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

Nicholas Gray[1], Scott Ferson[1], and Vladik Kreinovich[2]
[1]*Institute for Risk and Uncertainty*
University of Liverpool
Liverpool L69 7ZF, UK
Nicholas.Gray@liverpool.ac.uk, sandp8@gmail.com
[2]*Department of Computer Science*
University of Texas at El Paso
El Paso, TX 79968, USA, vladik@utep.edu

Abstract. The quality of a testing method is usually measured by using sensitivity, specificity, and/or precision. To compute each of these three characteristics, we need to know the ground truth, i.e., we need to know which objects actually have the tested property. In many applications (e.g., in medical diagnostics), the information about the objects comes from experts, and this information comes with uncertainty. In this paper, we show how to take this uncertainty into account when gauging the quality of testing methods.

Keywords: sensitivity, specificity, precision, unknown ground truth

1. Formulation of the Problem

Formulation of the problem: in brief. In many practical situations, algorithms help us recognize the situation. Let us give a few examples.

- In medicine, algorithms use symptoms and measurement results to provide a diagnosis.
- In engineering, algorithms take the results of measurements and observations and, based on these results, decide whether, e.g., a road is expected to seriously deteriorate 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, etc.

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 situations when we eventually learn the ground truth, 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. 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.

To get 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 paper.

Before we describe how we propose to do it, let us first describe the problem in more detail.

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 corresponding disease.

To gauge the quality of a testing method – and to compare the quality of different testing methods – several characteristics are used. The most widely used are sensitivity, specificity, and precision; see, e.g., (Yerushalmy, 1947; Altman and Bland, 1994; Boyko, 1994; Loong, 2003; Macmillan and Creelman, 2004; Parikh et al., 2008; Powers, 2011; Sheskin, 2011; Ting, Sammut, and Webb, 2011; Tharwat, 2018). In order to describe these characteristics, let us introduce the corresponding notations.

Notations and comments.

- Let P denote the set of all the objects from the tested sample that *actually have* the tested property (e.g., 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 (e.g., 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 (e.g., the set of all the people who the test classified as having the tested disease).
- Let S_- denote the set of all the objects for which the test *concluded that they do not have* the tested property (e.g., the set of all the people who the test classified as having the tested disease).

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, as we have mentioned, 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. In general, the number of elements in a set S will be denoted by $|S|$.

Let us now describe the usual characteristics.

Sensitivity. The first of the three characteristics is *sensitivity*. It is also known as *recall* or *True Positive Rate* – TPR for short. In the formulas in this paper, we will use the abbreviation TPR to describe sensitivity.

Sensitivity is defined as the proportion, among all the objects with the tested property, of the ones that were correctly classified by the test: e.g., the proportion of sick people for which the test recognized the disease.

In terms of our notations, the set of objects that have the tested property is P. The number of elements in this set is $|P|$. Among these objects, the set of all objects that have been correctly classified by the testing method is the intersection $P \cap S_+$ of the set P and the set S_+ of all the objects that were classified by the testing method as having the property. The number of such objects is equal to $|P \cap S_+|$. Thus, the sensitivity is equal to

$$\text{TPR} = \frac{|P \cap S_+|}{|P|}. \tag{1}$$

Specificity. The second of the three most used characteristics is *specificity*. It is also known as *True Negative Rate* – TNR, for short. In the formulas in this paper, we will use the abbreviation TNR to describe specificity.

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: e.g., the proportion of healthy people that this test classified as healthy.

In terms of our notations, the set of objects that do not have the tested property is N. The number of elements in this set is $|N|$. Among these objects, the set of all objects that have been correctly classified by the method is the intersection $N \cap S_-$ of the set N and the set S_- of all the objects that were classified by the testing method as not having the property. The number of such objects is equal to $|N \cap S_-|$. Thus, the specificity is equal to

$$\text{TNR} = \frac{|N \cap S_-|}{|N|}. \tag{2}$$

Precision. The final of the three characteristics is *precision*. It is also known as Positive Predictive Value – PPV, for short. In the formulas in this paper, we will use the abbreviation PPV to describe precision.

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 – e.g., the proportion of sick people among those that the test classified as sick.

In terms of our notations, the set of objects that were classified as having the property is S_+. The number of elements in this set is $|S_+|$. Among these objects, the set of all objects that actually have the tested property is the intersection $P \cap S_+$ of the set S_+ and the set P of all the objects that actually have the tested property. The number of such objects is equal to $|P \cap S_+|$. Thus, the precision is equal to

$$\text{PPV} = \frac{|P \cap S_+|}{|S_+|}. \tag{3}$$

How can we use these characteristics to compare different testing methods. For each of the three characteristics, the larger the value of the characteristic, the better – and in the prefect 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.

Comment. Of course, to make a definite conclusion about which testing method is better, we 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.

So, strictly speaking, 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 statistically 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.

Important problem: often, we do not know the "ground truth". The formulas for computing the above three characteristics assume that we know know the "ground truth", i.e., that we know exactly:

- which objects have the tested property and
- which objects do not have this property.

In the above example, which patients have the tested disease.

In practice, however, this information often comes from experts – e.g., from medical doctors – and 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?

What we do in this paper. This is the problem that we address in this paper: we sow how to take the expert's uncertainty into account when estimating the above characteristics of the testing method.

2. How to Describe Expert's Uncertainty

Need for subjective probability. 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.

Since the whole procedure is based on statistics, it is reasonable to try to gauge the expert's degree of certainty c_i in his or her statement by a probability value. Once we know this degree of certainty, then:

– 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 $p_i = c_i$. Correspondingly, 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. Correspondingly, the probability that the object i *has* the tested property is equal to $p_i = 1 - c_i$.

Comment. Probability values describing expert's degree of confidence are known as *subjective probabilities* – to distinguish them from usual (*objective*) probabilities, that describe the frequency with which certain events occur. 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.

How do we get subjective probabilities? Where can we get the subjective probabilities from? A natural idea is to ask the experts. Sometimes, they are able to gauge their own degrees of certainty by providing the corresponding number.

What is the expert cannot provide such probabilities – but we have a record of the expert's past estimates. In many cases, the expert cannot meaningfully provide the corresponding subjective probabilities. How can we then gauge the expert's uncertainty?

One possible approach is to use the above analogy between subjective and objective probabilities. If we have a record of past estimates of the same expert, estimates for which we actually 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 $c_i = 0.8$.

Sometimes, we cannot do this for each individual expert, but we can estimate the overall subjective probability c of experts. The confidence c is usually close to 1, to it makes sense to represent it as $c = 1 - \varepsilon$ for some small $\varepsilon > 0$. In this case, we take $p_i = c = 1 - \varepsilon$ is the experts believe that the i-th object has the tested property, and $p_i = 1 - c = \varepsilon$ if they don't.

What can we do in all other cases? But what if an expert cannot estimate his/her degree of confidence by a number, and we do not have the record of this expert's past estimates. How can we then estimate the expert's degree of confidence?

To do that, we can use standard techniques from decision theory; see, e.g., (Fishburn, 1969; Luce and Raiffa, 1989; Raiffa, 1997; Nguyen, Kosheleva, and Kreinovich, 2009; Kreinovich, 2014). According to decision theory, to estimate the expert's certainty in a statement S, we can ask this expert to compare, for different values p from the interval $[0, 1]$, the following two 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.

Clearly:

– If the expert prefers the first alternative, this means that his/her subjective probability of S is smaller than p.

– If the expert prefers the second alternative, this means that his/her subjective probability of S is larger than p.

We can use following bisection procedure to find the corresponding subjective probability. In the beginning, all we know about the subjective probability p is that it is located somewhere in the interval $[\underline{p}, \overline{p}] = [0, 1]$. 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 $[\underline{p}, \overline{p}]$. Then, on the next stage, we compute the midpoint

$$p_m = \frac{\underline{p} + \overline{p}}{2} \tag{4}$$

and ask the expert to compare the alternative "reward with probability p_m" with the alternative "reward if S is true".

– If the expert prefers the alternative "reward with probability p_m", this means that his/her subjective probability is smaller than p_m. Since we already know that the subjective probability p is in the interval $[\underline{p}, \overline{p}]$ and is, thus, larger than or equal to $\underline{p}$, we can thus conclude that p is in the interval $[\underline{p}, p_m]$.

– If the expert prefers the alternative "reward is S is true", this means that his/her subjective probability is larger than p_m. Since we already know that the subjective probability p is in the interval $[\underline{p}, \overline{p}]$ and is, thus, smaller than or equal to $\overline{p}$, we can thus conclude that p is in the interval $[\underline{p}_m, \overline{p}]$.

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, ..., and in k steps, we get an interval of width 2^{-k}. If we take a midpoint of this interval, this midpoint 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 particular:

– in 3 steps – i.e., by asking 3 questions to the expert – we estimate the subjective probability with accuracy $2^{-4} = \frac{1}{16} < 10\%$;

– in 6 steps – i.e., by asking 6 questions to the expert – we estimate the subjective probability with accuracy $2^{-7} = \frac{1}{128} < 1\%$; and

– in 9 steps – i.e., by asking 9 questions to the expert – we estimate the subjective probability with accuracy $2^{-10} = \dfrac{1}{1024} < 0.1\%$.

Summarizing. By using one of the above methods, we can estimate, for each object i, the expert's degree of confidence c_i in his or her statement about this object. Depending on whether this statement was positive or negative, we can then estimate the expert's subjective probability p_i that the i-th object has the tested property:

– if the expert believes that most probably the object has the tested property, then we take $p_i = c_i$, and

– if the expert believes that most probably the object does not have the tested property, then we take $p_i = 1 - c_i$.

In some cases, instead of individual values p_i for each i, we only know the overall degree of confidence $c = 1 - \varepsilon$ in the expert's statement. In this case:

– if the expert believes that most probably the object has the tested property, then we take $p_i = c = 1 - \varepsilon$, and

– if the expert believes that most probably the object does not have the tested property, then we take $p_i = 1 - c = \varepsilon$.

Let us now show how we can use these subjective probabilities p_i.

3. How to Take Expert's Uncertainty into Account: General Analysis

Notations. Let us first introduce some additional notations.

Let us denote:

– by E_+ the set of all the objects that, according to the experts, most probably have the desired property, and

– by E_- the set of all the objects that, according to the experts, most probably do not have the desired property.

In general, due to the expert uncertainty:

– the set E_+ may be different from the set of P of the objects that actually have the tested property, and

– the set E_- may be different from the set of N of the objects that actually do not have the tested property.

Let n denote the overall number of tested objects. In terms of our previous notations,

$$n = |P| + |N| = |E_+| + |E_-| = |S_-| + |S_+|. \tag{5}$$

Let us enumerate these objects by numbers from 1 to n. In these notation, all the sets that considered earlier – namely, the sets P, N, S_-, and S_+ and their intersections – become subsets of the sample $\{1, \ldots, n\}$.

Let $\chi_P(i)$ denote the characteristic function of the set P of all the objects that actually have the tested property, i.e.:

- if the object i has the tested property, then $\chi_P(i) = 1$, and
- if the object i does not have the tested property, then $\chi_P(i) = 0$.

General analysis of the problem. We consider situations in which we do not know for sure whether the i-th object has the tested property or not. 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 $\chi_P(i) = 0$.

In statistics, for each random variable η, a reasonable idea is to compute its mean $E[\eta]$ and its variance $V[\eta]$– and, as we will see later, this is useful in our case as well. For the random variable $\chi_P(i)$, we have

$$E\left[\chi_P(i)\right] = p_i \cdot 1 + (1 - p_i) \cdot 0 = p_i \tag{6}$$

and

$$V\left[\chi_P(i)\right] = E\left[\left(\chi_P(i) - E\left[\chi_P(i)\right]\right)^2\right] = p_i \cdot (1 - p_i)^2 + (1 - p_i) \cdot (0 - p_i)^2 =$$

$$p_i \cdot (1 - p_i)^2 + (1 - p_i) \cdot p_i^2 = p_i \cdot (1 - p_i) \cdot [(1 - p_i) + p_i] = p_i \cdot (1 - p_i). \tag{7}$$

We can use these results to answer, e.g., a question of how many objects actually have the desired property, i.e., what is the number of elements $|P|$ in the set P. 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 – which is also equivalent to adding $\chi_P(i)$. So, we can describe the above procedure as simply adding all the values $\chi_P(i)$ corresponding to all n objects. As a result, we get the value

$$|P| = \sum_{i=1}^{n} \chi_P(i). \tag{8}$$

To be able to get a good estimate of the test's quality, we need to test a sufficiently large number of objects. Thus we can conclude that the number n is large. So, the above sum (8) is the sum of a large number of small independent random variables.

It is reasonable to assume that the estimates corresponding to different objects – and often produced by different experts – are statistically independent. It is known that, due to the Central Limit Theorem, the distribution of such sums is close to Gaussian; see, e.g., (Sheskin, 2011). Thus, it is reasonable to assume that PPV is normally distributed. Its mean is equal to the sum of the means, i.e.,

$$E[|P|] = \sum_{i=1}^{n} p_i. \tag{9}$$

For the sum of independent random variables, the variance is equal to the sum of the variables, so we have

$$V[|P|] = \sum_{i=1}^{n} p_i \cdot (1 - p_i). \tag{10}$$

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.

4. Estimating Precision: Analysis of the Problem

General case. According to the formula (3), precision PPV is defined as the ratio of $|P \cap S_+|$ to $|S_+|$. The set S_+ of all the objects that the test classifies as having the tested property does not depend on expert estimates. The only thing that, in this formula, depends on the expert estimates, is the value $|P \cap S_+|$ – since it depends on which objects actually have this property or not, and our only information about this comes from the experts.

Similarly to the previous section, we can conclude that

$$|P \cap S_+| = \sum_{i \in S_+} \chi_P(i). \tag{11}$$

Thus, we have

$$E\left[|P \cap S_+|\right] = \sum_{i \in S_+} p_i \tag{12}$$

and

$$V\left[|P \cap S_+|\right] = \sum_{i \in S_+} p_i \cdot (1 - p_i). \tag{13}$$

When we divide a random variable by a constant (in this case, by $|S_+|$), then the mean value divides by the same constant, while the variance divides by the square of this constant. So, we have

$$E[\mathrm{PPV}] = \frac{1}{|S_+|} \cdot \sum_{i \in S_+} p_i \tag{14}$$

and

$$V[\mathrm{PPV}] = \frac{1}{|S_+|^2} \cdot \sum_{i \in S_+} p_i \cdot (1 - p_i). \tag{15}$$

Case when we only know the overall degree of confidence $c = 1-\varepsilon$ in expert statements. In this case, we have $p_i = 1 - \varepsilon$ if $i \in E_+$ and $p_i = \varepsilon$ if $i \in E_-$. Thus:

$$\sum_{i \in S_+} p_i = \sum_{i \in S_+ \cap E_+} (1-\varepsilon) + \sum_{i \in S_+ \cap E_-} \varepsilon = |S_+ \cap E_+| \cdot (1-\varepsilon) + |S_+ \cap E_-| \cdot \varepsilon. \tag{16}$$

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

$$\sum_{i \in S_+} p_i = |S_+ \cap E_+| \cdot (1-2\varepsilon) + |S_+| \cdot \varepsilon. \tag{17}$$

Therefore,

$$E[\mathrm{PPV}] = (1-2\varepsilon) \cdot \frac{|S_+ \cap E_+|}{|S_+|} + \varepsilon. \tag{18}$$

Similarly, we have $p_i \cdot (1-p_i) = \varepsilon \cdot (1-\varepsilon)$, so $\sum\limits_{i \in S_+} p_i \cdot (1-p_i) = |S_+| \cdot \varepsilon \cdot (1-\varepsilon)$, and the formula (15) takes the form

$$V[\mathrm{PPV}] = \frac{\varepsilon \cdot (1-\varepsilon)}{|S_+|}. \tag{19}$$

Based on precision, when can we say that one testing method is better than the other one? To compare two different methods, with means $E[\mathrm{PPV}_1]$ and $E[\mathrm{PPV}_2]$ and variances $V[\mathrm{PPV}_1]$ and $V[\mathrm{PPV}_2]$, we can use the usual technique for comparing two random variables, and conclude that the difference between PPV_1 and PPV_2 if

$$E[\mathrm{PPV}_1] - E[\mathrm{PPV}_2] \geq t \cdot \sqrt{V[\mathrm{PPV}_1] + V[\mathrm{PPV}_2]}, \tag{20}$$

for an appropriate t (the value t depends on the desired confidence level).

5. Estimating Precision: Results

General conclusion. If we take into account expert uncertainty, then PPV – as well as two other characteristics – becomes a random variable.

General case. If we know, for each object i, the subjective probability p_i that this object has the tested property, then the mean and variance of PPV can be determined by using formulas (14) and (15).

Case when we only know the overall degree of confidence $c = 1-\varepsilon$ in expert statements. In this case, we can estimate the mean nd variance of PPV by using formulas (18) and (19).

Comments.

– The formula (18) can be reformulated as follows: 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 ε to the resulting product.

– Strictly speaking, to the variance values estimated by using formulas (15) and (19), we should add the variances caused by the fact that we are estimate PPV based on the finite sample. Since the expert uncertainty and the uncertainty caused by the finiteness of the sample are independent, to get the overall variance, we should simply add the new expression to the variance to the known expressions corresponding to sample finiteness.

How do we decide which testing method is better. To decide whether one testing method is statistically significantly better than another one, we use the formula (20) with an appropriate value t.

6. Estimating Sensitivity: Analysis of the Problem

General case. In terms of the values $\chi_P(i)$, the actual value of the sensitivity is

$$\mathrm{TPR} = \frac{\sum\limits_{i \in S_+} \chi_P(i)}{\sum\limits_{i=1}^{n} \chi_P(i)}. \tag{21}$$

Here,

$$\sum_{i=1}^{n} \chi_P(i) = \Sigma_+ + \Sigma_-, \tag{22}$$

where we denoted $\Sigma_+ \stackrel{\text{def}}{=} \sum\limits_{i \in S_+} \chi_P(i)$ and $\Sigma_- \stackrel{\text{def}}{=} \sum\limits_{j \in S_-} \chi_P(j)$. Since these two sums contain different random variables $\chi_P(i)$ and $\chi_P(j)$, and we assumed that all the variables $\chi_P(i)$ an $\chi_P(j)$ are independent, the sums Σ_+ and Σ_- are independent as well. Thus

$$\mathrm{TPR} = \frac{\Sigma_+}{\Sigma_+ + \Sigma_-}. \tag{23}$$

Here, similarly to the case of precision, we can conclude that both Σ_+ and Σ_- are independent (approximately) Gaussian random variables, with means

$$E[\Sigma_+] = \sum_{i \in S_+} p_i \text{ and } E[\Sigma_-] = \sum_{j \in S_-} p_j \tag{24}$$

and variances

$$V[\Sigma_+] = \sum_{i \in S_+} p_i \cdot (1 - p_i) \text{ and } V[\Sigma_-] = \sum_{j \in S_-} p_j \cdot (1 - p_j). \tag{25}$$

We can thus find the mean and standard deviation of TPR if we simulate two Gaussian random variables Σ_+ and Σ_-, then compute the ratio (23) for each simulation, and compute the mean and average of these simulation results.

Case when we only know the overall degree of confidence $c = 1-\varepsilon$ in expert statements. In this case, the formulas (24) and (25) take the following form:

$$E[\Sigma_+] = (1-\varepsilon)\cdot|S_+ \cap E_+| + \varepsilon\cdot|S_+ \cap E_-| \text{ and } E[\Sigma_-] = (1-\varepsilon)\cdot|S_- \cap E_+| + \varepsilon\cdot|S_- \cap E_-|; \quad (26)$$

$$V[\Sigma_+] = \varepsilon\cdot(1-\varepsilon)\cdot|S_+| \text{ and } V[\Sigma_-] = \varepsilon\cdot(1-\varepsilon)\cdot|S_-|. \quad (27)$$

How do we decide which testing method is better. For each characteristic X, we can say – similarly to the formula (20) – that the first testing method is better if

$$E[X_1] - E[X_2] \geq t\cdot\sqrt{V[X_1] + V[X_2]}, \quad (28)$$

for an appropriate t; the value t depends on the desired confidence level.

7. Estimating Sensitivity: Resulting Algorithm

How to estimate sensitivity of a testing method based on the testing results. First, depending on whether we know all the values p_i for each i or only one value ε, we use either the formulas (24)–(25) or the formulas (26)–(27) to find the values of the mean and variance of Σ_+ and Σ_-.

Then, several (K) times we run a usual random number generator for normally distributed random variables to get N simulated values $\Sigma_+^{(k)}$ and $\Sigma_-^{(k)}$. Based on these simulated values, we use the formula (23) to estimate the simulated values of TPR as

$$\mathrm{TPR}^{(k)} = \frac{\Sigma_+^{(k)}}{\Sigma_+^{(k)} + \Sigma_-^{(k)}}. \quad (29)$$

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

$$E[\mathrm{TPR}] = \frac{1}{K}\cdot\sum_{k=1}^{K}\mathrm{TPR}^{(k)} \text{ and } V[\mathrm{TPR}] = \frac{1}{K-1}\cdot\sum_{k=1}^{K}\Big(\mathrm{TPR}^{(k)} - E[\mathrm{TPR}]\Big)^2. \quad (30)$$

How do we decide which testing method is better. We say that the first testing method is better if

$$E[\mathrm{TPR}_1] - E[\mathrm{TPR}_2] \geq t\cdot\sqrt{V[\mathrm{TPR}_1] + V[\mathrm{TPR}_2]}, \quad (31)$$

for an appropriate t; the value t depends on the desired confidence level.

8. Estimating Specificity: Analysis of the Problem and the Resulting Algorithm

Analysis of the problem. In terms of the values $\chi_P(i)$, the actual value of the sensitivity is

$$\mathrm{TNR} = \frac{\sum\limits_{i \in S_-} (1 - \chi_P(i))}{\sum\limits_{j=1}^{n} (1 - \chi_P(j))} = \frac{|S_-| - \sum\limits_{j \in S_-} \chi_P(j)}{n - \sum\limits_{i=1}^{n} \chi_P(i)} = \frac{|S_-| - \Sigma_-}{n - \Sigma_+ - \Sigma_-}. \tag{32}$$

We already know that Σ_+ and Σ_- can be viewed as independent normally distributed random variables, with known means and variances. Thus, we arrive at the following algorithm.

How to estimate sensitivity of a testing method based on the testing results. First, depending on whether we know all the values p_i for each i or only one value ε, we use either the formulas (24)–(25) or the formulas (26)–(27) to find the values of the mean and variance of Σ_+ and Σ_-.

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

Up to now, we perform the same computation steps as when estimating sensitivity. Now, the computations differ. To estimate specificity, based on the simulated values $\Sigma_+^{(k)}$ and $\Sigma_-^{(k)}$, we use the formula (32) to estimate the simulated values of TNR as

$$\mathrm{TNR}^{(k)} = \frac{|S_-| - \Sigma_-^{(k)}}{n - \Sigma_+^{(k)} - \Sigma_-^{(k)}}. \tag{33}$$

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

$$E[\mathrm{TNR}] = \frac{1}{K} \cdot \sum_{k=1}^{K} \mathrm{TNR}^{(k)} \text{ and } V[\mathrm{TNR}] = \frac{1}{K-1} \cdot \sum_{k=1}^{K} \Big(\mathrm{TNR}^{(k)} - E[\mathrm{TNR}]\Big)^2. \tag{34}$$

How do we decide which testing method is better. We say that the first testing method is better if

$$E[\mathrm{TNR}_1] - E[\mathrm{TNR}_2] \geq t \cdot \sqrt{V[\mathrm{TNR}_1] + V[\mathrm{TNR}_2]}, \tag{35}$$

for an appropriate t; the value t depends on the desired confidence level.

9. Conclusion

A usual way of gauging the quality of a testing method is to compare the results of this method 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, but 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. To get a more accurate description of a quality of a testing method, it is desirable to explicitly take into account the degree of expert's certainty.

In this paper, we provide methods that explicitly take into account these degrees of certainty when estimating the quality of a testing method.

Acknowledgements

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) and HRD-1242122 (Cyber-ShARE Center of Excellence). 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.

The authors are thankful to all the participants of the First International Conference on Data and Information Fusion DIR'2019 (Santa Fe, New Mexico, August 21–23, 2019) for valuable discussions, and to the anonymous referees for useful advice.

References

Altman, D. G., and J. M. Bland, "Diagnostic tests. 1: Sensitivity and specificity", *BMJ*, 308(6943): 1552, 1994.

Boyko, E. J., "Ruling out or ruling in disease with the most sensitive or specific diagnostic test: short cut or wrong turn?", *Medical Decision Making*, 14(2): 175–179, 1994.

Fishburn, P. C., *Utility Theory for Decision Making*, John Wiley & Sons Inc., New York, 1969.

Kreinovich, V., Decision making under interval uncertainty (and beyond)", In: Guo, P., and W. Pedrycz (eds.), *Human-Centric Decision-Making Models for Social Sciences*, Springer Verlag, 2014, pp. 163–193.

Loong, T. W., "Understanding sensitivity and specificity with the right side of the brain", *BMJ*, 327(7417): 716–719, 2003.

Luce, R. D., and R. Raiffa, *Games and Decisions: Introduction and Critical Survey*, Dover, New York, 1989.

Macmillan, N. A., and C. D. Creelman, *Detection Theory: A User's Guide*, Psychology Press, Hove, East Sussex, United Kingdom, 2004.

Nguyen, H. T., O. Kosheleva, and V. Kreinovich, "Decision making beyond Arrow's 'impossibility theorem', with the analysis of effects of collusion and mutual attraction", *International Journal of Intelligent Systems*, 24(1): 27–47, 2009.

Parikh, R., A. Mathai, S. Parikh, G. Chandra Sekhar, and R. Thomas, "Understanding and using sensitivity, specificity and predictive values", *Indian Journal of Ophthalmology*, 56(1): 45–50, 2008.

Powers, D. M. W., "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation", *Journal of Machine Learning Technologies*, 2(1): 37–63, 2011.

Raiffa, H., *Decision Analysis*, McGraw-Hill, Columbus, Ohio, 1997.

Sheskin, D. J., *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.

Tharwat, A., "Classification assessment methods", *Applied Computing and Informatics*, 2018 doi:10.1016/j.aci.2018.08.003.

Ting, K. M., C. Sammut, and G. I. Webb, (eds.), *Encyclopedia of Machine Learning*, Springer, Cham, Switzerland, 2011.

Yerushalmy, J., "Statistical problems in assessing methods of medical diagnosis with special reference to X-ray techniques", *Public Health Reports*, 62(2): 1432–1439, 1947.