Supplementary Information

Problem description

Given is a haystack that consists of N haybales. We are interested in whether there is a needle in the haystack. For getting a rather quick answer to this question, we do not check all of the haybales present in the haystack, but only n out of the total N haybales. Due to the fact that we do not check all of the haybales, we will sometimes miss a needle in a haystack. In the following, we denote with TPR and FPR the global true positive and false positive rate, respectively, for the classification of the whole haystack. Analogously, TNR and FNR denote the global true and false negative rate.

One-phase search strategy

Assuming no error in classification of individual haybales

Assuming that we do not make any errors in the classification of the individual haybales, the TPR is given as

$$TPR = 1 - FNR \tag{1}$$
$$= 1 - (1-p)^n, \tag{2}$$

where p is the probability that there is a needle in an individual haybale and $(1-p)^n$ is the probability that there is no needle in any of the n haybales that we check. For $p \ll 1$, the total true positive rate can be approximated by $TPR \approx n \cdot p$.

For the FPR we only consider haystacks that do not contain any needle. Given that we do not make any errors in the classification of the individual haybales, we therefore will not find any needle in the whole search. Hence, $FPR = 0$.

Considering errors in classification of individual haybales

Let us now consider that we will make some errors in the classification of individual haybales. Let fpr and fnr denote the false positive and false negative rate, respectively, for the classification of a single haybale.

Assuming that there is a needle in the haystack, there are two ways for classifying a haybale as containing a needle: Either there is a needle and we correctly identify it, or there is no needle, but we incorrectly classify the haybale as containing one. The TPR hence is calculated as

$$TPR = 1 - FNR \tag{3}$$
$$= 1 - \Big(1 - \big((1-fnr)\cdot p + fpr \cdot (1-p)\big)\Big)^n, \tag{4}$$

where $(1-fnr)\cdot p + fpr\cdot(1-p)$ is the probability for a positive classification. Assuming, that we rarely make an error in the classification of a relevant case, i.e. the false negative rate equals zero, we obtain

$$TPR = 1 - \Big(1 - \big(p + fpr \cdot (1-p)\big)\Big)^n \tag{5}$$
$$\approx 1 - \Big(1 - p - fpr\Big)^n. \tag{6}$$

The above approximation is feasible, as both p and fpr are assumed to be very small and hence, their product is negligible.

Assuming that there is no needle in the haystack, we only get a positive result for the overall test if we incorrectly classify a haybale containing no needle as a relevant case, which happens with probability fpr. Thus, the FPR is given as

$$FPR = 1 - TNR \tag{7}$$
$$= 1 - \big(1 - fpr\big)^n. \tag{8}$$

Two-phase search strategy

In the case of an extremely low number of relevant instances, i.e. needles in the haystack, the search can be performed more efficiently using a two-phase approach that immensely reduces the time needed for the search. In a first quick test round, TPR is optimized without spending too much effort in keeping fpr low. Out of the n_1 haybales that are superficially screened, we disregard the negatives and keep the positives for a second phase, in which we thoroughly recheck the hits of the first round. Thus, we also achieve a low number of false positive haybales in the overall search process.

The global true positive rate TPR for the search of the haystack is equal to $1 - FNR$. The FNR is the probability of finding no needle in any of the n_1 checked haybales, given that there is at least one needle present in the whole haystack. For the two-phase search strategy, no needle is found in a haybale in the following cases: Given that there is a needle in the haybale we check, we could already miss it in the first round, or we could detect it in the first round, but miss it in the second round. Given that there is no needle present in the haybale we check, we could correctly identify it as an irrelevant case in the first round, or we could incorrectly identify it as a relevant case in the first round, but correctly identify it as irrelevant in the second round. In total, we thus get

$$TPR = 1 - FNR \tag{9}$$
$$= 1 - \Big(fnr_1 \cdot p + (1 - fnr_1) \cdot fnr_2 \cdot p + (1 - fpr_1) \cdot (1 - p) + fpr_1 \cdot (1 - fpr_2) \cdot (1 - p)\Big)^{n_1}, \tag{10}$$

where fnr_1, fnr_2 denote the false negative rates for checking a single haybale in the first and second round, respectively, and fpr_1, fpr_2 denote the respective false positive rates. Assuming, that we rarely make an error in the classification of a relevant case, i.e. the false negative rates equal zero, we can simplify the above equation:

$$TPR = 1 - \Big(1 - p - fpr_1 \cdot fpr_2 \cdot (1 - p)\Big)^{n_1} \tag{11}$$
$$\approx 1 - \Big(1 - p - fpr_1 \cdot fpr_2\Big)^{n_1}. \tag{12}$$

This approximation is feasible as fpr_1, fpr_2 and p are assumed to be rather small and hence, their product is negligible.

The global false positive rate FPR is the probability of incorrectly claiming to find a needle in a haystack that actually does not contain any needle. If there is no needle present in the whole haystack, none of the individual haybales will contain a needle either. We will correctly classify a

single haybale as irrelevant case if either we already correctly classify it as containing no needle in the first round, or we incorrectly claim to find a needle in the first round, but correctly classify it as an irrelevant case in the second round. Thus, the FPR results in

$$\begin{aligned} FPR &= 1 - TNR && (13)\\ &= 1 - \Big((1 - fpr_1) + fpr_1 \cdot (1 - fpr_2)\Big)^{n_1} && (14)\\ &= 1 - \Big(1 - fpr_1 \cdot fpr_2\Big)^{n_1}. && (15) \end{aligned}$$

Time comparison

An important factor we need to consider when choosing a suitable search strategy is the time required to perform the search. Let Δt denote the time needed to check a single haybale. The overall search time T can then be calculated as $T = n \cdot \Delta t$, where n is the number of checked individual haybales. In order to achieve a high global true positive rate TPR we need a considerable number of tested haybales n, which increases the time needed for the search. The overall time required can be reduced by decreasing the time to check an individual haybale. However, this will result in a higher number of false classifications and hence, higher false positive rates.

For the sake of simplicity, let us assume that the time for testing a single haybale scales with fpr^{-1}, and hence the total time T for testing n haybales is proportional to

$$T = \frac{n \cdot \Delta t}{fpr}, \tag{16}$$

i.e. x-fold lower local false positive rates require x-fold longer test times. If we want to avoid a high number of false positives, we need to spend a considerable amount of time for the search.

However, we can significantly speed up the search by using the two-phase search strategy. For $n = n_1$ and $fpr = fpr_1 \cdot fpr_2$, it holds that TPR and FPR of the one-phase and two-phase search approach are approximately the same (compare equations (6), (8), (12), (15)). The difference between the two search strategies, however, is the total overall time required to perform the search. For the two-phase search strategy, the first phase has a duration of

$$T_{\text{one-phase}} = \frac{n_1 \cdot \Delta t}{fpr_1}\,. \tag{17}$$

For the second phase, however, only the positives of the first phase are considered, which are given by

$$n_2 = (p + fpr_1 \cdot (1 - p)) \cdot n_1, \tag{18}$$

assuming that $fnr_1 = 0$. For the time required for the second phase, it holds that

$$T_{\text{two-phase}} = \frac{n_2 \cdot \Delta t}{fpr_2}\,. \tag{19}$$

Taken together, we obtain

$$\begin{aligned} T &= T_1 + T_2 && (20)\\ &= n_1 \cdot \Delta t \cdot \Big(\frac{1}{fpr_1} + \frac{p}{fpr_2} + \frac{fpr_1 \cdot (1 - p)}{fpr_2}\Big)\,. && (21) \end{aligned}$$

With equations (16) and (20) we now can calculate the ratio of the times needed for the one-phase and two-phase strategy:

$$\frac{T_{\text{two-phase}}}{T_{\text{one-phase}}} = fpr_2 + p \cdot fpr_1 + (1-p)(fpr_1)^2 \tag{22}$$

$$\approx fpr_2 + p \cdot fpr_1 + (fpr_1)^2. \tag{23}$$

The two-phase search approach is thus massively faster than the one-phase approach, if we compare same figures of merit for TPR and FPR.