Performance measures - Confidence Intervals For Multiclass F1 Scores

Students Adam Vinestock, Idan Cohen

Introduction

This report presents a comprehensive overview of the paper titled 'Confidence Interval for Micro-Averaged F1 and Macro-Averaged F1 Scores' in the field of performance measures. The F1 score, which is defined as the harmonic mean of precision and recall, plays a crucial role in evaluating models due to its favorable characteristics, particularly when dealing with imbalanced datasets. Despite its notable properties, the F1 score is rarely used in diagnostic studies within the medical field. As a single performance measure, the F1 score may be preferred over specificity and accuracy, as the latter measures may be artificially high even for a poor classifier with a high false negative rate. This is particularly evident in low disease prevalence scenarios. While statistical methods for inference have been developed for the F1 score in binary classification problems, extending these methods to the multi-class classification problem remains an open challenge. To address this knowledge gap, the paper proposes novel methods based on the large sample multivariate central limit theorem. These methods enable the estimation of F1 scores with confidence intervals, providing a framework to quantify the uncertainty associated with the scores. This approach facilitates a more comprehensive evaluation of model performance in the multi-class classification setting.

Relevant background

To evaluate a multi-class classifier, a single summary measure is often sought. As extensions of the F1 score for the binary classification, there exist two types of such measures: a micro-averaged F1 score and a macro-averaged F1 score. Although a single summary measure is highly desired, comparing point estimates without taking into account their uncertainty can be problematic. They neglect sampling variability, overlook the uncertainty of model performance and can lead to unreliable assessments. For example, consider the results of an analysis reported by Dong et al. [2], in this analysis the authors concluded a classifier outperformed others by comparing point estimates without taking into account their uncertainty.
In this section we introduce notations and definitions as well as the key components necessary for comprehending the methods employed in the paper. Given a multi-class classification task, with $r > 2$ classes We can compute the confusion matrix using the observed data. We denote cell probabilities p_{ij} and marginal probabilities $p_{i\cdot}, p_{\cdot i}$ for rows and columns respectively.

The micro-averaged precision (miP) and micro-averaged recall (miR) are defined as:

$$miP = \frac{\sum_{i=1}^{r} TP_i}{\sum_{i=1}^{r}(TP_i + FP_i)} = \frac{\sum_{i=1}^{r} p_{ii}}{\sum_{i=1}^{r} p_{i\cdot}} = \sum_{i=1}^{r} p_{ii}$$

$$miR = \frac{\sum_{i=1}^{r} TP_i}{\sum_{i=1}^{r}(TP_i + FP_i)} = \frac{\sum_{i=1}^{r} p_{ii}}{\sum_{i=1}^{r} p_{\cdot i}} = \sum_{i=1}^{r} p_{ii}$$

Micro-averaged F1 - is defined as the harmonic mean of miP, miR:

$$miF_1 = 2 \times \frac{miP \times miR}{miP + miR} = \sum_{i=1}^{r} p_{ii}$$

Macro-averaged F1 - is defined as the simple arithmetic mean of per class F1:

$$F_{1i} = 2 \times \frac{P_i \times R_i}{P_i + R_i} = 2\frac{p_{ii}}{p_{i\cdot} + p_{\cdot i}}$$

$$maF_1 = \frac{1}{r}\sum_{i=1}^{r} F_{1i} = \frac{2}{r}\sum_{i=1}^{r} \frac{p_{ii}}{p_{i\cdot} + p_{\cdot i}}$$

For Variance estimation and computing confidence intervals the paper incorporates several essential components:
Maximum Likelihood Estimation (MLE) - Finds the distribution parameter values that maximize the likelihood given

observed data. In this study, it is utilized on the observed confusion matrix to calculate the estimated parameters of the F1 distribution.

Central Limit Theorem (CLT) - The average of many independent observations, regardless of the underlying distribution approximates to a normal distribution. This theorem is particularly relevant as it enables the computation of the variance and confidence interval of the F1 score. The paper uses an extension of this theorem called the Delta-Method which is essential for our specific use case. The Delta-Method is a technique used to approximate the distribution of a function of random variables using a first-order Taylor expansion. It is commonly applied when dealing with transformations or functions of random variables. The Delta Method allows for approximating the distribution of the transformed random variable using the mean and variance of the original random variable. This approach precisely meets our requirement to approximate the miF1 and maF1 scores, which are functions of the observed confusion matrix.

Confidence Intervals (CI) - Defines a plausible range of values for the true parameter by considering the uncertainty resulting from sampling, based on the observed data. This ultimately allows for a comprehensive comparison of models performance while taking into account their uncertainty.

The Paper's Key Points

In order to derive the confidence interval for miF_1, maF_1, and maF_1^*[1], the authors consider the event frequencies as a multinomial distribution with a sample size of n and probabilities $p = (p_{11}, \ldots, p_{1r}, p_{21}, \ldots, p_{2r}, \ldots, p_{r1}, \ldots, p_{rr})^T$, where $(n_{11}, n_{12}, \ldots, n_{rr}) \sim Multinomial(n; \mathbf{p})$. Now, using MLE, we can calculate $\hat{p}_{ij} = \frac{n_{ij}}{n}$.

Applying the multivariate central limit theorem, we have $\sqrt{n}(\hat{p} - p) \dot{\sim} Normal\left(0_{r^2}, diag(\mathbf{p}) - \mathbf{p}\mathbf{p}^T\right)$. Thus, as n increases, this expression converges to a normal distribution centered around 0 (0_{r^2} is a r^2X1 vector whose elements are all 0) with the given covariance matrix. Using the multivariate delta-method, we will get that $\widehat{maF_1}$ and $\widehat{miF_1}$ are approximately normally distributed as $\widehat{maF_1} \dot{\sim} \text{Normal}(maF_1, Var(\widehat{maF1}))$ and $\widehat{miF_1} \dot{\sim} \text{Normal}(miF_1, Var(\widehat{miF1}))$, while $Var\left(\widehat{miF_1}\right) = \left(\sum_{i=1}^{r} p_{ii}\right)\left(1 - \sum_{i=1}^{r} p_{ii}\right)/n$ and
$Var\left(\widehat{maF_1}\right) = \frac{2}{r^2}\left\{\sum_{i=1}^{r} \frac{F_{1i}(p_{i\cdot}+p_{\cdot i}-2p_{ii})}{(p_{i\cdot}+p_{\cdot i})^2}\left(\frac{p_{i\cdot}+p_{\cdot i}-2p_{ii}}{p_{i\cdot}+p_{\cdot i}} + \frac{F_{1i}}{2}\right) + \sum_{i=1}^{r}\sum_{j\neq i} \frac{p_{ij}F_{1i}F_{1j}}{(p_{i\cdot}+p_{\cdot i})(p_{j\cdot}+p_{\cdot j})}\right\}/n$. Finally, after completing the calculation of variances, we can proceed to substitute the computed values into the corresponding formulas:

$$\text{miF1 Confidence Interval: } \widehat{miF_1} \pm Z_{1-\frac{\alpha}{2}} \times \sqrt{\widehat{\text{Var}}(\widehat{miF_1})}$$

$$\text{maF1 Confidence Interval: } \widehat{maF_1} \pm Z_{1-\frac{\alpha}{2}} \times \sqrt{\widehat{\text{Var}}(\widehat{maF_1})}$$

where $\widehat{Var}\left(\widehat{miF_1}\right)$ is $Var\left(\widehat{miF_1}\right)$ with p_{ii} replaced by $\hat{p}_{ii}$, and Z_p denotes the 100 p-th percentile of the standard normal distribution.

Implications for the field of machine learning

This paper fills the gap in computing confidence intervals for F1 scores in multi-class classification tasks, providing a valuable tool to quantify the uncertainty associated with them. Upon examining subsequent research papers published since its inception in 2021, it has garnered a substantial number of 50 citations. However, we find it important to highlight that while certain subsequent studies have incorporated and employed its fundamental elements [3], a significant portion of other studies did not fully exploit the method's inherent strength, namely the computation of a confidence interval. Instead, their primary focus has been on computing the miF1 and maF1 scores [4], [5]. While we would expect researchers measuring model performance to adopt this method and report on confidence intervals, we find that the current implications remain somewhat limited and haven't fully exploited the potential so far.

References

[1] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management*, vol. 45, no. 4, pp. 427–437, 2009.

[2] H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews, and Y. Guo, "Mixed neural network approach for temporal sleep stage classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 2, pp. 324–333, 2017.

[1] maF_1^* is an alternative definition for Macro-averaged F_1 score, defined by Sokolova et al.[1]. In our presentation, we've deliberately chose to omit it from our review.

[3] W. Ju and W. Jiang, "A note on comparison of f-measures," *arXiv preprint arXiv:2112.04677*, 2021.

[4] B. R. Bergo and E. Bjørne-Larsen, "Deparadoxifying strategic decisions: An integrated approach utilizing machine learning and natural language processing,"

[5] A. Turchin, S. Masharsky, and M. Zitnik, "Informatics in medicine unlocked,"