Visualizing the operating range of a classification system

George Hripcsak

ABSTRACT

The performance of a classification system depends on the context in which it will be used, including the prevalence of the classes and the relative costs of different types of errors. Metrics such as accuracy are limited to the context in which the experiment was originally carried out, and metrics such as sensitivity, specificity, and receiver operating characteristic area—while independent of prevalence—do not provide a clear picture of the performance characteristics of the system over different contexts. Graphing a prevalence-specific metric such as F-measure or the relative cost of errors over a wide range of prevalence allows a visualization of the performance of the system and a comparison of systems in different contexts.

When comparing two systems, it is often desirable to summarize the performance of the systems into a single metric so that the systems can be ranked. This goal is elusive, however, because the performance of a system usually depends on the context in which it will operate. For a classification system, one of the most important contextual features is the prevalence of the classes. A system that determines whether or not a disease is present may perform well for common diseases but not as well for screening rare diseases because prevalence determines the number and types of errors. This paper addresses how to quantify and visualize performance with respect to prevalence.

Assume first a simple system that classifies any case as positive or negative for some condition, either by directly classifying the case into these two classes or by producing a continuous score that can be compared with a threshold, with cases exceeding the threshold being considered positive. Given a reference standard against which the researcher can judge performance, the result is the familiar two-by-two table, which is shown in table 1 along with the definitions of true positive (TP), false negative (FN), false positive (FP) and true negative (TN), which correspond to the four possible classification outcomes.

Table 2 summarizes common properties and performance metrics. Prevalence is the proportion of the population that truly has the condition. Accuracy is the proportion of cases that are correctly classified. Accuracy is a straightforward and understandable measure, but it is linked to the specific prevalence that happened to be in the experiment. If the system is used with a different prevalence, the accuracy will change.

Sensitivity is the proportion of cases with the condition that the system correctly identifies as positive, and specificity is the proportion of cases without the condition that the system correctly identifies as negative. These parameters do not vary with prevalence. Given a new context with a known prevalence, one can calculate the number of errors (FP and FN) expected. The error rate is commonly expressed as the positive predictive value and negative predictive value, in which the positive predictive value is the proportion of cases that the system deems positive that are truly positive, and the negative predictive value is the analogous measure for negative cases.

When comparing two systems without knowing the target prevalence, errors cannot be calculated. One can compare sensitivity and specificity directly, but one system may have higher sensitivity and the other may have higher specificity, leaving it ambiguous which is better. These two measures can be summarized using a receiver operating characteristic (ROC) curve, which is simply a graph of sensitivity versus specificity. Given a system that produces a continuous score, by varying the threshold over which cases are considered positive, one can produce a range of sensitivities and corresponding specificities. These sensitivity—specificity pairs can be plotted on the ROC curve (eg, figure 1). The area under such a curve quantifies performance. The area has the following interpretation: it is the probability that, given two cases in which one happens to have the condition and one lacks it, that the numeric score for the one with the condition will exceed the other’s score. For systems that do not produce a continuous score, the single sensitivity—specificity pair can be plotted as a point on the graph, and an area can be derived by taking the area under lines drawn from the point to the (0, 1) and (1, 0) corners of the graph.

Unfortunately, the ROC area is not always a reasonable summary of performance. It favors moderate values of sensitivity and specificity (eg, 0.1 to 0.9), which are effective for moderate prevalences but not for less common conditions. For example, in a sample of 100 000 cases, on average, a system with sensitivity 0.9 and specificity 0.9 performs well at a prevalence of 0.1, capturing 9000 of 10 000 positive cases with 1000 FN, and capturing 9000 additional FP errors. It thus captures most of the cases with the condition, and half of the cases that the system identifies as having the condition turn out to be truly positive. If the system is used for screening, then half the cases that undergo more extensive or manual review would have the condition. When the prevalence is 0.001, the system will capture 90 of 100 positive cases but it also capture 9000 FP, which swamp the 90 TP. If this system were used for screening,
Table 1 Two-by-two table

<table>
<thead>
<tr>
<th>Reference standard</th>
<th>System +</th>
<th>System −</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>−</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

FN, false negative; FP, false positive; TN, true negative; TP, true positive.

Figure 1 Receiver operating characteristic (ROC) curve. ROC curves for the four systems illustrated in figures 2 and 3. System 1 has no predictive value, system 2 is the arc of a circle, and systems 3 and 4 are single points on the curve.

We consider two prevalence-dependent performance metrics here: the F-measure and the expected cost of the FN and FP.

To illustrate the graphs, we consider four systems. One is a system with no predictive value. Its ROC curve is a diagonal line, and its ROC area is 0.5 (figure 1). It illustrates the minimum performance that one needs to settle for. The second one is an artificial system with an ROC curve, which is always above the diagonal, defined as the arc of a circle (figure 1):

$$F_{pr} = \max_{se, sp} \frac{2 \cdot se \cdot pr}{pr + se \cdot pr + (1 - sp)(1 - pr)}$$

where se is sensitivity and sp is specificity.

The last two systems are real systems published in the literature. Both are adverse event detection systems, and both produce a single classification. One has a sensitivity of 0.28 and specificity of 0.9996, and the other has a sensitivity of 0.77 and specificity of 0.94. The goal of the exercise is to demonstrate where each system should excel within which contexts.

In figure 2, the performance of each system is quantified as the system’s F-measure, which is calculated over a range of prevalences (plotted on a logit scale). For each prevalence, the plotted F-measure is the maximum of all the possible F-measures at that prevalence (formally, it is the supremum instead of the maximum because the endpoints with sensitivity or specificity equal to zero may not be defined).

Then only approximately one in 100 cases that undergo extensive review would have the condition. A different system, with sensitivity 0.5 and specificity 0.999, might perform better at the lower prevalence: it would identify 50 of 100 cases with the condition, but produce only 10 FP. Which system is better depends on the cost of missing cases with the condition (FN) compared with the cost of identifying non-cases (FP).

When classification systems are used for information retrieval, their performance is often quantified as recall and precision. Recall is generally defined to be sensitivity in these cases, and precision is defined to be the positive predictive value (although in some experiments, more complex definitions are used to account for partial matches). This combination of measures is particularly useful when the number of TN is either unknown or undefined. For example, in a study of a text processing system that identifies phrases, it may not make sense to count all possible phrases that the system did not identify, because they are overlapping.

Researchers frequently summarize recall and precision as the F-measure, which is the harmonic mean of recall and precision, with an optional weighting factor. In the case of a balanced (unweighted) F-measure, it has been shown that the F-measure is identical to the positive specific agreement. This is advantageous because the positive specific agreement can be shown to be a simple probability. It is the probability that if one rater (the reference standard or the system) is chosen randomly and if a case is said to be positive by that one, it will be identified as positive by the other. It is thus a measure of agreement on positive cases, but unlike sensitivity it accounts for both errors on positive cases and errors on negative cases; in this sense, it is a summary metric. One can correlate a set of sensitivity, specificity, and prevalence points with the expected positive specific agreement.

METHODS AND RESULTS

We propose to visualize the operating range of a system with respect to prevalence by graphing performance versus prevalence on a logarithmic axis. Use of a logarithmic axis highlights performance for screening when the prevalence of the condition is low.

Table 2 Performance metrics

<table>
<thead>
<tr>
<th>Prevalence</th>
<th>(TP + FN)/(TP + FN + FP + TN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>(TP + TN)/(TP + FN + FP + TN)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>TP/(TP + FN)</td>
</tr>
<tr>
<td>Specificity</td>
<td>TN/(TN + FP)</td>
</tr>
<tr>
<td>PPV</td>
<td>TP/(TP + FP)</td>
</tr>
<tr>
<td>NPV</td>
<td>TN/(TN + FN)</td>
</tr>
<tr>
<td>Recall</td>
<td>Sensitivity (expanded definitions exist for complicated experiments)</td>
</tr>
<tr>
<td>Precision</td>
<td>PPV (expanded definitions exist for complicated experiments)</td>
</tr>
<tr>
<td>ROC area</td>
<td>Area under the curve of sensitivity versus specificity from 0 to 1</td>
</tr>
<tr>
<td>F-measure</td>
<td>(1 + β2) precision-recall/(β2 precision + recall)</td>
</tr>
<tr>
<td>Positive specific agreement</td>
<td>2 TP/(2 TP + FN + FP) = F-measure when β = 1</td>
</tr>
</tbody>
</table>

FN, false negative; FP, false positive; NPV, negative predictive value; PPV, positive predictive value; ROC, receiver operating characteristic; TN, true negative; TP, true positive.
The performance of the hybrid system will describe reveals the relative performance (and proportions of FP and FN cases, with an optional weight to here is to minimize error. The error is de... to over 0.6.

maximizes precision. Therefore, for screening rare cases, this system will excel. System 4 is the best at intermediate prevalence, because the gain in sensitivity offsets the loss in specificity. The crossover between the two systems occurs around a prevalence of 5%. This crossover is not at all obvious from merely looking at the raw sensitivity and specificity values. System 2 does best at the highest prevalence, always exceeding system 1 because it exceeds system 1 at every point on the ROC curve.

Even without system 2, note that system 1, which has no predictive value, also exceeds both systems 3 and 4 at the highest prevalence. This is because when prevalence is high enough, it is better to review every case rather than use the system. That is, system 1 includes the endpoints (0, 1) and (1, 0). At sensitivity 1 and specificity 0, system 1 accepts all cases. At a prevalence near 1, the errors induced by systems 3 or 4 missing some truly positive cases exceed the error resulting from system 1 accepting all cases.

The researcher has the option to plot systems that have only one point either as a single point (as in figure 2) or as an ROC curve, which is composed of the two lines that connect that single point to the corners (0, 1) and (1, 0). This works because given any two systems each represented as a point on an ROC graph, a hybrid system can be created by randomly choosing one system or the other.4 The performance of the hybrid system will fall along a line connecting the two points, where the location depends on the probability of selecting one system or the other. If the researcher’s intent is actually to create a hybrid by randomly picking between the system and one of the corners, then calculating the maximum performance over the lines may make sense.

The second metric is cost, and is shown in figure 3. The goal here is to minimize error. The error is defined as the sum of the proportions of FP and FN cases, with an optional weight to reflect the relative cost of FP versus FN. For systems that produce a continuous score, the cost is minimized over all possible sensitivity—specificity pairs:

$$C_{pr,w} = (1/w) \sum_{i \in [0,1]} (1 - sp) \cdot (1 - pr)(1 - sp),$$

where w is the relative cost of FP compared with FN and $C_{pr,w}$ is the cost at prevalence, pr, and cost, w. Sox et al13 describe a method for finding the ideal point on the ROC curve given the relative cost of FP compared with FN and the prevalence, but it assumes a smooth, well-behaved curve. Empirically measured curves are often imperfect, so an alternative is to use numeric approximation to find the minimum cost.

Figure 3 shows system performance with a relative cost of 3 units for missing a positive case compared with 1 unit for choosing a negative case (w=1/3). Note that system 3 again performs best (lowest error) at low prevalence, and system 4 performs best at intermediate prevalence, with the crossover between the systems at approximately 4%. System 2 excels at the highest prevalence, and again, even without system 2, system 1 exceeds systems 3 and 4 at the highest prevalence. Changing w shifts the curves with respect to prevalence, but the overall profile remains the same.

Of the two graphs, the cost-based one reflects the trade-offs that the user is going to experience in practice, but the F-measure graph can be used when the future costs are not known. If the prevalence is known but not the cost, then the cost graph can be plotted as cost versus w.

Recent informatics publications of system performance include a mixture of publications of full ROC curves and single points on the curve. A prevalence graph (see supplementary data, available online only) of the five text mining systems published by Botstis et al13 reveals the relative performance (and similarity) of the systems, whereas relative performance is somewhat obscured by a table of numbers or even a plot on an ROC curve. A prevalence graph is less useful when the prevalence is more or less fixed, such as is the case for a search filter study by van der Glind et al14 because everyone shares the same Medline and thus the context is constant. A prevalence graph of the results of Wei et al15 confirms that the systems with higher specificities work better for low-prevalence screening, but the
study also illustrates that an ROC curve remains useful. The study’s Naïve Bayes system has significantly lower ROC area, but the curve itself reveals that the Naïve Bayes system performs similarly to the other two when it operates, but it has a significantly lower ROC area because it produces a single point whereas the others produce a curve. In fact, a non-parametric estimate of ROC area16 shows the three systems to be similar. This fact is obscured by both the original ROC area estimate and the prevalence graph.

Investigators have been studying how best to compare systems for many years.17 It is often assumed that the curves are well behaved, such that they are smooth and two ROC curves do not cross each other.17 If two curves do not cross each other, then the higher one will exceed the other over the entire prevalence graph. The graph thus differentiates systems when either the empirical measurement of the systems produces curves that cross, or when at least one of the systems can only be operated at a single sensitivity–specificity point.

Hand4 carries out an extensive analysis of the ROC area and suggests an alternative to the ROC area. Ultimately, however, the search for the perfect single measure is elusive. The performance of a system cannot actually be summarized in a single number. There are other alternatives to simple ROC curves for visualizing performance. Detection error trade-off curves, which use a logarithmic plot of the FP versus the FN rate, emphasize performance at the extremes of prevalence.18 The best approach to judging performance depends on the context. The prevalence graph may be useful in teaching ROC concepts, and it may be useful to grasp relative performance quickly in systems designed for screening.

CONCLUSION

In summary, graphing performance versus prevalence can illustrate the relative performance of several systems in different contexts. For a system that produces a continuous score, the metric has to be optimized (best threshold) at each prevalence. Unlike simple accuracy or F-measure, the graph is not limited to the context of the original experiment, and unlike sensitivity, specificity, and ROC area, the graph illustrates the trade-offs in different contexts.

Funding This work was funded by National Library of Medicine grant ‘Discovering and applying knowledge in clinical databases’ (R01 LM006910).

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES