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

Pouria Ramazi[1] and Arash Mari Oriyad[2]

[1]Department of Mathematics and Statistics, Brock University, St. Catharines, L2S 3A1, ON, Canada
[2]Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan 84156-83111, Iran

Accuracy [2], recall [3], and specificity [1] are three of the most used 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 can be calculated as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
$$recall = \frac{TP}{TP + FN} \tag{1}$$
$$specificity = \frac{TN}{TN + FP}$$

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 and specificity) threshold for a given set of instances is a real value between 0 and 1, resulting in the highest accuracy (resp. recall and specificity).

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 ILP formulation for finding the optimal threshold for accuracy, recall, specificity, 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 predicted by a classifier for the ith instance. Then, $\tilde{y}_i$ would be a binary decision variable that represents the label of ith instance after applying the threshold t to $\hat{y}_i$ and can be calculated as below.

$$\tilde{y}_i = \begin{cases} 0, & \hat{y}_i < t \\ 1, & \hat{y}_i \geq t \end{cases} \tag{2}$$

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

$$accuracy = 1 - \frac{1}{N} \sum_{i=1}^{N} |y_i - \tilde{y}_i|$$
$$recall = \frac{1}{N^+} \sum_{i=1}^{N} y_i \tilde{y}_i \tag{3}$$
$$specificity = \frac{1}{N^-} \sum_{i=1}^{N} (1 - y_i)(1 - \tilde{y}_i)$$

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

$$\text{maximize} \quad \sum_{i=1}^{N} -\frac{\alpha}{N}|y_i - \tilde{y}_i| + \frac{\beta}{N^+}y_i\tilde{y}_i + \frac{\gamma}{N^-}(1-y_i)(1-$$

$$\text{subject to:} \quad \tilde{y}_i = \begin{cases} 0, & \hat{y}_i < t \\ 1, & \hat{y}_i \geq t \end{cases} \quad \forall 1 \leq i \leq N$$

$$\tilde{y}_i \in \{0,1\} \quad \forall 1 \leq i \leq N$$

$$0 \leq t \leq 1$$

$$(4)$$

To make the objective function linear, we introduced new decision variable z_i corresponding to the non-linear term $|y_i - \tilde{y}_i|$ along with 2 new constraints for each i shown in equation 5. Moreover, to linearize $\tilde{y}_i$ constraints in equation 4, we combined the conditions and replace them with two new constraints for each i in equation 5.

$$\text{maximize} \quad \sum_{i=1}^{N} -\frac{\alpha}{N}z_i + \frac{\beta}{N^+}y_i\tilde{y}_i + \frac{\gamma}{N^-}(1-y_i)(1-\tilde{y}_i)$$

$$\text{subject to:}$$

$$z_i \geq y_i - \tilde{y}_i \quad \forall 1 \leq i \leq N \quad (1)$$

$$z_i \geq \tilde{y}_i - y_i \quad \forall 1 \leq i \leq N \quad (2)$$

$$\hat{y}_i(1-\tilde{y}_i) < t \quad \forall 1 \leq i \leq N \quad (3)$$

$$\tilde{y}_i(\hat{y}_i - 1) \geq t - 1 \quad \forall 1 \leq i \leq N \quad (4)$$

$$\tilde{y}_i \in \{0,1\} \quad \forall 1 \leq i \leq N$$

$$z_i \in \mathbb{R} \quad \forall 1 \leq i \leq N$$

$$0 \leq t \leq 1$$

$$(5)$$

More precisely, constraints (1) and (2) force the decision variable z_i to be greater than $|y_i - \tilde{y}_i|$. Hence, by maximizing $-z_i$, we are maximizing $-|y_i - \tilde{y}_i|$ too. Constraints (3) and (4) are formulating the relation between $\tilde{y}_i$ and t represented in equation 2. For example, if $\hat{y}_i = 0.7$ and $t = 0.5$ 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 $\tilde{y}_i(0.7-1) \geq 0.6 - 1$.

Note that the proposed 0-1 ILP formulation in equation 5 can not output $t = 0$ as the optimal threshold and this case must be considered separately.

To further improve the 0-1 ILP formulation, we considered a case where number of unique ordered pairs $(y_i, \hat{y}_i)$ for $0 \leq i \leq N$ is significantly less than total number of instances, N. Suppose M is
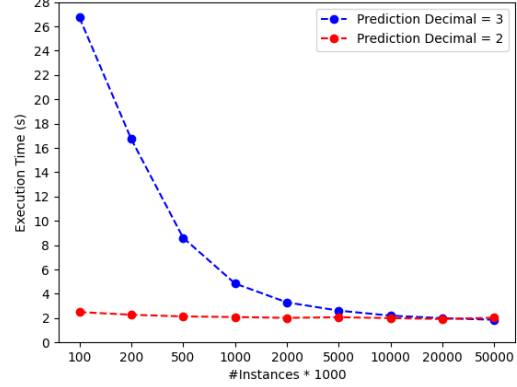

Figure 1: Comparison between 0-1 ILP and exhaustive search methods in terms of average execution time. Each curve demonstrates the proportion of average execution time using the ILP approach respect to the search method. All experiments are repeated 100 times.

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 0-1 ILP formulation is equivalence to the previous one but empirically more efficient.

$$\text{maximize} \quad \sum_{i=1}^{M} m_i\left(-\frac{\alpha}{N}z_i + \frac{\beta}{N^+}y_i\tilde{y}_i + \frac{\gamma}{N^-}(1-y_i)(1-\tilde{y}_i)\right)$$

$$\text{subject to:}$$

$$z_i \geq y_i - \tilde{y}_i \quad \forall 1 \leq i \leq M$$

$$z_i \geq \tilde{y}_i - y_i \quad \forall 1 \leq i \leq M$$

$$\hat{y}_i(1-\tilde{y}_i) < t \quad \forall 1 \leq i \leq M$$

$$\tilde{y}_i(\hat{y}_i - 1) \geq t - 1 \quad \forall 1 \leq i \leq M$$

$$\tilde{y}_i \in \{0,1\} \quad \forall 1 \leq i \leq M$$

$$z_i \in \mathbb{R} \quad \forall 1 \leq i \leq M$$

$$0 \leq t \leq 1$$

$$(6)$$

The proposed 0-1 ILP formulation is compared with a search based method (Appendix 1) in terms of execution time (Fig. 1).

Conclusion/Discussion ...

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. 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. URL: https://ojs.aaai. org/index.php/AAAI/article/view/5193.