How to Gauge the Quality of a Testing Method When Ground Truth Is Known with Uncertainty

Nicholas Gray\textsuperscript{1}, Scott Ferson\textsuperscript{1}, and Vladik Kreinovich\textsuperscript{2}

\textsuperscript{1}Institute for Risk and Uncertainty
University of Liverpool, Liverpool L69 7ZF, UK
Nicholas.Gray@liverpool.ac.uk, sandp8@gmail.com

\textsuperscript{2}Department of Computer Science, University of Texas at El Paso
500 W. University, El Paso, Texas 79968, USA, vladik@utep.edu
1. Formulation of the Problem

- In many practical situations, algorithms help us recognize the situation.

- In medicine, algorithms use symptoms and measurement results to provide a diagnosis.

- In engineering, algorithms use the results of measurements and observations to decide, e.g.:
  - whether a road may fail in the nearest future (and thus, repairs are needed now),
  - or it can stay in working condition until the next year’s testing.

- In military applications, algorithms help us decide whether a radar signal indicates:
  - an incoming enemy plane
  - or an innocent flock of birds.
2. Formulation of the Problem (cont-d)

• In many applications, the available algorithms are not perfect: sometimes, they lead to a wrong result:
  – a medical system can misdiagnose,
  – a military system can mistakenly classify an innocent object as an enemy attack, etc.

• In many situations, we eventually learn the ground truth.

• In such situations, we can gauge the quality of a testing method by comparing its results with the ground truth.

• Based on the results of this comparison, we can estimate how good is the testing method.

• The challenge is that in many application areas, we do not always know the ground truth.
3. Formulation of the Problem (cont-d)

- For example, in medical diagnostics, the ground truth is supposed to come from medical doctors.

- However, in many cases, the doctors themselves are not 100% confident in their diagnoses.

- The existing techniques for gauging the quality of testing methods:
  - either ignore such uncertain diagnoses altogether,
  - or, vice versa, ignore the corresponding uncertainty and treat all the diagnoses as the ground truth.

- We want a better understanding of the quality of different testing methods.

- It is therefore desirable to explicitly take the experts’ uncertainty into account.

- This is what we do in this talk.
4. Quality of Testing Methods

- For many properties – e.g., for different diseases – we have different testing methods.
- These methods are rarely perfect.
- For example, for medical tests:
  - sometimes, the test missed a disease, and
  - sometimes, the test return an alarming result even when the patient does not have the disease.
- Several characteristics are used:
  - to gauge the quality of a testing method, and
  - to compare the quality of different testing methods.
- The most widely used are sensitivity, specificity, and precision.
5. Quality of Testing Methods (cont-d)

• Let $P$ denote the set of all the objects from the tested sample that actually have the tested property.

• Example: the set of all the people in the sample who actually have the tested disease.

• Let $N$ denote the set of all the objects from the tested sample that do not have the tested property.

• Example: the set of all the people in the sample who do not have the tested disease.

• Let $S_+$ denote the set of all the objects for which the test concluded that they have the tested property.

• Let $S_-$ denote the set of all the objects for which the test concluded that they do not have the property.
6. Quality of Testing Methods (cont-d)

- A perfect test should classify:
  - all the objects that actually have this property – and only these objects
  - as having the tested property.
- So, for a perfect test, we should have \( P = S_+ \), and, correspondingly, \( N = S_- \).
- In reality, tests are not perfect, so we may have misclassified objects.
- The usual characteristics for gauging the quality of a testing method use:
  - the numbers of objects with or without the tested property
  - that were classified correctly or incorrectly.
7. Quality of Testing Methods (cont-d)

- In general, the number of elements in a set \( S \) will be denoted by \( |S| \).

- One of the characteristics is sensitivity.

- It is also known as recall or True Positive Rate – TPR for short.

- Sensitivity is defined as the proportion:
  - among all the objects with the tested property,
  - of the ones that were correctly classified by the test:

  \[
  \text{TPR} = \frac{|P \cap S_+|}{|P|}.
  \]

- Another characteristics is specificity.

- It is also known as True Negative Rate – TNR, for short.
8. Quality of Testing Methods (cont-d)

- Specificity is defined as the proportion:
  - among the objects that do not have the tested property,
  - of the ones that were correctly classified by the test:

\[
TNR = \frac{|N \cap S_-|}{|N|}.
\]

- Example: the proportion of healthy people that this test classified as healthy.

- One more characteristic is *precision*.

- It is also known as Positive Predictive Value – PPV, for short.
9. Quality of Testing Methods (cont-d)

- Precision is defined as the proportion:
  - among object that the test classified as having the tested property,
  - of the objects who actually have this property:
    \[
    \text{PPV} = \frac{|P \cap S_+|}{|S_+|}.
    \]

- Example: the proportion of sick people among those that the test classified as sick.

- For each of the three characteristics, the larger the value of the characteristic, the better.
10. Quality of Testing Methods (cont-d)

• In the perfect case, all three characteristics are equal to 1.

• From this viewpoint:
  – a reasonable way to compare different testing methods is
  – to compare the values of one or more of the three characteristics.

• If for one of the methods, the corresponding value is larger, this means that:
  – from the viewpoint of this characteristic,
  – this method is better.
11. Comment

- To make a conclusion about which testing method is better, we also need to take into account that:
  - the values of each characteristic come from a finite sample and
  - are, thus, only an approximate representation of the actual quality of a testing method.

- For the same method:
  - for different random samples,
  - we can get slightly larger or slightly smaller values of the corresponding characteristic.
12. Comment (cont-d)

• So:
  – to make a definite conclusion that one of the testing methods is better,
  – we need to check that the difference between the values of the characteristic is stat. significant.

• There are known statistical procedures for checking this.

• This is especially important to take into account when the sample sizes are small.

• When the sample sizes are large, the corresponding randomness becomes very small.
13. Often, We Do Not Know the “Ground Truth”

- The formulas for computing the above three characteristics assume that we know the “ground truth”:
  - which objects have the tested property and
  - which objects do not have this property.
- In the above example, we know which patients have the tested disease.
- In practice, however, this information often comes from experts – e.g., from medical doctors.
- Experts are often not 100% sure about their statements and their diagnoses.
- How can we take this expert uncertainty into account when gauging the quality of a test?
14. How to Describe Expert’s Uncertainty

• For each object $i$, an expert makes:
  – either a statement that the object has the tested property,
  – or a statement that the object does not have the tested property.

• In both cases, the expert is usually not absolutely confident in his/her statement.

• The whole procedure is based on statistics.

• So, it is reasonable to gauge the expert’s degree of certainty $c_i$ in his/her statement by a probability value.
15. Describing Expert’s Uncertainty (cont-d)

- If the expert believes that the object $i$ most probably has the tested property, then:
  - the probability $p_i$ that this object has the tested property is equal to $p_i = c_i$; and
  - the probability that the object $i$ does not have the tested property is equal to $1 - c_i$.

- If the expert believes that the object $i$ most probably does not have the tested property, then:
  - the probability that this object does not have the tested property is equal to $c_i$; and
  - the probability that the object $i$ has the tested property is equal to $p_i = 1 - c_i$. 
16. Describing Expert’s Uncertainty (cont-d)

- Probability values describing expert’s degree of confidence are known as *subjective probabilities*.

- This distinguishes them from usual (*objective*) probabilities, that describe the frequencies.

- For example, the fact that the probability $1/2$ of the coin falling heads means that:
  
  – in general,
  
  – the coin will fall heads in half of the cases.
17. How Do We Get Subjective Probabilities?

• A natural idea is to ask the experts themselves.
• In many cases, the expert cannot meaningfully provide the corresponding subjective probabilities.
• How can we then gauge the expert’s uncertainty?
• Sometimes, we have a record of past estimates of the same expert, for which we know the ground truth.

Then, for this expert:

– we can estimate our degree of confidence $c_i$ in this expert’s statement
– as the proportion of cases in which the expert turned out to be right.

• For example, if in the past, the medical doctor was right 80% of the time, we take $c_i = 0.8$. 
18. How To Get Subjective Probabilities (cont-d)

- Sometimes, we cannot do this for each individual expert.
- However, we can estimate the overall subjective probability $c$ of experts.
- The confidence $c$ is usually close to 1, so it makes sense to represent it as $c = 1 - \varepsilon$ for some small $\varepsilon > 0$.
- In this case:
  - we take $p_i = c = 1 - \varepsilon$ if the experts believe that the $i$-th object has the tested property, and
  - we take $p_i = 1 - c = \varepsilon$ if they don’t.
- What if we do not have the record of this expert’s past estimates?
- To do that, we can use standard techniques from decision theory.
19. How To Get Subjective Probabilities (cont-d)

- Namely, we can ask this expert to compare, for different \( p \in [0, 1] \), the following alternatives:
  - getting a certain reward (e.g., $100) with probability \( p \), or
  - getting the exact same reward if the statement \( S \) turned out to be true.
- If the expert prefers the 1st alternative, this means that his/her subjective probability of \( S \) is smaller than \( p \).
- If the expert prefers the 2nd alternative, this means that his/her subjective probability is larger than \( p \).
- We can use the following bisection procedure to find the corresponding subjective probability.
- In the beginning, all we know about the subjective probability \( p \) is that it is somewhere in \([p, \overline{p}] = [0, 1] \).
20. How To Get Subjective Probabilities (cont-d)

- At each stage of this process, we will decrease the size of this interval by half.

- This can be done as follows.

- Suppose that at some stage, we have an interval \([p, \bar{p}]\).

- Then, on the next stage, we:
  
  - compute the midpoint \(p_m = \frac{p + \bar{p}}{2}\) and
  
  - ask the expert to compare “reward with probability \(p_m\)” with “reward if \(S\) is true”.

- If the expert prefers the alternative “reward with probability \(p_m\)”, this means that \(p < p_m\).

- We already know that the subjective probability \(p\) is in the interval \([\underline{p}, \bar{p}]\) and, thus, \(\underline{p} \leq p\).

- So, we conclude that \(p\) is in the interval \([\underline{p}, p_m]\).
21. How To Get Subjective Probabilities (cont-d)

- If the expert prefers the alternative “reward is $S$ is true”, this means that $p_m < p$.

- So, we conclude that $p$ is in the interval $[p_m, \bar{p}]$.

- In both cases, we get an interval of half-size that contains the actual subjective probability.

- We start with an interval of width 1.

- In the first step, we decrease the width of the interval to 1/2, in 2 steps to 1/4, …

- In $k$ steps, we get an interval of width $2^{-k}$. 
22. How To Get Subjective Probabilities (cont-d)

- The midpoint of this interval represents the subjective probability with accuracy $2^{-(k+1)}$.
- This way, after a small number of iterations, we get the subjective probability with a reasonably high accuracy.
- In 3 steps – i.e., by asking 3 questions to the expert – we estimate $p$ with accuracy $2^{-4} = \frac{1}{16} < 10\%$.
- In 6 steps – i.e., by asking 6 questions to the expert – we estimate $p$ with accuracy $2^{-7} = \frac{1}{128} < 1\%$.
- In 9 steps – i.e., by asking 9 questions to the expert – we estimate $p$ with accuracy $2^{-10} = \frac{1}{1024} < 0.1\%$. 
23. How to Take Expert’s Uncertainty into Account: General Analysis

• Let $E_+$ be the set of all the objects that, according to the experts, most probably have the desired property.

• Let $E_-$ be the set of all the objects that, according to the experts, most probably don’t have the property.

• In general, due to the expert uncertainty, $E_+ \neq P$ and $E_- \neq N$.

• Let $n$ denote the overall number of tested objects:

$$n = |P| + |N| = |E_+| + |E_-| = |S_-| + |S_+|.$$

• Let us enumerate these objects by numbers from 1 to $n$.

• Then, all the sets of objects become subsets of the sample $\{1, \ldots, n\}$. 


24. Taking Expert’s Uncertainty into Account (cont-d)

- Let $\chi_P(i)$ denote the characteristic function of the set $P$ of all the objects that have the tested property, i.e.:
  - if the object $i$ has the property, then $\chi_P(i) = 1$, and
  - if the object $i$ does not have the tested property, then $\chi_P(i) = 0$.

- We consider situations in which we do not know for sure whether the $i$-th object has the tested property.

- All we know, based on the expert’s estimate, is that this happens with probability $p_i$.

- In other words, the value $\chi_P(i)$ is a random variable:
  - with probability $p_i$, we have $\chi_P(i) = 1$, and
  - with the remaining probability $1 - p_i$, we have $\chi_P(i) = 0$. 
25. Taking Expert’s Uncertainty into Account (cont-d)

• In statistics, for each random variable $\eta$, a reasonable idea is to compute its mean $E[\eta]$ and its variance $V[\eta]$.

• For the random variable $\chi_P(i)$, we have

\[
E[\chi_P(i)] = p_i \cdot 1 + (1 - p_i) \cdot 0 = p_i \quad \text{and}
\]

\[
V[\chi_P(i)] = E \left[ (\chi_P(i) - E[\chi_P(i)])^2 \right] =
\]

\[
p_i \cdot (1 - p_i)^2 + (1 - p_i) \cdot (0 - p_i)^2 =
\]

\[
p_i \cdot (1 - p_i)^2 + (1 - p_i) \cdot p_i^2 = p_i \cdot (1 - p_i) \cdot [(1 - p_i) + p_i] = p_i \cdot (1 - p_i).
\]

• What is the number of elements $|P|$ in the set $P$?
26. Taking Expert’s Uncertainty into Account (cont-d)

- This number can be obtained if:
  - we consider all the elements from the sample \{1, \ldots, n\} one by one, and
  - add 1 every time we have an element from the set \( P \), i.e., every time when \( \chi_P(i) = 1 \).

- If the element \( i \) does not belong to the set \( P \) (i.e., when \( \chi_P(i) = 0 \)), then we do not add anything.

- This is also equivalent to adding \( \chi_P(i) \).

- So, we can simply add all the values \( \chi_P(i) \) corresponding to all \( n \) objects: \( |P| = \sum_{i=1}^{n} \chi_P(i) \).

- To be able to get a good estimate of the test’s quality, we need to test a sufficiently large number of objects.
27. Taking Expert’s Uncertainty into Account (cont-d)

- Thus we can conclude that the number \( n \) is large.
- It is reasonable to assume that the estimates corresponding to different objects are statistically independent.
- So, the above sum is the sum of a large number of small independent random variables.
- It is known that, due to the Central Limit Theorem, the distribution of such sums is close to Gaussian.
- Thus, it is reasonable to assume that \( |P| \) is normally distributed.
- Its mean is equal to the sum of the means, i.e.,

\[
E[|P|] = \sum_{i=1}^{n} p_i.
\]
28. Taking Expert’s Uncertainty into Account (cont-d)

- For the sum of independent random variables, the variance is equal to the sum of the variables, so we have
  \[ V[|P|] = \sum_{i=1}^{n} p_i \cdot (1 - p_i). \]

- Now, we are ready to analyze how the expert’s uncertainty affect the values of the three characteristics.

- We will start with the case of precision, which turns out to be the easiest to analyze.
29. Estimating Precision

- Precision PPV is defined as the ratio $|P \cap S_+|/|S_+|$.

- $S_+$ is the set of all the objects that the test classifies as having the tested property.

- This set does not depend on expert estimates.

- The only thing that, in this formula, depends on the expert estimates, is the value $|P \cap S_+|$; so:

$$E[PPV] = \frac{1}{|S_+|} \cdot \sum_{i \in S_+} p_i,$$

$$V[PPV] = \frac{1}{|S_+|^2} \cdot \sum_{i \in S_+} p_i \cdot (1 - p_i).$$

- Strictly speaking, to this variance, we should add the variance caused by the finiteness of sample.
30. Case When We Only Know the Overall Degree of Confidence $C = 1 - \varepsilon$ in Experts

- We have $p_i = 1 - \varepsilon$ if $i \in E_+$ and $p_i = \varepsilon$ if $i \in E_-$, so:

$$\sum_{i \in S_+} p_i = \sum_{i \in S_+ \cap E_+} (1-\varepsilon) + \sum_{i \in S_+ \cap E_-} \varepsilon = |S_+ \cap E_+| \cdot (1-\varepsilon) + |S_+ \cap E_-| \cdot \varepsilon.$$

- Here, $|S_+ \cap E_-| = |S_+| - |S_+ \cap E_-|$, so

$$\sum_{i \in S_+} p_i = |S_+ \cap E_+| \cdot (1 - 2\varepsilon) + |S_+| \cdot \varepsilon.$$

- Therefore, $E[PPV] = (1 - 2\varepsilon) \cdot \frac{|S_+ \cap E_+|}{|S_+|} + \varepsilon$.

- So, we take the value that we would have obtained if we did not take expert uncertainty into account:
  - multiply it by $1 - 2\varepsilon$, and
  - add $\varepsilon$ to the resulting product.
31. Case When We Only Know the Overall Degree of Confidence (cont-d)

- Similarly, we have \( p_i \cdot (1 - p_i) = \varepsilon \cdot (1 - \varepsilon) \), so

\[
\sum_{i \in S_+} p_i \cdot (1 - p_i) = |S_+| \cdot \varepsilon \cdot (1 - \varepsilon),
\]

and we get

\[
V[PPV] = \frac{\varepsilon \cdot (1 - \varepsilon)}{|S_+|}.
\]
32. Based on Precision, When Is One Testing Method Better Than the Other?

- What if we have two different methods, with:
  - means $E[\text{PPV}_1]$ and $E[\text{PPV}_2]$ and
  - variances $V[\text{PPV}_1]$ and $V[\text{PPV}_2]$.

- We can use the usual technique for comparing two Gaussian random variables.

- We conclude that the first method is better if
  $$E[\text{PPV}_1] - E[\text{PPV}_2] \geq t \cdot \sqrt{V[\text{PPV}_1] + V[\text{PPV}_2]}.$$

- Here, the value $t$ depends on the desired confidence level.
33. Estimating Sensitivity (TPR)

- In terms $\chi_P(i)$: $TPR = \frac{\sum_{i \in S_+} \chi_P(i)}{\sum_{i=1}^{n} \chi_P(i)}$.

- Here, $\sum_{i=1}^{n} \chi_P(i) = \Sigma_+ + \Sigma_-$, where we denoted

$$\Sigma_+ \text{ def } = \sum_{i \in S_+} \chi_P(i) \text{ and } \Sigma_- \text{ def } = \sum_{j \in S_-} \chi_P(j).$$

- These two sums contain different random variables $\chi_P(i)$ and $\chi_P(j)$.

- We assumed that all the variables $\chi_P(i)$ an $\chi_P(j)$ are independent.

- So, the sums $\Sigma_+$ and $\Sigma_-$ are independent as well.
34. Estimating Sensitivity (cont-d)

- Here, \( TPR = \frac{\Sigma_+}{\Sigma_+ + \Sigma_-} \); so:
  - similarly to the case of precision,
  - we can conclude that \( \Sigma_+ \) and \( \Sigma_- \) are independent (approximately) Gaussian random variables, with
    \[
    E[\Sigma_+] = \sum_{i \in S_+} p_i \quad \text{and} \quad E[\Sigma_-] = \sum_{j \in S_-} p_j,
    \]
    \[
    V[\Sigma_+] = \sum_{i \in S_+} p_i \cdot (1-p_i) \quad \text{and} \quad V[\Sigma_-] = \sum_{j \in S_-} p_j \cdot (1-p_j).
    \]
- We can thus find the mean and standard deviation of TPR if we:
  - simulate Gaussian random variables \( \Sigma_+ \) and \( \Sigma_- \),
  - then compute the ratio for each simulation, and
  - compute the mean and average of these simulation results.
35. Estimating Sensitivity (cont-d)

- If we only know the overall degree of confidence \( c = 1 - \varepsilon \) in expert statements, then

  \[
  E[\Sigma_+] = (1 - \varepsilon) \cdot |S_+ \cap E_+| + \varepsilon \cdot |S_+ \cap E_-|,
  \]

  \[
  E[\Sigma_-] = (1 - \varepsilon) \cdot |S_- \cap E_+| + \varepsilon \cdot |S_- \cap E_-|,
  \]

  \[
  V[\Sigma_+] = \varepsilon \cdot (1 - \varepsilon) \cdot |S_+|, \text{ and } V[\Sigma_-] = \varepsilon \cdot (1 - \varepsilon) \cdot |S_-|.
  \]

- For each characteristic \( X \), the 1st testing method is better if:

  \[
  E[X_1] - E[X_2] \geq t \cdot \sqrt{V[X_1] + V[X_2]}.
  \]
36. Estimating Specificity

- In terms of the values $\chi_P(i)$, sensitivity is:

$$TNR = \frac{\sum_{i \in S_+} (1 - \chi_P(i))}{\sum_{j=1}^{n} (1 - \chi_P(j))} = \frac{|S_-| - \sum_{j \in S_-} \chi_P(j)}{n - \sum_{i=1}^{n} \chi_P(i)} = \frac{|S_-| - \Sigma_-}{n - \Sigma_+ - \Sigma_-}.$$ 

- We already know that $\Sigma_+$ and $\Sigma_-$ can be viewed as independent normally distributed random variables.

- We know their means and variances.

- Thus, we arrive at the following algorithm.

- First, we find the values of the mean and variance of $\Sigma_+$ and $\Sigma_-$. 
37. Estimating Specificity (cont-d)

- Then, several ($K$) times:
  - we run a usual random number generator for normally distributed random variables
  - to get $K$ simulated values $\Sigma^{(k)}_{+}$ and $\Sigma^{(k)}_{-}$.

- We estimate specificity as

$$\text{TNR}^{(k)} = \frac{|S_{-}| - \Sigma^{(k)}_{-}}{n - \Sigma^{(k)}_{+} - \Sigma^{(k)}_{-}}.$$

- Finally, based on these simulated values, we estimate the mean and variance of TNP in the usual way, as:

$$E[\text{TNR}] = \frac{1}{K} \cdot \sum_{k=1}^{K} \text{TNR}^{(k)} \quad \text{and}$$

$$V[\text{TNR}] = \frac{1}{K - 1} \cdot \sum_{k=1}^{K} \left(\text{TNR}^{(k)} - E[\text{TNR}]\right)^2.$$
38. Estimating Specificity (cont-d)

- How do we compare two methods?
- We say that the first testing method is better if

\[ E[TNR_1] - E[TNR_2] \geq t \cdot \sqrt{V[TNR_1] + V[TNR_2]} \]
39. Conclusion

- A usual way of gauging the quality of a testing method is to compare its results with ground truth.
- However, in many practical situations, we do not always know the ground truth.
- For example, we may want to gauge the quality of a medical diagnostic system.
- However, for some patients, the medical doctors are not 100% sure what is the correct diagnosis.
- Usually, we:
  - either ignore such cases,
  - or simply ignore the uncertainty and consider the most probable diagnosis as the ground truth.
40. Conclusion (cont-d)

- We want a more accurate description of a quality of a testing method.
- So, it is desirable to explicitly take into account the degree of expert’s certainty.
- In this talk, we provide methods that:
  - explicitly take into account these degrees of certainty
  - when estimating the quality of a testing method.
41. Acknowledgments

This work was supported in part by the National Science Foundation grants:

- 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science);
- HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.