

Optimal threshold problem: A novel 0-1 ILP-based approach

Pouria Ramazi¹ and Arash Mari Oriyad²

¹Department of Mathematics and Statistics, Brock University, St. Catharines, L2S 3A1, ON, Canada

²Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan 84156-83111, Iran

Abstract

Accuracy, recall, specificity, and precision are four of the most commonly used performance measures to evaluate binary classification models. To calculate these measures, probabilities generated by classifiers must be converted into deterministic labels using a threshold. An optimal threshold that maximizes performance measures can be challenging to find, especially when the number of instances is significantly large. We propose an integer linear programming (ILP) formulation to find a threshold maximizing any linear function of accuracy, recall, specificity, and precision. Computational results demonstrate apparent advantages in using ILP approach compared to a trivial exhaustive search.

Accuracy [2], recall [3], specificity [1], and precision are four of the most commonly-used performance measures to evaluate binary classification models. Considering TP , TN , FP , and FN as the number of true-positive, true-negative, false-positive, and false-negative instances, respectively, these measures are defined as follows:

$$\begin{aligned} accuracy &= \frac{TP + TN}{TP + TN + FP + FN} \\ recall &= \frac{TP}{TP + FN} \\ specificity &= \frac{TN}{TN + FP} \\ precision &= \frac{TP}{TP + FP} \end{aligned} \tag{1}$$

While calculating performance measures introduced in equation 1 require deterministic labels, classification models often produce probabilistic outputs. Hence, one must apply a threshold to the probabilities generated by a classifier to obtain deterministic labels. An optimal accuracy (resp. recall, specificity, and precision) threshold for a given set of instances is a real value between 0 and 1, resulting in the highest accuracy (resp. recall, specificity, and precision).

A trivial approach to determine such a threshold is an exhaustive search over all possible thresholds to obtain the best one. As an alternative method,

we introduce a 0-1 integer linear programming (ILP) formulation to the problem of finding the optimal threshold for accuracy, recall, specificity, precision, and any linear combination of them.

Assume a set of N instances with a binary target variable y . We call N^+ and N^- as the number of positive and negative instances. Also, suppose \hat{y}_i is the probability generated by a binary classifier for the i th instance. Then, \tilde{y}_i would be a binary decision variable that represents the label of i th instance after applying the threshold τ to \hat{y}_i . More precisely, if $\hat{y}_i < \tau$, then \tilde{y}_i will be 0 and vice versa. The goal is to find a value for τ which maximize a linear function of performance measures.

Using the introduced notations and according to equation 1, accuracy, recall, specificity, and precision can be calculated as follows:

$$\begin{aligned} accuracy &= 1 - \frac{1}{N} \sum_{i=1}^N |y_i - \tilde{y}_i| \\ recall &= \frac{1}{N^+} \sum_{i=1}^N y_i \tilde{y}_i \\ specificity &= \frac{1}{N^-} \sum_{i=1}^N (1 - y_i)(1 - \tilde{y}_i) \\ precision &= \frac{\sum_{i=1}^N y_i \tilde{y}_i}{\sum_{i=1}^N \tilde{y}_i} \end{aligned} \quad (2)$$

Put them all together, the following optimization formulation is achieved to find the optimal threshold maximizing $\alpha \times accuracy + \beta \times recall + \gamma \times specificity + \zeta \times precision$, where α, β, γ , and ζ are arbitrary coefficients set by user.

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^N \left(-\frac{\alpha}{N} |y_i - \tilde{y}_i| + \frac{\beta}{N^+} y_i \tilde{y}_i + \frac{\gamma}{N^-} (1 - y_i)(1 - \tilde{y}_i) \right) + \zeta \frac{\sum_{i=1}^N y_i \tilde{y}_i}{\sum_{i=1}^N \tilde{y}_i} \\ \text{subject to:} \quad & \tilde{y}_i = \begin{cases} 0, & \hat{y}_i < \tau \\ 1, & \hat{y}_i \geq \tau \end{cases} \quad \forall 1 \leq i \leq N \\ & \tilde{y}_i \in \{0, 1\} \quad \forall 1 \leq i \leq N \\ & 0 \leq \tau \leq 1 \end{aligned} \quad (3)$$

To make the term $|y_i - \tilde{y}_i|$ linear, we introduced new decision variable z_i along with 2 new constraints for each i shown in equation 4. More precisely, constraints (1) and (2) force the decision variable z_i to be greater than $|y_i - \tilde{y}_i|$, and hence, by maximizing $-z_i$, we are maximizing $-|y_i - \tilde{y}_i|$ too.

Linearizing the term $\frac{\sum_{i=1}^N y_i \tilde{y}_i}{\sum_{i=1}^N \tilde{y}_i} \dots$

Moreover, to linearize \tilde{y}_i constraints in equation 3, we combined the conditions and replace them with two new constraint sets (3) and (4) for each i in equation 4. These constraints are formulating the relation between \tilde{y}_i and τ . For example, if $\hat{y}_i = 0.7$ and $\tau = 0.5$ for an arbitrary i , then \tilde{y}_i is forced to be 1 according to the constraints $0.7(1 - \tilde{y}_i) < 0.6$ and $\tilde{y}_i(0.7 - 1) \geq 0.6 - 1$.

Figure 1: Comparison between ILP and exhaustive search methods in terms of average execution time. Each curve demonstrates the proportion of obtained average execution time by the ILP approach respect to the search method. All experiments were repeated 100 times and predictions were generated using the continuous uniform distribution $U[0, 1]$.

$$\begin{aligned}
& \text{maximize} \quad \sum_{i=1}^N \left(-\frac{\alpha}{N} z_i + \frac{\beta}{N^+} y_i \tilde{y}_i + \frac{\gamma}{N^-} (1 - y_i)(1 - \tilde{y}_i) \right) + \zeta \frac{\sum_{i=1}^N y_i \tilde{y}_i}{\sum_{i=1}^N \tilde{y}_i} \\
& \text{subject to:} \\
& \quad z_i \geq y_i - \tilde{y}_i \quad \forall 1 \leq i \leq N \quad (1) \\
& \quad z_i \geq \tilde{y}_i - y_i \quad \forall 1 \leq i \leq N \quad (2) \\
& \quad \hat{y}_i(1 - \tilde{y}_i) < \tau \quad \forall 1 \leq i \leq N \quad (3) \\
& \quad \tilde{y}_i(\hat{y}_i - 1) \geq \tau - 1 \quad \forall 1 \leq i \leq N \quad (4) \\
& \quad \tilde{y}_i \in \{0, 1\} \quad \forall 1 \leq i \leq N \\
& \quad z_i \in \mathbb{R} \quad \forall 1 \leq i \leq N \\
& \quad 0 \leq \tau \leq 1
\end{aligned} \tag{4}$$

Note that the proposed ILP formulation in equation 4 cannot output $t = 0$ as the optimal threshold and this case must be considered separately.

To further improve the ILP formulation, we considered a case where number of unique ordered pairs (y_i, \hat{y}_i) for $1 \leq i \leq N$ is significantly less than total number of instances, N . Suppose M is the total number of distinct ordered pairs, and m_i is the repetition number of a particular ordered pair in the original instance set. Hence, the following ILP formulation is equivalence to the previous one but empirically more efficient when $M \ll N$.

$$\begin{aligned}
& \text{maximize} \quad \sum_{i=1}^M m_i \left(-\frac{\alpha}{N} z_i + \frac{\beta}{N^+} y_i \tilde{y}_i + \frac{\gamma}{N^-} (1 - y_i)(1 - \tilde{y}_i) \right) + \zeta \frac{\sum_{i=1}^M m_i y_i \tilde{y}_i}{\sum_{i=1}^M m_i \tilde{y}_i} \\
& \text{subject to:} \\
& \quad z_i \geq y_i - \tilde{y}_i \quad \forall 1 \leq i \leq M \\
& \quad z_i \geq \tilde{y}_i - y_i \quad \forall 1 \leq i \leq M \\
& \quad \hat{y}_i(1 - \tilde{y}_i) < \tau \quad \forall 1 \leq i \leq M \\
& \quad \tilde{y}_i(\hat{y}_i - 1) \geq \tau - 1 \quad \forall 1 \leq i \leq M \\
& \quad \tilde{y}_i \in \{0, 1\} \quad \forall 1 \leq i \leq M \\
& \quad z_i \in \mathbb{R} \quad \forall 1 \leq i \leq M \\
& \quad 0 \leq \tau \leq 1
\end{aligned} \tag{5}$$

The proposed ILP-based approach is compared with a simple exhaustive search method with time complexity of $O(N^2)$ in terms of average execution time where $\alpha = 1$, $\beta = 0$, $\gamma = 0$, and $\zeta = 0$ (Fig. 1).

Results demonstrate that by increasing the number of instances (N), the ILP method eventually overcomes the exhaustive search in terms of execution time.

References

- [1] Alan G. Glaros and Rex B. Kline. “Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model”. In: *Journal of Clinical Psychology* 44.6 (1988), pp. 1013–1023. DOI: [https://doi.org/10.1002/1097-4679\(198811\)44:6<1013::AID-JCLP2270440627>3.0.CO;2-Z](https://doi.org/10.1002/1097-4679(198811)44:6<1013::AID-JCLP2270440627>3.0.CO;2-Z). eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/1097-4679%28198811%2944%3A6%3C1013%3A%3AAID-JCLP2270440627%3E3.0.CO%3B2-Z>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/1097-4679%28198811%2944%3A6%3C1013%3A%3AAID-JCLP2270440627%3E3.0.CO%3B2-Z>.
- [2] Jin Huang and Charles X Ling. “Using AUC and accuracy in evaluating learning algorithms”. In: *IEEE Transactions on knowledge and Data Engineering* 17.3 (2005), pp. 299–310.
- [3] Brendan Juba and Hai S. Le. “Precision-Recall versus Accuracy and the Role of Large Data Sets”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 4039–4048. DOI: [10.1609/aaai.v33i01.33014039](https://doi.org/10.1609/aaai.v33i01.33014039). URL: <https://ojs.aaai.org/index.php/AAAI/article/view/5193>.

$$\begin{aligned}
& \text{maximize} \quad \sum_{i=1}^N \left(-\frac{\alpha}{N} z_i + \frac{\beta}{N^+} y_i \tilde{y}_i + \frac{\gamma}{N^-} (1 - y_i)(1 - \tilde{y}_i) \right) + \zeta \frac{\sum_{i=1}^N y_i \tilde{y}_i}{\sum_{i=1}^N \tilde{y}_i} \\
& \text{subject to:} \\
& \quad z_i \geq y_i - \tilde{y}_i \quad \forall 1 \leq i \leq N \quad (1) \\
& \quad z_i \geq \tilde{y}_i - y_i \quad \forall 1 \leq i \leq N \quad (2) \\
& \quad \hat{y}_i(1 - \tilde{y}_i) < \tau \quad \forall 1 \leq i \leq N \quad (3) \\
& \quad \tilde{y}_i(\hat{y}_i - 1) \geq \tau - 1 \quad \forall 1 \leq i \leq N \quad (4) \\
& \quad \tilde{y}_i \in \{0, 1\} \quad \forall 1 \leq i \leq N \\
& \quad z_i \in \mathbb{R} \quad \forall 1 \leq i \leq N \\
& \quad 0 \leq \tau \leq 1
\end{aligned} \tag{6}$$