5-1-2015

Estimating the Accuracy of Automated Classification Systems Using Only Expert Ratings that are Less Accurate than the System

Paul E. Lehner

The MITRE Corporation, plehner@mitre.org

Estimating the Accuracy of Automated Classification Systems Using Only Expert Ratings that are Less Accurate than the System

Estimating the Accuracy of Automated Classification Systems Using Only Expert Ratings that are Less Accurate than the System

Paul E. Lehner
The MITRE Corporation
McLean, VA, USA

A method is presented to estimate the accuracy of an automated classification system based only on expert ratings on test cases, where the system may be substantially more accurate than the raters. In this method an estimate of overall rater accuracy is derived from the level of inter-rater agreement, Bayesian updating based on estimated rater accuracy is applied to estimate a ground truth probability for each classification on each test case, and then overall system accuracy is estimated by comparing the relative frequency that the system agrees with the most probable classification at different probability levels. A simulation analysis provides evidence that the method yields reasonable estimates of system accuracy under diverse and predictable conditions.

Keywords: Inter-rater reliability, Kappa, artificial intelligence

Introduction

Information technology is advancing to develop systems that address problems of increasing sophistication and complexity. A quick scan of programs sponsored by research funding agencies (e.g., www.nih.gov, www.nsf.gov, www.darpa.mil, www.iarpa.gov) showed new systems being developed to address complex problems as diverse as automated medical and clinical diagnoses, technology readiness evaluation, detection of emerging technologies, classification of the behavioral contents of unstructured video segments, recognition and classification of metaphors used in natural language text and many others.

The complexities of the problems that these advanced systems address make it difficult to evaluate the accuracy of such systems. It is usually necessary to

---

Dr. Lehner is a Consulting Scientist with The MITRE Corporation. Email him at plehner@mitre.org.

resort to using expert raters to assign ground truth for test cases. However, the complexity of these problems also challenge to the expert raters. Raters often disagree as to which is the correct category. Furthermore as future systems address problems of ever increasing sophistication and complexity, it seems likely that the experts will be even more challenged and exhibit even lower levels of agreement. Ground truth data sets based on expert assignments are fallible and are likely to become more so in the future.

Using expert raters to assign ground truth to test cases is a well-established practice. For classification problems, which are the focus of this paper, a statistic such as Kappa is used to measure inter-rater agreement; and then the rating process is refined until a satisfactory level of agreement is reached. Once the agreement threshold is reached, assignments of individual raters or collaborating teams of raters are treated as truth and system accuracy is measured by the level of agreement with the assigned ground truth (See Gwet, 2010 for review).

For several reasons, this common scientific practice does not adequately meet the needs of advanced system evaluation. First, the level of agreement amongst raters will rarely meet a satisfactory level. The problems that these systems address are simply too complex. About the only way to increase the level of agreement is to select relatively simple and therefore non-representative test cases.

Second, estimating system accuracy by measuring the level of agreement with expert raters makes the *de facto* assumption that the experts are more accurate than the system. This assumption runs contrary to a substantial body of empirical research where it is often found that simple algorithms outperform human experts in complex judgments (Dawes, 1979; Grove, Zald, Lebow, Snitz, & Nelson 2001; Tetlock, 2005). It should not be presumed that the experts are more accurate than the system.

Third, there is considerable evidence to suggest that for a wide variety of judgment tasks collaborative team judgments are not substantially more accurate than the judgments of randomly selected individual team member (e.g., Surowiecki, 2005; Armstrong, 2006). In judgment tasks, where there is no obvious correct answer, it should not be presumed that collaboration will reliably lead the raters to converge to the correct answer.

Finally, when evaluating a classification system the statistic of greatest interest is the accuracy of the system - the proportion of system assignments that are correct. Unfortunately there is an unclear relationship between inter-rater reliability statistics such as Kappa, the probability of correct ground truth

assignments and the accuracy of any systems tested against error-prone ground truth assignments.

A different approach is presented here to using expert ratings to estimate the accuracy of classification systems. Rather than treat expert ratings as a surrogate for ground truth, expert ratings are treated as error prone estimates of ground truth where independent ratings are fused to estimate ground truth probabilities, and the ground truth probabilities are then used to estimate system accuracy.

One practical instantiation of this estimation approach is described below. In addition simulation test results are provided that support several claims. First, under diverse conditions, this approach reliably yields estimates of system accuracy that are approximately correct. If a system is 90% accurate then this approach will yield an estimate of system accuracy that is close to 90%. Second, the accuracy of the estimate of system accuracy is largely independent of whether the expert raters are more or less accurate than the system. If a system is in fact 90% accurate, and the raters are individually 60% accurate, then the estimate of system accuracy will still be approximately 90%. Third, reliable estimates of system accuracy can often be obtained with a reasonably small number of test cases (e.g. fifty test cases with three expert raters). In complex domains it is important to keep sample sizes as small as possible, since it may be time consuming and costly to obtain expert ratings. Fourth, and importantly, the conditions under which the above three claims may break down are predictable. Therefore test data sets can be intentionally constructed to ensure that the conditions are met that are needed for accurate estimation of system accuracy.

Estimating the accuracy of system classifications

The method for estimating accuracy described below was derived from the following assumptions.

AA1. For each case there is a unique correct classification.

AA2. For each case raters independently assign classifications.

AA3. Expected agreement between raters increases as expected rater accuracy increases.

Assumption AA3 refers to *expected* agreement and accuracy. Here "accuracy" refers to the total proportion of correct classifications made by all the raters, irrespective of which raters are making correct and incorrect classifications. And "agreement" refers to the total proportion of pairwise agreement among all of

the raters and cases. For any particular set of cases, accuracy may be low yet agreement high (the raters made the same mistakes), but AA3 asserts that *in general* there is an expected positive relationship between accuracy and agreement.

Theorem 1:

AA1-AA3 are ensured if and only if the raters behave as though their selection for each case is determined by a single confusion matrix where the conditional probability of correct assignment is constant and the conditional probability of all incorrect assignments is equal.

That is to say all raters on all problems are selecting from a single confusion matrix with a structure such as shown in Table 1.

The proof of this theorem is found in the Appendix. The general structure of the proof shows that if the raters are assigning classifications using any process other than selecting from a common confusion matrix with the structure illustrated in Table 1, then it is always possible to construct a classification process with lower expected accuracy and higher expected agreement, or higher accuracy and lower agreement; thereby violating the assumed monotonic relationship between expected accuracy and expected agreement.

Table 1. Implied Structure of Rater Confusion Matrices for Four Category Problem (A to D are true categories and "A" to "D" are selected categories.)

	"A"	"B"	"C"	"D"
A	P_c	$(1-P_c)/3$	$(1-P_c)/3$	$(1-P_c)/3$
B	$(1-P_c)/3$	P_c	$(1-P_c)/3$	$(1-P_c)/3$
C	$(1-P_c)/3$	$(1-P_c)/3$	P_c	$(1-P_c)/3$
D	$(1-P_c)/3$	$(1-P_c)/3$	$(1-P_c)/3$	P_c

AA1 through AA3 also seem to be assumed implicitly in many contexts where the Kappa statistic is applied. Indeed it is AA3 that would seem to warrant the common practice of using expert ratings as surrogates for ground truth when high levels of inter-rater agreement are found. Consequently it is reasonable to claim that the estimation method described below is derived from assumptions implicit in the Kappa statistic and how Kappa is often used. Because of this relationship to the Kappa statistic, in the remainder of this paper AA1-AA3 will be referred to as *K-assumptions*. Furthermore, the properties of equal rater

accuracy, equal error probabilities and equal problem difficulty that are implied by the *K-assumptions* will be referred to as *K-properties*.

Table 2. Sample data of expert ratings and system assignments for 10 test cases

Case #	Rater 1	Rater 2	Rater 3	Rater 4	System
1	"C"	"D"	"C"	"C"	"A"
2	"B"	"D"	"C"	"C"	"C"
3	"C"	"C"	"D"	"C"	"C"
4	"B"	"B"	"D"	"D"	"B"
5	"A"	"B"	"B"	"B"	"B"
6	"C"	"B"	"D"	"A"	"A"
7	"A"	"A"	"A"	"A"	"A"
8	"A"	"D"	"B"	"C"	"C"
9	"D"	"B"	"A"	"A"	"D"
10	"A"	"D"	"A"	"B"	"B"

The estimation method is straightforward to explain in the context of an example. Consider the test data in Table 2. There are 10 test cases, 4 categories, 4 raters and the system's proposed answers. When referring to ground truth the four categories are labeled *A, B, C, D*; when referring to rater and system assignments they are labeled "A", "B", "C", "D".

As described below the estimation method is composed of four basic steps.

Estimate rater accuracy

Given that each rater has an identical confusion matrix, with the structure shown in Table 1, the probability that two raters will agree on any one case is

$$P_a = P_c^2 + \frac{(1-P_c)^2}{N-1} \tag{1}$$

Here P_a is the probability of agreement, P_c is the probability that a rater will make the correct assignment, and N is the number of categories. Solving for P_c yields

$$P_c = \left(\frac{1}{N}\right) + \sqrt{\left(\frac{(N-1)*P_a - \dfrac{N-1}{N}}{N}\right)} \qquad (2)$$

Eq. 2 is used to estimate rater accuracy. In the 10 cases in Table 1 there was 33% agreement (20 pairs out of 60). Setting P_a to .33 and solving for P_c yields $P_c = 0.5$; which is the estimate of rater accuracy.

Estimate base rates

The probability that a rater will assert a category, say "A", is as follows:

$$P(\text{"A"}) = P(\text{"A"}|A)*P(A) + \left(1 - \frac{P(\text{"A"}|A)}{N-1}\right)*(1-P(A)) \qquad (3)$$

Here $P(\text{"A"})$ is the marginal probability that the rater asserts "A", $P(\text{"A"}|A)$ is the conditional probability that the rater will assert "A" if A is true, and $P(A)$ is the marginal probability of A. Solving for $P(A)$ yields

$$P(A) = \frac{(N-1)*P(\text{"A"}) - 1 + P(\text{"A"}|A)}{N*P(\text{"A"}|A) - 1} \qquad (4)$$

Setting $P(\text{"A"})$ to be the observed relative frequency of "A", and $P(\text{"A"}|A)$ to be the estimate of P_c from above, yields

$$P(A) = \frac{(N-1)*P(\text{"A"}) - 1 + P_c}{N*P_c - 1} \qquad (5)$$

Eq. 5 is used to estimate the base rate for each category by setting P_c to be the estimate from above and $P(\text{"X"})$ to be the observed relative frequency across all raters and ratings that category X was assigned. In Table 1 there are 11 instances of each of the categories; so the estimated base rate is 0.325 for category A. Applying Eq. 5 to the other categories yields base rates of 0.25, 0.25 and 0.175 for B, C and D respectively.

Estimate ground truth probabilities

Use Bayes rule, assuming conditional independence for each rater, to estimate ground truth probabilities. For example, in case 1 above the raters selected "CCDC". So for each possible ground truth value calculate $P(\ldots|\text{"CDCC"})$ and normalize.

$$P(A|\text{"CDCC"}) \sim P(A) * P(\text{"C"}|A) * P(\text{"D"}|A) * P(\text{"C"}|A) * P(\text{"C"}|A)$$
$$= .325 * .167 * .167 * .167 * .167 = .00025 \rightarrow .041$$
$$P(B|\text{"CDCC"}) \sim P(B) * P(\text{"C"}|B) * P(\text{"D"}|B) * P(\text{"C"}|B) * P(\text{"C"}|B)$$
$$= .25 * .167 * .167 * .167 * .167 = .00019 \rightarrow .032$$
$$P(C|\text{"CDCC"}) \sim P(C) * P(\text{"C"}|C) * P(\text{"D"}|C) * P(\text{"C"}|C) * P(\text{"C"}|C)$$
$$= .25 * .5 * .167 * .5 * .5 = .00521 \rightarrow .860$$
$$P(D|\text{"CDCC"}) \sim P(D) * P(\text{"C"}|D) * P(\text{"D"}|D) * P(\text{"C"}|D) * P(\text{"C"}|D)$$
$$= .175 * .167 * .5 * .167 * .167 = .00041 \rightarrow .067$$

Repeating this step for the other 9 cases yields the estimated probability distributions shown in Table 3.

Table 3. Estimated ground truth probabilities for sample data

Case #	Ground Truth Probability				System Answer
	A	B	C	D	
1	0.041	0.032	0.860	0.067	"A"
2	0.084	0.195	0.584	0.136	"C"
3	0.041	0.032	0.860	0.067	"C"
4	0.074	0.511	0.057	0.358	"B"
5	0.120	0.828	0.031	0.021	"B"
6	0.325	0.250	0.250	0.175	"A"
7	0.975	0.009	0.009	0.006	"A"
8	0.325	0.250	0.250	0.175	"C"
9	0.657	0.169	0.056	0.118	"D"
10	0.657	0.169	0.056	0.118	"B"

Estimate system accuracy

Assume any probability distribution over the categories for each test case. For any test case, let P_g be the probability of the classification with the highest probability,

P_s be the probability that the system will assign the correct answer, P_a be the probability that the system will assign the same classification as the highest ground truth probability. It follows that

$$P_a = P_g * P_s + \left(1 - P_g\right) * \frac{1 - P_s}{N - 1} \qquad (6)$$

Note that this relationship holds whether or not the classification with the highest probability is correct. Solving for P_s yields

$$P_s = \frac{\left(N - 1\right) * P_a - 1 + P_g}{N * P_g - 1} \qquad (7)$$

Eq. 7 is used to estimate system accuracy as follows. First separate the test cases into bins with approximately the same highest estimated ground truth probability. In this paper the ranges (.9, 1.0], (.8, .9], (.7, .8], etc. are used. For example, in Table 3 there is one case in the (.9, 1.0] range, 3 cases in the (.8, .9] range, 2 cases in the (.6, .7] range, etc. Second for each bin calculate the average ground truth probability within the bin; record the proportion of system assignments that agree with the most probable answer; then estimate system accuracy for each bin using equation Eq. 7. Third estimate overall system accuracy by taking the average of the estimated accuracy in each bin weighted by the number of cases in each bin. This is shown in Table 4.

Table 4. Estimate of System Accuracy for Sample Data

Probability Bin	Average Ground Truth Probability	Number in Bin	Proportion of Agreement	Estimated Accuracy
.9 – 1.0	0.975	1	1.000	1.000
.8 - .9	0.849	3	0.667	0.776
.6 - .7	0.657	2	0.000	0.000
.5 - .6	0.548	2	0.333	0.452
.2 - .3	0.325	2	0.500	1.000
			Weighted Average =	0.731

The reader may be curious as to why the estimate of system accuracy is not simply the average of the estimated ground truth probabilities for the system answers. The reason is that taking the average will consistently underestimate

system accuracy; because the system's answer is itself additional evidence for each category. So, for example, if the system answer is "C" and the estimated ground truth probability for C is 0.6; then a better estimate for C would be somewhat higher than .6. But until system accuracy is estimated it cannot be determined how much more than .6 is appropriate. In the above example, the average estimated ground truth probability of the system answers is .466, but the estimate of system accuracy in Table 4 is 0.731.

Note that the value of Kappa (using 1/number-categories to determine random agreement) for the data in Table 2 is

$$\text{Kappa} =$$
$$= \left(\text{Observed Agreement - Random Agreement}\right)/\left(1.0 - \text{Random Agreement}\right)$$
$$= \left(.333 - .25\right)/\left(1 - .25\right) = 0.11$$

Standard thresholds normally require a level of $\text{Kappa} = 0.7$ before the expert ratings are considered usefully reliable (Gwet 2010). $\text{Kappa} = 0.11$ is considered "slight agreement" and is far too low for the ratings to be considered useful for establishing ground truth.

Overall then, in the sample data provided in Table 2; inter-rater agreement is "slight" ($\text{Kappa} = 0.11$), estimated rater accuracy is 0.50, and estimated system accuracy is 0.731.

Performance and robustness

The above example illustrates how to estimate system accuracy for classification problems even when inter-rater agreement and estimated rater accuracy are very low. This section examines the accuracy of estimates of system accuracy, and the robustness of those estimates, through a series of simulations.

All of the simulations described below use the following procedure to assign the confusion matrix for each rater and the system, based on values set to four parameters: an initial probability of correct assignment (IPC), a problem difficulty adjustment (PDA), degree of asymmetric dispersion (AD), and a proportional error range (PER).

Each confusion matrix is constructed as follows:

1. Initially assign the conditional probability of a correct classification to be IPC for all categories.

2. Add PDA to the conditional probabilities of correct assignment.
3. For each category distribute the remaining probability (1 - IPC - PDA) to the incorrect classifications in a manner that is proportional to the distance from the correct classification, where the probability of a classification that is M steps removed from the correct classifications is AD times more likely than a classification that is M+1 steps removed.
4. For each conditional probability of incorrect assignment (IC) set the range to be [IC - PER*IC, IC + PER*IC], then randomly select a new probability by uniform sampling over this range.
5. Normalize the modified confusion matrix after the random changes in step 4 so that expected accuracy is equal to IPC + PDA.

For example, if there are five categories and (IPC, PDA, AD, PER) = (.6, 0, 1.0, 0), then the resulting confusion matrix is shown in Table 5.

Table 5. Confusion matrix where (IPC, PDA, AD, PER) = (0.6, 0, 1.0, 0)

Correct Category	Classification				
	"A"	"B"	"C"	"D"	"E"
A	0.6	0.1	0.1	0.1	0.1
B	0.1	0.6	0.1	0.1	0.1
C	0.1	0.1	0.6	0.1	0.1
D	0.1	0.1	0.1	0.6	0.1
E	0.1	0.1	0.1	0.1	0.6

On the other hand, if (IPC, PDA, AD, PER) = (.6, -.2, 2.0, 1.0), then the confusion matrix after the first three steps would be as shown in Table 6.

Table 6. Confusion matrix where (IPC, PDA, AD, PER) = (0.6, -0.2, 2.0, 0)

Correct Category	Classification				
	"A"	"B"	"C"	"D"	"E"
A	0.400	0.320	0.160	0.080	0.040
B	0.218	0.400	0.218	0.109	0.055
C	0.100	0.200	0.400	0.200	0.100
D	0.055	0.109	0.218	0.400	0.218
E	0.040	0.080	0.160	0.320	0.400

Then after adding random variation around the incorrect probability assignments in step 4, and renormalizing in step 5, the resulting confusion matrix would look something like the randomly generated confusion matrix shown in Table 7.

Table 7. Example of randomly generated confusion matrix where (IPC, PDA, AD, PER) = (0.6, -0.2, 2.0, 1.0)

Correct Category	Classification				
	"A"	"B"	"C"	"D"	"E"
A	0.349	0.438	0.106	0.082	0.025
B	0.015	0.439	0.291	0.183	0.073
C	0.034	0.225	0.377	0.301	0.064
D	0.107	0.088	0.085	0.512	0.207
E	0.010	0.008	0.098	0.469	0.415

For a selected sample size, N, a "simulation run" executes the following:

1. Randomly select the base rate probability for each classification
2. Generate the confusion matrices for each rater and the system
3. Use the base rate probability and confusions matrices to randomly generate N cases.
4. Estimate system accuracy (using method described above)
5. Compare estimated system accuracy to "true" system accuracy, where there are two measures of true system accuracy
   a. Expected accuracy (i.e. $P(A)*P(\text{"A"}|A) + P(B)*P(\text{"B"}|B) + \ldots)$
   b. Proportion correct in sample

When *K-Assumptions* are satisfied

This section examines circumstances where the assumptions implicit in Kappa are satisfied. That is to say where the raters are selecting from a single confusion matrix of the structure shown in Table 1 and where the system confusion matrix also has the same well-behaved structure.

Illustrated in Figure 1 is the asymptotic behavior of the estimation method. The simulation results depicted in Figure 1 had five categories, three experts each with 60% accuracy, 5000 test cases for each run, and where there are 10 runs each with system accuracy set to .1, .3, .5, .7 and .9 respectively.
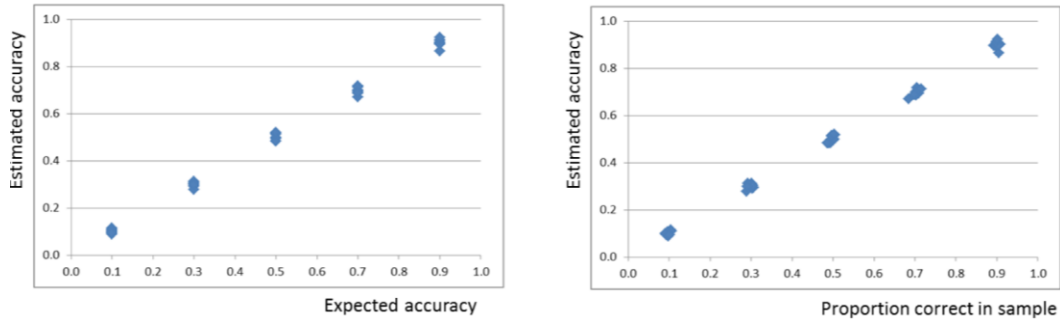
Figure 1. Estimated vs. true system accuracy from simulations with accuracy of three experts each at 0.6, sample size at 5000, with equal error probabilities and equal problem difficulty. (Kappa = 0.251)

The results depicted in Figure 1 indicate that estimates of system accuracy cluster tightly around true system accuracy. When true system accuracy is 0.1, which is less accurate than random guessing (0.2), estimates of system accuracy cluster tightly around 0.1. When true system accuracy is 0.9, which is far better than the raters' accuracy (0.6), estimates of system accuracy cluster tightly around 0.9. Across all fifty simulation runs the average value of Kappa was just 0.251.

The results below depict what happens when sample size and rater accuracy are varied. Figures 2-4 depict the results of fifty simulation runs with a sample size of 200 per run and rater expert accuracy is set to .4, .6 and .8 respectively.
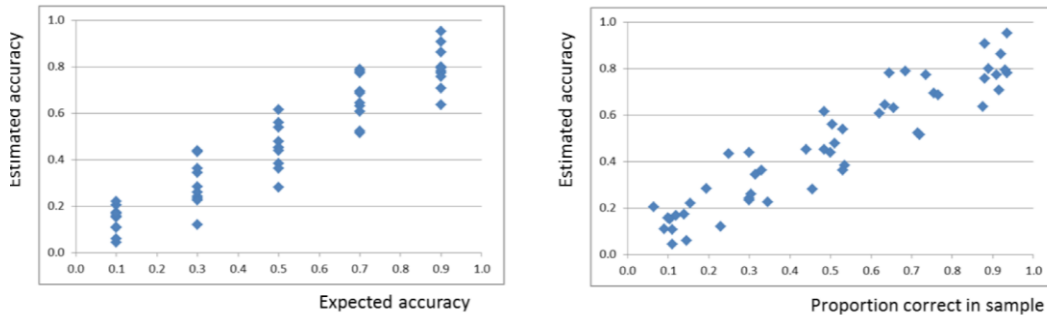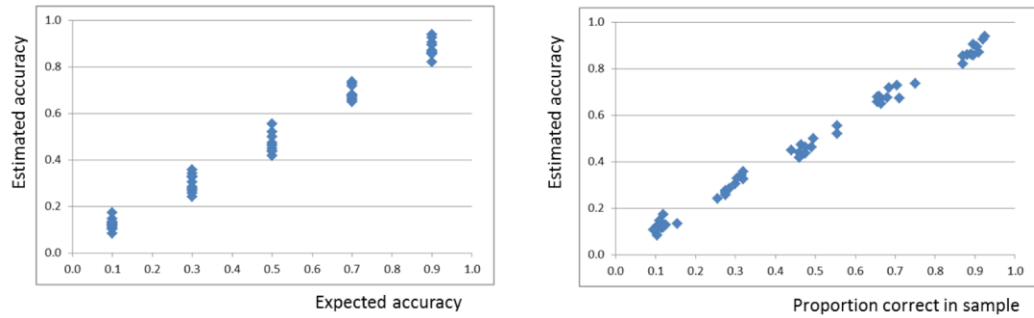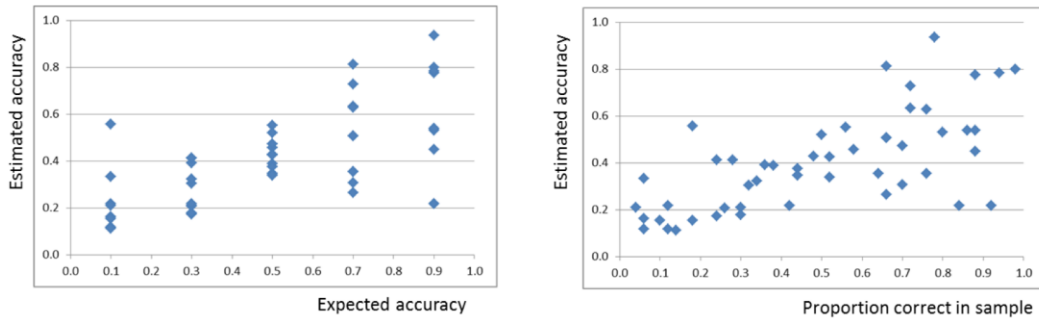

Figure 2. Estimated vs. true system accuracy from simulations with accuracy of three raters each at 0.4, sample size at 200, with equal error probabilities and equal problem difficulty. (Kappa = .065)
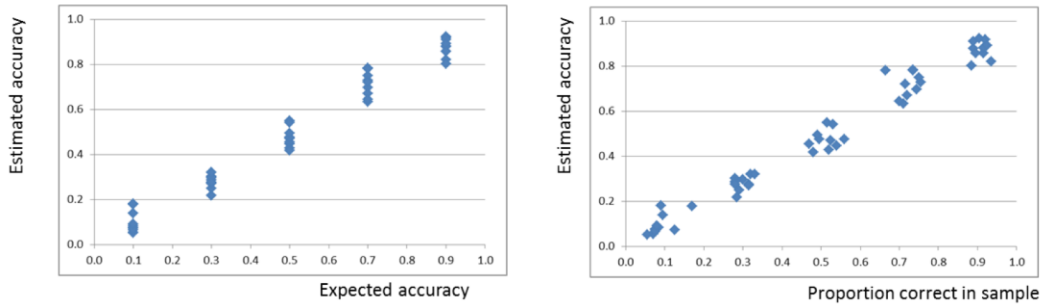
Figure 3. Estimated vs. true system accuracy from simulations with accuracy of three raters each at 0.6, sample size at 200, with equal error probabilities and equal problem difficulty. (Kappa = .255)

Figure 4. Estimated vs. true system accuracy from simulations with accuracy of three raters each at 0.8, sample size at 200, with equal error probabilities and equal problem difficulty. (Kappa = .562)

The results shown in Figures 2-4 indicate that the correspondence between estimated and true system accuracy improves rapidly as rater accuracy improves. Even when the raters are just 60% accurate, estimates of system accuracy are consistently within ± 0.1 of true system accuracy.

Figures 5-7 depict results when sample size is further reduced to just 50 cases per run. When rater accuracy is 0.4 there is little correspondence between estimated and true system accuracy. However when rater accuracy is 0.6 and 0.8 this correspondence improves quickly.

Figure 5. Estimated vs. true system accuracy from simulations with accuracy of three experts each at 0.4, sample size at 50, with equal error probabilities and equal problem difficulty. (Kappa = .060)
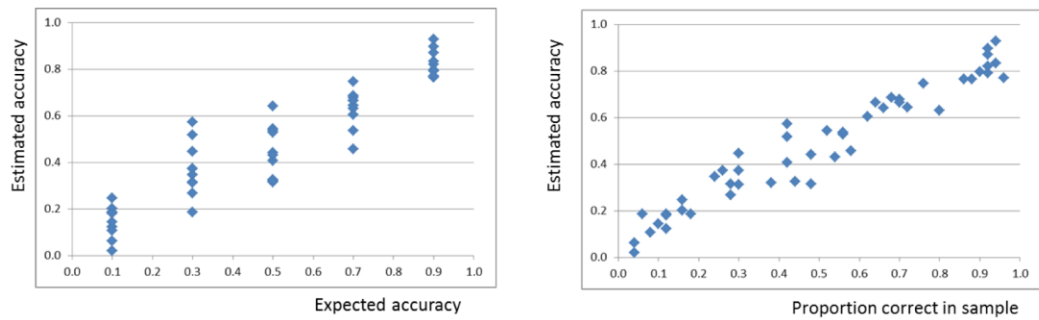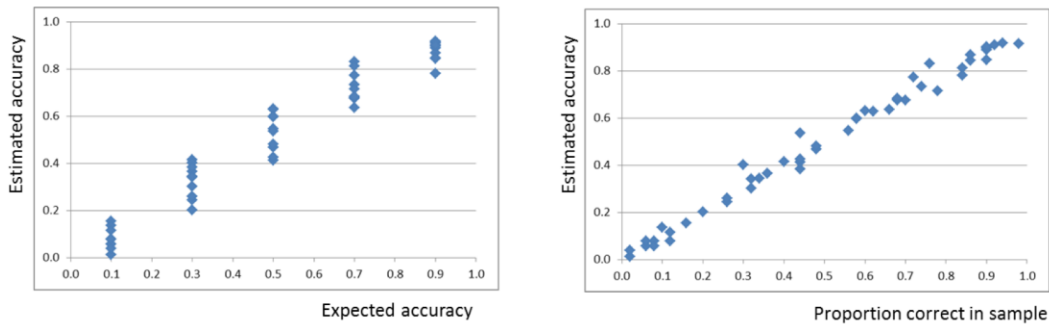

Figure 6. Estimated vs. true system accuracy from simulations with accuracy of three experts each at 0.6, sample size at 50, with equal error probabilities and equal problem difficulty. (Kappa = .244)

Note that in Figures 6 and 7 the two measures of true system accuracy yield slightly different results. Estimated accuracy corresponds more closely to proportion correct in sample than to expected accuracy. This occurs because the proportion correct in a sample varies according to a binomial distribution defined by system accuracy. So even if there is perfect correspondence between estimated accuracy and proportion correct (as is the case when rater accuracy is set to 1.0), the standard deviation of the estimate around expected accuracy (E_a) would still be equal to $(E_a \cdot (1-E_a)/N)^{1/2}$.

Figure 7. Estimated vs. true system accuracy from simulations with accuracy of three experts each at 0.8, sample size at 50, with equal error probabilities and equal problem difficulty. (Kappa = .546)

In summary, when the *K-assumptions* are satisfied, the estimation method exhibits an orderly relationship between estimated and true system accuracy. Estimates of system accuracy are unbiased, and the correspondence between true and estimated system accuracy improve rapidly as rater accuracy and sample size increase.

When *K-Assumptions* are substantially violated

In practice it is difficult to imagine a circumstance where the *K-assumptions* and the implied *K-properties* are satisfied. All raters are not equally accurate; some are typically more experienced and expert than others. All types of errors are not equally probable; this property is certainly false when the categories are naturally ordered or when the raters have some idea of which categories have the highest base rates. And all problems are not equally difficult; unless the test cases are carefully pre-selected and therefore unrepresentative of real world diversity.

In this section the behavior of the estimation method is examined in cases where the *K-properties* are violated. In all of the simulation runs summarized below the *K-properties* of equal rater accuracy, equal problem difficulty, and equal error probabilities are substantially violated. Specifically:

Rater accuracy (IPC) was varied by .1. For example, instead of three raters with .6 accuracy, initial accuracy would be set to .5, .6 and .7 respectively.

Problem difficulty (PDA) was varied by .2. For about a third of the test cases rater and system accuracy were reduced by .2 (or set to a minimum of 0.0) and for about another third accuracy was increased by .2 (or set to the maximum of 1.0).

136

Asymmetric dispersion (AD) was set to 2.0. An incorrect answer that is 'next to' the correct answer is twice as likely as one two steps removed and 4 times as likely as one 3 steps removed, etc.

Error probabilities were randomly varied by up to 100% (PER=1.0). For example, if the error probability is initially set to .2 then that error probability would be randomly selected from the range [0, .4]. This random variation is done independently for each error probability.

To appreciate the magnitude of impact of these parameter settings consider again Tables 5 and 7 above. Table 5 is exactly the confusion matrix that results when initial rater accuracy is set to .6 and the *K-properties* are satisfied. Table 7 is representative of about 1/3 of the cases when initial rater accuracy is set to .6 but with the above parameter settings. It seems fair to characterize Table 7 as a substantial variation from Table 5.

All of the simulation runs in this section use the above parameter settings to systematically and then randomly vary the rater and system confusion matrices. The results shown in Figure 8 illustrate the asymptotic behavior of the estimation method when the *K-properties* are substantially violated. Note that when system accuracy is preset to .1 and .9, expected accuracy is .133 and .867 respectively. This occurs because problem difficulty is varied plus and minus 0.2, but accuracy can be no lower than 0.0 or higher than 1.0. So for example when system accuracy is preset to 0.1, one third of the problems have system accuracy reset to 0.3, one third stay at 0.1 and the remaining third are reset to 0.0; then averaged expected system accuracy is then .133.

There is a linear relationship between estimated and true accuracy. There is also some bias in the estimates; estimated accuracy is too high when true system accuracy is low and estimated accuracy is to low when true system accuracy is high. Note though that when the system was more accurate than the raters the estimates of system accuracy were still consistently higher than the raters' accuracy. The estimate of system accuracy may be conservative, but it is not bounded by the raters' accuracy.
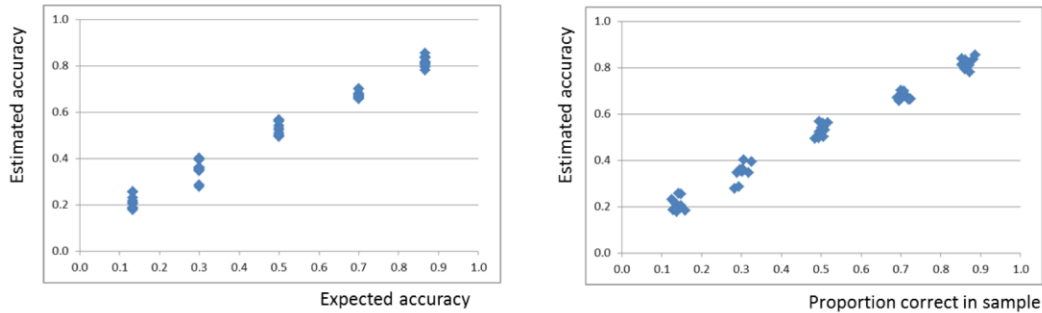
Figure 8. Estimated vs. true system accuracy from simulations with accuracy of three raters at .5, .6 and .7; sample size at 5000 and confusion matrices systematically then randomly varied. (Kappa = 0.305)

There is a straightforward explanation for this estimation bias. The violations of the *K-properties* inflated inter-rater agreement. Because inter-rater agreement is used to estimate rater accuracy, as per Eq. 2, this leads to a slightly inflated estimate of rater accuracy. Inflated estimates of rater accuracy in turn lead to overestimates of the ground truth probabilities for the categories with the highest estimated ground truth probabilities. Finally given the equation for deriving system accuracy from the ground truth probabilities (Eq. 7) this leads to the estimation bias. In comparing Figures 1 and 8, note that Kappa was .251 and .305 respectively; and the average estimated accuracy for the runs in Figure 1 was exactly 0.60 and the average estimated rater accuracy for the runs in Figure 8 was 0.64.

In general violations of the *K-properties* will inflate expected inter-rater agreement with one exception. Differences between rater accuracy decreases rather than increases expected inter-rater agreement, but the net effect is small when compared to the larger opposite effect of the other violations. For example, if overall rater accuracy is set to .6 and then varied by .2 (i.e. rater accuracy set to .4, .6, .8 respectively) and true system accuracy is 0.9 then estimated accuracy will be approximately 0.924 – a 0.024 overestimate. But if instead problem difficulty is varied by the same amount (.4, .6, .8 respectively) then system accuracy will be approximately 0.857 – a 0.043 underestimate. Varying dispersion by 100% around the error probabilities results in an approximate 0.036 underestimate, and setting asymmetric dispersion to 2.0 results in a 0.068 underestimate.

In Figures 9-11 the sample size is 200 cases per run and expected rater accuracy is set to .4, .6 and .8 respectively. In Figures 12-14 sample size is

reduced to 50 cases per run. Except for the bias toward underestimating high system accuracy (and overestimating low system accuracy) these results are similar to the results with the matrices that satisfy the *K-properties*. Increasing rater accuracy and sample size both decrease the variance of the estimate. The estimation bias is pronounced when rater accuracy is very low (0.4), noticeable when rater accuracy is moderate (0.6), and appears negligible when rater accuracy is high (0.8).

In practice, most efforts to evaluate system accuracy address systems that are hypothesized to perform well. For such evaluations the estimates derived from this method become increasingly conservative as the ratings of the experts are increasingly suspect.
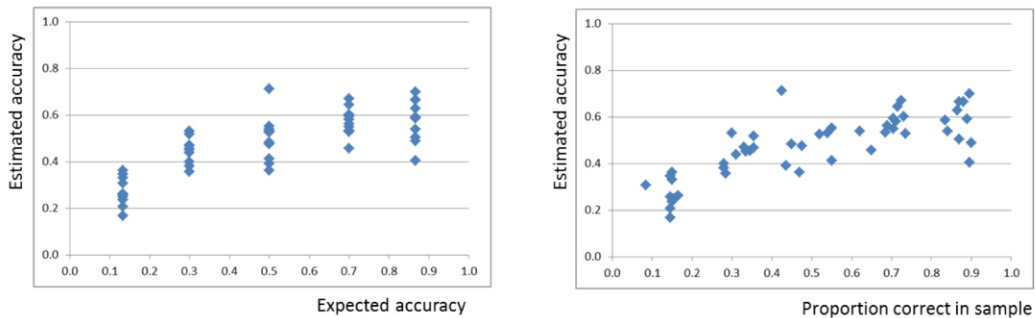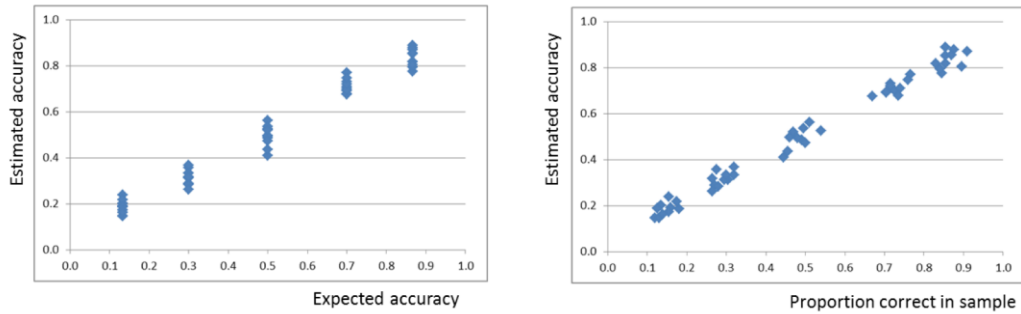
Figure 9. Estimated vs. true system accuracy from simulations with accuracy of three raters at .3, .4 and .5; sample size at 200 and confusion matrices systematically then randomly varied. (Kappa = .142)
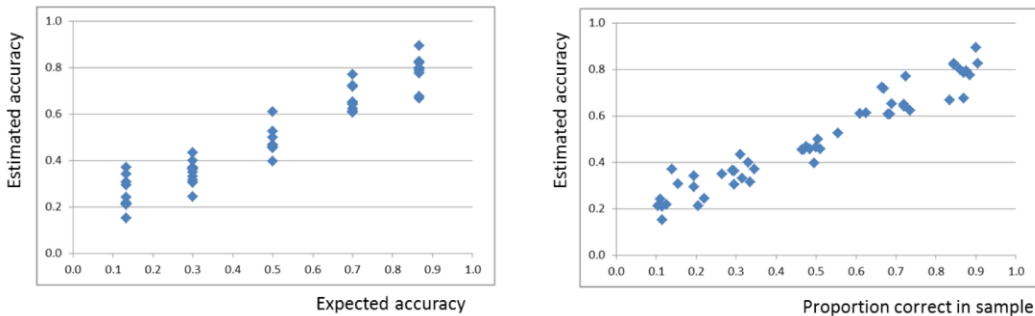
Figure 10. Estimated vs. true system accuracy from simulations with accuracy of three raters at .5, .6 and .7; sample size at 200 and confusion matrices systematically then randomly varied. (Kappa = .306)

Figure 11. Estimated vs. true system accuracy from simulations with accuracy of three raters at .7, .8 and .9; sample size at 200 and confusion matrices systematically then randomly varied. (Kappa = .578)
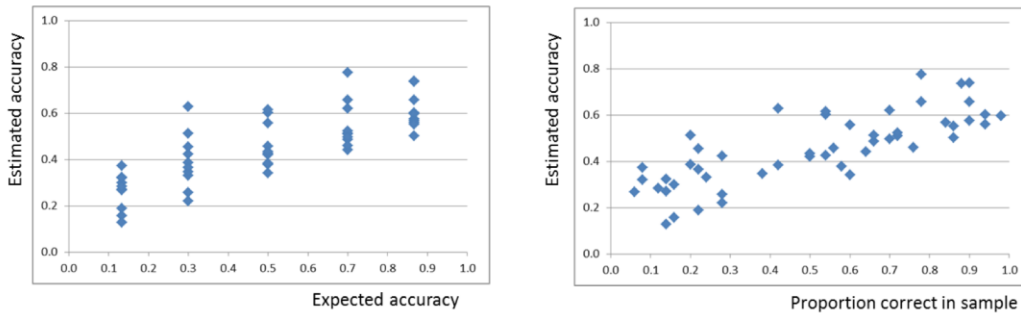

Figure 12. Estimated vs. true system accuracy from simulations with accuracy of three raters at .3, .4 and .5; sample size at 50 and confusion matrices systematically then randomly varied. (Kappa = .144)

Figure 13. Estimated vs. true system accuracy from simulations with accuracy of three raters at .5, .6 and .7; sample size at 50 and confusion matrices systematically then randomly varied. (Kappa = .311)
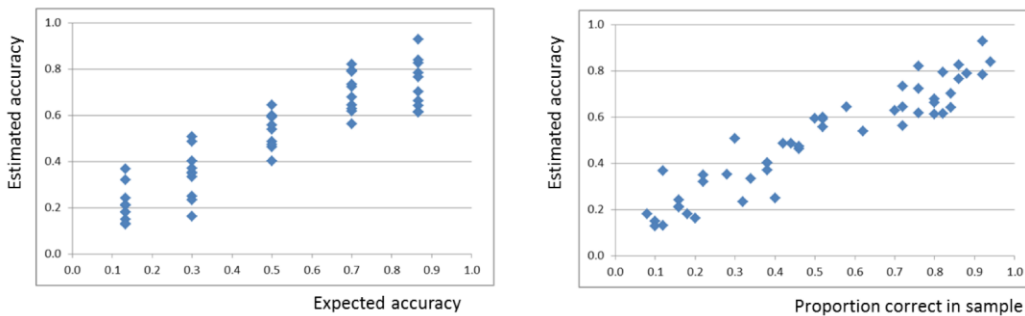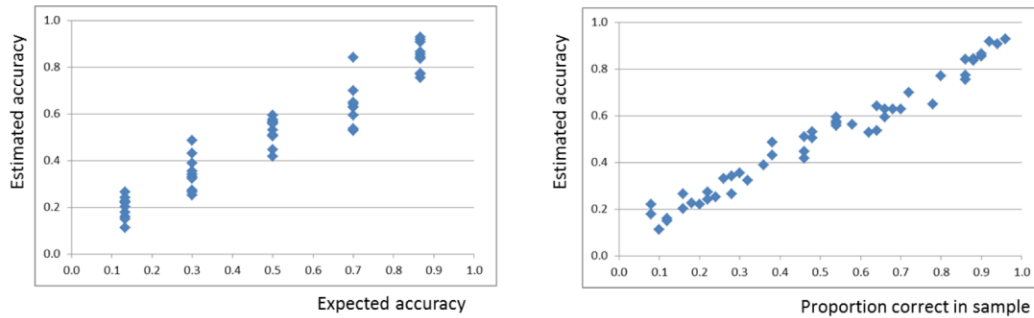
Figure 14. Estimated vs. true system accuracy from simulations with accuracy of three raters at .7, .8 and .9; sample size at 50 and confusion matrices systematically then randomly varied. (Kappa = .586)

Discussion

The objective in this study was to demonstrate that it is feasible to reliably estimate the accuracy of system classifications when ground truth can only be estimated with fallible expert ratings. The simulation results described herein provide evidence for the claims stated in the introduction, namely that reliable estimates of system accuracy can be obtained from fallible expert ratings under a diverse conditions, that the reliability of these estimates is approximately the same whether the system is more or less accurate than the expert raters, and that the conditions under which these accuracy estimates become unreliable are predictable (e.g., inter-rater agreement is low and sample size is small).

In the estimation method the level of inter-rater agreement is used to estimate the overall accuracy of the expert ratings, Bayesian updating based on the estimated expert accuracy is used to estimate a "ground truth" probability for each classification, and finally system accuracy is estimated by comparing the relative frequency that the system assignment agrees with the most probable classification at different probability levels.

Although the estimation method was derived from assumptions that are implicit in the Kappa statistic (and how it is often used), a simulation analysis shows that the accuracy of the estimates of system accuracy are robust against substantial variations from the rater behavior implied by those assumptions. The accuracy of the estimates of system accuracy is driven primarily by overall rater accuracy (which can be estimated from inter-rater agreement) and sample size.

Recommended use and uses to avoid

The simulation results presented herein suggest an overall data collection and estimation approach where measured inter-rater agreement is used to determine the number of test cases needed to obtain high confidence in system accuracy estimates. For example for five category problems with three raters if initial data collection indicates that Kappa is around .3 then data collection should continue for at least 200 cases. This would be a sufficient number of cases to obtain 90% "confidence" that estimated accuracy is within .1 of true accuracy. On the other hand, if Kappa is around .55 then a sample size of 100 cases is sufficient to ensure the same "confidence interval." As the number of raters and categories varies, so does the parametric relationship between sample size and confidence in estimates of system accuracy; so additional simulation runs such as those shown here would be needed to determine sample size requirements.

In this approach all test cases are useable, even ones where raters substantially disagree. This makes it feasible to randomly select test cases from the population of problems from which the system is likely to be applied which in turn should facilitate the ability generalize test results to practice.

As noted above, violations of the *K-properties* (equal rater accuracy, problem difficulty and error probabilities) will bias the estimate of system accuracy. The magnitude of this bias interacts with overall rater accuracy. If system accuracy is high and rater accuracy low then the estimation procedure described herein will likely substantially under estimate system accuracy. In the above simulations, for example, on five category problems when true system accuracy was .9 and rater accuracy was .4 the estimate of system accuracy was around .6. Consequently when Kappa is very low (e.g. less than .2) it would be helpful to examine the inter-rater agreement data for patterns that suggest violations of the *K-properties*. For example, the K-property of equal error probabilities implies that all pairwise disagreements are equally likely (e.g. "AB" as likely as "AE") and a statistical test can be performed to help determine if this pattern is violated. If it is, then the estimate of system accuracy can be adjusted upwards. There is much work to be done to determine how and when such adjustments should be made, but doing so seems feasible.

The estimation method described herein is specifically intended for cases where each rater is an independent measure of ground truth classifications. The procedure assumes the causal structure shown in Figure 15-10a.
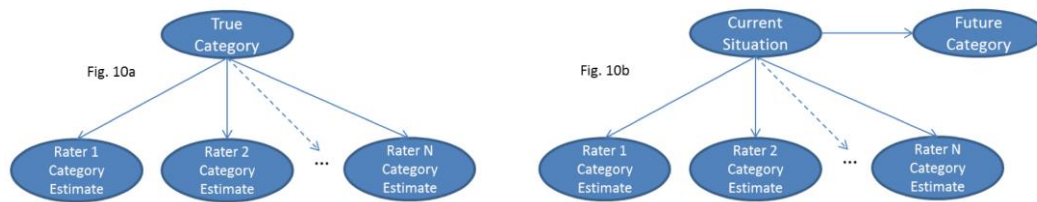
Figure 15. Assumed causal relationship between ground truth and expert ratings vs. causal structure of forecasting tasks

There are many applications that involve aggregation of independent estimates from multiple individuals but do not have the causal structure shown in Figure 15-10a. For many such applications use of the estimation method described here would be inappropriate. For example, it is becoming common practice in forecasting to systematically combine the ratings of multiple independent forecasters (e.g. Surowieki, 2005). Although the estimation method presented here could be mechanically applied to such forecasting tasks, such an application may yield spurious results. Forecasting tasks do not have the causal structure shown in Figure 15-10a, but have a causal structure closer to the one shown in Figure 15-10b where expert ratings are not in any sense direct measures of the future outcomes. On the other hand the estimation method can and has been used to retrospectively assess whether a forecasted outcome actually occurred. For example Lehner et al. (2012) examined the accuracy of the imprecise forecasts typically found in published forecasts by using multiple raters to retrospectively assess whether the forecasted outcome occurred and then using an estimation method similar to the one presented here to estimate the accuracy of a collection of forecasts. Similarly Levitt and Lehner (2011) applied a variation of this method to resolve disagreeing historical judgments as to the timeframe when key developments occurred in the maturation of new technologies.

The distinction between Figures 15-10a and 15-10b is essentially the distinction between medical diagnosis and medical prognosis. It would be appropriate to apply the method to estimate the accuracy of a new diagnostic system by comparing system diagnoses to those of medical professionals, but it would be inappropriate to use it to estimate the accuracy of a new system's prognoses by comparing them to the prognoses of medical professionals.

In general it is important that the causal structure relating the rater and system selections to ground truth match the structure assumed by the estimation method. The process of collecting ratings from the experts should be engineered

143

to ensure this causal structure; such as by ensuring that the expert ratings are independent and to the extent possible having available the same data for each rater for each test case.

The estimation method presented here was developed to address test and evaluation of an automated classification system after development. However it does seem feasible to also employ this approach during system development. Specifically the estimation method could be used to develop training data sets with a probability distribution of correct classifications for each training case.

Related and future research

The research presented in this paper had the very specific goal of demonstrating that it is feasible to reasonably estimate system accuracy using fallible expert ratings even when the system is substantially more accurate than the experts. Nothing in this paper would support a claim that the estimation method presented here is in any sense optimal. There are many opportunities for improvement. Three suggestions are offered below.

First, the estimation method was designed for use with classification problems for which there is no natural ordering to the categories. The simulation results suggest that the method is robust even when there is a natural ordering, but the accuracy of estimates of system accuracy would likely be improved if the method is modified to specifically account for the fact that certain types of errors are more likely than others. For example, if the natural ordering is *A, B, C, D, E*, then a rating of "A" should be more evidence for category B than for category E. The method presented here treats *B* and *E* equally.

Second, as noted above, it should be feasible to develop statistical procedures to estimate whether and to what degree *K-properties* are violated. From these estimates it should be also feasible to adjust the system accuracy estimates to correct for bias. This area is unexplored.

Third, the estimation method presented here is entirely algebraic. Everything is derived directly from some percent-of-agreement statistics. No effort was made to estimate base rates and confusion matrices that represent a "best fit" to the inter-rater agreement data. But there are best fit methods that could be used for this purpose. For example, the non-linear optimization methods in Latent Class Analysis (McCutcheon, 1987) could be used to find maximum likelihood estimates for the base rate and confusion matrix probabilities. Both Uebersax (1988) and Carpenter (2008) applied this approach to binary classification problems; and Carpenter also used Bayes inference to aggregate ratings and

estimate classification probabilities. Similarly one could use non-linear optimization to find base rates and confusion matrix probabilities that minimize the difference between expected and observed relative frequency of each inter-rater pair (relative frequency of "AA", "AB", "AC" …). It remains an open and interesting question as to whether use of such optimization methods would yield better results.

Acknowledgements

References

Armstrong, J. S. (2006). How to make better forecasts and decisions: Avoid face-to-face meetings. *The International Journal of Applied Forecasting*, *5*, 3-15.

Bishop, M. A., & Trout, J. D. (2002). 50 years of successful prediction modeling should be enough. *Philosophy of Science*, *69*(S3), S197-S208.

Carpenter, B. (2008). *Multi-level Bayesian models of categorical data annotation*. Manuscript found at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.174.1374&rep=rep1&type=pdf.

Dawes, R. M. (1979). The robust beauty of improper linear models. *American Psychologist*, *34*, 571- 582.

Grove, W. H., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2001). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19-30.

Gwet, K. L. (2010). *Handbook of inter-rater reliability* (2nd Ed.). Advanced Analytics, LLC.

Lehner, P., Michelson, A., Adelman, L., & Goodman, A. (2012). Using Inferred Probabilities to Measure the Accuracy of Imprecise Forecasts. *Judgment and Decision Making*, *7*(6), 728-740.

Levitt, T. & Lehner, P. (2011) Baseline Judgment Estimation and Challenge Question Answer Assignment for the FUSE Program. Technical Report, The MITRE Corporation, December 28, 2011.

McCutcheon, A. L. (1987). *Latent class analysis*. Thousand Oaks, California: Sage Publications.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor Books.

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton, N.J.: Princeton University Press.

Uebersax, J. S. (1988). Validity inferences from inter observer agreement. *Psychological Bulletin*, *104*(3), 405-416.

Appendix

Proof of Theorem 1

Restating the assumptions:

AA1. For each case there is a unique correct classification

AA2. For each case raters independently assign classifications

AA3. Expected agreement between raters increases as expected rater accuracy increases.

Begin with a few definitions.

Definition of *correct classification* in AA1: For each case there is a vector $<c_1, c_2 \dots c_n>$ where for some index i, $c_i = 1$ and the remaining values are 0.

Definition of *independent assignment* in AA3: For each case, the probability that a rater will select a class is conditionally independent of the other raters' selections.

Independent assignments allow the description of each rater's selection behavior as a probability vector. That is to say, for each case each rater has a selection probability for each category. These will be called *selection vectors*.

Definition of *rater accuracy* in AA3: For M raters and N cases, rater accuracy is defined as the total proportion of correct selections.

For example, if there are 10 cases and three raters who make correct assignments in 7, 5 and 9 of the cases respectively, then rater accuracy = 0.7.

The three lemmas below all use the same proof strategy. Begin with any two selection vectors that are not identical. Construct a selection vector that is the average of the two. The average vector will necessarily have the same expected accuracy but a different level of expected agreement than the original two vectors. If the average vector has higher/lower expected agreement, then create a new

vector by slightly reducing/increasing the probability of correct assignment in the average vector. When the change is sufficiently small the new vector will have higher/lower expected accuracy and lower/higher expected agreement than the original two vectors. Most of the algebraic complexity in these proofs is the result of showing one way to calculate a change that is always "sufficiently small".

Lemma 1:

To ensure AA1-AA3 within each case all raters must behave as though they are selecting a category using the same selection vector.

Proof: Let $<p_{11}, p_{12} \ldots p_{1n}>$ and $<p_{21}, p_{22} \ldots p_{2n}>$ be the selection vectors of 2 raters for a specific case; where some probabilities do not agree (e.g. $p_{11} \neq p_{21}$). For purposes of the proofs below, assume that category 1 is the correct category. (The arguments below apply no matter which category is correct.)

Below it is shown how to construct from two different selection vectors a common selection vector for both raters where expected accuracy is lower but expected agreement higher. Consequently unless the two raters have the same selection vector, there will always be another pair of vectors with lower expected accuracy and higher expected agreement – violating AA3.

Set $\quad p_i = \left(p_{1i} + p_{2i} \right)/2 \quad , \quad e_i = \left(p_{1i} - p_{2i} \right)/2 \quad , \quad d = \left(e_1^2 / \left(2 * \left(p_2 - p_1 \right) \right) \right) \quad$,if $p_1 < p_2$, $d = -\left(e_1^2 / \left(2 * \left(p_2 - p_1 \right) \right) \right)$, and $d = 0$ if $p_1 = p_2$

For selection vectors $< p_{11}, p_{12} \ldots p_{1n} >$ and $< p_{21}, p_{22} \ldots p_{2n} >$

Expected accuracy $= p_1$

Expected agreement $= p_{11} * p_{21} + p_{12} * p_{22} + \ldots + p_{1n} * p_{2n}$

$$= \left(p_1 + e_1 \right) * \left(p_1 - e_1 \right) + \left(p_2 + e_2 \right)$$
$$* \left(p_2 - e_2 \right) + \ldots + \left(p_n + e_n \right) * \left(p_n - e_n \right)$$
$$= p_1^2 + p_2^2 + \ldots + p_n^2 - e_1^2 - e_2^2 - \cdots - e_n^2$$

(A1)

For selection vectors $< p_1, p_2 \ldots p_n >$ and $< p_1, p_2 \ldots p_n >$

Expected accuracy $= p_1$

Expected agreement $= p_1^2 + p_2^2 + \ldots + p_n^2$

(A2)

For selection vectors $< p_1, p_2 \ldots p_n >$ and $< p_1, p_2 \ldots p_n >$

Expected accuracy $= p_1$ (A3)

Expected agreement $= p_1^2 + p_2^2 + \ldots + p_n^2$

Expected accuracy in (A1) is higher than in (A3), but expected agreement is lower; where the common selection vector in (A3) was constructed from a difference between the vectors in (A1). Consequently, whenever there is a difference between the selection vectors of two raters a selection probability vector for the two raters can be constructed with lower expected accuracy but high expected agreement.

Within each case if the selection vectors of the raters differ AA3 is not guaranteed.   ***

Lemma 2:

To ensure AA1-AA3 within each case the error probability is the same for all incorrect categories.

Proof: From Lemma 1 it is known that AA1-AA3 imply that for each case all raters have the same selection vector. Let that vector be $<p_1, p_2 \ldots p_n>$. Assume category 1 is the correct assignment and that the remaining probabilities are not all equal.

Below it is shown how to construct selection vector, with equal probability for all incorrect assignments, where expected accuracy is higher but expected agreement lower. Consequently the error probabilities are unequal, there will always be a vector with higher expected accuracy and lower expected agreement – violating AA3.

Set $\quad p_e = (p_2 + \ldots + p_n)/(n-1) \quad , \quad e_i = (p_i - p_e) \quad$ for all $\quad i > 1, \quad$ set $e_{min} = \min(|e_2| \ldots |e_n|)$ and $d = e_{min}^2 / 2$.

Note that $(e_2 + \ldots e_n) = 0$ and that there are at least 2 e_i that are not zero.

For the vector $< p_1, p_2 \ldots p_n >$,

Expected accuracy $= p_1$

$$
\begin{aligned}
\text{Expected agreement} &= p_1^2 + p_2^2 + \ldots + p_n^2 \\
&= p_1^2 + (p_e + e_2)^2 + \ldots + (p_e + e_n)^2 \quad\quad\text{(A4)} \\
&= p_1^2 + p_e^2 + \ldots + p_e^2 + 2p_e\left(\begin{array}{c} e_2 + e_3 \\ + \ldots + e_n \end{array}\right) + e_2^2 + e_3^2 + \ldots e_n^2 \\
&= p_1^2 + (n-1)p_e^2 + e_2^2 + e_3^2 + \ldots e_n^2
\end{aligned}
$$

For the vector $< p_1, p_e \ldots p_e >$

$$\text{Expected accuracy} = p_1 \quad\quad\text{(A5)}$$

$$\text{Expected agreement} = p_1^2 + (n-1)p_e^2$$

For the vector $< p_1 + d, \ p_e - d, \ p_e \ldots p_e >$

$$\text{Expected accuracy} = p_1 + d = p_1 + e_{\min}^2/2$$

$$
\begin{aligned}
\text{Expected agreement} &= (p_1 + d)^2 + (p_e - d)^2 + p_e^2 + \ldots + p_e^2 \\
&= p_1^2 + (n-1)p_e^2 + 2p_1 d - 2p_e d + 2d^2 \quad\quad\text{(A6)} \\
&= p_1^2 + (n-1)p_e^2 + 2d(p_1 - p_e) + 2d^2 \\
&= p_1^2 + (n-1)p_e^2 + e_{\min}^2 * (p_1 - p_e)) + e_{\min}^4/2
\end{aligned}
$$

Since $e_{\min}^2*(p_1 - p_e)) + e_{\min}^4/2 < e_{\min}^2 + e_{\min}^2 <= e_2^2 + e_3^2 + \ldots e_n^2$, expected agreement in (A4) is higher than expected agreement in (A6) even though expected accuracy is lower.

Consequently, whenever the probability of incorrect assignment is unequal, there will always be a selection vector with higher expected accuracy and lower expected agreement, violating AA3.

Within each case and selection vector if the error probabilities are unequal AA3 is not guaranteed.

 ***

Lemma 3:

To ensure AA1-AA3 the selection vector must be the same across all cases.

Proof: Lemmas 1 and 2 show that AA1-AA3 imply that for each case the raters have identical selection vectors of the form <p_e ... p_c ... p_e> where p_c is the probability of assigning the correct category and $p_e = (1-p_c)/(n-1)$ where n is the number of categories.

Below it is shown that across different cases the selection vectors must have the same values for p_c (and therefore p_e) else a violation of AA3 can be constructed.

Let p_{c1} and p_{c2} be the probability of correct assignment on two different cases, and p_{e1} and p_{e2} the corresponding error probabilities. For each case, order the cases such that the correct assignment is first. So for all raters the probability vector is <$p_{c1}, p_{e1}, ... p_{e1}$> for case 1 and <$p_{c2}, p_{e2}, ..., p_{e2}$> for case 2, but the categories may be in a different order. The proof below makes no reference to matching categories across cases so this ordering does not affect the proof.

Set $p_c = (p_{c1} + p_{c2})/2$, $p_e = (p_{e1} + p_{e2})/2$, $e_c = (p_{c1} - p_c)$, $e_e = (p_{e1} - p_e)$, $e_{min} = \min(|e_c|, |e_e|)$, $d = e_{min}^2/2$

For two cases with accuracy $p_{c1} \neq p_{c2}$

$$\text{Expected accuracy} = p_c$$

$$
\begin{aligned}
\text{Expected agreement} &= \left(p_{c1}^2 + (n-1)\, p_{e1}^2 + p_{c2}^2 + (n-1)\, p_{e2}^2 \right)/2 \\
&= \left(\frac{(p_c + e_c)^2 + (n-1)(p_e + e_e)^2}{+ (p_c - e_c)^2 + (n-1)(p_e - e_e)^2} \right) \Big/ 2 \qquad (A7) \\
&= \left(2p_c^2 + 2(n-1)\, p_e^2 + 2e_c^2 + 2(n-1)\, e_e^2 \right)/2 \\
&= p_c^2 + (n-1)\, p_e^2 + e_c^2 + (n-1)\, e_e^2
\end{aligned}
$$

For two cases with accuracy $p_{c1} = p_{c2}$

$$\text{Expected accuracy} = p_c \qquad (A8)$$

$$\text{Expected agreement} = p_c^2 + (n-1)\, p_e^2$$

For two cases with accuracy vectors $< p_c + d, \, p_e - d, \, p_e \ldots p_e) >$

$$\text{Expected accuracy} = p_c + d$$

$$
\begin{aligned}
\text{Expected agreement} &= (p_c + d)^2 + (p_e - d)^2 + (n-2) p_e^2 \\
&= p_c^2 + (n-1) p_e^2 + 2 p_c d + d^2 - 2 p_e d + d^2 \\
&= p_c^2 + (n-1) p_e^2 + 2d (p_c - p_e) + 2d^2 \\
&= p_c^2 + (n-1) p_e^2 + 2 (e_{min}^2 / 2) \\
& \quad (p_c - p_e) + 2 (e_{min}^2 / 2)^2 \\
&= p_c^2 + (n-1) p_e^2 + (p_c - p_e) e_{min}^2 + e_{min}^4 / 2
\end{aligned}
\tag{A9}
$$

Since $e_{min}^2 * (p_c - p_e)) + e_{min}^4 / 2 < e_{min}^2 + e_{min}^2 <= e_c^2 + e_e^2$, expected agreement in (A7) is higher than expected agreement in (A9) even though expected accuracy is lower.

Consequently, whenever the probability of correct assignment across cases is unequal, there will always be a probability vector that is the same across cases with higher expected accuracy and lower expected agreement, violating AA3. Across cases, if the selection vectors differ then AA3 is not guaranteed.   ***

Theorem 1:

AA1-AA3 are ensured if and only if the raters behave as though their selection for each case is determined by a single confusion matrix where the conditional probability of correct assignment is constant and the conditional probability of all incorrect assignments is equal.

Proof: The "only if" necessity portion follows directly from Lemmas 1-3. Sufficiency follows the fact that with a constant conditional probability of correct assignment (P_c) and incorrect assignments (P_e), expected accuracy is P_c and expected agreement is $P_c^2 + (n-1) P_e^2 = P_c^2 + (1 - P_c)^2 / (n-1)$. Clearly expected agreement increases monotonically with P_c.   ***