Never mind the metrics—what about the uncertainty? Visualising binary confusion matrix metric distributions to put performance in perspective

David Lovell [1] **Dimity Miller** [2] **Jaiden Capra** [3] **Andrew P. Bradley** [4]

Abstract

There are strong incentives to build classification systems that show outstanding performance on various datasets and benchmarks. This can encourage a narrow focus on models and the performance metrics used to evaluate and compare them—resulting in a growing body of literature to evaluate and compare *metrics*. This paper strives for a more balanced perspective on binary classifier performance metrics by showing how uncertainty in these metrics can easily eclipse differences in empirical performance. We emphasise the discrete nature of confusion matrices and show how they can be well represented in a 3D lattice whose cross-sections form the space of receiver operating characteristic (ROC) curves. We develop novel interactive visualisations of performance metric contours within (and beyond) ROC space, showing the discrete probability mass functions of true and false positive rates and how these relate to performance metric distributions. We aim to raise awareness of the substantial uncertainty in performance metric estimates that can arise when classifiers are evaluated on empirical datasets and benchmarks, and that performance claims should be tempered by this understanding.

1. Introduction

Today's algorithmic modeling culture (Breiman, 2001) rewards those whose models outperform all others. Performance optimisation is central to statistical, ML and AI models (Thomas & Uminsky, 2020), and numerical metrics are often regarded as objective, valid indicators of performance; a model's performance on a test dataset or benchmark task is seen as indicative of its ability to perform "in the wild" on new, real-world data. This paper strives for a more balanced perspective on classifier performance metrics by visualising their distributions under different models of uncertainty. We aim to draw attention towards the role of more data—and data that is more *representative* of a classifier's intended application—to characterise and improve different classifiers, rather than judging them through performance shoot-outs alone.

Confusion matrices summarise the empirical performance of classifiers. Simplest are those that tally the four outcomes of *binary decisions* (Figure 1). A desire to further summarise each matrix with one number—which, necessarily loses information because the matrices have three degrees of freedom—has led to many metrics with different meanings, interpretations and ranges (Appendix A).

These different performance metrics produce different classifier performance rankings, and there are often strong incentives to top these ranks (Maier-Hein et al., 2018). Those seeking to develop or promote "the best classifier" may wonder which is "the best performance metric"; several studies argue the merits of different metrics, even though what is "best" in practice depends on a classifier's specific real-world application (Hand, 2006; Rudin & Radin, 2019). Others have sought to understand and characterise the behaviour and interpretation of these different metrics; we follow this philosophy to reveal further insight.

In essence, we show how uncertainty in (discrete) confusion matrices manifests in various (continuous) performance. Our aim is to encourage more attention towards reducing uncertainty in performance estimates before attempting to argue the merits of a particular classifier or metric.

2. Prior motivating work

2.1. Studies that argue the merits of specific metrics

Our research was stimulated by a series of papers whose titles suggest that Matthews Correlation Coefficient (MCC) is better than other metrics (Chicco & Jurman, 2020; Chicco

[1]School of Computer Science and Centre for Data Science, Queensland University of Technology, Australia. [2]School of Electrical Engineering and Robotics, Queensland University of Technology, Australia. [3]Queensland University of Technology, Australia. [4]School of Science, Technology and Engineering, University of the Sunshine Coast, Australia. Correspondence to: David Lovell <David.Lovell@qut.edu.au>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Total $N = p + n$ *predicted class*	*actual class* positive $p = a + c$	negative $n = b + d$
positive $\hat{p} = a + b$	TP True Positives (a)	FP False Positives (b)
negative $\hat{n} = c + d$	FN False Negatives (c)	TN True Negatives (d)

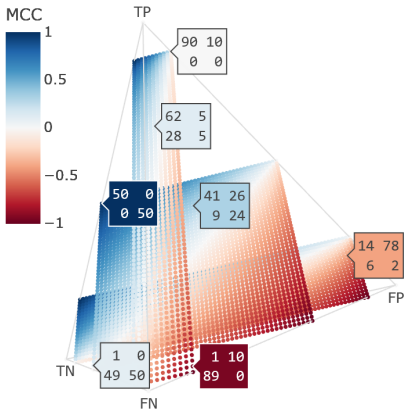


Figure 1: (Left) A binary confusion matrix shows the counts of a classifier's predictions in response to a set of examples whose actual classes are known. With fixed total (N), these four-element matrices have 3 degrees of freedom.
(Right) 3D projections of confusion matrices with $N = 100$. Each point corresponds to a unique confusion matrix and is coloured by its Matthews Correlation Coefficient. We label the four extreme points ($\mathrm{TP} = 100, \mathrm{FN} = 100$, etc.) of the regular simplex. Rather than show all 176 851 possible points, we show three slices corresponding to matrices where $p = 20, 50, 90$, from back to front. This graphic is a composite from our interactive visualisationour interactive visualisation (see Appendix D.1).

et al., 2021a;b). While the story in these papers is more nuanced than their titles suggest, their high citation rates suggest that they have caught people's attention, especially in bioinformatics (Chicco, 2017). A recent opinion piece may see MCC's popularity rise in robotic vision (Chicco & Jurman, 2022). However, Zhu (2020) challenges the idea that MCC should be *"generally regarded as a balanced measure which can be used even if the classes are of very different sizes"*. Note that Chicco et al. (2021b) clearly state that MCC is perhaps not the best measure in all situations.

Instead of arguing for specific metrics, we seek to enable practitioners to explore and understand the behaviour of performance metrics and the uncertainty in their empirical values. *"Several rates that summarize... the confusion matrix exist nowadays; none of them, however, has reached consensus in the computer science"* (Chicco et al., 2021a). The idea of "the best" metric oversimplifies the fact that *"each of the indicators serves a different purpose"* (Glas et al., 2003). Practitioners have a responsibility to understand the strengths and limitations of different indicators for the task at hand, and that metrics only *summarise* empirical performance; they do not provide more precise estimates of performance—that demands more data, representative of the context into which a classifier would be deployed.

2.2. Studies that relate various performance metrics

Powers (2011) describes algebraic relationships between several metrics before using simulation to explore their behaviour; Ferri et al. (2009) use cluster analysis to group metrics that behave similarly. It is difficult to draw strong conclusions from the values metrics take on simulated or experimental data, especially when comparisons are averaged or clustered across different datasets.

Other authors seek to characterise performance metrics more directly by exploring how they satisfy various properties. Sokolova & Lapalme (2009) consider how the values of different metrics change as the counts in a confusion matrix change; Brzezinski et al. (2018) consider ten desirable properties of metrics on confusion matrices with *fixed* totals, but different prevalence. Gösgens et al. (2021) consider further desirable properties and show that some of them are incompatible. Luque et al. (2019) characterise different performance metrics as class balance changes.

2.3. Studies that visualise performance metric geometry

Broadly speaking, there have been two main approaches to visualising the values that performance metrics take over the space of possible binary confusion matrices. The first uses the *rates* that characterise different confusion matrices (e.g., true positive rate, false positive rate, prevalence (Flach, 2003; Luque et al., 2019)). This yields visualisations that are essentially *"a collection of stacked-up ROC spaces, with the z-coordinate corresponding to the proportion of the positive class"* (Brzezinski et al., 2018).

The second approach uses projections that preserve *distances* between confusion matrices. To explain the notion of distance here, consider the confusion matrix in Figure 1. By adding 1 to one element and subtracting 1 from another, we can construct 12 different adjacent confusion matrices, equidistant from the original matrix. Barycentric projection

(Brzezinski et al., 2018) ensures these adjacent confusion matrices are equidistant in 3D projection. Rather than producing a cubic stack of square ROC spaces, this projection maps confusion matrices to points that are tetrahedrally packed—different class imbalance ratios yield rectangular cross-sections of this tetrahedron.

2.4. Studies that consider uncertainty

Classifier performance is estimated using finite amounts of test data; the more data we use, and the more representative that data is, the more certain we are about the classifier's performance on new data. Also, in multinomial classification (Lovell et al., 2021), it is important to consider the amounts of test data we have *for each class of interest*; while the total amount of test data can be large, a specific class can be rare, and our estimates of a classifier's ability to correctly detect it become less certain.

Confusion matrices summarise the performance of binary classifiers by counting TP, the number of positive examples (p) correctly classified, and TN, the number of negative examples (n) correctly classified. Bayesian statistics provides an elegant framework for incorporating prior belief into modeling the predictive distribution of these counts (i.e., the distribution of counts we may expect to see in future trials) through the beta-binomial model (Murphy, 2012; Navarro & Perfors, 2010; Agresti, 2013). This approach is used by Tötsch & Hoffmann (2021) who refine Caelen's (2017) Dirichlet-multinomial model of confusion matrices.

Having reviewed prior relevant research, we now present our approach to visualising the distributions of these metrics so that practitioners can put empirical classifier performance statistics into perspective.

3. Viewing binary confusion matrices in 3D

When the four elements of binary confusion matrices sum to a fixed total, these matrices have only three degrees of freedom and can therefore be represented in three dimensions. Chicco et al. (2021a) do this by dividing each element of the confusion matrix by its total and using three of these ratios (TP/N, TN/N and FP/N) to project the confusion matrix into a *confusion tetrahedron* with vertices $A = (1, 0, 0)$, $B = (0, 1, 0)$, $C = (0, 0, 1)$ and $O = (0, 0, 0)$ Note that this is not an isometric projection. More importantly, this projection carries only *relative* information. By dividing by N, this projection loses information about how many times we have observed a classifier make predictions, and thus, how certain we can be about its predictive performance. With these issues in mind, we propose an isometric projection that *preserves information* about the underlying counts:

$$\begin{bmatrix} a & b & c & d \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \end{bmatrix} = \begin{bmatrix} x & y & z \end{bmatrix}. \quad (1)$$

We refer to the projected points as a *confusion simplex*, to distinguish it from the confusion tetrahedron of (Chicco et al., 2021a). The vertices of a confusion simplex for matrices of size N are $\mathrm{TP} = (N, 0, 0), \mathrm{FP} = (0, N, 0), \mathrm{FN} = (0, 0, N)$ and $\mathrm{TN} = (-\frac{N}{3}, -\frac{N}{3}, -\frac{N}{3},)$ as shown in Figure 1(b). The Euclidean distance between each pair of vertices is $N\sqrt{2}$. This projection preserves information about the counts of confusion matrices and, hence, how certain we can be about classifier performance.

While drafting this paper, we learned that a similar isometric projection had been previously proposed by Brzezinski et al. (2018) which maps the extreme confusion matrices with a given total (i.e., the matrices where $\mathrm{TP} = N$, $\mathrm{FP} = N$, *etc.*) to four of the corners of the cube $[-1, +1]^3$. Like the confusion tetrahedron, Brzezinski et al.'s projection carries only *relative* information about confusion matrices but could be rescaled to provide a rotated, translated version of Equation (1).

4. Slicing the 3D confusion simplex into ROCs

Classifiers are often evaluated and compared by presenting them with a fixed set of N examples, p positive, and n negative. Interest centres on how many actual positives are correctly identified (TP) and how many actual negatives are correctly identified (TN). Thus, the performance of classifiers can be represented in two dimensions using a rectangular lattice of $(p + 1) \times (n + 1)$ points.

Figure 1(b) shows perspective views of three different 2D slices (rectangular lattices) of points in a confusion simplex: a tall, skinny lattice ($p = 90$, $n = 10$); a square lattice where classes are *balanced* ($p = 50$, $n = 50$); and a short, wide lattice ($p = 20$, $n = 80$). These slices can be shown in rectangular 2D orthographic projections (Figure 2(a)). When the axes of these projections are scaled so that they plot the true positive rate (TPR) against the false positive rate (FPR) of each point, we see these points in the space of the receiver operating characteristic (ROC) curve (Figure 2).

ROC curves provide a compact summary of empirical true and false positive rates but, by using rates, they omit information about the underlying numbers of positive and negative examples classified. These numbers indicate how certain we are about a classifiers' performance: the more positive and negative cases the classifier has seen, the more certain we are about its performance: true and false positive rates alone do not carry this information. To inject

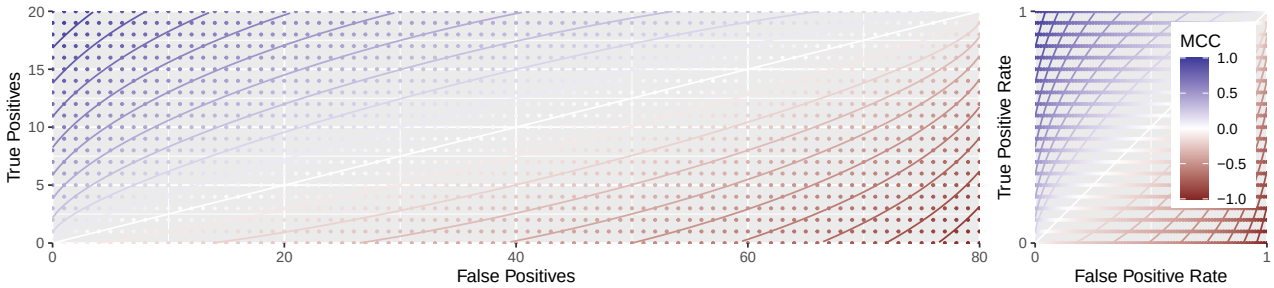


Figure 2: (a) Orthographic projection of the slice of points from the confusion simplex of Figure 1(b) where $p = 20$ and $n = 80$, coloured by MCC value. Continuous lines indicate the contours of MCC, ranging from $-0.9, \ldots, 0.9$. While MCC can be calculated for continuous arguments, confusion matrices map to $(p + 1) \times (n + 1)$ discrete points in this 2D lattice. (b) ROC curves plot a classifier's true positive *rate* against its false positive *rate*. This is equivalent to re-scaling the x-axis of (a) by a factor of $\frac{1}{n}$ and the y-axis by $\frac{1}{p}$.

that information back into visualisations, we could use orthographic projection (Figure 2(a)) so that the aspect ratio clearly shows the balance of positive and negative examples. However, when classes are highly imbalanced, this approach becomes challenging to print and inspect. To deal with this, we suggest faintly plotting all possible points in the ROC space for reference (as in Figure 2(b)) to reveal when there are few positive or negative examples. This approach can also be used with precision-recall plots (Davis & Goadrich, 2006) and we provide an interactive visualisation at to show how reference points in ROC and Precision-Recall plots relate (Appendix D.2). When p or n are so large that the lattice of ROC points appears continuous, we suggest labeling the axis ticks with the *numbers* (not *rates*) of false and true positives to show the amount of underlying data (as in Figure 5).

Having visualised the discrete lattice of points achievable in ROC space, we now show how to visualise continuous performance metrics in that space, and beyond

5. The geometry of performance metrics

There are many confusion matrix performance metrics to make sense of (Appendix A). It is not obvious how these metrics behave when the four entries of the confusion matrix are not zero, which is typically of interest when people want to rank or compare the performance of different classifiers.

One way to better understand these metrics is to plot their *contours*, the isolines along which they take a particular value k. We have used this to illustrate the curved contours of the MCC above (Figure 2) and we go further by deriving algebraic expressions for the exact contours of many popular metrics—both prevalence-dependent (Appendix B) and independent (Appendix C). We provide interactive visualisations of these contours (Appendix D.3) and animations that show how contours change with prevalence for Accuracy, Balanced Accuracy, F_1 score and Matthews Correlation Coefficient (Appendix D.4). The advantage of these algebraic expressions is that they can show the exact performance metric contours for any size of confusion matrix.

The algebraic expressions of these contours also allow us to appreciate the geometry of performance metrics *beyond* ROC space. For example, MCC can be understood as a set of elliptical contours, whose eccentricity depends on class imbalance (Figure B.2). These contours show how performance metrics vary (and also their symmetries (Brzezinski et al., 2018; Luque et al., 2019)) in a way that our visual system can readily apprehend. To complete the picture, we need to model how confusion matrices are likely to vary…

6. Modeling uncertainty in confusion matrices

Using the notation in Figure 1, suppose a classifier is presented with p_1 positive examples to classify and that it gets a_1 of these correct (true positives) and the remainder $c_1 = p_1 - a_1$ incorrect (false negatives). On the basis of these observations, what do we believe is the probability (θ_a) that this classifier correctly identifies positive examples? A frequentist approach would estimate that $\theta_a = a_1/p_1 = \text{TPR}$ is the empirical true positive rate of the classifier and that the distribution of a future correct classifications with p new examples is

$$a|p, \theta_a \sim \text{Binomial}(p, \theta_a) \tag{2}$$

so that

$$P(a|p, \theta_a) = \binom{p}{a} \theta_a^a (1 - \theta_a)^{p-a}. \tag{3}$$

A Bayesian approach (Navarro & Perfors, 2010; Murphy, 2012; Agresti, 2013) allows us to express our prior uncertainty about a classifier's true positive rate—this is particularly important when we have small amounts of data. We can assign a prior distribution to θ_a, and it is mathematically convenient to do that with a beta distribution:

$$\theta_a|u,v \;\sim\; \mathrm{Beta}(u,v).$$

As this prior is conjugate to the binomial distribution, after observing a_1 true positive (and c_1 false negative) classifications of p_1 examples, the posterior distribution of θ_a remains a beta distribution:

$$\theta_a|a_1,p_1,u,v \;\sim\; \mathrm{Beta}(u+a_1,v+c_1)$$

and the posterior predictive distribution of seeing a correct classifications of p further examples is

$$\mathrm{P}(a \mid p,a_1,c_1,u,v) = \binom{p}{a}\frac{\mathrm{Beta}(u+a_1+a,v+c_1+c)}{\mathrm{Beta}(u+a_1,v+c_1)}. \tag{4}$$

a is distributed according to a *beta-binomial* distribution:

$$a \mid p,a_1,c_1,u,v \;\sim\; \mathrm{BB}(a,p,u+a_1,v+c_1). \tag{5}$$

Equations (2) and (3) give us the basis of a *binomial* model for uncertainty in confusion matrices and Equations (4) and (5) give us the basis of a *beta-binomial* model. Using the same logic, we can express models for the number of true negatives (d) returned by classifier presented with n negative examples when the probability of the classifier correctly identifying these negative examples is θ_d.

Using beta-binomial models for true positives and true negatives demands that we declare our prior uncertainty about θ_a and θ_d. For demonstration, and in the absence of other relevant information, we choose the uninformative uniform priors: $\theta_a \;\sim\; \mathrm{Beta}(1,1)$ and $\theta_d \;\sim\; \mathrm{Beta}(1,1)$.

A common scenario in classifier development and evaluation is to present a classifier with a test set of p_1 positive and n_1 negative examples. In this situation, we can assume that the probabilities of correctly classifying a further a positive and d negative examples are statistically independent, so that

$$P(a,d|a_1,p_1,d_1,n_1) = P(a|a_1,p_1)\cdot P(d|,d_1,n_1) \tag{6}$$

for the binomial model, where d_1 is the number of true negative classifications observed. The joint distribution of true positives and negatives under a beta-binomial model can be factored similarly, enabling confusion matrix uncertainty to be visualised. We note that there are potential situations where this independence assumption may not hold, for example in cytopathology analysis, if apparently normal cells are extracted from a sample obtained from a positive patient in addition to abnormal cells (Burger et al., 1981).

7. Visualising confusion matrix variation

The ideas set out in the previous section are the same as those set out by Tötsch & Hoffmann (2021) who proceed to use *simulation* to estimate the posterior distribution of various performance metrics under beta-binomial models of true positives and negatives. However, as Tötsch & Hoffmann point out, *"the posterior distribution can be derived analytically. There is no need for Markov chain Monte Carlo sampling"*. So, rather than using time-consuming simulation, we have used the pmfs of Equations (3) and (4) to develop an interactive visualisation of the impact of uncertainty on confusion matrices and their performance metrics in ROC space and in Precision-Recall space (Appendix D.5).

Figure 3 shows two screenshots from this visualisation. Unlike the histograms of samples from the posterior predictive distributions of true positive and true negative rates in Tötsch & Hoffmann (2021), Figure 3 exposes the underlying discreteness of ROC space. Note that the pmfs of the beta-binomial model are broader than those of the binomial, reflecting the additional uncertainty in true and false positive rates embodied in this Bayesian model. Appendix E uses a real-world example to illustrate the differences between these models when data are scarce (Rodrigues et al., 2013).

8. Visualising performance metric variation

We can now visualise the distribution of a given performance metric by summing the probability masses that lie along each contour of the performance metric in ROC space. The geometry of performance metric contours in conjunction with the layout of the $(n+1)\times(p+1)$ possible points in ROC space determines which probability masses are summed together (as illustrated in Appendix F).

Figure 4 shows the posterior predictive pmfs of MCC, BA and F_1 values, given observations of $\left[\begin{smallmatrix}16 & 8\\ 4 & 32\end{smallmatrix}\right]$. Note the spread of these pmfs about the *maximum a posteriori* (MAP) value of each performance metric. This corresponds to the distributions of true and false positive rates shown in Figure 3 and puts performance comparisons into perspective: MCC, BA and F_1 each show substantial uncertainty about their observed values.

Note also the shape of the pmfs in Figure 4 and that both BA and F_1 have a few points lying above that bell-curve. These occur where several ROC points lie on the same performance metric contour (e.g., $\mathrm{F}_1 = \frac{4}{10}$ and $\mathrm{F}_1 = \frac{2}{3}$, Figure F.1). The numbers of points on each contour are shown in the bottom row of Figure 4. Adding one extra negative example markedly changes the confluence of ROC points and performance metric contours for MCC and BA, less so for F_1 (Figure F.2). The discrete nature of confusion matrices can lead to jumps in performance metrics pmfs,

even though those metrics are smooth continuous functions.

The more labelled data evaluated by a classifier, the more certain we can be about its true and false positive rates. Figure 5 shows this with the posterior predictive pmfs of confusion matrices of increasing totals but constant true and false positive rates: more data yields more precise estimates. Roughly speaking, under the binomial and beta-binomial models of uncertainty, the standard deviation of the true and false positive rate pmfs will be proportional to $\frac{1}{\sqrt{N}}$.

Our interactive visualisations aim to foster understanding of how uncertainty in the discrete domain of confusion matrices gives rise to distributions of continuous performance metrics. Armed with that understanding, practitioners can then use more compact summaries (e.g., box plots, violin plots) to report estimated performance metric distributions. The main point to note is that there can be significant uncertainty in performance metrics, uncertainty which depends on the amount of data used in evaluating empirical performance, not the performance metrics themselves. To reduce that uncertainty requires more data that is representative of the cases the classifier will see in production.

9. Discussion

Our goal is to ensure that practitioners can understand, visualise and put into perspective the magnitude and nature of uncertainty in classifier performance estimates to inform more meaningful discussion of the strengths and limitations of specific classification systems. This work has its limitations. We appreciate that performance evaluation and model selection can involve a host of competing considerations beyond predictive performance, such as model transparency and fairness to different groups affected by model predictions; these are not within the scope of our work here. Nor do we address the common assumption that evaluation data is representative, i.e., that future data are expected to come from the same process as past data and have roughly the same range of values (McElreath, 2016).

Related to that last point is the question of how data diversity and representativeness may manifest in confusion matrix uncertainty. Obviously, the more representative data we can use for evaluation the more accurate and precise our classifier performance estimates (Figure 5). But whether more diverse data results in higher or lower performance depends on how separable the two classes are by the classification system at hand. Here both data quality and quantity are important: creating synthetic data points through perturbing or interpolating real data (Chawla et al., 2002)will certainly increase the quantity of of data, it won't necessarily increase its quality, i.e., how well it represents future examples.

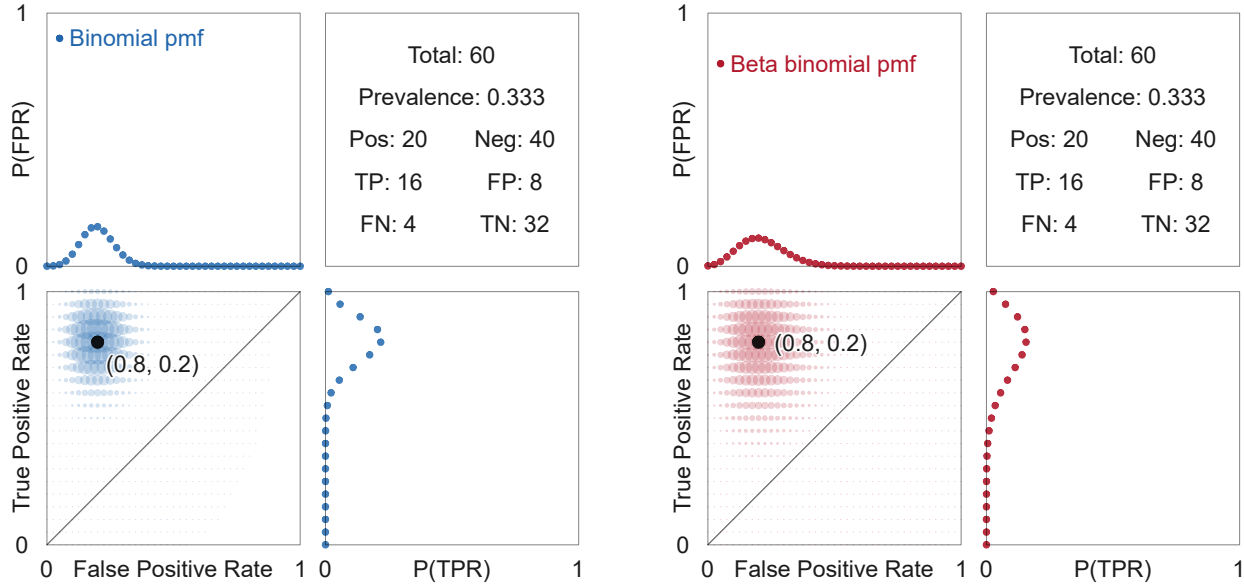
The Bayesian approach we have described combines prior beliefs with observed data to form posterior distributions of plausible classifier performance metrics. A frequentist approach would hold that uncertainty is a consequence of sampling variation, leading to a focus on the *sampling distribution* of classifier performance metrics. Murphy (2012) contrasts and critiques these two paradigms, mentioning the bootstrap as a means to approximate the sampling distribution and confidence intervals as a way to characterise its spread. The Bayesian approach however is ideal for the situation we want to highlight: where a lack of data creates substantial uncertainty about the future performance of a classifier. Frequentist approaches are challenged by small data sets, as illustrated in Appendix E. Furthermore the Bayesian approach can yield a the functional form of a posterior (e.g., beta-binomial) enabling us to visualise its pmf precisely, whereas a frequentist approach would involve empirical approximations to the sampling distribution.

We began by mentioning the strong incentives to build classification systems that show outstanding performance and we acknowledge the frequent use of evaluation metrics selecting the "best" model. Unfortunately, uncertainty in performance metrics does not make this task any easier. Rather, it forces us to confront the limitations arising from restricted data, and the need to balance multiple competing considerations, including issues beyond predictive performance such as system fairness and transparency.

While we clearly don't believe that there is one "best" performance metric, one could argue that the best classifier minimises the expected *costs* (or maximizes the *benefits*) of its future decisions (Section B.5). This is a conceptually appealing way to describe classifier performance with a single number but, in practice, adds further uncertainty into the evaluation process, i.e., uncertainty about the costs or benefits of different decisions. Eliciting and quantifying these costs is difficult, subjective and generally ignored.

Much attention has been given to the "problem" of class imbalance (Luque et al., 2019; Mullick et al., 2020; Lovell et al., 2021). We use quotes to emphasise that class imbalance is mainly problematic to those trying to build automated decision making systems—the real problem of rare but highly adverse cases (e.g., life threatening disease) is that they occur at all; we would not want them to happen more frequently. Also, class imbalance is a necessary consequence of multinomial classification: with C possible classes and N examples, at least one class will have equal or fewer than N/C examples.

There are two types of confusion matrix performance metric: *prevalence-independent* (or *balanced*) metrics, whose values depend only on rates of true positives and negatives (Appendix C); and *prevalence-dependent* metrics whose values depend on these rates and prevalence (Appendix B). (Note that balanced versions of prevalence dependent metrics can be derived—see Luque et al. (2019) and Section C.12 .)



(a) Posterior predictive pmfs: binomial model.

(b) Posterior predictive pmfs: beta-binomial model

Figure 3: Joint and marginal posterior predictive pmfs under different models of uncertainty for the confusion matrix $\begin{bmatrix} 16 & 8 \\ 4 & 32 \end{bmatrix}$. These are the distributions we would expect to see if the classifier that produced that confusion matrix was given a further 20 positive and 40 negative examples to classify. Top left panels show the pmfs of the false positive rate; bottom right panels show the pmfs of the true positive rate; bottom left panels show the joint pmfs of true and false positive rates in ROC space. Point areas in the joint distribution plots are proportional to the probability masses they are centred on. These are screenshots from our interactive visualisation (see Appendix D.5).

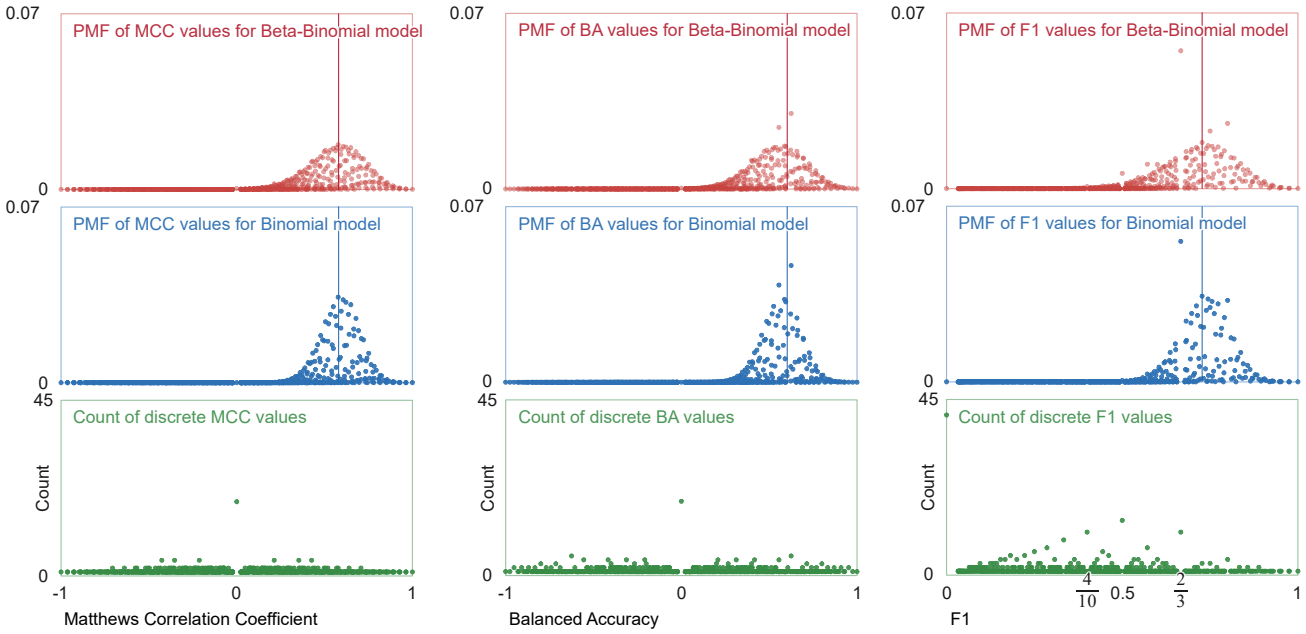


Figure 4: There is potential for substantial variation about the performance metric values of the observed confusion matrix as shown by these posterior predictive pmfs of MCC, BA, F_1 (left to right) under beta-binomial (red) and binomial models (blue) of uncertainty, given the observed confusion matrix used in Figure 3. Vertical lines show the MAP performance metric values. The green points show the number of times each value of a particular performance metric is observed: these counts are proportional to the prior pmfs of performance metric values. F_1 values of $\frac{4}{10}$ and $\frac{2}{3}$ are observed 11 times (Figure F.1). These are screenshots from our interactive visualisation (Appendix D.5).

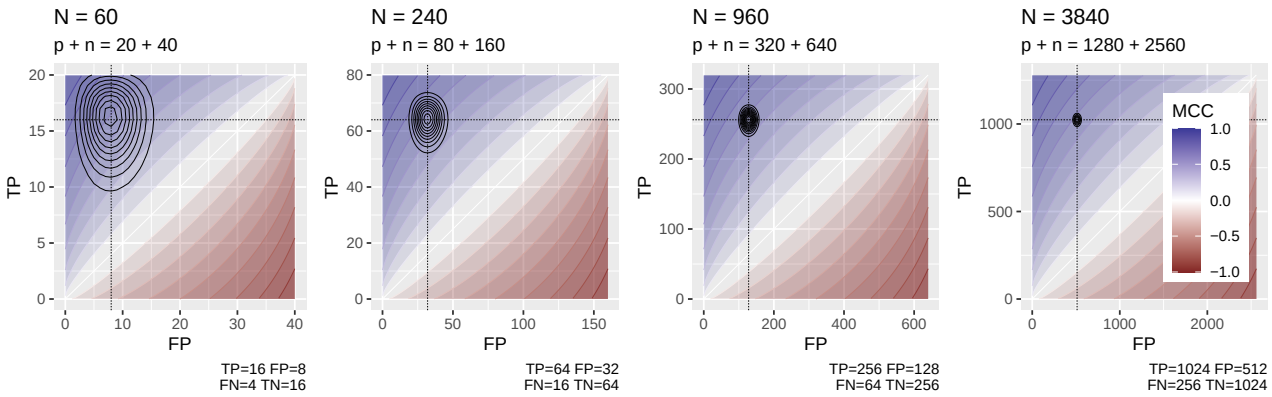


Figure 5: Reducing uncertainty in performance metrics requires more data to increase the precision of the predictive distribution of confusion matrices. These contour plots show the posterior predictive pmfs (under a beta-binomial model of uncertainty) after observing confusion matrices of increasing size but with the same false and true positive rates (0.2, 0.8). As confusion matrix totals (N) increase by a factor of 4, the heights and widths of the contours decrease by a factor of $\frac{1}{2}$.

This has led to discussion about which metrics are best for scenarios where classes are balanced or imbalanced. However, neither balanced nor prevalence-dependent performance metrics reduce the uncertainty arising from finite amounts of evaluation data—only additional data can do that. We note that augmenting examples of rare classes to assist with learning and performance evaluation will distort class prior probabilities, and the trained classifier's outputs will have to be adjusted to provide accurate predictions for real-world class abundance (Saerens et al., 2002).

The ideas we have presented for binary confusion matrices could be extended to multinomial classification by treating a $C \times C$ multinomial confusion matrix as a set of C binary classification problems. This *one-versus-all* strategy is common in multinomial classification and has been used to provide compact, informative visualisations of confusion matrices in terms of their prior and posterior odds (Lovell et al., 2021). The models of uncertainty we have presented could be applied to each class versus all others, yielding posterior predictive pmfs of various performance metrics. One advantage of this approach would be to reveal the degree of uncertainty associated with each class so that those developing classification systems could consider where best to direct their attention in making improvements. An alternate approach would be to change the probabilistic model of confusion matrix distributions from binomial (or beta-binomial) to multinomial (or Dirichlet-multinomial) (Murphy, 2012). But with $C > 2$ classes, we can no longer visualise the space of $C \times C$ confusion matrices in 3 dimensions, nor their performance metrics. Handling uncertainty in multinomial or multi-label (Görtler et al., 2022) classification scenarios is certainly an open challenge.

10. Conclusion

Publications advocating specific performance metrics have motivated us to study the (continuous) geometry of performance metrics and the (discrete) geometry of the space of confusion matrices they are applied to. Through this, we have gained a clearer understanding of the cause and effect of uncertainty in performance metrics: the primary cause is uncertainty about the confusion matrices that will be produced by a trained classifier; this depends on the numbers of actual positive and negative examples we have observed the classifier determine. Using binomial and (more conservative) beta-binomial models of uncertainty, we can calculate the exact pmfs of the predictive distribution of confusion matrices, given the classifications we have observed.

Using the contours of various performance metrics, we have demonstrated how the posterior predictive pmfs of confusion matrices can be transformed into posterior predictive pmfs of different performance metrics. We have provided a range of static and interactive visualisations for researchers and practitioners to explore uncertainty in confusion matrices and various performance metrics derived from them. These contributions aim to put performance metrics and their uncertainty into perspective, specifically, when observations of positive or negative classes are few, uncertainty in performance metrics can easily eclipse differences in performance between classifiers. Arguments about which classifier performs best, or which performance metric is best, need to take this uncertainty into account, and the visualisations we provide enable researchers to do that.

Our work also provides a useful perspective on performance evaluation where classes are imbalanced—a common scenario in binary decision making and a necessary situation in

multinomial classification. Some metrics depend on class imbalance; others do not; and "balanced" metrics can be created from prevalence-dependent ones. However, the fundamental issue in performance evaluation is not so much the *relative* abundance of different classes as their *absolute* abundance: our estimates of a classifier's ability to correctly detect a class will be highly uncertain when there are few instances that class, regardless of balance. Performance metrics can't address this: ore data is needed and, until it arrives, we must acknowledge the uncertainty in our findings.

Finally, and with a view to the ethical dimensions of this work, we acknowledge that classifier performance evaluation goes far beyond purely quantitative metrics. It is heartening to see the breadth of issues in performance evaluation, benchmarking and datasets gaining more attention. Still, quantitative metrics will always play a prominent role in considering the strengths and limitations of classification systems. We hope that this visualisation of confusion matrix performance metrics and their uncertainties will inform their use in practice.

References

Agresti, A. *Categorical Data Analysis*. Number 792 in Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 3rd ed edition, 2013. ISBN 978-0-470-46363-5.

Balayla, J. Prevalence threshold (Φe) and the geometry of screening curves. *PLOS ONE*, 15(10):e0240215, 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0240215.

Breiman, L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, August 2001. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1009213726.

Brzezinski, D., Stefanowski, J., Susmaga, R., and Szczęch, I. Visual-based analysis of classification measures and their properties for class imbalanced problems. *Information Sciences*, 462:242–261, September 2018. ISSN 0020-0255. doi: 10.1016/j.ins.2018.06.020.

Burger, G., Jutting, U., and Rodenacker, K. Changes in benign cell populations in cases of cervical cancer and its precursors. *Analytical and Quantitative Cytology*, 3(4): 261–271, December 1981. ISSN 0190-0471.

Caelen, O. A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, 81(3):429–450, December 2017. ISSN 1573-7470. doi: 10.1007/s10472-017-9564-8.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. ISSN 1076-9757. doi: 10.1613/jair.953.

Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Mining*, 10(1):35, December 2017. ISSN 1756-0381. doi: 10.1186/s13040-017-0155-3.

Chicco, D. and Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21 (1):6, January 2020. ISSN 1471-2164. doi: 10.1186/s12864-019-6413-7.

Chicco, D. and Jurman, G. An Invitation to Greater Use of Matthews Correlation Coefficient in Robotics and Artificial Intelligence. *Frontiers in Robotics and AI*, 9, 2022. ISSN 2296-9144. URL `https://www.frontiersin.org/article/10.3389/frobt.2022.876814`.

Chicco, D., Starovoitov, V., and Jurman, G. The Benefits of the Matthews Correlation Coefficient (MCC) Over the Diagnostic Odds Ratio (DOR) in Binary Classification Assessment. *IEEE Access*, 9:47112–47124, 2021a. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3068614.

Chicco, D., Tötsch, N., and Jurman, G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, 14(1):13, February 2021b. ISSN 1756-0381. doi: 10.1186/s13040-021-00244-z.

Davis, J. and Goadrich, M. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 233–240, New York, NY, USA, June 2006. Association for Computing Machinery. ISBN 978-1-59593-383-6. doi: 10.1145/1143844.1143874.

Desmos, Inc. Desmos — Let's learn together., n.d. URL `https://www.desmos.com/`.

Ferri, C., Hernández-Orallo, J., and Modroiu, R. An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1):27–38, 2009. ISSN 0167-8655. doi: 10.1016/j.patrec.2008.08.010.

Flach, P. A. The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pp. 194–201, Washington, DC, USA, August 2003. AAAI Press. ISBN 978-1-57735-189-4.

Fowlkes, E. B. and Mallows, C. L. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983. ISSN 0162-1459. doi: 10.2307/2288117.

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., and Bossuyt, P. M. M. The diagnostic odds ratio: A single indicator of test performance. *Journal of Clinical Epidemiology*, 56(11):1129–1135, November 2003. ISSN 0895-4356. doi: 10.1016/S0895-4356(03)00177-X.

Görtler, J., Hohman, F., Moritz, D., Wongsuphasawat, K., Ren, D., Nair, R., Kirchner, M., and Patel, K. Neo: Generalizing Confusion Matrix Visualization to Hierarchical and Multi-Output Labels. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pp. 1–13, New York, NY, USA, April 2022. Association for Computing Machinery. ISBN 978-1-4503-9157-3. doi: 10.1145/3491102.3501823.

Gösgens, M., Zhiyanov, A., Tikhonov, A., and Prokhorenkova, L. Good Classification Measures and How to Find Them. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17136–17147. Curran Associates, Inc., 2021.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, May 2017. ISSN 0957-4174. doi: 10.1016/j.eswa.2016.12.035.

Hand, D. J. Classifier Technology and the Illusion of Progress. *Statistical Science*, 21(1):1–14, February 2006. ISSN 0883-4237, 2168-8745. doi: 10.1214/088342306000000060.

Lovell, D., McCarron, B., Langfield, B., Tran, K., and Bradley, A. P. Taking the Confusion Out of Multinomial Confusion Matrices and Imbalanced Classes. In Xu, Y., Wang, R., Lord, A., Boo, Y. L., Nayak, R., Zhao, Y., and Williams, G. (eds.), *Data Mining*, Communications in Computer and Information Science, pp. 16–30, Singapore, 2021. Springer. ISBN 9789811685316. doi: 10.1007/978-981-16-8531-6_2.

Luque, A., Carrasco, A., Martín, A., and de las Heras, A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231, July 2019. ISSN 0031-3203. doi: 10.1016/j.patcog.2019.02.023.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., Feldmann, C., Frangi, A. F., Full, P. M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B. A., März, K., Maier, O., Maier-Hein, K., Menze, B. H., Müller, H., Neher, P. F., Niessen, W., Rajpoot, N., Sharp, G. C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A. A., van der Sommen, F., Wang, C.-W., Weber, M.-A., Zheng, G., Jannin, P., and Kopp-Schneider, A. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 9(1):5217, December 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07619-7.

McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Number 122 in Chapman & Hall/CRC Texts in Statistical Science Series. CRC Press/Taylor & Francis Group, Boca Raton, 2016. ISBN 978-1-4822-5344-3.

Mullick, S. S., Datta, S., Dhekane, S. G., and Das, S. Appropriateness of performance indices for imbalanced data classification: An analysis. *Pattern Recognition*, 102:107197, June 2020. ISSN 0031-3203. doi: 10.1016/j.patcog.2020.107197.

Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, MA, 2012. ISBN 978-0-262-01802-9.

Navarro, D. and Perfors, A. An introduction to the Beta-Binomial model, 2010. URL `https://compcogsci-3016.djnavarro.net/technote_betabinomial.pdf`.

Powers, D. M. W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology*, 2(1):37–63, 2011. URL `http://arxiv.org/abs/2010.16061`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

Rodrigues, N. V. S., Cardoso, E. M., Andrade, M. V. O., Donnici, C. L., and Sena, M. M. Analysis of seized cocaine samples by using chemometric methods and FTIR spectroscopy. *Journal of the Brazilian Chemical Society*, 24:507–517, March 2013. ISSN 0103-5053, 1678-4790. doi: 10.5935/0103-5053.20130066.

Rudin, C. and Radin, J. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. *Harvard Data Science Review*, 1(2), 2019. doi: 10.1162/99608f92.5a8a3a3d.

Saerens, M., Latinne, P., and Decaestecker, C. Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. *Neural Computation*, 14(1):21–41, January 2002. ISSN 0899-7667, 1530-888X. doi: 10.1162/089976602753284446.

Saito, T. and Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3):e0118432, March 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0118432.

Sievert, C. *Interactive Web-Based Data Visualization with R, Plotly, and Shiny*. Chapman and Hall/CRC, 2020. ISBN 978-1-138-33145-7. URL `https://plotly-r.com`.

Sokolova, M. and Lapalme, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, July 2009. ISSN 0306-4573. doi: 10.1016/j.ipm.2009.03.002.

Thomas, R. and Uminsky, D. The Problem with Metrics is a Fundamental Problem for AI. *arXiv:2002.08512 [cs]*, February 2020. URL `http://arxiv.org/abs/2002.08512`.

Tötsch, N. and Hoffmann, D. Classifier uncertainty: Evidence, potential impact, and probabilistic treatment. *PeerJ Computer Science*, 7:e398, March 2021. ISSN 2376-5992. doi: 10.7717/peerj-cs.398.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Zhu, Q. On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. *Pattern Recognition Letters*, 136:71–80, August 2020. ISSN 0167-8655. doi: 10.1016/j.patrec.2020.03.030.

A. Performance metrics and other definitions

Many sums, products, ratios and functions of the four elements of a binary confusion matrix have been defined. Unfortunately, the notation varies between authors as does the row and column layout of the confusion matrix itself. We define performance metrics using the notation and layout of the confusion matrix in Figure 1 in the main paper:

Total $N = p + n$ *predicted class*	*actual class* positive $p = a + c$	negative $n = b + d$
positive $\hat{p} = a + b$	**TP** True Positives (a)	**FP** False Positives (b)
negative $\hat{n} = c + d$	**FN** False Negatives (c)	**TN** True Negatives (d)

The following equations define performance metrics and other quantities used in this paper. We use an asterisk (*) to indicate quantities that depend on *prevalence* (Eq. (12)).

Here are the definitions of the row, column and overall totals of the confusion matrix:

$$\text{Total} \qquad N = \mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN} \tag{7}$$

$$\text{Condition Positive} \qquad p = \mathrm{TP} + \mathrm{FN} \tag{8}$$

$$\text{Condition Negative} \qquad n = \mathrm{FP} + \mathrm{TN} \tag{9}$$

$$\text{Predicted Positive}^* \qquad \hat{p} = \mathrm{TP} + \mathrm{FP} \tag{10}$$

$$\text{Predicted Negative}^* \qquad \hat{n} = \mathrm{FN} + \mathrm{TN}. \tag{11}$$

Prevalence refers to the proportion of positive cases in a dataset:

$$\text{Prevalence}^* \qquad \mathrm{Prev} = \frac{p}{N}. \tag{12}$$

True and False Positive rates form the Positive Likelihood Ratio:

$$\text{True Positive Rate} \qquad \mathrm{TPR} = \frac{\mathrm{TP}}{p} = 1 - \mathrm{FNR} \qquad \text{Sensitivity, Recall} \tag{13}$$

$$\text{False Positive Rate} \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{n} = 1 - \mathrm{TNR} \tag{14}$$

$$\text{Positive Likelihood Ratio} \qquad \mathrm{LR}_+ = \frac{\mathrm{TPR}}{\mathrm{FPR}} \tag{15}$$

while True and False Negative rates form the Negative Likelihood Ratio:

$$\text{True Negative Rate} \qquad \mathrm{TNR} = \frac{\mathrm{TN}}{n} \qquad \text{Specificity} \tag{16}$$

$$\text{False Negative Rate} \qquad \mathrm{FNR} = \frac{\mathrm{FN}}{p} \tag{17}$$

$$\text{Negative Likelihood Ratio} \qquad \mathrm{LR}_- = \frac{\mathrm{FNR}}{\mathrm{TNR}} \tag{18}$$

and these Likelihood Ratios form the Diagnostic Odds Ratio:

$$\text{Diagnostic Odds Ratio} \qquad \mathrm{DOR} = \frac{\mathrm{LR}_+}{\mathrm{LR}_-} = \frac{\mathrm{TP} \cdot \mathrm{TN}}{\mathrm{FP} \cdot \mathrm{FN}}. \tag{19}$$

To help visualise LR_+, LR_- and DOR in comparison to other performance metrics, we introduce the following scaled versions of their logarithms

$$\text{scaled log } \mathrm{LR}_+ \qquad \mathrm{slLR}_+ = \frac{1}{\log(n(p-1)/p)} \log\left(\frac{\mathrm{TP}}{p}\frac{n}{\mathrm{FP}}\right) \qquad \text{Section C.4} \tag{20}$$

$$\text{scaled log } \mathrm{LR}_- \qquad \mathrm{slLR}_- = \frac{1}{\log(pn/(n-1))} \log\left(\frac{\mathrm{FN}}{p}\frac{n}{TN}\right) \qquad \text{Section C.5} \tag{21}$$

$$\text{scaled log DOR} \qquad \mathrm{slDOR} = \frac{1}{\log(p-1)(n-1)} \log\left(\frac{\mathrm{TP}\cdot\mathrm{TN}}{\mathrm{FP}\cdot\mathrm{FN}}\right) \qquad \text{Section C.8} \tag{22}$$

True Positive and True Negative rates are the basis of the following prevalence-independent performance metrics:

$$\text{Balanced Accuracy} \qquad \mathrm{BA} = \frac{\mathrm{TPR}+\mathrm{TNR}}{2} = \frac{\mathrm{BM}+1}{2} \tag{23}$$

$$\text{Bookmaker Informedness} \qquad \mathrm{BM} = \mathrm{TPR}+\mathrm{TNR}-1 \qquad \text{Youden's } J\text{, Delta P} \tag{24}$$

$$\text{Geometric Mean} \qquad \mathrm{GM} = \sqrt{\mathrm{TPR}\cdot\mathrm{TNR}} \tag{25}$$

$$\text{Prevalence Threshold} \qquad \mathrm{PT} = \frac{\sqrt{\mathrm{FPR}}}{\sqrt{\mathrm{TPR}}+\sqrt{\mathrm{FPR}}} \qquad \text{See (Balayla, 2020)} \tag{26}$$

Next come ratios that relate to a classifier's predictions. These depend on prevalence, but (Luque et al., 2019) have proposed prevalence-independent ("balanced") versions:

$$\text{Positive Predictive Value}^* \qquad \mathrm{PPV} = \frac{\mathrm{TP}}{\hat{p}} \qquad \text{Precision} \tag{27}$$

$$\text{Balanced PPV} \qquad \mathrm{PPV}_{bal} = \frac{\mathrm{TPR}}{1+\mathrm{TPR}-\mathrm{TNR}} \qquad \text{See (Luque et al., 2019)} \tag{28}$$

$$\text{Negative Predictive Value}^* \qquad \mathrm{NPV} = \frac{\mathrm{TN}}{\hat{n}} \qquad \text{See (Powers, 2011)} \tag{29}$$

$$\text{Balanced NPV} \qquad \mathrm{NPV}_{bal} = \frac{\mathrm{TNR}}{1+\mathrm{TNR}-\mathrm{TPR}}. \qquad \text{See (Luque et al., 2019)} \tag{30}$$

and these predictive values form Markedness and its balanced version:

$$\text{Markedness}^* \qquad \mathrm{MK} = \mathrm{PPV}+\mathrm{NPV}-1 \tag{31}$$

$$\text{Balanced Markedness} \qquad \mathrm{MK}_{bal} = \mathrm{PPV}_{bal}+\mathrm{NPV}_{bal}-1. \qquad \text{See (Luque et al., 2019)} \tag{32}$$

Accuracy refers to the proportion of correctly classified examples and is prevalence-dependent, unlike BA and BM:

$$\text{Accuracy}^* \qquad \mathrm{Acc} = \frac{\mathrm{TP}+\mathrm{TN}}{N}. \tag{33}$$

Like Accuracy, both F_1 and Threat Score involve ratios of sums of confusion matrix elements:

$$\mathrm{F}_1{}^* \qquad \mathrm{F}_1 = \frac{2\mathrm{PPV}\cdot\mathrm{TPR}}{\mathrm{PPV}+\mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \qquad \text{Sørensen–Dice coefficient} \tag{34}$$

$$\text{Balanced } \mathrm{F}_1 \qquad \mathrm{F}_{1bal} = \frac{2\mathrm{TPR}}{2+\mathrm{TPR}-\mathrm{TNR}} \qquad \text{See (Luque et al., 2019)} \tag{35}$$

$$\text{Threat Score}^* \qquad \mathrm{TS} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}} \qquad \text{Jaccard index, Critical Success index} \tag{36}$$

$$\text{Balanced Threat Score} \qquad \mathrm{TS}_{bal} = \frac{\mathrm{TPR}}{2-\mathrm{TNR}}. \qquad \text{Section C.13} \tag{37}$$

We define Decision Benefits as a weighted sum of confusion matrix elements:

$$\text{Decision Benefits}^* \qquad \text{DB} = \beta_a \cdot \text{TP} + \beta_b \cdot \text{FP} + \beta_c \cdot \text{FN} + \beta_d \cdot \text{TN}. \qquad \text{Section B.5} \tag{38}$$

The remaining metrics combine the four elements of the confusion matrix in more complex ways:

$$\text{Matthews Correlation Coefficient}^* \qquad \text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{\hat{p} \cdot p \cdot n \cdot \hat{n}}} \qquad \text{phi coefficient} \tag{39}$$

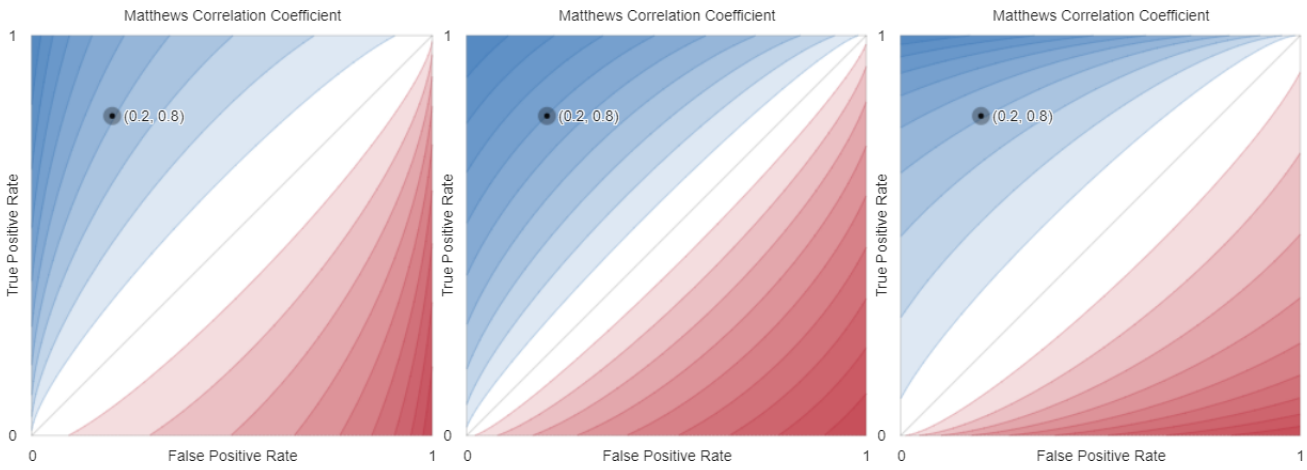
$$= \text{sgn}(\text{BM}) \sqrt{\text{BM} \cdot \text{MK}} \qquad \text{See (Powers, 2011)} \tag{40}$$

$$\text{Balanced MCC} \qquad \text{MCC}_{bal} = \frac{\text{BM}}{\sqrt{1 - (\text{TPR} - \text{TNR})^2}} \qquad \text{See (Luque et al., 2019)} \tag{41}$$

$$\text{Fowlkes-Mallows Index}^* \qquad \text{FM} = \sqrt{\text{PPV} \cdot \text{TPR}} \qquad \text{See (Fowlkes \& Mallows, 1983)} \tag{42}$$

$$\text{Balanced FM} \qquad \text{FM}_{bal} = \frac{\text{TPR}}{\sqrt{1 + \text{TPR} - \text{TNR}}} \qquad \text{Section C.12} \tag{43}$$

$$\text{Cohen's kappa}^* \qquad \kappa = \frac{2(\text{TP} \cdot \text{TN} - \text{FN} \cdot \text{FP})}{\hat{p} \cdot n + p \cdot \hat{n}}. \qquad \text{Section C.16} \tag{44}$$



(a) MCC with Prev = 0.1 (b) MCC with Prev = 0.5 (c) MCC with Prev = 0.9

Figure B.1: Contours of the Matthews Correlation Coefficient in the ROC space.

B. Performance metric contours that depend on prevalence

B.1. Matthews Correlation Coefficient

Using the notation of Figure 1, the Matthews Correlation Coefficient (Eq. (39)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned}\mathrm{MCC}(a,b,c,d) &= \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \\ &= \frac{ad-(n-d)(p-a)}{\sqrt{(a+n-d)pn(p-a+d)}}.\end{aligned}$$

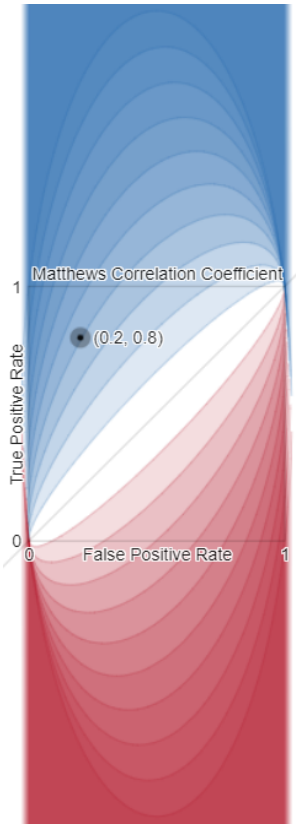
For given numbers of positives (p) and negatives (n), this performance metric achieves a value of $-1 \leq k \leq 1$ along the contour lines with

$$a(k,p,n,d) = \begin{cases} \dfrac{1}{2(k^2p+n)}\left(+\sqrt{\dfrac{k^2p(n+p)^2(4d(n-d)+k^2np)}{n}} + 2dp(k^2-1)+k^2p(p-n)+2np\right) & k \geq 0 \\ \dfrac{1}{2(k^2p+n)}\left(-\sqrt{\dfrac{k^2p(n+p)^2(4d(n-d)+k^2np)}{n}} + 2dp(k^2-1)+k^2p(p-n)+2np\right) & k < 0 \end{cases} \tag{45}$$

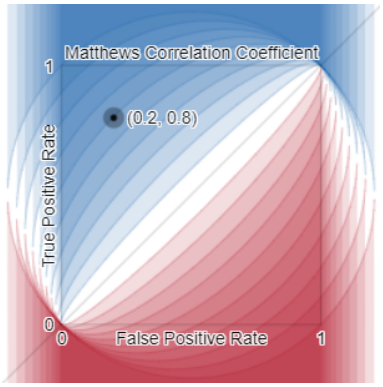
or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k,p,n,\delta) = \begin{cases} \dfrac{1}{2p(k^2p+n)}\left(+\sqrt{\dfrac{k^2p(n+p)^2(4n^2\delta(n-\delta)+k^2np)}{n}} + 2n\delta p(k^2-1)+k^2p(p-n)+2np\right) & k \geq 0 \\ \dfrac{1}{2p(k^2p+n)}\left(-\sqrt{\dfrac{k^2p(n+p)^2(4n^2\delta(n-\delta)+k^2np)}{n}} + 2n\delta p(k^2-1)+k^2p(p-n)+2np\right) & k < 0 \end{cases} \tag{46}$$

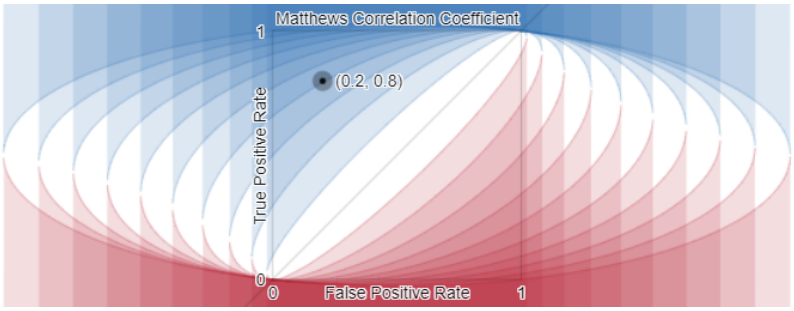
in which case, all contours where $k < 0$ intersect at $(\alpha, \delta) = (0, 1)$ and all contours where $k > 0$ intersect at $(\alpha, \delta) = (1, 0)$, as is the case for Markedness. These intersections become apparent when we visualise the contours of the Matthews Correlation Coefficient within and beyond the ROC space (Figure B.2) which reveals that the contours describe a series of concentric ellipses whose eccentricity depends on prevalence.



(a) MCC with Prev=0.1

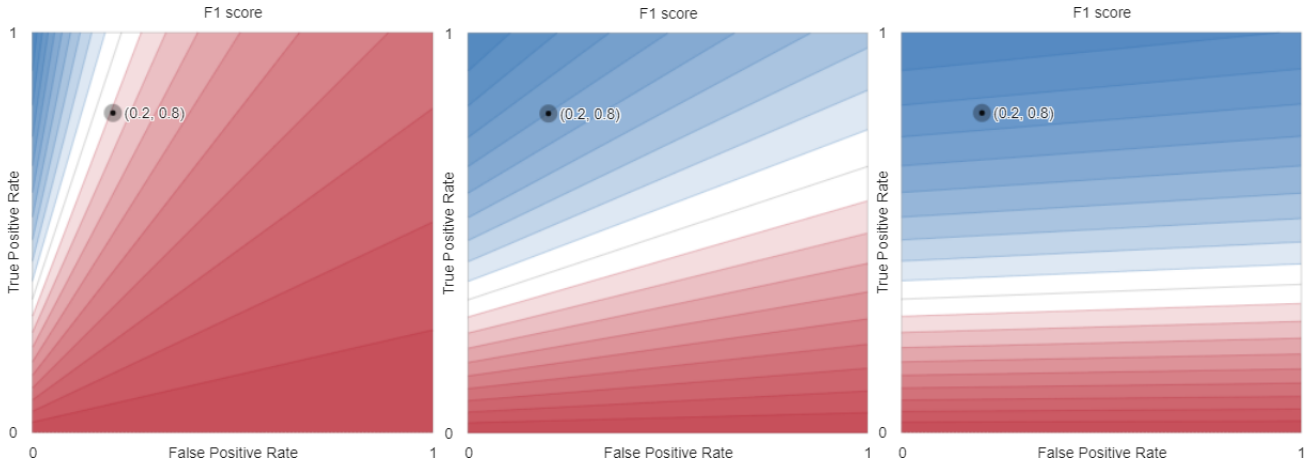


(b) MCC with Prev=0.5



(c) MCC with Prev=0.9

Figure B.2: Contours of the Matthews Correlation Coefficient within and beyond the ROC space describe a series of concentric ellipses whose eccentricity depends on prevalence and which intersect where $(\mathrm{FPR}, \mathrm{TPR}) = (0, 0)$ and $(\mathrm{FPR}, \mathrm{TPR}) = (1, 1)$



(a) F_1 score with Prev = 0.1 (b) F_1 score with Prev = 0.5 (c) F_1 score with Prev = 0.9

Figure B.3: Contours of the F_1 score in the ROC space.

B.2. F_1 Score

Using the notation of Figure 1, the F_1 score (Eq. (34)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned}\mathrm{F}_1(a,b,c,d) &= \frac{2a}{2a+b+c} \\ &= \frac{2a}{2a+n-d+p-a}.\end{aligned}$$

For given numbers of positives (p) and negatives (n), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k,p,n,d) = \frac{k(d-n-p)}{k-2} \tag{47}$$

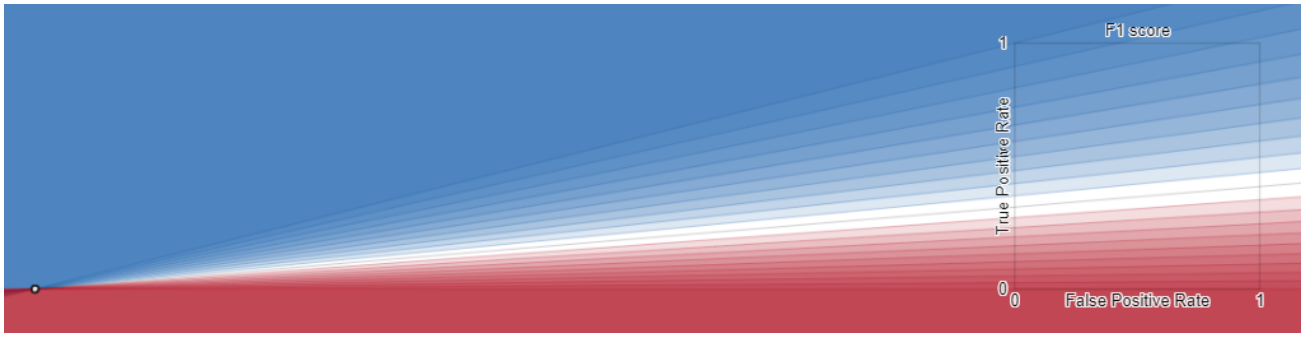
or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k,p,n,\delta) = \frac{k(n\delta-n-p)}{p(k-2)} \tag{48}$$

in which case, all contours intersect at $(\alpha, \delta) = (0, (p+n)/n)$. These intersections become apparent when we visualise the contours of the F_1 score within and beyond the ROC space (Figure B.4) which reveals how the slopes of the linear contours depend on prevalence, similar to those of the Threat Score.

(a) F_1 with Prev=0.1

(b) F_1 with Prev=0.5



(c) F_1 with Prev=0.8

Figure B.4: Contours of the F_1 score in and beyond the ROC space are straight lines that intersect at $(\alpha, \delta) = (0, (p+n)/n)$ or, equivalently $(\mathrm{FPR}, \mathrm{TPR}) = (1 - (p+n)/n, 0)$.

(a) Threat Score with Prev = 0.1 (b) Threat Score with Prev = 0.5 (c) Threat Score with Prev = 0.9

Figure B.5: Contours of the Threat Score (Jaccard Index, Critical Success Index) in the ROC space.

B.3. Threat Score (Jaccard Index, Critical Success Index)

Using the notation of Figure 1, the Threat Score (Eq. (36)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned} \mathrm{TS}(a, b, c, d) &= \frac{a}{a + b + c} \\ &= \frac{a}{p + n - d}. \end{aligned}$$

For given numbers of positives (p) and negatives (n), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k, p, n, d) = k(p + n - d) \tag{49}$$

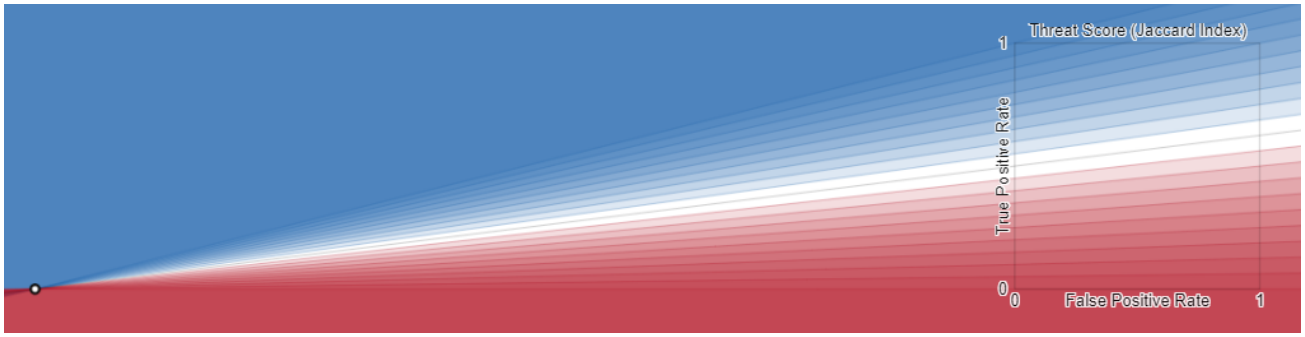
or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, p, n, \delta) = k(p + n(1 - \delta))/p \tag{50}$$

in which case, all contours intersect at $(\alpha, \delta) = (0, (p + n)/n)$. These intersections become apparent when we visualise the contours of the F_1 score within and beyond the ROC space (Figure B.6) which reveals how the slopes of the linear contours depend on prevalence, similar to those of the F_1 Score.

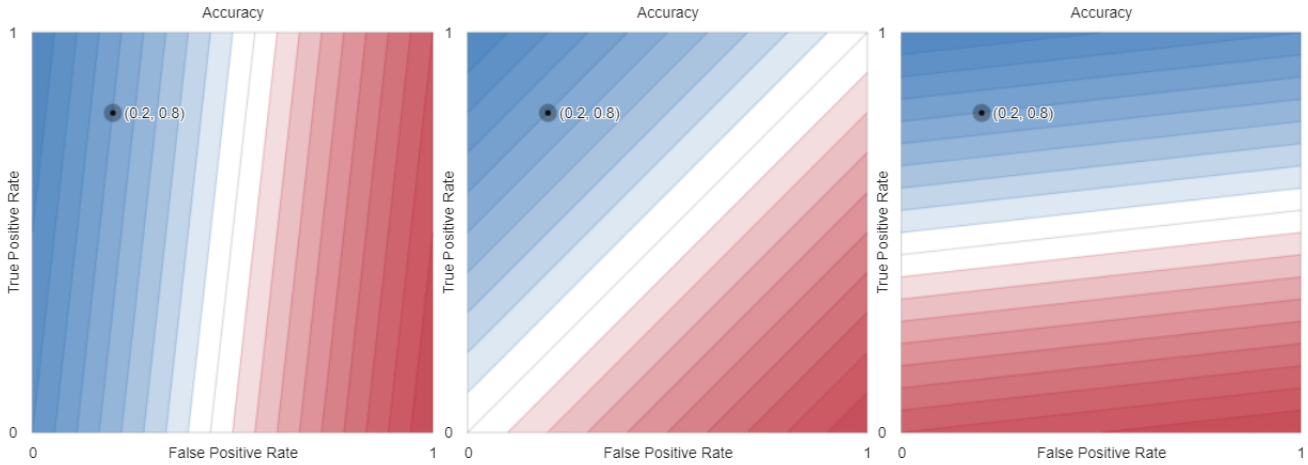
(a) Threat Score with Prev=0.1

(b) Threat Score with Prev=0.5



(c) Threat Score with Prev=0.8

Figure B.6: Contours of the Threat Score score in and beyond the ROC space are straight lines that intersect at $(\alpha, \delta) = (0, (p+n)/n)$ or, equivalently $(\mathrm{FPR}, \mathrm{TPR}) = (1-(p+n)/n, 0)$.



(a) Accuracy with Prev = 0.1 (b) Accuracy with Prev = 0.5 (c) Accuracy with Prev = 0.9

Figure B.7: Contours of accuracy in the ROC space.

B.4. Accuracy

Using the notation of Figure 1, the Accuracy (Eq. (33)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned} \mathrm{TS}(a,b,c,d) &= \frac{a+d}{a+b+c+d} \\ &= \frac{a+d}{p+n}. \end{aligned}$$

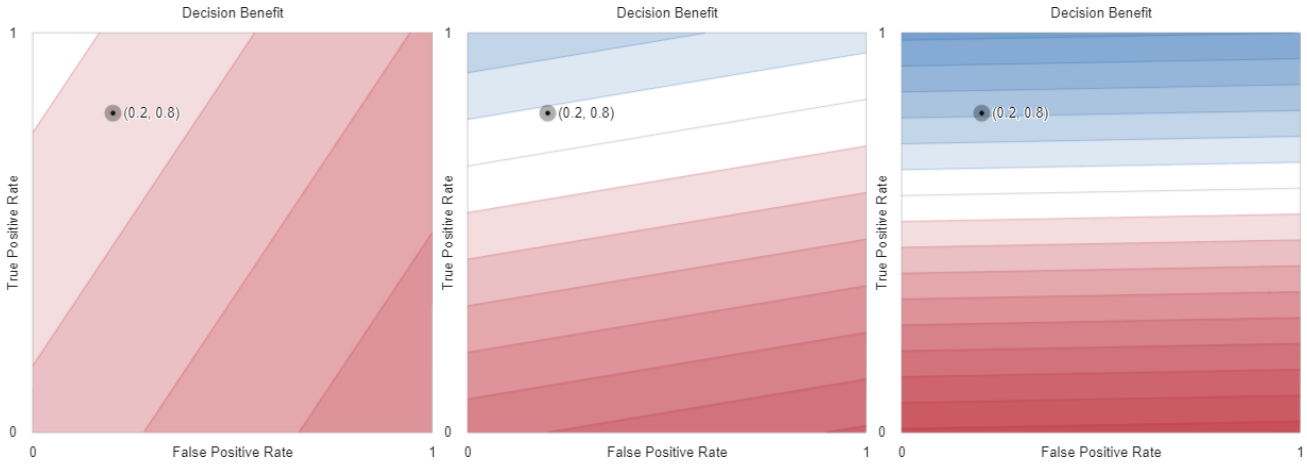
For given numbers of positives (p) and negatives (n), this performance metric achieves a value of $0 \leq k < 1$ along the contour lines with

$$a(k,p,n,d) = k(p+n) - d \tag{51}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

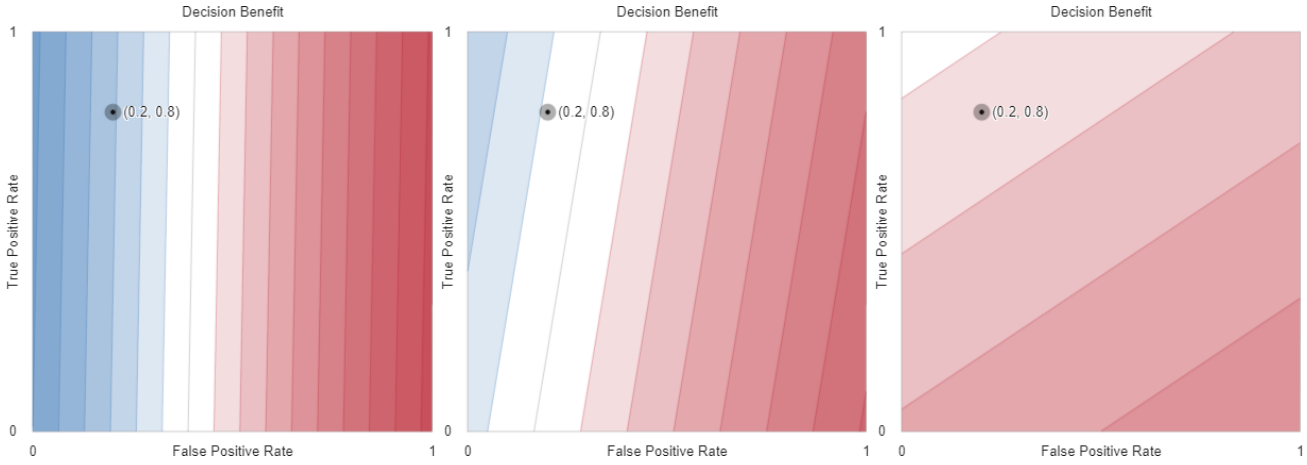
$$\alpha(k,p,n,\delta) = \frac{k(p+n) - n\delta}{p}. \tag{52}$$

These contour lines are parallel and planar.



(a) Decision Benefits with Prev = 0.1 (b) Decision Benefits with Prev = 0.5 (c) Decision Benefits with Prev = 0.9

Figure B.8: Contours of Decision Benefits in the ROC space using $\boldsymbol{\beta} = \begin{bmatrix} 7 & 3 \\ 1 & 4 \end{bmatrix}$.



(a) Decision Benefits with Prev = 0.1 (b) Decision Benefits with Prev = 0.5 (c) Decision Benefits with Prev = 0.9

Figure B.9: Contours of Decision Benefits in the ROC space using $\boldsymbol{\beta} = \begin{bmatrix} 4 & 1 \\ 1 & 7 \end{bmatrix}$.

B.5. Decision Costs (Benefits)

While it is common to refer to the *costs* of correct and incorrect classifications, we work here equivalently in terms of *benefits* in keeping with other performance metrics in this paper where "bigger is better" and to ensure the colour scales used in our graphics can be interpreted consistently ("blue is better").

We define the matrix of benefits associated with the confusion matrix of Figure 1 as

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_a & \beta_b \\ \beta_c & \beta_d \end{bmatrix}.$$

Literature that concentrates on decision costs for a *fixed prevalence* treats this matrix as having only two degrees of freedom, e.g., the ratio (or difference) of costs for true positive and false negative classifications, and the ratio (or difference) of costs for true negative and false positive classifications. This paper considers confusion matrices with fixed totals, but *varying prevalence*, so our parameterisation of the benefits matrix affords three degrees of freedom.

Using the notation of Figure 1, the Decision Benefits (Eq. (38)) can be rewritten in terms of a, d, p, n and benefits $\boldsymbol{\beta}$ as

$$\begin{aligned}\mathrm{DB}(a, b, c, d, \boldsymbol{\beta}) &= a \cdot \beta_a + b \cdot \beta_b + c \cdot \beta_c + d \cdot \beta_d \\ &= a \cdot \beta_a + (n - d)\beta_b + (p - a)\beta_c + d \cdot \beta_d \\ &= a(\beta_a - \beta_c) + d(\beta_d - \beta_b) + p \cdot \beta_c + n \cdot \beta_d\end{aligned}$$

For given numbers of positives (p) and negatives (n), this performance metric achieves a value of k along the contour lines with

$$a(k, p, n, d, \boldsymbol{\beta}) = -d\frac{\beta_d - \beta_b}{\beta_a - \beta_c} + \frac{k - p \cdot \beta_c - n \cdot \beta_d}{\beta_a - \beta_c} \tag{53}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, p, n, \delta, \boldsymbol{\beta}) = -\delta\frac{n(\beta_d - \beta_b)}{p(\beta_a - \beta_c)} + \frac{k - p \cdot \beta_c - n \cdot \beta_d}{p(\beta_a - \beta_c)}. \tag{54}$$

These contour lines are parallel and planar.

To help with plotting the contours for all confusion matrices of size N, we use a shifted and scaled version of the benefits matrix, $\boldsymbol{\beta}^*$, whose minimum element is 0 and whose maximum element is $\frac{1}{N}$

$$\boldsymbol{\beta}^* = \frac{1}{N} \cdot \frac{\boldsymbol{\beta} - \min(\boldsymbol{\beta})}{\max(\boldsymbol{\beta})}.$$

This ensures feasible contours range between $[0, 1]$. In interactive plotting, we enforce the constraints

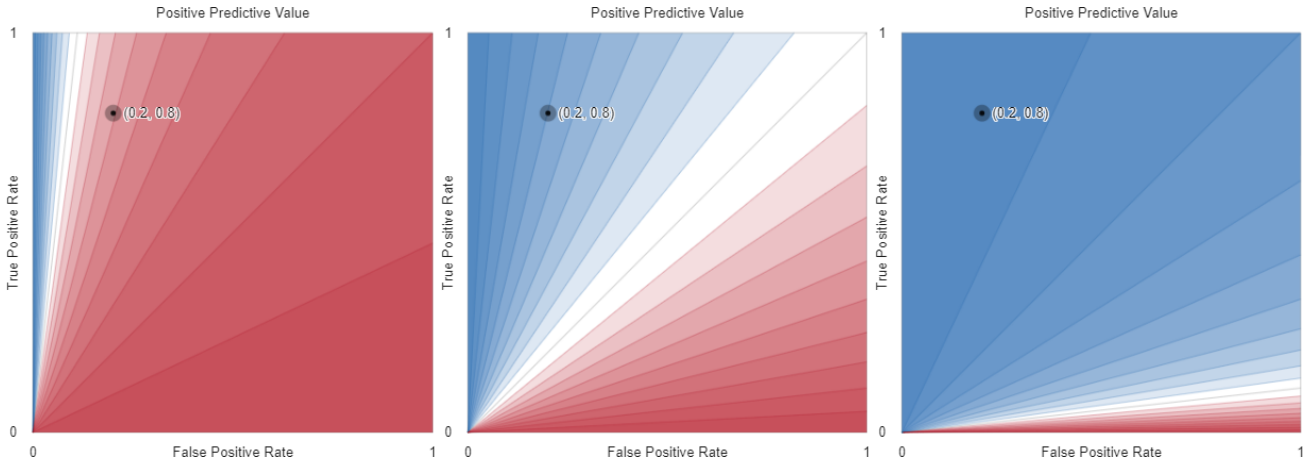
$$\begin{aligned}\beta_a > \beta_c \geq 0 \\ \beta_d > \beta_b \geq 0\end{aligned}$$

to ensure that the contours have finite, positive slope.

When

$$\boldsymbol{\beta} = \begin{bmatrix}\beta & 0 \\ 0 & \beta\end{bmatrix}$$

for some positive β, the Decision Cost contours are the same as those of Accuracy.



(a) Positive Predictive Value with Prev = 0.1 (b) Positive Predictive Value with Prev = 0.5 (c) Positive Predictive Value with Prev = 0.9

Figure B.10: Contours of Positive Predictive Value in the ROC space.

B.6. Positive Predictive Value (Precision)

Using the notation of Figure 1, the Positive Predictive Value (Eq. (27)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned}\mathrm{PPV}(a,b,c,d) &= \frac{a}{a+b} \\ &= \frac{a}{a+n-d}.\end{aligned}$$

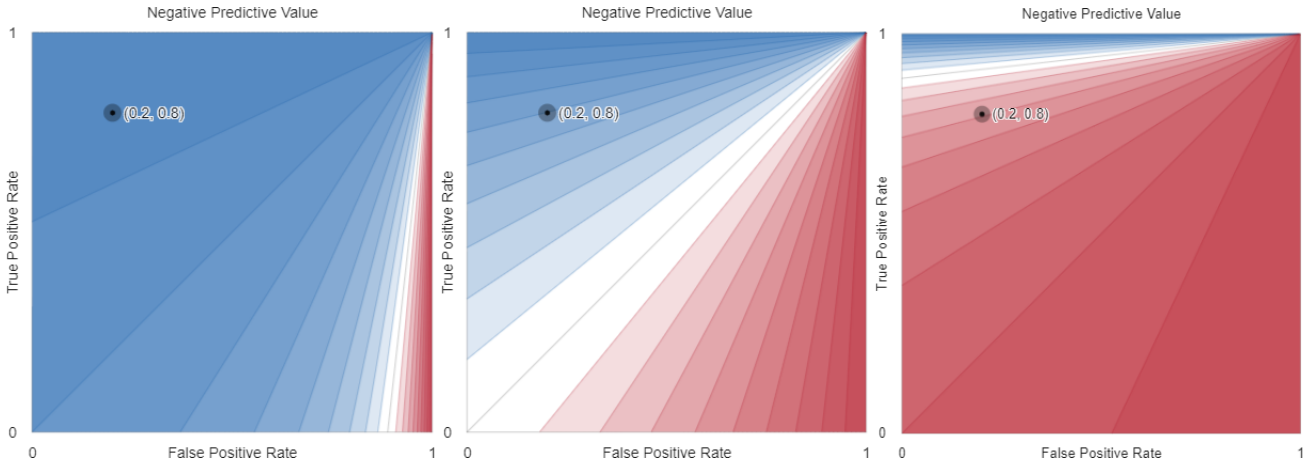
For given numbers of positives (p) and negatives (n), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k,p,n,d) = \frac{k(d-n)}{k-1} \tag{55}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k,p,n,\delta) = \frac{kn(\delta-1)}{p(k-1)} \tag{56}$$

in which case, all contours intersect at $(\alpha, \delta) = (0, 1)$.



(a) Negative Predictive Value with Prev = 0.1 (b) Negative Predictive Value with Prev = 0.5 (c) Negative Predictive Value with Prev = 0.9

Figure B.11: Contours of Negative Predictive Value in the ROC space.

B.7. Negative Predictive Value

Using the notation of Figure 1, the Negative Predictive Value (Eq. (27)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned} \mathrm{NPV}(a, b, c, d) &= \frac{d}{c+d} \\ &= \frac{d}{p-a+d}. \end{aligned}$$

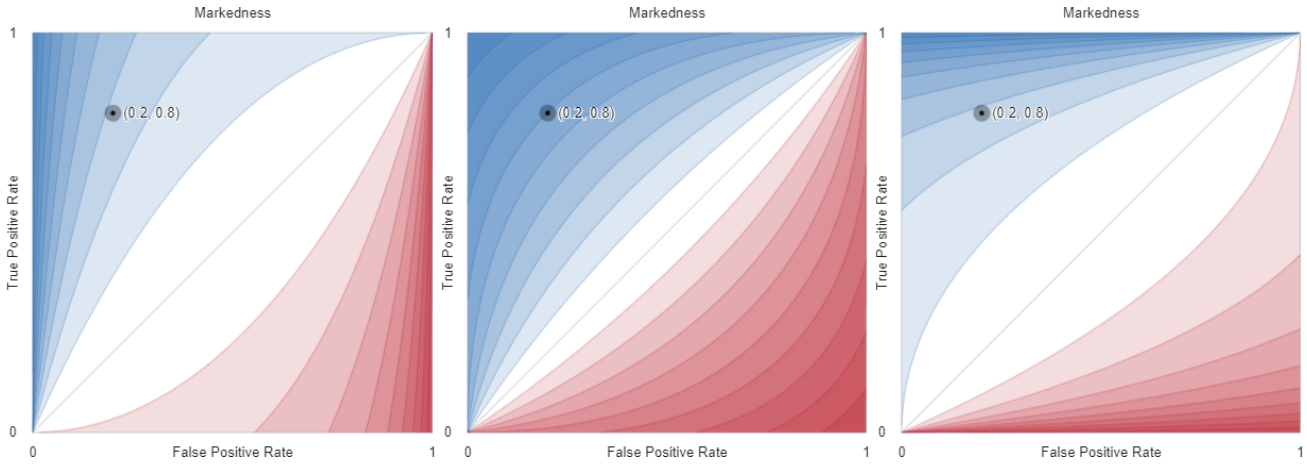
For given numbers of positives (p) and negatives (n), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k, p, n, d) = p + \frac{d(k-1)}{k} \tag{57}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, p, n, \delta) = 1 + \frac{n\delta(k-1)}{pk} \tag{58}$$

in which case, all contours intersect at $(\alpha, \delta) = (1, 0)$.



(a) Markedness with Prev = 0.1 (b) Markedness with Prev = 0.5 (c) Markedness with Prev = 0.9

Figure B.12: Contours of Markedness in the ROC space.

B.8. Markedness

Using the notation of Figure 1, Markedness (Eq. (31)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned}\mathrm{MK}(a,b,c,d) &= \frac{a}{a+b} + \frac{d}{c+d} - 1 \\ &= \frac{a}{a+n-d} + \frac{d}{p-a+d} - 1.\end{aligned}$$

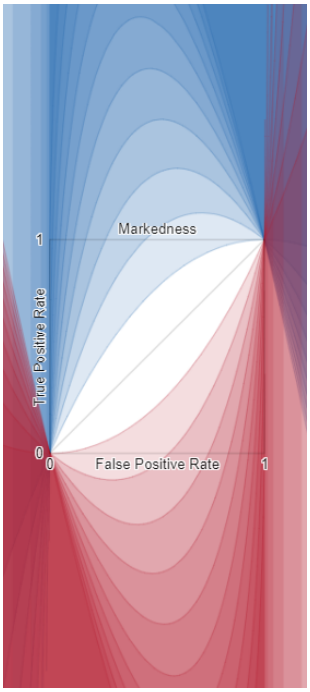
For given numbers of positives (p) and negatives (n), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k,p,n,d) = \begin{cases} \dfrac{\sqrt{(kn+kp+n)^2 - 4dk(n+p)} + 2dk - kn + kp - n}{2k} & k \geq 0 \\ -\dfrac{\sqrt{(kn+kp+n)^2 - 4dk(n+p)} - 2dk + kn - kp + n}{2k} & k < 0 \end{cases} \tag{59}$$

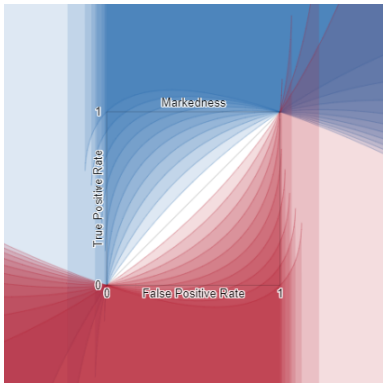
or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k,p,n,\delta) = \begin{cases} \dfrac{\sqrt{(kn+kp+n)^2 - 4n\delta k(n+p)} + 2n\delta k - kn + kp - n}{2pk} & k \geq 0 \\ -\dfrac{\sqrt{(kn+kp+n)^2 - 4n\delta k(n+p)} - 2n\delta k + kn - kp + n}{2pk} & k < 0 \end{cases} \tag{60}$$

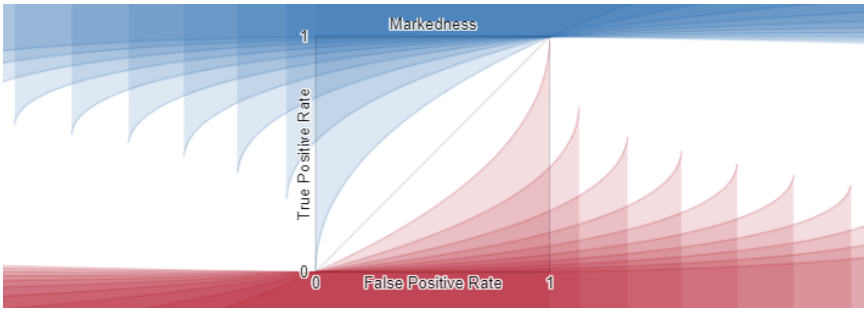
in which case, all contours where $k < 0$ intersect at $(\alpha, \delta) = (0, 1)$ and all contours where $k > 0$ intersect at $(\alpha, \delta) = (1, 0)$, as is the case for the Matthews Correlation Coefficient.



(a) Markedness with Prev=0.1

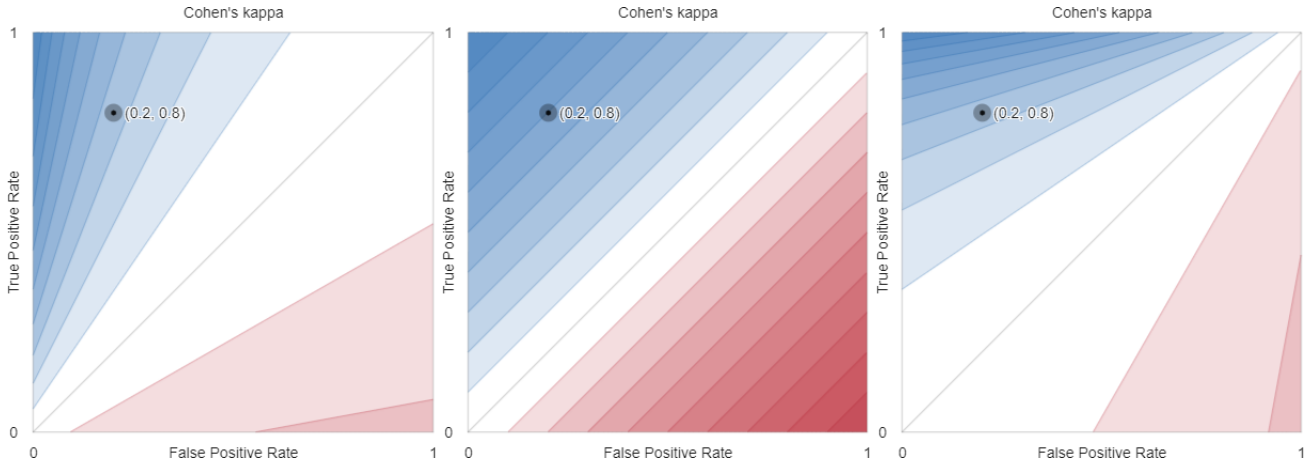


(b) Markedness with Prev=0.5



(c) Markedness with Prev=0.9

Figure B.13: Contours of Markedness within and beyond the ROC space describe a series of polynomial curves whose shape depends on prevalence and which intersect where $(\mathrm{FPR}, \mathrm{TPR}) = (0, 0)$ and $(\mathrm{FPR}, \mathrm{TPR}) = (1, 1)$



(a) Cohen's kappa with Prev = 0.1 (b) Cohen's kappa with Prev = 0.5 (c) Cohen's kappa with Prev = 0.9

Figure B.14: Contours of Cohen's kappa in the ROC space.

B.9. Cohen's kappa

Using the notation of Figure 1, Cohen's kappa (Eq. (44)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned}\kappa(a,b,c,d) &= \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)} \\ &= \frac{2(ad-(n-d)(p-a))}{(a+n-d)n+p(p-a+d)}.\end{aligned}$$

For given numbers of positives (p) and negatives (n), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k,p,n,d) = \frac{dk(n-p)+2dp-k(n^2+p^2)-2np}{(k-2)n-kp} \tag{61}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

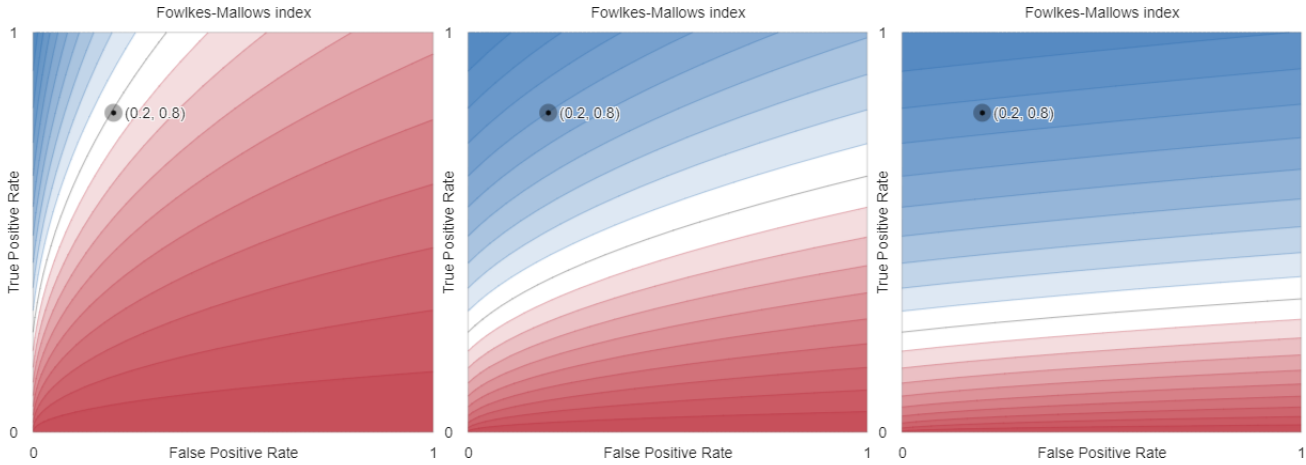
$$\alpha(k,p,n,\delta) = \frac{n\delta k(n-p)+2n\delta p-k(n^2+p^2)-2np}{p((k-2)n-kp)} \tag{62}$$

$$= \frac{n\delta k(n-p)-2np(1-\delta)-k(n^2+p^2)}{p(k(n-p)-2n)} \tag{63}$$

in which case, all contours intersect at

$$(\alpha,\delta) = \left(\frac{n}{n-p}, 1-\frac{n}{n-p}\right)$$

when $p \neq n$.



(a) Fowlkes-Mallows index with Prev = 0.1 (b) Fowlkes-Mallows index with Prev = 0.5 (c) Fowlkes-Mallows index with Prev = 0.9

Figure B.15: Contours of Fowlkes-Mallows index in the ROC space.

B.10. Fowlkes-Mallows index

Using the notation of Figure 1, the Fowlkes-Mallows Index (Eq. (42)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned}\mathrm{FM}(a,b,c,d) &= \sqrt{\frac{a}{a+b}\cdot\frac{a}{a+c}} \\ &= \sqrt{\frac{a}{a+n-d}\cdot\frac{a}{p}}\end{aligned}$$

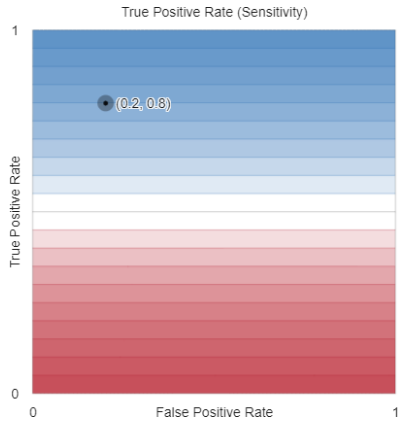
For given numbers of positives (p) and negatives (n), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k,p,n,d) = \frac{1}{2}\left(\sqrt{k^2p(-4d+k^2p+4n)} + k^2p\right) \tag{64}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k,p,n,\delta) = \frac{1}{2p}\left(\sqrt{k^2p(-4nd+k^2p+4n)} + k^2p\right). \tag{65}$$

Figure C.1: Contours of True Positive Rate in the ROC space.



C. Performance metric contours that are independent of prevalence

C.1. True Positive Rate (Sensitivity)

Using the notation of Figure 1, the True Positive Rate (Eq. (13)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned}\mathrm{TPR}(a,b,c,d) &= \frac{a}{a+c} \\ &= \frac{a}{p}.\end{aligned}$$

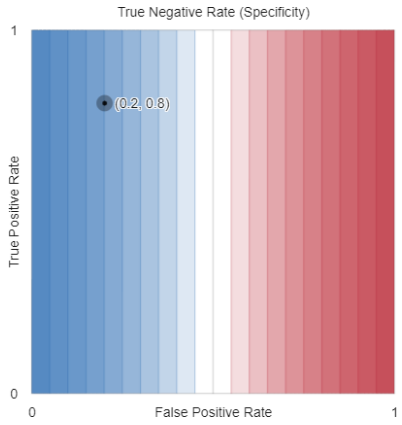
For given numbers of positives (p), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k,p) = pk \tag{66}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k) = k \tag{67}$$

Figure C.2: Contours of True Negative Rate in the ROC space.



C.2. True Negative Rate (Specificity)

Using the notation of Figure 1, the True Negative Rate (Eq. (16)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned} \mathrm{TNR}(a, b, c, d) &= \frac{d}{b+d} \\ &= \frac{d}{n}. \end{aligned}$$

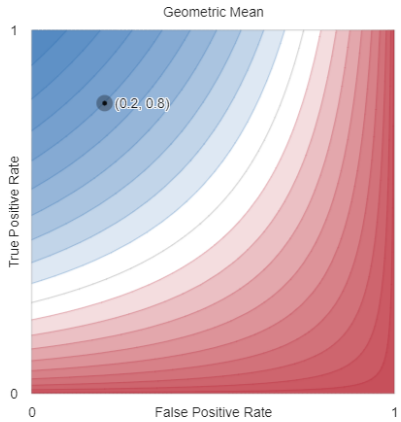
For given numbers of negatives (n), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$d(k, n) = nk \tag{68}$$

or, in terms of true negative rate $\delta = d/n$

$$\delta(k) = k \tag{69}$$

Figure C.3: Contours of the geometric mean of true positive rate (sensitivity) and true negative rate (specificity) in the ROC space.



C.3. Geometric Mean

Using the notation of Figure 1, the Geometric Mean (Eq. (25)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned}\mathrm{GM}(a,b,c,d) &= \sqrt{\frac{a}{a+c}\cdot\frac{d}{b+d}} \\ &= \sqrt{\frac{a}{p}\cdot\frac{d}{n}}.\end{aligned}$$

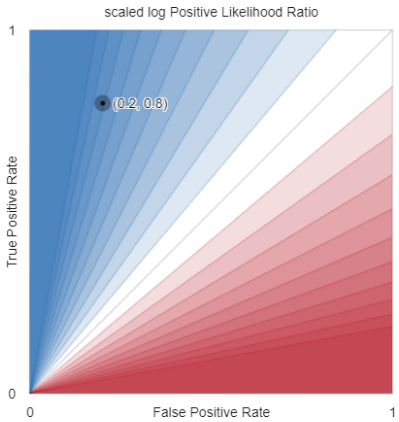
For given numbers of positives (p) and negatives (n), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k,p,n,d) = \frac{k^2 np}{d} \tag{70}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k,\delta) = \frac{k^2}{\delta} \tag{71}$$

Figure C.4: Contours of the scaled logarithm of the positive likelihood ratio in the ROC space.



C.4. Positive Likelihood Ratio (LR_+) and scaled log Positive Likelihood Ratio

Using the notation of Figure 1, the logarithm of the Likelihood Ratio of a positive outcomes (Eq. (15)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned}\log \mathrm{LR}_+(a,b,c,d) &= \log\left(\frac{a}{p}\frac{n}{b}\right) \\ &= \log\left(\frac{a}{p}\frac{n}{n-d}\right).\end{aligned}$$

For given numbers of positives (p) and negatives (n), this performance metric achieves a value of k along the contour lines with

$$a(k,p,n,d) = \frac{e^k p(n-d)}{n} \tag{72}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k,\delta) = e^k p(1-\delta) \tag{73}$$

$\log \mathrm{LR}_+$ has range $(-\infty, \infty)$. To visualise the finite values of this function, it is convenient to work with a scaled version of this function whose contours lie between $[-1, 1]$. The largest finite value of LR_+ is $n(p-1)/p$ so we can produce a scaled version of Eq. 72 by dividing $\log \mathrm{LR}_+(a,b,c,d)$ by

$$M = \log(n(p-1)/p) \tag{74}$$

to give

$$\text{scaled}\log \mathrm{LR}_+(a,b,c,d) = \frac{1}{\log(n(p-1)/p)} \log\left(\frac{a}{p}\frac{n}{b}\right) \tag{75}$$

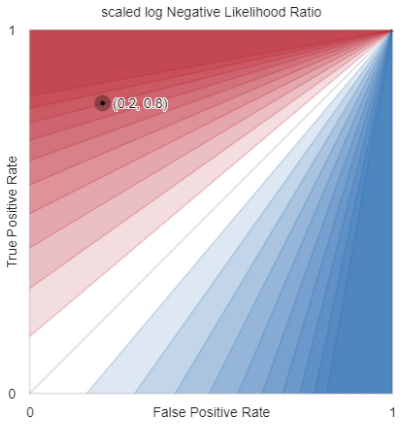
which yields the contour equations

$$a(k,p,n,d) = \frac{e^{Mk} p(n-d)}{n} \tag{76}$$

and

$$\alpha(k,\delta) = e^{Mk} p(1-\delta) \tag{77}$$

Figure C.5: Contours of the scaled logarithm of the negative likelihood ratio in the ROC space.



C.5. Negative Likelihood Ratio (LR_-) and scaled log Negative Likelihood Ratio

Using the notation of Figure 1, the logarithm of the Likelihood Ratio of a negative outcomes (Eq. (18)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned}\log \mathrm{LR}_-(a,b,c,d) &= \log\left(\frac{c}{p}\frac{n}{d}\right) \\ &= \log\left(\frac{p-a}{p}\frac{n}{d}\right).\end{aligned}$$

For given numbers of positives (p) and negatives (n), this performance metric achieves a value of k along the contour lines with

$$a(k,p,n,d) = p - \frac{de^k p}{n} \tag{78}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k,\delta) = 1 - \delta e^k \tag{79}$$

$\log \mathrm{LR}_-$ has range $(-\infty, \infty)$. To visualise the finite values of this function, it is convenient to work with a scaled version of this function whose contours lie between $[-1, 1]$. The largest finite value of LR_- is $pn/(n-1)$ so we can produce a scaled version of Eq. 78 by dividing $\log \mathrm{LR}_-(a,b,c,d)$ by

$$M = \log(pn/(n-1)) \tag{80}$$

to give

$$\text{scaled}\log \mathrm{LR}_-(a,b,c,d) = \frac{1}{\log(pn/(n-1))}\log\left(\frac{c}{p}\frac{n}{d}\right) \tag{81}$$

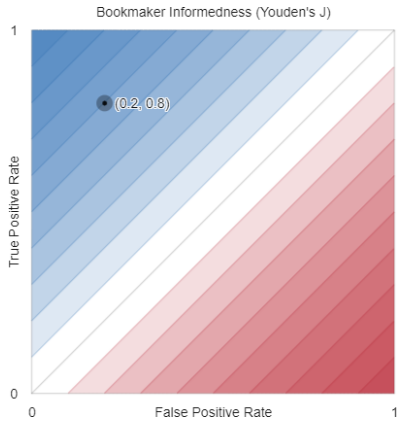
which yields the contour equations

$$a(k,p,n,d) = p - \frac{de^{Mk}p}{n} \tag{82}$$

and

$$\alpha(k,\delta) = 1 - \delta e^{Mk} \tag{83}$$

Figure C.6: Contours of Bookmaker Informedness in the ROC space.



C.6. Bookmaker Informedness, BA

Using the notation of Figure 1, the Bookmaker Informedness (Eq. (24)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned}\mathrm{BM}(a,b,c,d) &= \frac{a}{a+c} + \frac{d}{b+d} - 1 \\ &= \frac{a}{p} + \frac{d}{n} - 1.\end{aligned}$$

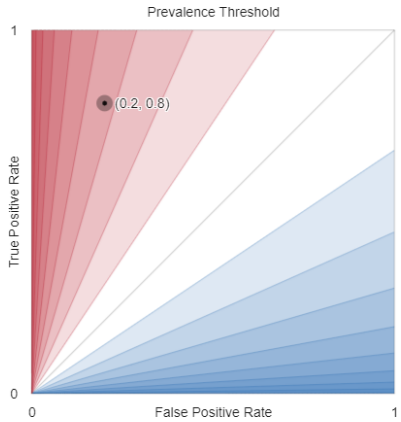
For given numbers of positives (p) and negatives (n), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k,p,n,d) = p\left(k + 1 - \frac{d}{n}\right) \tag{84}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k,\delta) = k + 1 - \delta \tag{85}$$

Figure C.7: Contours of Prevalence Threshold in the ROC space.



C.7. Prevalence Threshold

Using the notation of Figure 1, the Prevalence Threshold (Eq. (26)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned}\mathrm{PT}(a,b,c,d) &= \frac{\sqrt{b/(b+d)}}{\sqrt{a/(a+c)}+\sqrt{b/(b+d)}} \\ &= \frac{\sqrt{(n-d)/n}}{\sqrt{a/p}+\sqrt{(n-d)/n}}.\end{aligned}$$

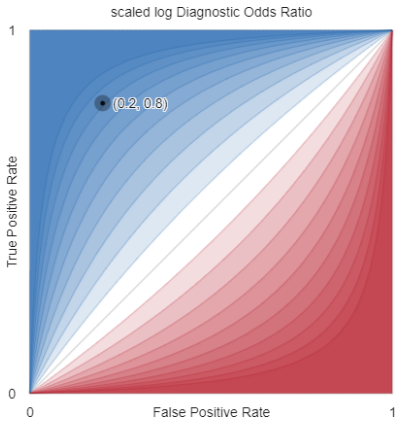
For given numbers of positives (p) and negatives (n), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k,p,n,d) = \frac{(k-1)^2 p(n-d}{k^2 n} \tag{86}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k,\delta) = \frac{(k-1)^2(1-\delta)}{k^2} \tag{87}$$

Figure C.8: Contours of scaled log Diagnostic Odds Ratio in the ROC space.



C.8. log Diagnostic Odds Ratio and scaled log Diagnostic Odds Ratio

Using the notation of Figure 1, the logarithm of the Diagnostic Odds Ratio (Eq. (19)) can be rewritten in terms of a, d, p, n as

$$\begin{aligned} \log \mathrm{DOR}(a,b,c,d) &= \log\left(\frac{ad}{bc}\right) \\ &= \log\left(\frac{ad}{(p-a)(n-d)}\right). \end{aligned} \tag{88}$$

For given numbers of positives (p) and negatives (n), this performance metric achieves a value of k along the contour lines with

$$a(k,p,n,d) = \frac{e^k(d-n)p}{d(e^k-1)-e^k n} \tag{89}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k,\delta) = \frac{e^k(\delta-1)}{\delta(e^k-1)-e^k} \tag{90}$$

$\log \mathrm{DOR}$ has range $(-\infty, \infty)$. To visualise the finite values of this function, it is convenient to work with a scaled version of this function whose contours lie between $[-1, 1]$. The largest finite value of DOR is $(p-1)(n-1)$ so we can produce a scaled version of Eq. 90 by dividing $\log \mathrm{DOR}(a,b,c,d)$ by

$$M = \log(p-1)(n-1) \tag{91}$$

to give

$$\text{scaled}\log \mathrm{DOR}(a,b,c,d) = \frac{1}{\log(p-1)(n-1)} \log\left(\frac{ad}{bc}\right) \tag{92}$$

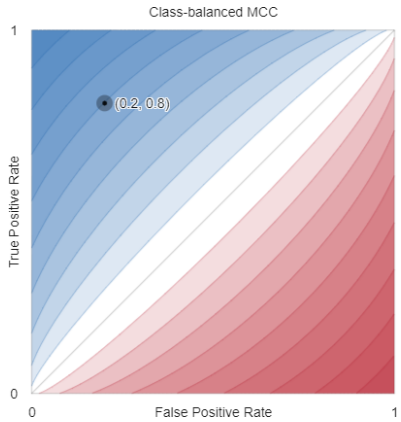
which yields the contour equations

$$a(k,p,n,d) = \frac{e^{Mk}(d-n)p}{d(e^{Mk}-1)-e^{Mk}n} \tag{93}$$

and

$$\alpha(k,\delta) = \frac{e^{Mk}(\delta-1)}{\delta(e^{Mk}-1)-e^{Mk}}. \tag{94}$$

Figure C.9: Contours of the balanced Matthews Correlation Coefficient in the ROC space.



C.9. Balanced Matthews Correlation Coefficient

Using the notation of Figure 1, the balanced Matthews Correlation Coefficient (Eq. (41)) can be rewritten in terms of a, d, p, n as

$$\mathrm{MCC}_{bal} = \frac{a/p + d/n - 1}{\sqrt{(1 - \left(\frac{a}{p} \cdot \frac{d}{n}\right)^2}} \tag{95}$$

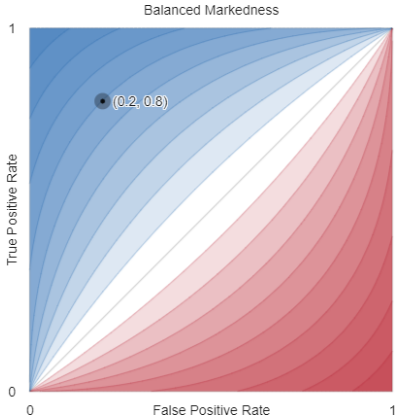
For given numbers of positives (p) and negatives (n), this performance metric achieves a value of k along the contour lines with

$$a(k, p, n, d) = \frac{p}{n} \cdot \frac{k\sqrt{-4d^2 + 4dn + k^2n^2} + n + d(k^2 - 1)}{k^2 + 1} \tag{96}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = \frac{k\sqrt{-4\delta^2 + 4d\delta + k^2} + 1 + \delta(k^2 - 1)}{k^2 + 1} \tag{97}$$

Figure C.10: Contours of the balanced Markedness in the ROC space.



C.10. Balanced Markedness

Using the notation of Figure 1, balanced Markedness (Eq. (32)) can be rewritten in terms of a, d, p, n as

$$\mathrm{MK}_{bal} = \frac{a/p}{a/p + 1 - d/n} + \frac{d/n}{d/n + 1 - a/p} - 1 \tag{98}$$

$$= \frac{a}{a + p - pd/n} + \frac{d}{d + n - na/p} - 1 \tag{99}$$

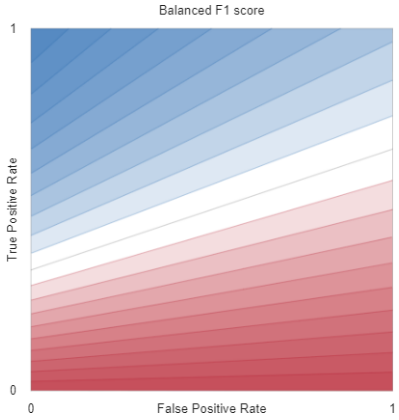
For given numbers of positives (p) and negatives (n), this performance metric achieves a value of k along the contour lines with

$$a(k, p, n, d) = p\left(\frac{d}{n} + \frac{\sqrt{(2k+1)^2 - 8dk/n} - 1}{2k}\right) \tag{100}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = \delta + \frac{\sqrt{(2k+1)^2 - 8\delta k} - 1}{2k} \tag{101}$$

Figure C.11: Contours of $\mathrm{F}_{1\,bal}$ in the ROC space.



C.11. Balanced F_1 ($\mathrm{F}_{1\,bal}$)

Using the notation of Figure 1, balanced F_1 (Eq. (35)) can be rewritten in terms of a, d, p, n as

$$\mathrm{F}_{1\,bal} = \frac{2a/p}{2 + a/p - d/n} \tag{102}$$

For given numbers of positives (p) and negatives (n), this performance metric achieves a value of k along the contour lines with

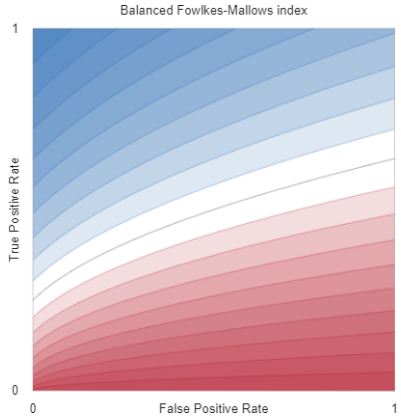
$$a(k, p, n, d) = \frac{kp(d - 2n)}{(k - 2)n} \tag{103}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = \frac{k(\delta - 2)}{k - 2} \tag{104}$$

in which case, all contours intersect at $(\alpha, \delta) = (0, 2)$.

Figure C.12: Contours of FM_{bal} in the ROC space.



C.12. Balanced Fowlkes-Mallows index

Using the notation of Figure 1 and the same approach as (Luque et al., 2019), a balanced version of the Fowlkes-Mallows Index (Eq. (43)) can be rewritten in terms of a, d, p, n as

$$\mathrm{FM}_{bal}(a, b, c, d) = \frac{\frac{a}{p}}{\sqrt{1 + \frac{a}{p} - \frac{d}{n}}}$$

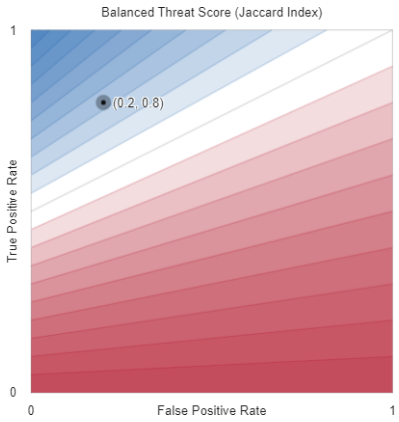
For given numbers of positives (p) and negatives (n), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k, p, n, d) = \frac{kp}{2}\left(\sqrt{k^2 + 4 - 4d/n} + k\right) \tag{105}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = \frac{k}{2}\left(\sqrt{k^2 + 4 - 4\delta} + k\right). \tag{106}$$

Figure C.13: Contours of TS_{bal} in the ROC space.



C.13. Balanced Threat Score

Using the notation of Figure 1 and the same approach as (Luque et al., 2019), a balanced version of the Threat Score (Eq. (37)) can be rewritten in terms of a, d, p, n as

$$\mathrm{TS}_{bal}(a, b, c, d) = \frac{\frac{a}{p}}{2 - \frac{d}{n}}$$

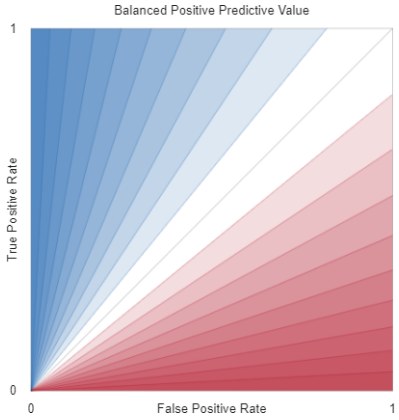
For given numbers of positives (p) and negatives (n), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k, p, n, d) = kp\left(2 - \frac{d}{n}\right) \tag{107}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = k\,(2 - \delta) \tag{108}$$

Figure C.14: Contours of PPV_{bal} in the ROC space.



C.14. Balanced Positive Predictive Value

Using the notation of Figure 1, the balanced Positive Predictive Value (Eq. (28)) can be rewritten in terms of a, d, p, n as

$$\mathrm{PPV}_{bal}(a, b, c, d) = \frac{\frac{a}{p}}{\frac{a}{p} + \left(1 - \frac{d}{n}\right)}$$

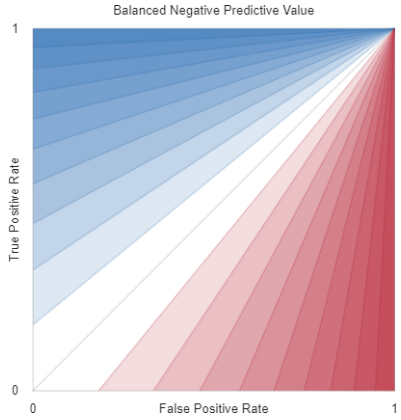
For given numbers of positives (p) and negatives (n), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k, p, n, d) = \frac{kp(d - n)}{(k - 1)n} \tag{109}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = \frac{k(\delta - 1)}{k - 1}. \tag{110}$$

Figure C.15: Contours of NPV_{bal} in the ROC space.



C.15. Balanced Negative Predictive Value

Using the notation of Figure 1, the balanced Negative Predictive Value (Eq. (30)) can be rewritten in terms of a, d, p, n as

$$\mathrm{NPV}_{bal}(a, b, c, d) = \frac{\frac{d}{n}}{\frac{d}{n} + \left(1 - \frac{a}{p}\right)}$$

For given numbers of positives (p) and negatives (n), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k, p, n, d) = \frac{dp(k-1)}{kn} + p \tag{111}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = \frac{\delta(k-1)}{k} + 1. \tag{112}$$

C.16. Balanced Cohen's Kappa is Bookmaker Informedness

Using the relationships:

$$\mathrm{TP} = p \cdot \mathrm{TPR} \qquad \mathrm{FP} = n(1 - \mathrm{TNR})$$
$$\mathrm{FN} = p(1 - \mathrm{TPR}) \qquad \mathrm{TN} = n \cdot \mathrm{TNR}$$

we can rewrite Cohen's Kappa (Eq. (44)) as

$$\kappa = \frac{2(p \cdot \mathrm{TPR} \cdot n \cdot \mathrm{TNR} - p(1 - \mathrm{TPR})n(1 - \mathrm{TNR}))}{(p \cdot \mathrm{TPR} + n(1 - \mathrm{TNR}))n + p(p(1 - \mathrm{TPR}) + n \cdot \mathrm{TNR})}.$$

Using the same approach as (Luque et al., 2019), we can create a class-balanced version of this metric by setting $p = n$ to give an expression in terms of the true positive and true negative rate alone:

$$\begin{aligned}
\kappa_{bal} &= \frac{2(\mathrm{TPR} \cdot \mathrm{TNR} - (1 - \mathrm{TPR})(1 - \mathrm{TNR}))}{\mathrm{TPR} + 1 - \mathrm{TNR} + 1 - \mathrm{TPR} + \mathrm{TNR}} \\
&= \frac{2(\mathrm{TPR} \cdot \mathrm{TNR} - 1 + \mathrm{TPR} + \mathrm{TNR} - \mathrm{TPR} \cdot \mathrm{TNR})}{2} \\
&= \mathrm{TPR} + \mathrm{TNR} - 1
\end{aligned}$$

which is the same as Bookmaker Informedness (Eq. (24)).

D. Links to interactive visualisations, animations and source code

We have used R and its `plotly` and `tidyverse` libraries (R Core Team, 2020; Sievert, 2020; Wickham et al., 2019), as well as Desmos' Graphing Calculator (Desmos, Inc., n.d.) to provide interactive visualisations for several key concepts in this paper. These visualisations are described in this section along with links to specific figures in the main paper. Desmos automatically ensures that the underlying code is available to copy and develop further and we provide RMarkdown for all other figures.

Source code (Rmarkdown) is available from Github at `https://github.com/DavidRLovell/Never-mind-the-metrics` under the GNU General Public License v3.0.

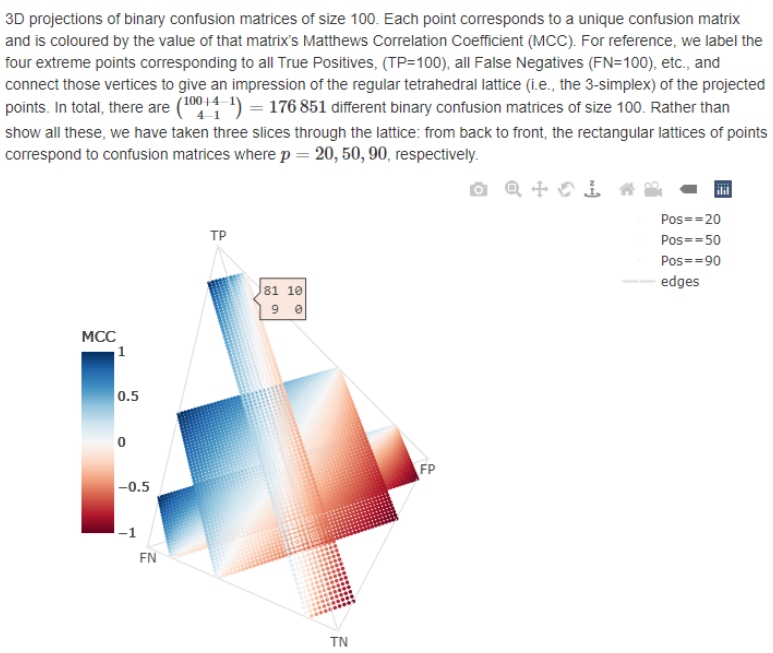


Figure D.1: Screenshot of the interactive 3D confusion simplex (`http://bit.ly/see-confusion-simplex`).

D.1. Interactive 3D confusion simplex

`http://bit.ly/see-confusion-simplex` shows an interactive visualisation of the 3D projection of binary confusion matrices of size 100. Each point corresponds to a unique confusion matrix and is coloured by the value of that matrix's Matthews Correlation Coefficient (MCC). In total, there are $\binom{100+4-1}{4-1} = 176\,851$ different binary confusion matrices of size 100. Rather than show all of these, we have taken three slices through the lattice: from back to front, the rectangular lattices of points correspond to confusion matrices where $p = 20, 50, 90$, respectively.

Users can mouse over the tetrahedron, then click and drag to change its orientation. Clicking on the text 'Pos==20' will toggle that slice of the confusion matrix.

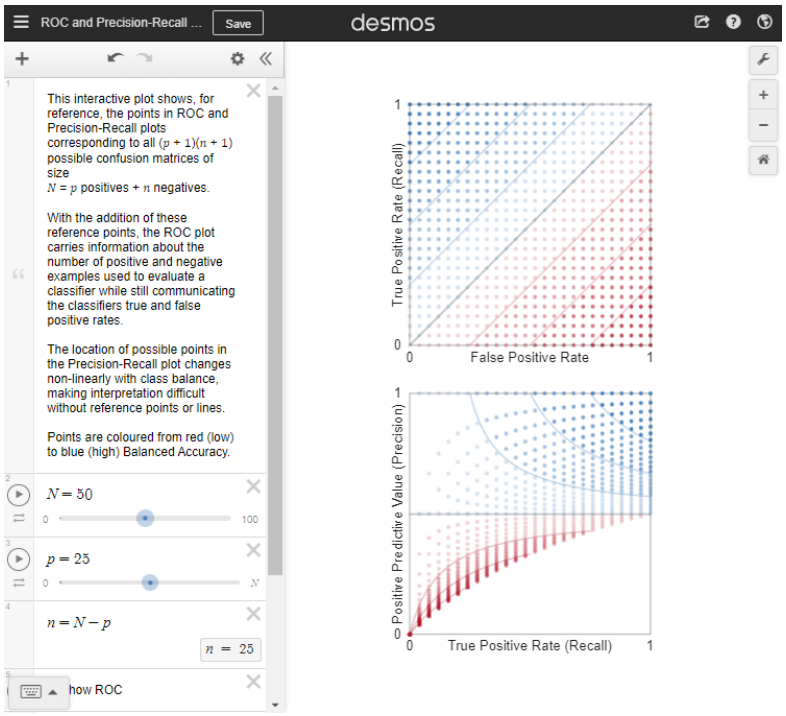


Figure D.2: Screenshot of the Desmos visualisation of possible points in ROC and Precision-Recall spaces (`http://bit.ly/see-ROC-reference-points`).

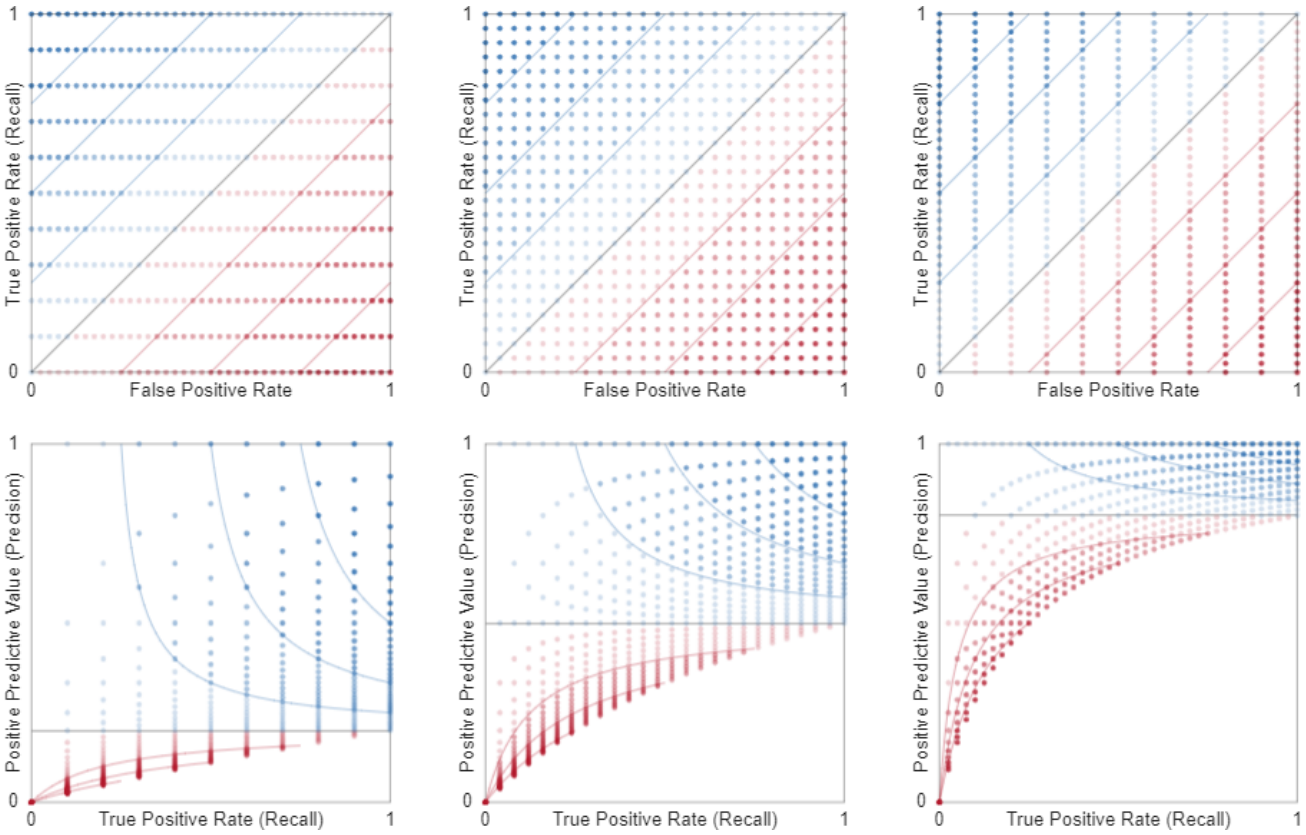
D.2. All possible ROC and Precision-Recall reference points

`http://bit.ly/see-ROC-reference-points` shows all possible $(p+1) \times (n+1)$ points in ROC and Precision-Recall spaces corresponding to confusion matrices of size $N = p + n$, coloured from red (low) to blue (high) Balanced Accuracy. Users can change N and p by adjusting the sliders in the left hand side of the Desmos window.

This pointillist approach can be used with ROC curves and Precision-Recall plots (Davis & Goadrich, 2006) which map $(\mathrm{FPR}, \mathrm{TPR})$ points in ROC space to $(\mathrm{TPR}, \mathrm{PPV})$ according to

$$(x, y) \mapsto \left(y, \left(1 + \frac{n}{p} \cdot \frac{x}{y} \right)^{-1} \right), \tag{113}$$

a mapping which clearly depends on class balance through the factor $\frac{n}{p}$. While (Saito & Rehmsmeier, 2015) regard these precision-recall plots as more informative than ROC curves because their achievable shape depends on prevalence, we find them hard to interpret without reference points.



(a) All possible ROC and Precision-Recall points with 10 positive and 40 negative cases (Prev = 0.2).

(b) All possible ROC and Precision-Recall points with 25 positive and 25 negative cases (Prev = 0.5).

(c) All possible ROC and Precision-Recall points with 40 positive and 10 negative cases (Prev = 0.8).

Figure D.3: Using reference points in the background of ROC and Precision-Recall plots can indicate the number of positive and negative examples used in a confusion matrix while maintaining a 1:1 plot aspect ratio. The points on these plots correspond to confusion matrices of size 50. We have used colours and reference contours corresponding to Balanced Accuracy (see Section C.6) to show the mapping between ROC and Precision-Recall plots, but when plotting actual data, these reference points could be made faint and unobtrusive. These screenshots are taken from our interactive visualisation of ROC and Precision-Recall reference points.

Figure D.4: Screenshot of the Desmos visualisation of confusion matrix performance metric contours (`http://bit.ly/see-confusion-metrics`). There are many things that users can switch on and off in this visualisation by clicking on the small round circles at the left edge of the screen.

D.3. Confusion matrix performance metric contours

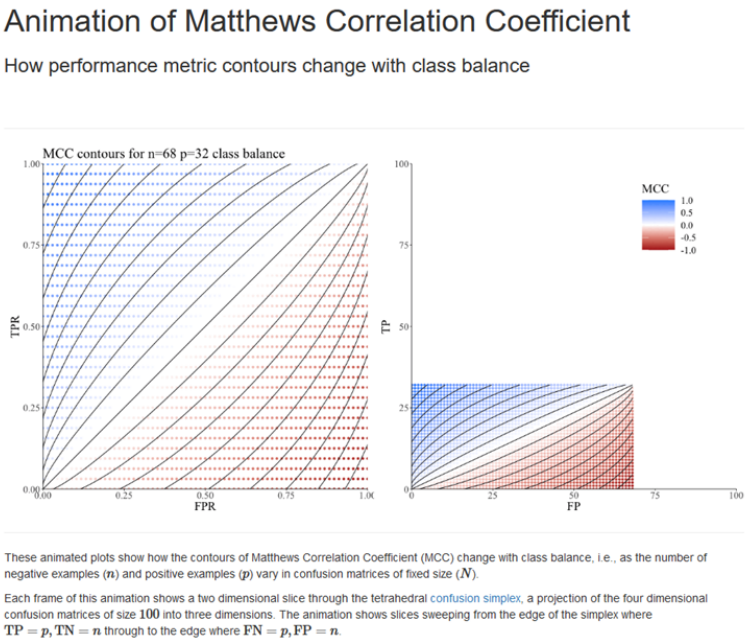
`http://bit.ly/see-confusion-metrics` enables us to interactively visualise a range of confusion matrix performance metrics by plotting their contours, coloured from red (low) to white (middle) to blue (high). This visualisation was used to produce all of the figures in Appendices B and C.

Users can change N and p by adjusting the sliders in the left hand side of the Desmos window, and can set the position of a test point by adjusting the a_1 and d_1 sliders. There are many things that users can turn on and off by clicking on the small round circles at the left edge of the screen:

Contours of prevalence-dependent and prevalence independent metrics. These switches are titled Show Accuracy, Show MCC, through to Show Geometric Mean and, when activated, display the contours of the chosen performance metrics

Additional information and decoration switches allow users to show all possible ROC points; a movable test point whose corresponding confusion matrix and performance metric values can be displayed; and various titles. Importantly, users can toggle the limits of what is displayed, so that performance metric contours *beyond* ROC space can be visualised (as in Figure B.2).

Figure D.5: Screenshot of animation of Matthews Correlation Coefficient performance metric contours (`http://bit.ly/see-animated-MCC`).



D.4. Animated performance metric contours

These animated plots show how the contours of various performance metrics change with class balance, i.e., as the number of negative examples (n) and positive examples (p) vary in confusion matrices of fixed size (N).

We have created animations of

- Accuracy: `https://bit.ly/see-animated-accuracy`
- Balanced Accuracy: `https://bit.ly/see-animated-BA`
- F_1 Score: `https://bit.ly/see-animated-F1`
- Matthews Correlation Coefficient: `https://bit.ly/see-animated-MCC`

Each animation frame shows a two dimensional slice through the tetrahedral confusion simplex, a projection of the four dimensional confusion matrices of size 100 into three dimensions. The animation shows slices sweeping from the edge of the simplex where $\mathrm{TP} = p, \mathrm{TN} = n$ through to the edge where $\mathrm{FN} = p, \mathrm{FP} = n$.

Each coloured point corresponds to a specific confusion matrix in which

$$\begin{bmatrix} \mathrm{TP} & \mathrm{FP} \\ \mathrm{FN} & \mathrm{TN} \end{bmatrix} = \begin{bmatrix} \mathrm{TP} & \mathrm{FP} \\ p - \mathrm{TP} & n - \mathrm{FP} \end{bmatrix}$$

and $N = p + n = 100$. Hence, for a given p and n, we can plot the $(p + 1) \times (n + 1)$ points whose TP values range from 0 to p and whose FP values range from 0 to n while overlaying the contours of the ‘r metric.name‘ performance metric ranging from $-0.9, -0.8, \ldots, 0.9$.

Note that

- The contours of the performance metrics are defined continuously, but empirical confusion matrices can only take on values at the discrete points in these plots.
- The left hand plot shows these points and performance metric contours in ROC space in which a classifier's true positive *rate* is plotted against its false positive rate in the space of rational numbers from $[0, 1] \times [0, 1]$.

- The right hand plot shows these points and ‘r metric.name‘ contours as an orthographic projection of the slice of points from the confusion simplex.
- The left hand ROC plot is is equivalent to re-scaling the x-axis of the right hand plot by a factor of $\frac{1}{n}$ and the y-axis by $\frac{1}{p}$.

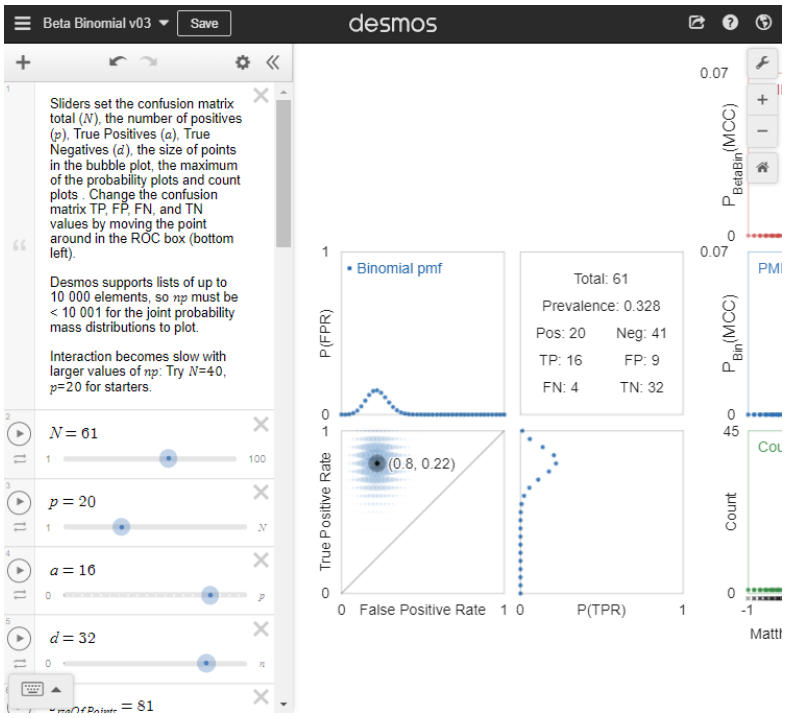


Figure D.6: Screenshot of the Desmos visualisation of confusion matrix uncertainty models (`http://bit.ly/see-confusion-uncertainty`). There are many things that users can switch on and off in this visualisation by clicking on the small round circles at the left edge of the screen.

D.5. Uncertainty in confusion matrices and their performance metrics

`http://bit.ly/see-confusion-uncertainty` enables interactive exploration of the posterior predictive pmfs of confusion matrices and three performance metrics (MCC, BA, F_1) under binomial and beta-binomial models of uncertainty. This visualisation was used to produce Figures 3, E.1, 4 and F.2.

Users can change N and p by adjusting the sliders in the left hand side of the Desmos window, and can set the position of a test point by adjusting the a and d sliders. There are many things that users can turn on and off by clicking on the small round circles at the left edge of the screen:

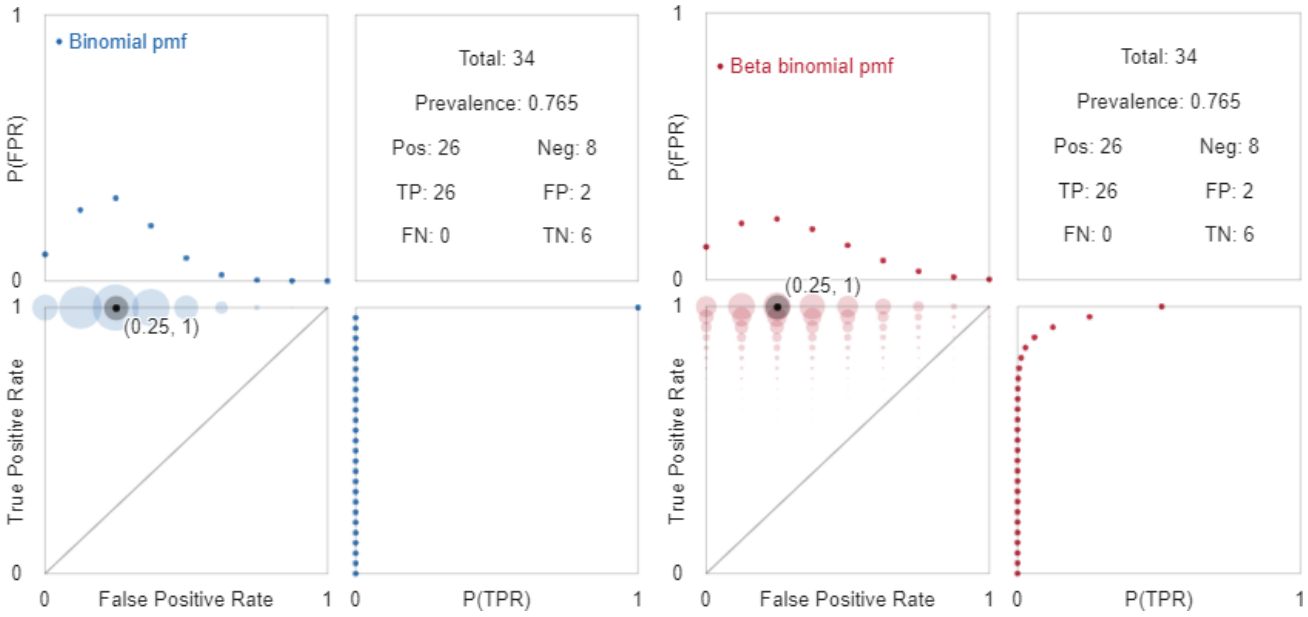
Marginal and joint pmfs of True and False Positive rates. Users can show these posterior predictive probability mass functions for confusion matrices of size $N = p + n$ under binomial and beta-binomial models of uncertainty, given that a True Positives and d True Negatives have been observed.

Posterior predictive pmfs of MCC, BA and F_1 can be shown using the Show PMF... switches for each performance metric. There are also switches to show the unique performance metric values (Show rug...), the number of times these unique values are observed (Show count...) and a histogram summary of the probability mass functions (Show histogram...).

Additional information and decoration switches allow users to show all possible ROC points; a movable test point whose corresponding confusion matrix and performance metric values can be displayed; and various labels.

Axis and point size scales are sliders that allow users to adjust the size of the points used in the joint pmf display, the maximum of the performance metric pmfs y-axis (P_{max}), and the maximum of the performance metric counts y-axis (C_{max}).

As noted on the visualisation, Desmos supports lists of up to $10\,000$ elements, so np must be $< 10\,001$ for the joint probability mass distributions to plot. This visualisation runs in your web browser and interaction becomes slow with larger values of np: we recommend starting with $N = 60, p = 20$.



(a) Posterior predictive pmfs: binomial model.

(b) Posterior predictive pmfs: beta-binomial model

Figure E.1: Joint and marginal posterior predictive probability mass functions (pmfs) of the cocaine purity classifier confusion matrix reported by (Rodrigues et al., 2013) and used by (Tötsch & Hoffmann, 2021) to illustrate confusion matrix uncertainty. These are screenshots from our interactive visualisation (see Appendix D.5).

E. Differences between binomial and beta-binomial models of uncertainty when data are scarce

To further illustrate the differences between the uncertainties conveyed by binomial and beta-binomial models, we consider the drug purity data from Rodrigues et al. (2013) used by Tötsch & Hoffmann (2021). Figure E.1 visualises the confusion matrix of 26 positive examples and 8 negative examples in which no false negatives were observed. The binomial model of the true positive rate places the entire probability mass at $\mathrm{TPR} = 1$ (Figure E.1a, bottom right), while the beta-binomial is more moderate, suggesting that a range of true positive values from 0.8–1.0 are plausible (Figure E.1b, bottom right). Tötsch & Hoffmann (2021)[Figure 4] explored uncertainty in this data by simulating 20 000 draws from the posterior predictive distribution of the beta-binomial model and using histograms to summarise the true positive and negative rates that were sampled. This work and the foundation that Caelen (2017) provided highlight the importance of modeling uncertainty in interpreting confusion matrices. We think that calculating and visualising the exact discrete probability mass functions conveys an even more meaningful and accurate appreciation of that uncertainty.

Do situations of data scarcity arise in the present era of "big data"? We believe so, and see two possible causes. The first is when the phenomenon of interest is genuinely (and often, mercifully) rare, as might occur in medical or health settings. The second arises in multinomial classification where the number of classes is relatively large in comparison to the total number of observations, leading to some classes with relatively few examples. Certainly, the number of publications devoted to learning from rare events and class-imbalanced data suggests that there are many researchers who are interested in working with scarce data (Haixiang et al., 2017). And, in the absence of additional information, the performance estimates of the classification models they build will be uncertain as a result of this scarce data, regardless of the performance metrics used.

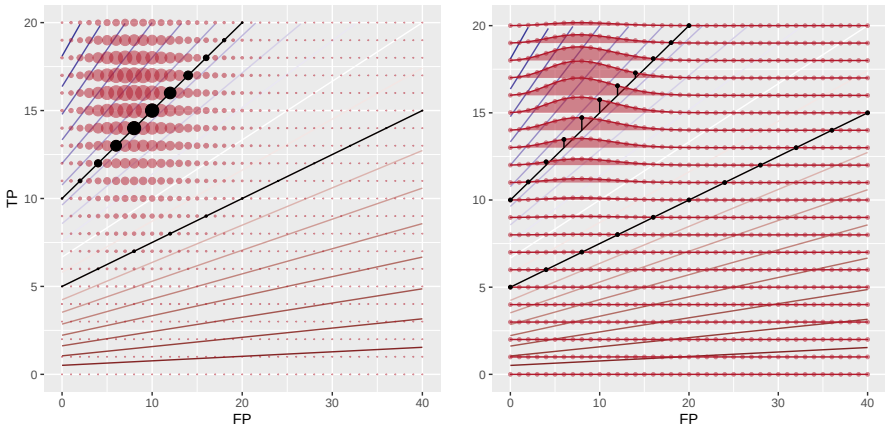


Figure F.1: Two ways to show confusion matrix pmfs and performance metric contours in ROC space. Both plots show the posterior predictive pmf of confusion matrices under a beta-binomial model of uncertainty for a classifier observed to produce the confusion matrix $\left[\begin{smallmatrix}16 & 8\\ 4 & 32\end{smallmatrix}\right]$ (the same as in Figure 4(b)). Like Figure 4, the left plot uses circle areas to represent probability mass; the right plot uses ridge lines. In the background are the contours of the F_1 performance metric and in black are the contours $\mathrm{F}_1 = \frac{4}{10}$ and $\mathrm{F}_1 = \frac{2}{3}$, along each of which lie 11 points in ROC space.

F. Visualising the conjunction of ROC points and metric contours

Having calculated and visualised the posterior predictive pmf of confusion matrices of size $N = p + n$ after observing a classifier's empirical performance, we can visualise the distribution of a given performance metric. Let $\mathbf{C}$ represent the confusion matrix random variable whose pmf is $P_{\mathbf{C}}(\mathbf{c})$, and let $M = \mu(\mathbf{C})$ represent the random variable we get by applying a performance metric function $\mu(\cdot)$ to a confusion matrix. The distribution of that performance metric is

$$P_M(m) = P(\mu(\mathbf{C}) = m) = \sum_{\mathbf{C}:\mu(\mathbf{C})=m} P_{\mathbf{C}}(\mathbf{c})$$

i.e., the probability mass at performance metric value m is sum of the probability masses where $\mu(\mathbf{C}) = m$. In other words, we find the pmf of the performance metric by summing the probability masses that lie along each contour of the performance metric in ROC space.

The geometry of performance metric contours in conjunction with the layout of the $(n + 1) \times (p + 1)$ possible points in ROC space determines which probability masses are summed together. Using F_1 for demonstration, Figure F.1 visualises the posterior predictive pmfs of confusion matrices in relation to performance metric contours. Plotting these probability masses against performance metric values gives posterior predictive distributions such as shown in the top and middle rows of Figures 4 and F.2.

Figure F.3 counts the number of ROC points that occur on the same performance metric contour. With confusion matrices of 20 positive and 40 negative examples (top row) certain MCC, BA and F_1 contours intersect multiple points in ROC space, most noticeably the 0 contours, which intersect 21 points with MCC and BA (along the $(0, 0), (1, 1)$ diagonal), and 41 points with F_1 (along the $(0, 0), (1, 0)$ horizontal).

One additional negative example (Figure F.3, bottom row) removes this confluence of points and contours in ROC space for MCC and BA, but not F_1. The MCC and BA pmfs change smoothly, following bell-shaped curves as we sweep across each row (i.e., true positive rate) of the joint pmfs in the bottom left panels of Figure 3 (a) and (b)—we emphasise this by using lines to connect these probabilities in Figure F.2, just like the ridge line plot in Figure F.1. The posterior pmf of F_1 remains much the same, due to the particular linear relationship between its contours in ROC space and the values of p and n (see Section B.2).

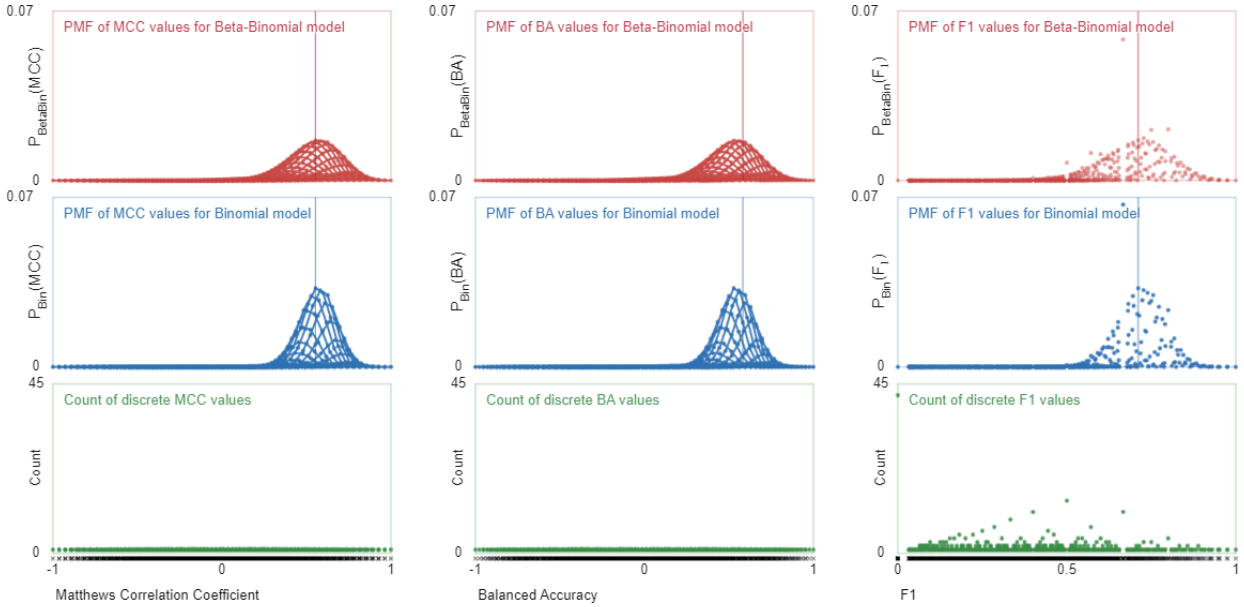


Figure F.2: These plots visualise the same quantities as Figure 4 but with a slightly different observed confusion matrix of 20 positive and *41* negative examples: $\begin{bmatrix} 16 & 9 \\ 4 & 32 \end{bmatrix}$, i.e., one more false negative. Note the changes in the counts of discrete MCC and BA values in comparison to Figure 4. To emphasise the now smoothly-changing pmf functions for MCC and BA, we use lines to connect points corresponding to the same true positive rate, (i.e., horizontal slices of the joint pmfs in Figure 3).

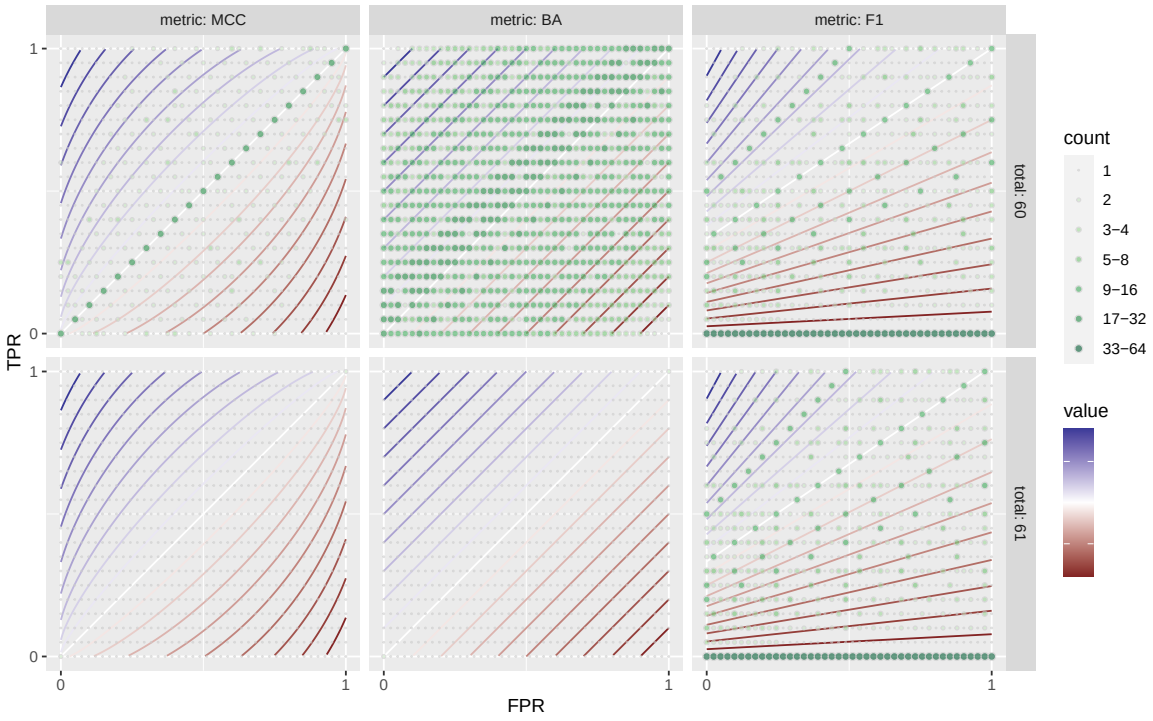


Figure F.3: Each panel shows all the possible points in the ROC space of confusion matrices of 20 positive and 40 negative examples (top row) and 20 positive and *41* negative examples (bottom row). Points are coloured by the number of times the performance metric value at that point is observed in the confusion matrices of those totals. Three different performance metrics are presented: MCC (left), BA (middle), F_1 (right). Performance metric contours are shown in the background, coloured by their value. Note that one additional negative example changes the configuration of possible points in ROC space so that each possible MCC and BA value is unique (bottom left and middle); the multiplicity of different F_1 values remains much the same. (bottom right).