negative makes our task in identifying the things in the case class easier for two reasons. Firstly, because there are fewer things remaining to process and, even if the cost is small, processing always costs something, both in resources and time. Secondly, because it raises the skew of the remaining population.

Definition The *mirror path* is the strategy of attempting to identify things of the null class; each true negative is a *step* on the path.

The most obvious way of pursuing the mirror path is to identify: (i) attributes that are indicative of the null class; and (ii) subpopulations that are loaded towards the null class.

The mirror path makes progress by generating true negatives and the two ratios that are therefore appropriate to measuring our progress along it are nyp and nyn, the anti-prospecting pair. It will be hard to pursue if we adopt a form of evaluation that does not count the true negatives at all, as in the prospecting pair.

We thus find ourselves in a situation in which a case can be made for all four ratios, the prospecting pair for the direct path and the anti-prospecting pair for the mirror path. To actually adopt all four would be a dimensional error in that there are two degrees of freedom in the system (four variables — the tallies of outcomes — with two constraints — the cardinality of the two classes) and no more than two indicators should be used to characterize it.

In choosing a pair, there is no reason to exclude any tally. Hence, we must choose between the achieving and correctness pairs. Achievement monitors the central concern of progress towards the identification of all the case class, pyn monitoring progress along the direct path and nyp progress along the mirror path. Meanwhile, correctness monitors the collateral damage that is suffered along the way, pyp for the direct path and nyn for the mirror path. Achievement is the principal's primary concern and thus it is natural that we should extract from the four the achieving pair. Thus, we have our third rule.

Rule 3 Determinate skew implies the achieving pair.

The argument immediately above suggests circumstances in which reverting to the default position of the prospecting pair will be less costly. First, the lower the skew, the longer the mirror path and the smaller the benefit that is accrued from each step on the path. Second, the lower the cost of processing a thing, the lower the advantage gained from excluding a thing from the null class. Third, we may struggle to find attributes that are indicative of the null class. Fourth, we may struggle to identify subpopulations loaded towards the null class.

Thus, the case in favor of the achieving pair is reduced in circumstances where the population is dominated by the null class, processing is cheap, no attributes indicative of the null class seem to be available and there is little variation in skew between identifiable subpopulations. Such circumstances do not eliminate the case in favor of the achieving pair; even if the benefits are small, why throw them away? Even if the mirror path turns out to be too difficult or expensive to pursue, such a decision should reflect investigation rather than the form of evaluation. But historic circumstances along these lines may help explain how some in the computing community, for example Davis and Goadrich (2006), came to the conclusion that severe imbalance in classes in itself justifies the use of the prospecting pair; compared to today's large language models, computerized models in earlier decades were small and training was cheap.

6.5 Indeterminate Skew

In the case of indeterminate skew, we need to consider three possibilities: that the null class only is indeterminate, that the case class only is indeterminate or that both classes are indeterminate.

Suppose the case class is determinate and the null class indeterminate. Then the mirror path evaporates in that any advantage gained by accurately identifying a thing in the null class is opaque and unquantifiable or, in the case of a null class with infinite cardinality, infinitely small. It is not possible to measure our achievement in excluding null class things. Thus, *nyp* is of no use to us and there are no grounds for adopting the achieving pair. We therefore should stick with our original preference for the prospecting pair.

Rule 4 If the case class is determinate and the null class indeterminate, adopt the prospecting pair.

Here is an example. Suppose we have a population where the case class has 1,000,000 members and the cardinality of the null class is indeterminate. Suppose the classifier then processes 1000 things, 300 of class 1 (of which it allocates 200 to class 1 and 100 to class 0) and 700 of class 0 (200 to class 1 and 500 to class 0). The classifier has processed 0.03 per cent of class 1. Its success in correctly allocating them the positive label is measured by *pyn* and we can compute that this is $2/3$. It is unclear what proportion of the class 0 subpopulation has been processed and hence it is unclear to what extent the processing of things in class 0 has reduced the task of identifying the things in class 1; progress along the mirror path is opaque. Hence computing *nyp* serves no purpose. On the other hand, we can measure the cost in false positives for each true positive that is identified. This is measured by *pyp*, which we can compute as $1/2$.

We have reached exactly the conclusion set out by Van Rijsbergen (1979, Chapter 7) in making the case for the use of precision and recall (*pyp* and *pyn*) in the field of information retrieval [with my annotations in square brackets]:

The situation may therefore be pictured as shown in Fig. 7.11, where *A* is the set of relevant documents [case class], *B* the set of retrieved documents [positives], and $A \cap B$ the set of retrieved documents which are relevant [true positives].

A reproduction of Fig. 7.11 in (Van Rijsbergen, 1979) is included in Fig. 1. In contrast to earlier set theoretic conceptions of the classificatory problem that had four subsets (corresponding to the four outcomes), this has only three subsets. The subset corresponding to the true negatives is missing altogether. Thus, information retrieval does not need to rely on an estimate of the number of documents that are *not* relevant, a set whose cardinality is hard to estimate and liable to be constantly changing.

Now suppose the case class is indeterminate and the null class determinate. *Pyp* is one of the default pair that the principal is naturally interested in. Both of its components, $|TP|$ and

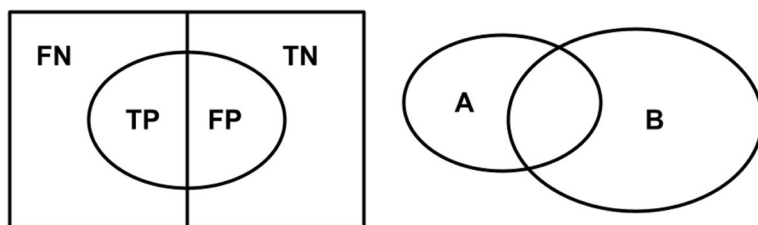


Fig. 1 Van Rijsbergen's conception of the nature of the classificatory problem in information retrieval (right) compared to the earlier conception (left)

[FP], are available to us. We should therefore adopt it as one of our pair. Since the case class is indeterminate, *pyn* is not available (the cardinality of the case class being its denominator). However, the mirror path exists and so achievement in the null class is a real sign of progress; we can measure this with *nyp*. Hence, we should adopt *pyp* and *nyp*, the avoidant pair.

Rule 5 If the case class is indeterminate and the null class determinate, adopt the avoidant pair.

Rule 6 If both classes are indeterminate, then the only ratio that can be computed is the one composed of [TP] and [FP], that is *pyp*, and this is therefore the only possible choice.

I am not aware of any existing usage of the avoidant pair or *pyp* on its own but that I think arises from a lack of understanding rather than an absence of reasons. It is certainly possible to have a case class of indeterminate cardinality. To see this, consider our experience with COVID-19 testing.

The cardinality of the case class (those who have COVID-19) can be constantly varying. However, through surveillance and modelling, the current (and short-term future) cardinality is not entirely mysterious in a country such as the UK. Also, the government's policy of interventions imposes some control on it. Thus, rather than being truly indeterminate, the cardinality of the COVID-19 case class in such a country is a more complex form of the bacteria-in-a-petri-dish scenario considered by Lampert & Gańczarski, that is where the cardinality of the case class is non-indeterminate-but-hard-to-evaluate.

Now, consider a state that lacks the UK's infrastructure of surveillance, modelling and control. If it wants to identify people with COVID-19, such a state *is* faced by a case class that is indeterminate. Here both the case class and null class (those who do not have COVID-19) are indeterminate, so that *pyp* is indicated.

Now consider a state that, perhaps via a census, has a good estimate of the size of its population of people who are legally entitled to reside in the country but has no good estimate of the number of illegal aliens in the country. Suppose this country wants to establish a system to identify illegal aliens among the population. Then, the case class is indeterminate, but the null class is determinate, and the avoidant pair is called for.

7 Conclusion

We can now get a high-level overview of the entire method. To apply the rules set out herein, any classifier must first be reduced to a pathway of binary classifiers and each classifier evaluated separately. The correct pair of ratios in each case then depends on decisiveness, perspective and, in the case of principal decisions, the determinacy of the classes (Table 5, Fig. 2).

It turned out that not all six pairs are required. We have no use for either the anti-avoidant or anti-prospecting pair, even in the case of principal decisions where these two pairs are not in principle superfluous. Instead, in one scenario we have been driven to resort to a single ratio.

It is possible that one or both classes will be indeterminate in the first two scenarios considered, non-decisiveness or the subject perspective. If so, it will be impossible to compute one or both of the necessary ratios.

If a wrong pair is chosen, we will find ourselves dealing with statistics that represent concerns that are not appropriate to the circumstances and people involved. Instead of

Table 5 Six rules are derived in this article for deciding between pairs of ratios, including cases in which only a single ratio is indicated. Four of them concern principal decisions

Rule	Circumstances	Pair
1	Non-decisive	Correctness
2	Subject perspective	Correctness
	Principal decisions	
3	Both classes determinate	Achieving
4	Null class indeterminate	Prospecting
5	Case class indeterminate	Avoidant
6	Both classes indeterminate	Pyp (one ratio only)

clarifying the trade-offs involved when making discardings, the quantification will confuse them. As Hand (2012) says, we will find ourselves answering the wrong question. This sets

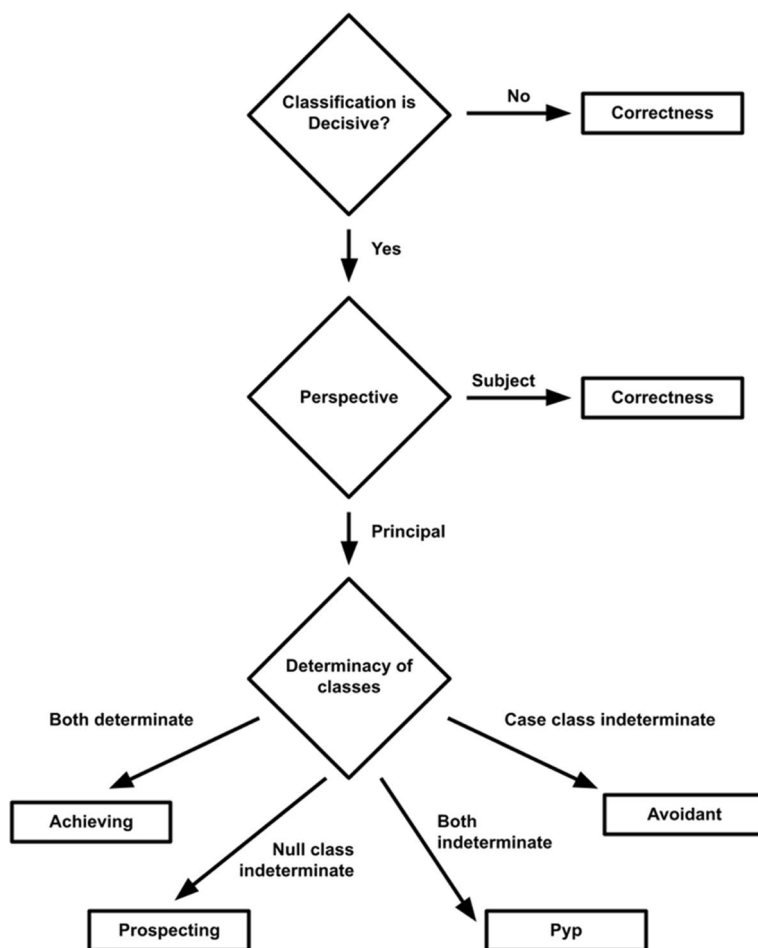


Fig. 2 The way in which context determines which pair of ratios (or in one case, a single ratio) should be used to evaluate a binary classifier can be summarized in a flow chart. This encapsulates the six rules set out in the article

the stage for flawed discardings and undermines the function of the statistics in helping people make choices. Two concrete examples of the problem stand out.

First, in the case where we neglect the question of determinacy or mistake an indeterminate class for a determinate one, we will be allowing ourselves to be guided by intrinsically unreliable measurements. This is a recipe for flawed discardings and reduced accuracy.

Second, in the case where the achieving pair is rejected in favor of the prospecting pair the benefits of the mirror path will be lost. One risk then is that discardings will again be flawed and accuracy will be reduced.

The argument in this article seeks justification in reasoning, not by demonstrating empirical results. This is not to deny that results are the ultimate test, but no results are produced in this article in support of the method set out. To a determined pragmatist, the only test is empirical results, and to such a reader I have no answer. A more moderate pragmatist, while accepting that reasoning may have something to offer, may still be troubled by the question of applicability: does the argument set out in this article apply in my field? Certainly, the goal of this article is to establish a method that is truly universal. This cannot be established through a formal proof nor through any number of examples, though I have tried to illustrate the argument with variety. Rather, one must look at the form the reasoning takes.

The spadework of reasoning in this article is accomplished with the help of some elementary ideas from the theory of measurement. This is as basic and universal as it gets if we are committed to an empirical understanding and cannot be avoided if one is engaged in a quantitative form of evaluation. However, the overall scheme of reasoning is provided by patiently distinguishing concerns. Once the concerns are distinguished, often the rationale for a methodological step stares us in the face. The concerns have been arrived at by distilling the common, telling features from a wide variety of different real-world scenarios, but only time can tell whether I have truly managed to encompass all possible circumstances. If your field has a concern that is not allowed for in this article, then this approach may not resolve the appropriateness question for you. Even so, I hope the underlying body of reasoning will help you find your way to the correct pair.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00357-024-09478-y>.

Data Availability In this foundational article, I do not analyze or generate any datasets.

Declarations

Ethics Approval This research does not contain any studies with human participations or animals performed by the author.

Conflict of Interest The author declares no competing interests.

References

- Barbin, E., & Marec, Y. (1987). Les recherches sur la probabilité des jugements de Simon-Denis Poisson. *Histoire & Mesure*, 2, 39–58. <https://doi.org/10.3406/hism.1987.1311>
- Cullerne Bown, W. (2018). The criminal justice system as a problem in binary classification. *The International Journal of Evidence & Proof*, 22, 363–391. <https://doi.org/10.1177/1365712718795548>
- Cullerne Bown, W. (2023). An epistemic theory of the criminal process, Part II: Packer, Posner and epistemic pressure. *Law, Probability and Risk*, 21, 61–83. <https://doi.org/10.1093/lpr/mgac014>

- Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning - ICML '06. Presented at the 23rd international conference* (pp. 233–240). Pittsburgh, Pennsylvania: ACM Press. <https://doi.org/10.1145/1143844.1143874>
- Dinga, R., Penninx, B.W.J.H., Veltman, D.J., Schmaal, L., Marquand, A.F., 2019. Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines. bioRxiv 743138. <https://doi.org/10.1101/743138>
- Ducharme, G.R. (2018). 'Critères de qualité d'un classifieur généraliste'. Working paper, University of Montpellier, France. <https://hal.umontpellier.fr/hal-01819793>. Retrieved 23 September 2019.
- Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179–188.
- Flach, P.A., (2003). The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics, in: *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, 194–201. Washington DC, USA: AAAI Press
- Franklin, J. (2002). *The Science of Conjecture: Evidence and Probability before Pascal, Annotated* (edition). The Johns Hopkins University Press.
- Hammond, K. R. (1996). *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error*. Oxford University Press.
- Hand, D. J. (2017). Measurement: A Very Short Introduction—Rejoinder to discussion. *Measurement: Interdisciplinary Research and Perspectives*, 15, 37–50. <https://doi.org/10.1080/15366367.2017.1360022>
- Hand, D. J. (2012). Assessing the Performance of Classification Methods. *International Statistical Review*, 80, 400–414. <https://doi.org/10.1111/j.1751-5823.2012.00183.x>
- Harnad, S. (2005). To Cognize is to Categorize: Cognition Is Categorization. In H. Cohen & C. Lefebvre (Eds.), *Handbook of Categorization in Cognitive Science* (pp. 19–43). Oxford: Elsevier Science Ltd. <https://doi.org/10.1016/B978-008044612-7/50056-1>
- Hossin, M., Sulaiman, M.N., (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2), 1–11.
- Jiao, Y., & Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant Biol*, 4, 320–330. <https://doi.org/10.1007/s40484-016-0081-2>
- Lampert, T. A., & Gançarski, P. (2014). The bane of skew. *Machine Learning*, 97, 5–32. <https://doi.org/10.1007/s10994-013-5432-x>
- Lever, J., Krzywinski, M., & Altman, N. (2016). Classification evaluation. *Nat Methods*, 13, 603–604.
- Liu, Y., Zhou, Y., Wen, S., & Tang, C. (2014). A Strategy on Selecting Performance Metrics for Classifier Evaluation. *IJCMCM*, 6, 20–35. <https://doi.org/10.4018/IJCMCM.2014100102>
- Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*, 2, 37–63.
- Škrabánek, P., Doležel, P. (2017). On Reporting Performance of Binary Classifiers'. Scientific Papers of the University of Pardubice. Series D, Faculty of Economics and Administration 2017, 41. <https://dk.upce.cz/handle/10195/69604>
- Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*. <https://doi.org/10.1016/j.aci.2018.08.003>
- Van Rijsbergen, C. J. (1979). *Information retrieval*. Butterworths.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.