A Literature Review on Rater Agreement Metrics

A review for finding a standard approach for comparing the performance of ML algorithms

Alireza Zolanvari*, Sepand Haghighi*, Sadra Sabouri *

September 2022

1 Introduction

In the machine learning ecosystem, especially in the field of supervised learning algorithms, there is a lack of a standard method to compare the performance of these algorithms. This issue is more tangible in problems that deal with imbalanced data or ordinal labels because common metrics such as accuracy, precision, and recall lose their effectiveness and even cause confusion. For example, in the research on breast cancer, more than 99% of subjects are from the negative class and less than 1% of them are in the positive class. For this problem, if on the one hand, a biased algorithm places the entire population in the negative class and on the other hand another algorithm places 97% of the population in the negative class and 3% in the positive class, the accuracy of algorithms will be 99% and 98% (the best case) respectively. Based on the accuracy value, it seems that the first algorithm has worked more successfully, but in reality, we know that this algorithm's performance has been poor and the accuracy metric is not a suitable tool to compare these two algorithms in this particular problem. In addition to the (im)balanced data that was shown in this example, various factors affect the way of evaluating and correctly comparing the performance of different algorithms. For this reason, it seems necessary to have a standard solution for the comprehensive comparison of the performance of machine learning algorithms.

This challenge can be modeled as an observer agreement problem. In statistics, to check the reliability of observation, a standard procedure is to ask two or more observers to independently examine the same group of units. Then, to measure the level of agreement between observers, metrics are defined that quantify this level of agreement from different aspects.[1] Using this definition, we can consider the output labels of an algorithm as the decision of an observer and measure its agreement with the actual labels which actually play the role of the second observer. In this way, the algorithm that has a higher agreement

*The authors are with the PyCM Development Team. This work is partly supported with a grant from the NLnet.

with the reference observer has shown a better performance.
The measurement of observer agreement has been comprehensively studied in both statistics and medical diagnosis, and different metrics have been introduced and investigated for this purpose. However, there is no consensus on which metric is more appropriate, what can be interpreted from the same value of them, and what statistical inference can be made according to them. In fact, the focus of each of these metrics is on a specific issue and they have usually followed a unique path without considering the generalization for other issues and establishing a meaningful relationship with other introduced metrics.
In the field of medical diagnosis, comprehensive studies have been conducted on the interpretation of each of these metrics, as well as their application and ability to be generalized. Some of these metrics can be used in multi-class problems, while some have been developed only for binary mode. Despite the fact that the metrics developed for multi-class problems have the ability to summarize the entire performance of the algorithm and its agreement with the reference observer in one number, the binary metrics also contain important information that makes us not ignore them. Therefore, In addition to the multi-class metrics, binary metrics are used for our purpose in such a way that we quantify and evaluate the degree of agreement of the algorithm with the vector of actual labels class-wise (One vs Rest).
In the following, summaries of the articles that present the main metrics introduced in both multi-class (overall) and binary (class-based) are presented, and then the studies that have studied the applied aspects of a number of these metrics are briefly discussed. We have tried to bring the articles that show both advantages and disadvantages of the introduced metrics to provide the readers with a comprehensive insight. It should be noted that this report is used to develop the "*Confusion matrices compare*" tool provided by PyCM [2] (a confusion matrix Python library) and will be more complete over time. The conclusion is left to the readers.

2 Overall statistics

Measuring Nominal Scale Agreement

The agreement measure introduced by Fleiss in [3] is unweighted and for each subject, a pair of observers rate the agreement. The point of this measure in comparison with kappa is that the observers for each subject can be different from the observers of other subjects. In other words, this paper generalizes kappa to the case where the same number of observers rate samples of each subject on a nominal scale, but the observers are not necessarily the same for each subject. Notably, the studied data type in this paper is categorical.

The Measurement of Observer Agreement

In [1], observers are considered as the source of error. Thus, they presented a statistical measure to quantify the reliability of the observers based on the kappa

value. This measure shows the extent to which the observers agree among themselves. For describing the level of strength of agreement between the observers, consistent labels are assigned to corresponding ranges of kappa. The focused data source in this study is multivariate categorical data. Although the authors introduced this measure as a general measure in terms of application, the case study in this paper is a clinical diagnosis.

Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology

Focusing on the interpretation of the results derived from psychological assessments and with the aim of providing comparable information from them, he proposed a guideline for determining levels of practical, fundamental, or clinical significance in [4]. Although the proposed guideline is compatible with different reliability measures such as Pearson and intraclass correlation, the author found the intraclass correlation more desirable for the application he focused on. His proposed guideline is very close to those developed by Fleiss in [3] and represented a version of those introduced by Landis& Koch in [1].

Alternatives to P Value: Confidence Interval and Effect Size

In [5], Lee introduced Cramer's benchmark to quantify the relationship between two variables (observations) based on "Effect Size" and "Confidence Interval." He used Cohen's d as the effect size to allow the comparison of statistical results resulting from different methods. In addition, he quantifies the error imposed on the effect size using a confidence interval. Although the Cramer's benchmark in this study is introduced to provide more comprehensive information about the magnitude of treatment effect in comparison with the P value, it can also provide useful information about the strength of agreement between the raters of a subject. the variables in this study are quantitative but they can easily be generalized to categorical variables.

3 Class-based statistics

Evaluation Measures Over Imbalanced Data Sets

Dealing with imbalanced data sets is one of the top 10 challenges of data mining. Here, in [6], a set of model assessment measures are provided for dealing with imbalanced data sets because the normal evaluation metrics such as normal accuracy are not effective anymore for such problems. The provided measures are categorized into two groups combined measures and graphical measures which are all suitable for binary classification problems. The mentioned combined measures are *G-means*, *positive* and *negative likelihood ratio*, *Discriminant Power*, *F-Measure*, *Balanced Accuracy*, *Youden index*, and *Matthews correlation coefficient*. *G-means* is the product of the prediction accuracies for both classes. In *positive* and *negative likelihood ratio*, a higher positive likelihood ratio and

a lower negative likelihood mean better performance in positive and negative classes respectively. *Discriminant Power* summarizes sensitivity and specificity in one measure. *F-Measure* is a harmonic mean of Precision and Recall and a high value of F-Measure indicates that the model performs better on the positive class. In addition, a modified version of this metric is also introduced in this paper in which the importance of precision versus recall in adjustable. The average of Sensitivity and specificity is termed *Balanced Accuracy*; *Youden index* evaluates the algorithm's ability to avoid failure, and *Matthews correlation coefficient* summarizes accuracies and error rates on both classes in a single measure.
On the hand, the introduced graphical measures are *ROC curve*, *Area Under Curve*, *Cumulative Gains Curve* and *Lift Chart*, and *Area Under Lift*. It is notable that, *Area Under Curve* and *Area Under Lift* are summary indicators of the ROC curve and Lift chart performance respectively. Although all these measures are helpful in evaluating the classification performance in an imbalanced problem, the interpretation table is only provided for *Discriminant Power*, *likelihood ratios*, and *Area Under Curve*.

Selecting and Interpreting Diagnostic Tests

Raslich et. al. in [7] investigate the components in selecting and interpreting clinical diagnostic test results. Throughout this article, it is shown that disease prevalence has a strong impact on the value of traditional measures such as accuracy, precision, and recall. To solve this problem, more reliable metrics which have been introduced in this article are less sensitive to factors such as disease prevalence. Likelihood ratios that express the magnitude by which the probability of disease in a specific patient is modified by the result of a test, are the first class of reliable metrics. An informative interpretation table is also provided for both positive and negative likelihood ratios. The other reliable introduced metric in this paper is the diagnostic odds ratio (DOR) which combines the strengths of sensitivity and specificity, as prevalence in dependent indicators, with the advantage of accuracy as a single indicator. However, an interpretation table is not suggested for this metric.

Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation

According to [8], most of the popular evaluation measures including Recall, Precision, and F-Measure are biased, they propagate the underlying marginal prevalence and biases, and, fail to take into account the chance level performance. Furthermore, more advanced metrics, such as Rand Accuracy and Cohen's Kappa, have some advantages but are nonetheless still biased measures. To tackle this problem, several concepts and measures that reflect the probability that prediction is informed versus chance are discussed in this article. In fact, the goal of this article is to measure the effectiveness of an empirical decision system or a scientific experiment, analyze and introduce new probabilistic and

information theoretic measures that overcome the problems with Recall, Precision and their derivatives like G-means and F-measure. to this end, first various forms of bias such as prevalence, bias, cost, and skew are presented to ensure that the readers are well-informed about all types of bias and their effects on performance measures. After that, the concepts of *informedness* and *markedness* are introduced. *Informedness/markedness* quantifies how informed/marked a predictor is for the specified condition, and specifies the probability that a prediction is informed/marked in relation to the condition (versus chance). It is also noted that the analog of markedness to regression coefficient, and that the G-Mean of the two measures is a dichotomous form of the Pearson correlation coefficient, termed Matthews' Correlation Coefficient (MCC), which is appropriate unless a continuous scale is being measured dichotomously in which case a Tetrachoric Correlation estimate would be appropriate. The interpretation of MCC is also provided in this article.

4 Reliability coefficient

Measurement of Reliability for Categorical Data in Medical Research

Article [9] is focused on the interpretability of the reliability measures as a special class of agreement level for categorical data. Categorical data can be "binary", "ordinal", or non of them, which here is termed just "categorical". Also, the term "consensus rating", by means of the "true/actual score" is proposed in this paper. Just as an example, in terms of *reproducibility* and *association of a single rating with the consensus score*, the interpretability of the *"kappa"* coefficient is investigated for binary data. Then, the analysis is generalized for categorical data. The main concern of the paper is to generalize the previous *ad hoc* approaches in medical research and clarify the assumptions for the ease of evaluating if a measure is appropriate to be adopted for real-world problems or not.

Reliability Procedures For Categorical Data in Performance Analysis

The focus of [10] is on the reliability assessment that is appropriate for the categorical data on the nominal scale. Three main sources of error are presented in this article that is Operational error: where the observer presses the wrong button to label an event, Observational errors: the observer fails to code an event and Definitional errors: the observer labels an event inappropriately. The first ad the third sources are not applicable to the case of comparing two confusion matrices. Two methods of comparison and agreement analysis are also described in this paper; intra-analyst test and inter-analyst test. In the first test, one person can perform the same task multiple times and in the second test, multiple people perform the task once. In the first test, the test result can only show if the rater is reliable or not. On the other hand, the result of the second test can show which rater is more reliable. According to this paper, any reliability study should report reliability statistics. To quantify the agreement level of two independent

tasks (regardless of the uniqueness of the rater), two measures are investigated in this article; Cohen's Kappa and Yule's Q. According to this paper, same as many other manuscripts, Kappa is a chance-corrected measure of agreement. However, it is not clear if taking chance into consideration is reasonable in performance analysis in all applications. Besides, as in this measure something that may or may not is presented, is removed, and it is difficult to interpret its value. On the other hand, Yule's Q is recommended as a more reliable measure of agreement level. The Yule's Q test is the odds ratio (OR) i.e. the odds of agreeing compared to not agreeing. Yule's Q is thus a test specifically for assessing the difference between concordant and discordant responses between two raters making dichotomous ratings. Since Yule's Q statistic produces a lower value when agreement levels fall below reasonable levels this may act as a better alert in comparison with Kappa.

A Comparison of Reliability Coefficients

There are some reliability coefficients developed specifically on a categorical scale as well as an interval scale. However, there is no specific measure focused on the ordinal scale. The article [11] compares seven different coefficients both analytically and by simulated and empirical data to show which coefficient is more reliable for the ordinal scale, how these coefficients are related, and whether the choice of coefficient matters. The analyzed set of coefficients consists of three kappa coefficients (Cohen's kappa, linearly weighted kappa, and quadratically weighted kappa) and four correlation coefficients (intraclass correlation, Pearson's correlation, Spearman's rho, and Kendall's tau-b). These groups are more common in quantifying reliability on a categorical scale and interval scale respectively. The analytical methods reveal the fact that differences between quadratic kappa and the Pearson and intraclass correlations are generally highly correlated and their differences increase if agreement becomes larger. Moreover, based on the simulated and empirical data, the increase of agreement between the raters results in more difference between all reliability coefficients. This paper concluded that the four correlation coefficients and quadratically weighted kappa have been highly correlated for the data in this study and finally, for this data, it does not really matter which of these five coefficients is used.

Reliability of Multi-category Rating Scales

[12] compares the linear and quadratic kappa coefficients for ordinal rating scales, as well as Pearson and Kendall correlation coefficients, using simulated data. They investigated whether a fixed value has the same meaning across reliability coefficients, and across rating scales with different numbers of categories. Results of this study include the following. There were usually less than 0.15 differences between quadratic kappa and Pearson and Kendall correlations. On the other hand, the value of linear kappa typically differed from quadratic kappa, Pearson, and Kendall correlations. The number of considered

categories also affects the differences between the coefficients. Differences tend to be smaller with two and three categories than with five or more categories. With two categories, the three kappa coefficients are identical.

References

[1] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[2] S. Haghighi, M. Jasemi, S. Hessabi, and A. Zolanvari, "Pycm: Multiclass confusion matrix library in python," *Journal of Open Source Software*, vol. 3, no. 25, p. 729, 2018.

[3] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[4] D. V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology." *Psychological assessment*, vol. 6, no. 4, p. 284, 1994.

[5] D. K. Lee, "Alternatives to p value: confidence interval and effect size," *Korean journal of anesthesiology*, vol. 69, no. 6, pp. 555–562, 2016.

[6] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced data sets," *J Inf Eng Appl*, vol. 3, no. 10, 2013.

[7] M. A. Raslich, R. J. Markert, S. A. Stutes *et al.*, "Selecting and interpreting diagnostic tests," *Biochemia Medica*, vol. 17, no. 2, pp. 151–161, 2007.

[8] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[9] H. C. Kraemer, "Measurement of reliability for categorical data in medical research," *Statistical methods in medical research*, vol. 1, no. 2, pp. 183–199, 1992.

[10] N. James, J. Taylor, and S. Stanley, "Reliability procedures for categorical data in performance analysis," *International Journal of Performance Analysis in Sport*, vol. 7, no. 1, pp. 1–11, 2007.

[11] A. de Raadt, M. J. Warrens, R. J. Bosker, and H. A. Kiers, "A comparison of reliability coefficients for ordinal rating scales," *Journal of Classification*, vol. 38, no. 3, pp. 519–543, 2021.

[12] R. I. Parker, K. J. Vannest, and J. L. Davis, "Reliability of multi-category rating scales," *Journal of School Psychology*, vol. 51, no. 2, pp. 217–229, 2013.