Case Studies in the use of ROC Curve Analysis for Sensor-Based Estimates in Human Computer Interaction

James Fogarty Ryan S. Baker Scott E. Hudson

Human Computer Interaction Institute
Carnegie Mellon University

Abstract

Applications that use sensor-based estimates face a fundamental tradeoff between true positives and false positives when examining the reliability of these estimates, one that is inadequately described by the straightforward notion of accuracy. To address this tradeoff, this paper examines the use of Receiver Operating Characteristic (ROC) curve analysis, a method that has a long history but is under-appreciated in the human computer interaction research community. We present the fundamentals of ROC analysis, the use of the A' statistic to compute the area under an ROC curve, and the equivalence of A' to the Wilcoxon statistic. We then present several case studies, framed in the context of our work on human interruptibility, demonstrating how ROC analysis can yield better results than analyses based on accuracy. These case studies compare sensor-based estimates with human performance, optimize a feature selection process for the area under the ROC curve, and examine end-user selection of a desirable tradeoff.

Key words: ROC curves, A' statistic, sensor-based estimates, context-aware computing, interruptibility.

1 Introduction and Motivation

Context-aware systems, intelligent environments, and adaptive interfaces offer potential advances that have drawn significant interest from the human-computer interaction research community. These advances are based in part on the creation of systems that sense low-level features of a situation, use models of people and the world to infer higher-level concepts, and take action based on these estimates. For example, our work has examined the creation of sensor-based statistical models of human interruptibility in office environments, showing models can estimate human interruptibility as well as or better than human observers [3, 4, 5, 11]. These models can be used to inform many different approaches to managing interruptions [13]. In a mediated approach, models could inform the timing of a notification delivery. In a negotiated approach, models could inform the salience of the presentation used for a pending interruption.

When investigating the reliability of sensor-based estimates, the straightforward and common notion of accuracy has substantial shortcomings. These arise from the fact that a simple measure of accuracy does not account for the different types of mistakes a model might make. For example, our models of human interruptibility can fail to detect that a person is "highly non-interruptible" or they can falsely report a person as "highly non-interruptible". Accuracy obscures the difference between these two errors and ignores a fundamental tradeoff that often exists between them.

This tradeoff arises because most models output a score representing the degree to which the model believes that some condition is true, such as a person being "highly non-interruptible." Applications then compare this score to a threshold. If a lower threshold is used, more true positives will be detected, as in our work when an application correctly detects that a person is "highly non-interruptible." However, a lower threshold also means that more false positives will be generated, as in our work when an application mistakenly believes that a person is "highly non-interruptible." This tradeoff between obtaining more true positives at the expense of additional false positives is not conveyed by accuracy.

Without the ability to understand how this tradeoff applies to a model, both researchers and practitioners face considerable challenges in trying to understand the effect that sensor-based estimates may have on a user. Additionally, applications that use sensor-based estimates but ignore or minimize this tradeoff miss an opportunity to provide end-users with more control over how they interact with sensor-based systems.

In order to provide some insight into this problem, this paper examines the use of Receiver Operating Characteristic (ROC) curve analysis, a technique with a long history in signal detection [7] and medical diagnostics [10, 14] that has more recently drawn attention from the machine learning community [1, 9]. While ROC analysis is not completely unknown to the human computer interaction research community, it is also not widely used and seems to be of increasing relevance as the community investigates context-aware systems, intelligent environments, and adaptive

interfaces. Beyond contributing to an increased awareness of ROC analysis, this paper also contributes several case studies of how ROC curves can yield better results than an accuracy measure when applied to relevant problems in human computer interaction research. These case studies are framed in the context of our work on sensor-based statistical models of human interruptibility. Presenting ROC curves in the context of our research gives us a set of concrete examples to use, and the improved results we obtain in these case studies are also a contribution.

The next section introduces ROC curves, their computation, and statistically principled comparisons of the area under curves. We then begin our case studies by examining how ROC curves allow the performance of a model to be compared with human estimates across a range of confidence levels, showing that our sensor-based statistical models of human interruptibility perform significantly better than human observers. Our second case study shows how feature selection based on optimizing the area under an ROC curve can yield significantly better models than optimizing for simple accuracy. Next is our final case study, using a dialog appropriate for end-user threshold selection to examine how optimizing feature selection for the area under the ROC curve can result in a better set of tradeoffs for presentation to end-users. We then highlight a measure for extending ROC analysis to multiple-class problems. This is followed by a short discussion and conclusion.

2 ROC Curve Overview

The tradeoff at different thresholds between obtaining more true positives at the expense of additional false positives is visualized in an ROC curve by plotting the tradeoff for every possible threshold. This yields a curve like that in Figure 1, which presents the tradeoff for a sensor-based naïve Bayes model of human interruptibility. As when estimating accuracy, this plot is obtained by building a model from a set of training data and then evaluating the model against a set of test data, often within a cross-validation process. The output of the model for each case in the test data is then compared against each possible threshold, producing a point for each threshold in the plot. These points are plotted in a unit square, with the vertical location of the point for each threshold corresponding to the percentage of positive cases in the test data that are correctly labeled as positive when using the model at that threshold. The horizontal location of the point for each threshold is the percentage of negative cases in the test data that are incorrectly labeled as positive when using the model at that threshold. Note that this means neither axis represents possible thresholds, but rather the possible thresholds are distributed along the length

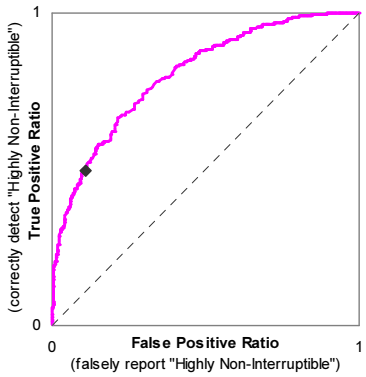

Figure 1 – ROC curve for a sensor-based statistical model of human interruptibility, with the .5 operating threshold highlighted.

of the curve. For example, we have marked the location of the .5 threshold in Figure 1, showing that classifying cases as positive when this model outputs a probability of .5 or greater detects 49.5% of "highly non-interruptible" situations, with false positives for 10.7% of negative cases.

Given this initial description, there are several characteristics of ROC curves worth nothing. All curves start in the bottom left corner, representing a threshold at which all cases are classified as negative, and end in the upper right corner, representing a threshold at which all cases are classified as positive. Better curves are closer to the upper-left corner (if one curve is above another at a given point on the horizontal axis, the higher curve is better at detecting true positives, while generating the same percentage of false positives as the lower curve). Curves should also always be above the diagonal (indicated as a dashed line in Figure 1), as a curve below the diagonal indicates that a model is generating more false positives than true positives (in which case, inverting the output of the model would provide a better model).

While the information presented in an ROC curve can help a researcher choose an appropriate threshold, ROC curves are especially appealing because they allow models to be compared independent of what threshold will be used in an application. When the curve of one model is completely above the curve of another model, it is clear that the model will perform better regardless of what threshold is used. But if two curves cross, the determination of which model is better again depends on what threshold will be used. While there is no single solution to this problem in the general case [8], many researchers have obtained good results using the area under the ROC curve as a single measure of the quality of a model. The area under an ROC curve also has very useful statistical properties, which we will discuss later in this section.

2.1 Computing an ROC Curve

While plotting a curve over every possible threshold may sound computationally expensive, the computation is actually very simple and inexpensive. A model is first evaluated against each case in the test data, outputting larger scores to indicate greater confidence that a case is positive. The cases are then sorted by their score. All of the points in the plot can then be computed in a single pass through the sorted cases. Each distinct score encountered in this pass represents a possible threshold. A point is plotted for that threshold based on what percentage of positive cases in the test data have scores greater than or equal to the threshold and what percentage of negative cases in the test data have scores greater than or equal to the threshold. Note that these counts of positive and negative cases can be maintained during the pass through the sorted cases, so they do not need to be computed from scratch at each threshold.

2.2 Area Under an ROC Curve

The area under an ROC curve is equal to the probability that a randomly selected positive case will receive a higher score than a randomly selected negative case.[1] In this section, we present the computation of this probability, and therefore the area under the ROC curve, using pair-wise comparisons. We focus on a tutorial presentation of the equations needed to work with and analyze ROC curves, and encourage readers interested in a more complete presentation of related theories to consult [10, 14]. Our presentation draws heavily from that in [10], though that presentation is in the context of medical diagnostics.

When using a set of test data to estimate the probability that a randomly selected positive case will receive a higher score than a randomly selected negative case, we compare the scores assigned by a model to each case in the test set. We define a function for comparing s_p, the score of a positive case, with s_n, the score of a negative case, as:

$$\mathrm{C}(s_p, s_n) = \begin{cases} 1 & \text{if } s_p > s_n \\ .5 & \text{if } s_p = s_n \\ 0 & \text{if } s_p < s_n \end{cases}$$

We then compute the average value of this comparison function over every pair of positive and negative cases:

$$A' = \frac{1}{n_p * n_n} \sum_{i=1}^{n_p} \sum_{j=1}^{n_n} \mathrm{C}(\mathrm{s}(P_i), \mathrm{s}(N_j))$$

where n_p and n_n are the number of positive and negative cases, s is the score of a case, and P and N are the sets of positive and negative test cases. The resulting estimate of the area under an ROC curve is known as A'.[2] As when plotting the ROC curve, A' can be computed in a single pass after sorting the cases in the test data by their scores.

Readers familiar with the Wilcoxon statistic, commonly used to compare the level of a quantitative variable in two populations, will recognize that the area under the ROC curve can be analyzed in terms of A' because A' is equivalent to the Wilcoxon statistic [10]. The Wilcoxon statistic is well-studied, and this equivalence means that a simple computation can be used to obtain the standard error for a given A', which we can then use to test the significance of a difference in the area under two ROC curves. Defining the terms D_p and D_n:

$$D_p = (n_p - 1)\left(\frac{A'}{2 - A'} - A'^2\right) \quad D_n = (n_n - 1)\left(\frac{2 * A'^2}{1 + A'} - A'^2\right)$$

the standard error for A' is:

$$\mathrm{SE}(A') = \sqrt{\frac{A'(1 - A') + \mathrm{D}_p + \mathrm{D}_n}{n_p * n_n}}$$

Given these formulas for A' and SE(A'), we can test the significance of a difference between the area under two ROC curves using a Z test, where Z is:

$$Z = \frac{A'_1 - A'_2.}{\sqrt{\mathrm{SE}(A'_1)^2 + \mathrm{SE}(A'_2)^2}}$$

In the case where we want to test whether a model is significantly more predictive than chance, we use $A'_2 = .5$ and $\mathrm{SE}(A'_2) = 0$. The significance of the Z value is then checked in a table.

2.3 ROC Curve Discussion

This section has presented ROC curves and A', the area under an ROC curve, together with statistical tests for examining the significance of A' and comparing values of A'. Presented visually, ROC curves allow inspection of a model's fundamental tradeoff between true

[1] In the interest of space, we do not prove this equality. Interested readers are encouraged to consult [7, 10]. For insight into the equality, consider that selecting a random negative case is equivalent to selecting a random location on the horizontal axis, which is equivalent to selecting a random threshold along the curve. At that threshold, the height of the curve is equal to the percentage of positive cases with a score greater than the threshold. Integrating across all values on the horizontal axis provides the probability across all cases.

[2] A related statistic, D', is sometimes used to estimate the area under an ROC curve. D' assumes the scores of positive and negative cases are being generated by overlapping normal distributions, while A' makes no assumptions about the underlying distributions.

positives and false positives, providing much more information than is conveyed by a straightforward notion of accuracy. When comparing two models, ROC curves make it clear that a curve entirely above another represents a model that will perform better regardless of what threshold is used. In the case where two curves cross, A' can be used as a measure of which model is better overall. We have also discussed how the relationship between A' and the Wilcoxon statistic allows us to use established methods to test whether a model is significantly more predictive than chance or than another model.

As an approach, ROC curve analysis has a history of being presented from one research community to another. While not completely unknown to the human computer interaction research community, ROC curve analysis is relatively uncommon and seems to be under-appreciated, so this section has presented what we believe are the most relevant properties of ROC analysis. In the coming sections, we illustrate how these properties can be applied to examining the reliability of sensor-based estimates.

3 Comparisons with Human Performance

A common problem when examining the reliability of sensor-based estimates is deciding what level of performance should be considered good. In the case of human interruptibility, it is clear that models should not be expected to perform perfectly. People cannot perfectly estimate interruptibility, and instead negotiate entry into interruptions after an initial judgment of whether an interruption is appropriate, with cues like eye contact avoidance or continuation of the task that would be interrupted indicating that an attempted interruption should be deferred [6]. When sensor-based estimates are similar in quality to estimates commonly made by people, comparing the performance of a model with human performance can support a compelling argument.

Because ROC curves are plotted on axes for the true and false positive rates, rather than with the range of possible scores along one axis, ROC curves support the comparison of estimates based on different ranges of scores. This property of ROC curves can be very useful to human computer interaction researchers, as statistical models typically output probabilities but people often find it difficult to effectively use a scale that contains more than five or seven values. The remainder of this section presents a case study comparison of the reliability of estimates made by human observers with the reliability of a sensor-based statistical model of human interruptibility. ROC curves support a comparison of the true positive versus false positive tradeoff in both types of estimates, providing a more complete comparison than the methods used in our prior work [3, 4]. This discussion also illustrates how human computer interaction researchers can use ROC curves to compare the performance of a model with the performance of people making a similar estimate.

	Human Observer Rating				
	Highly Non-Interruptible				Highly Interruptible
Self-Report	5	4	3	2	1
Other	**99** *(99)* *5.8%*	**219** *(318)* *18.8%*	**362** *(680)* *41.1%*	**516** *(1196)* *71.6%*	**498** *(1694)* *100.0%*
Highly Non-Interruptible	**250** *(250)* *35.4%*	**145** *(395)* *55.9%*	**101** *(496)* *70.3%*	**121** *(617)* *87.4%*	**89** *(706)* *100.0%*

Figure 2 – Observer estimates, by whether a self-report was "highly non-interruptible."

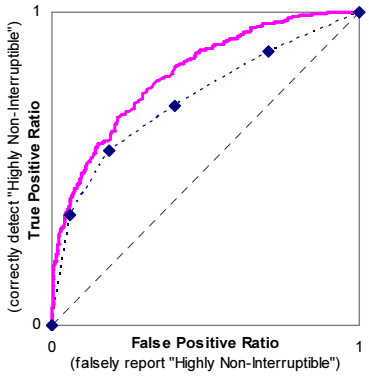

Figure 3 – Dashed curve for observer estimates and solid curve for a sensor-based statistical model.

3.1 Data Overview

Our work to examine human interruptibility has been largely based on collecting self-reports from office workers. At randomly selected times, the computer collecting our data played a recorded audio file prompting the participant to "rate your current interruptibility" on a five-point scale from "highly interruptible" to "highly non-interruptible." More than a third of self-reports have indicated that a participant was "highly non-interruptible," and so our work has focused on distinguishing "highly non-interruptible" situations from other situations [3, 4, 11].

To examine human estimates of interruptibility, we recruited observer subjects and showed them audio and video recordings from immediately before self-reports were collected. These observer subjects then estimated the interruptibility of the person in the recordings, using the same five-point scale [3]. Their estimates are shown in Figure 2. The top row shows estimates for situations self-reported as something other than "highly

non-interruptible," and the bottom row shows estimates for situations reported as "highly non-interruptible." The topmost number in each cell indicates how many times an observer rated each type of situation at that point on the five-point scale. The number in parentheses is a running sum from left to right as the rating scale decreases. The bottom number is ratio of the running sum to the total number of cases in the row. We include these ratios because they define the ROC curve for this data, with the percentages in first row providing horizontal values and the percentages in the second row providing vertical values, as seen in Figure 3, which we discuss next.

3.2 Analysis Comparison

In our prior work, we used accuracy to compare the performance of our statistical models with the performance of these human observers. The 1845 unshaded entries in Figure 2 correspond to a correct label of 5 for a "highly non-interruptible" situation or a correct label of a value other than 5 for a situation that was not "highly non-interruptible," for an accuracy of 76.9%. Using chi-squared tests, we have compared this accuracy to the accuracy of sensor-based statistical models [3, 4]. However, this comparison does not consider the relationship between human estimates and statistical models across the different possible thresholds.

Figure 3 presents a dashed ROC curve for the human observer estimates and a solid ROC curve for a sensor-based statistical model of the interruptibility of ten office workers with diverse responsibilities and working environments [4]. Note that the marked points on the human estimates curve correspond to the percentages shown in Figure 2, and the curve for the statistical model is entirely above the human estimates curve. These curves show that, for each point in the rating scale used by the human observers, the sensor-based statistical model used with the threshold that generates the same percentage of false positives will generate a higher percentage of true positives. Applying the statistical tests presented in the last section, we see that both the human observers ($A' = .724$, $Z = 18.7$, $p < .0001$) and the sensor-based statistical model ($A' = .804$, $Z = 19.0$, $p < .0001$) perform significantly better than chance. We can also see that the model is significantly more predictive than the human observers ($Z = 3.98$, $p < .0001$).

This case study has shown how ROC curves enable a more complete comparison of the performance of sensor-based estimates with human performance for similar estimates. We used a 5-point scale for human observers to indicate the degree to which a person was not interruptible, while our statistical models output probabilities. ROC curves allow these different types of scores to be compared, and A' allows us to test the significance of the difference in the area under two ROC curves. This ability to support comparisons of the performance of estimates based on different types of scores, and the resulting implications for making comparisons to human performance, seems especially useful to human computer interaction researchers.

4 ROC Curves in Feature Selection

By definition, statistical models are based on extracting correlations between dependent variables, generally referred to as features, and the variable being predicted, generally referred to as the class. A common approach is to create many potential features, then select the optimal subset of these potential features. To determine this optimal subset, a wrapper-based feature selection process starts with an empty set of features, then repeatedly adds or removes features until no change produces a better subset [12].

This feature selection process requires a metric for comparing the quality of potential feature subsets, and accuracy is commonly used. But optimizing for accuracy selects the best feature subset for the particular threshold used during selection, and using the selected features with a different threshold could result in sub-optimal performance. Because A', the area under an ROC curve, measures performance at all thresholds, selecting feature sets that optimize A' can be a better choice.

Figure 4 shows ROC curves for two sensor-based statistical models of human interruptibility, evaluated using our data collected from ten office workers with diverse responsibilities and working environments [4]. The solid curve is the same curve from Figures 2 and 3, representing a naïve Bayes model that was automatically created by `AmIBusy`, a system we are developing to support sensor-based statistical models of human interruptibility. The features in this model were selected in a wrapper-based optimization of A'. The dashed curve represents a model created by the same automatic process, except with a wrapper-based optimization of accuracy.

Considering performance at the .5 threshold, the model with a feature set optimized for accuracy correctly classifies 791, or 78.6%, of 1006 cases. The model with a feature set optimized for A' correctly classifies 771, or 76.6%, of 1006 cases. If we only considered accuracy at this threshold, we might decide that the first model was preferable. Inspecting the ROC curves in Figure 4, we see that the performance of the model optimized for accuracy is indeed slightly better at the .5 threshold (which we have marked), but it is also much worse for most of the curve. Using A' to

quantify this observation, we can see that the model optimized for accuracy has an A' of .740, significantly worse than the A' of .804 for the model optimized for A' ($Z = 2.68$, $p < .01$).

Considered in the terminology of ROC curves, optimizing feature selection for accuracy pulls a single point on an ROC curve towards the upper-left corner, but ignores every other point on the curve. The feature selection process also has no motivation to raise the score of a test case any higher than the threshold that is being used. In the case of feature selection with a .5 threshold, accuracy makes no distinction between a feature set that assigns a positive case a probability of .51 or a feature set that assigns the positive case a probability of .99. Slight differences in the data encountered when the model is deployed might then result in the .51 probability slipping below .5, unnecessarily reducing the reliability of the model. Optimizing A' allows feature selection to consider performance at all thresholds. A statistical model that assigns a positive case a probability of .99 will have a higher A' than if the model assigned a probability of .51, and so the feature selection process can choose features that maximize the separation between positive and negative cases. This case study has shown how these differences apply to human interruptibility, and the next section considers how they can manifest in an interface for end-user threshold selection.

5 Supporting End-User Threshold Selection

The fundamental tradeoff between true positives and false positives is inherent to applications that use sensor-based estimates, and applications that ignore or minimize this tradeoff are missing an opportunity to give end-users control over how they interact with sensor-based systems. By accounting and designing for this tradeoff, applications can enable end-user selection of the most desirable tradeoff. In the case of interruptibility, for example, many office workers might feel that they are interrupted too often, and so they might choose a threshold that aggressively minimized the salience of notifications delivered by an application (thus maximizing true positives). Other office workers might feel that the information conveyed by the application's notifications is sufficiently important that they prefer salient notifications, with notifications being deferred or presented more subtly only when it is very clear that they are not interruptible (thus minimizing false positives).

While ROC curves are a powerful tool for analyzing and understanding a model's tradeoff between true positives and false positives, they do not seem to be a good choice for presentation to end-users. ROC curves require significant explanation before they can be interpreted, as evidenced by the first several pages of this paper. ROC curves also do not convey the relative frequency of positive and negative cases in a data set[3]. In the case of interruptibility, knowing how often a model considers a person interruptible might inform that person's selection of a threshold.

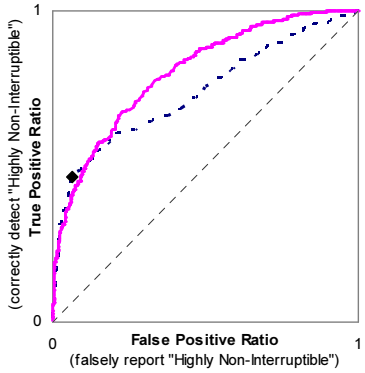

Figure 4 – Dashed curve for model optimized by accuracy and solid curve for model optimized by A'.

To examine how differences in models optimized for accuracy and models optimized for A' manifest in an interface for end-user threshold selection, we created the interface shown in Figures 5 and 6. This dialog is intended for a notification application using a mediated approach to manage interruptions caused by non-urgent notifications. Rather than using an ROC curve to present the tradeoff between preventing inappropriate interruptions (true positives) and unnecessarily delaying notifications (false positives), this dialog shows how many inappropriate interruptions are prevented at a given threshold, how many appropriate notifications are delivered, and the overall accuracy of the model for each threshold. The dialog presents both percentages and absolute scales, in order to convey the relative frequency of positive and negative cases from which the model has been constructed. Figure 5 presents this threshold selection dialog for the model optimized for accuracy in the previous section, while Figure 6 is for the model with a feature set optimized for A'.

As we might expect after analyzing the ROC curves presented in Figure 4, the model optimized for accuracy and shown in Figure 5 has an overall accuracy with a well-defined peak at the .5 threshold. An end-user unhappy with the tradeoff at the .5 threshold must accept a lower overall accuracy in order to adjust the tradeoff. In contrast, the model in Figure 6 with a feature set optimized for A' shows a plateau for overall

[3] This is an intentional characteristic of ROC curves, and not a shortcoming. To see why, consider assessing the reliability of a test for which the positive case occurs in less than one percent of a population. [10, 14].

accuracy over a relatively large range of thresholds. Within this plateau, different tradeoffs between preventing interruptions and delivering notifications result in the same overall accuracy. Even when this plateau drops off on the left side of the chart, the overall accuracy remains higher than in the model optimized for accuracy.

This section has presented a dialog appropriate for end-user threshold selection and examined how models optimized for A' and accuracy differ when presented in such a dialog. Designing for end-user threshold selection allows end-users to choose the tradeoff that they find most desirable, and models optimized for A' can provide a better set of alternatives.

6 Multiple Class ROC Analysis

Throughout this paper, we have focused on two-class problems, where sensor-based estimates are being used to choose between two possible alternatives. Two-class problems are very common, in part because existing machine learning algorithms tend to work best with two classes. However, applications sometimes require that a model choose from among three or more alternatives, commonly referred to as a multiple-class problem. This section presents some brief comments on using the measure M for multiple-class ROC analysis [9].

For a problem with c different class values, Hand and Till present the measure M:

$$M = \frac{1}{c(c-1)} \sum_{i=1}^{c} \sum_{j=i+1}^{c} A'(i \mid j) + A'(j \mid i)$$

where $A'(i \mid j)$ is A' for the subset of test cases with class i or j, using class i as the positive class. In the case of a two-class problem, $A'(i \mid j)$ is equal to $A'(j \mid i)$ and M is equal to A'. But $A'(i \mid j)$ and $A'(j \mid i)$ must both be used when examining multiple classes, as A' is not symmetrical when examining two classes from a multiple-class problem.

M has several good properties. As with A', M is computationally inexpensive to compute. Because M is based on the ability of a model to distinguish between pairs of classes, it is useful in the case where one or more classes in a multiple-class problem cannot be reliably detected (in contrast, a measure based on the ability of a model to detect every class could degenerate if one or more classes cannot be reliably separated from the other classes). Optimizing with respect to M will select features that maximize the separation between the classes that a model can detect. The major shortcoming of M relative to A' is the lack of an equation for the standard error of M, making it more difficult to test the significance of a difference between two values of M. Hand and Till suggest the use of

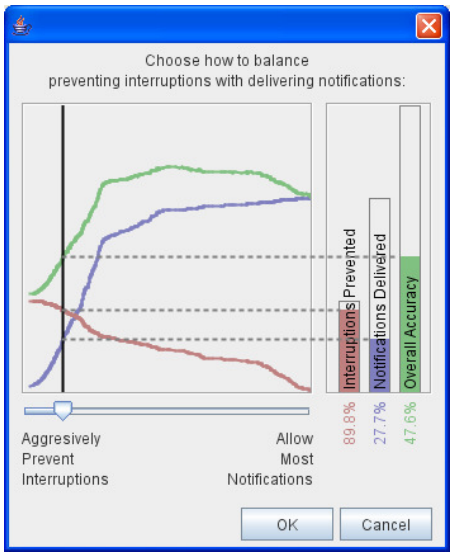

Figure 5 – Threshold selection dialog for a model optimized by accuracy.

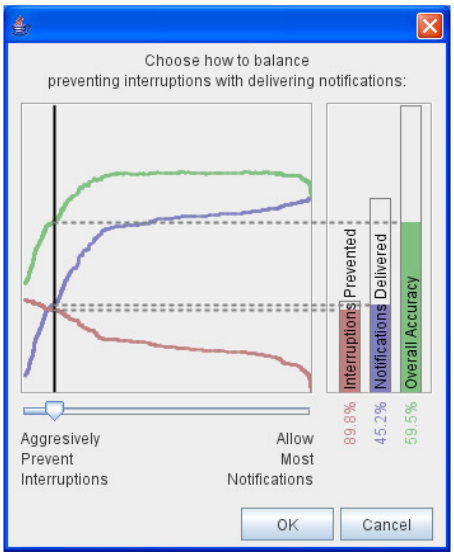

Figure 6 – Threshold selection dialog for a model optimized by A'.

bootstrap resampling methods [2]. Readers interested in further discussion of M and multiple-class ROC analysis are encouraged to use [9] as a starting point.

Discussion and Conclusion

This paper has presented ROC curves, the use of A' to analyze the area under an ROC curve, and how the equivalence of A' to the Wilcoxon statistic allows statistically principled comparison of the area under ROC curves. ROC analysis allows a principled examination of the reliability of sensor-based estimates across all possible thresholds, rather than the

single-threshold examination that is supported by accuracy. Considering performance across possible thresholds is important, as the tradeoff between obtaining more true positives at the expense of additional false positives is inherent to using a threshold with sensor-based estimates. Embracing and designing for this tradeoff, rather than ignoring or minimizing it, can also provide end-users with more control of how they interact with sensor-based estimates.

While some members of the human computer interaction research community are familiar with ROC analysis, its relevance seems to be under-appreciated as the community investigates context-aware systems, intelligent environments, and adaptive interfaces. The relationship between A' and the Wilcoxon statistic, which enables a straightforward test for the significance of a difference between the area under two ROC curves, seems to be particularly under-appreciated.

We have presented several case studies showing that ROC analysis, used with our work on sensor-based statistical models of human interruptibility, yields better results than we have previously obtained with accuracy. Comparing estimates of interruptibility made by human observers with estimates provided by our sensor-based statistical models, we have shown that our statistical models perform better at each of the confidence levels available to the human observers. Comparing the A' for the estimates collected from human observers and our sensor-based statistical model, we have shown that that our model is significantly more predictive than human observers. This is a more compelling comparison than our prior analysis using accuracy, as ROC analysis accounts for the varying degrees of certainty that people can have when making such estimates. Our second case study shows how a feature selection process that optimizes A' can yield better models than a process that optimizes accuracy. While accuracy considers the performance of a model at only a single threshold, the additional information available in A' can be used by a feature selection process to maximize the separation between scores assigned to positive and negative test cases. Because ROC curves can initially be difficult to interpret, our final case study examined how differences between A' and accuracy can manifest when data is examined in a dialog appropriate for end-user threshold selection. Each of these case studies contributes by improving upon the results we previously obtained with accuracy or by demonstrating how human computer interaction researchers can apply ROC curve analysis to relevant problems.

Acknowledgements

The authors would like to thank Judy Olson and Rachel Roberts for prompting their initial investigations into how ROC analysis can better address our research problems. We would also like to thank Darren Gergle for his comments at various points during the preparation of this paper. This work was funded in part by DARPA and by the National Science Foundation under Grants IIS-0121560, and IIS-0325351.

References

1. Bradley, A.P. (1997) The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, *30*. 1145-1159.
2. Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. Chapman & Hall, London.
3. Fogarty, J., Hudson, S., Atkeson, C., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J. and Yang, J. (2005) Predicting Human Interruptibility with Sensors. *To Appear, ACM Transactions on Computer-Human Interaction (TOCHI)*.
4. Fogarty, J., Hudson, S. and Lai, J. (2004) Examining the Robustness of Sensor-Based Statistical Models of Human Interruptibility. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2004)*, 207-214.
5. Fogarty, J., Ko, A.J., Aung, H.H., Golden, E., Tang, K.P. and Hudson, S.E. (2005) Examining Task Engagement in Sensor-Based Statistical Models of Human Interruptibility. *To Appear, Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2005)*.
6. Goffmann, E. On Facework. In Goffmann, E. ed. *Interaction Ritual*, Random House, New York, 1982, 5-45.
7. Green, D. and Swets, J. *Signal Detection Theory and Psychophysics*, John Wiley and Sons, New York, 1966, 45-49.
8. Hand, D.J. (1997) *Construction and Assessment of Classification Rules*. Wiley, Chichester.
9. Hand, D.J. and Till, R.J. (2001) A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, *45* (2). 171-186.
10. Hanley, J.A. and McNeil, B.J. (1982) The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, *143*. 29-36.
11. Hudson, S., Fogarty, J., Atkeson, C., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J. and Yang, J. (2003) Predicting Human Interruptibility with Sensors: A Wizard of Oz Feasibility Study. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2003)*, 257-264.
12. Kohavi, R. and John, G.H. (1997) Wrappers for Feature Subset Selection. *Artificial Intelligence*, *97* (1-2). 273-324.
13. McFarlane, D.C. (2002) Comparison of Four Primary Methods for Coordinating the Interruption of People in Human-Computer Interaction. *Human-Computer Interaction*, *17* (1). 63-139.
14. Metz, C.E. (1978) Basic Principles of ROC Analysis. *Seminars in Nuclear Medicine*, *8* (4). 283-298.