Consistency and classification of metrics for binary classifiers

K. Dirland
<***@***>

A. S. Lundervold
<***@***>

P.G.L. Porta Mana
<pgl@portamana.org>

(or any permutation thereof)

Draft. 4 March 2022; updated 29 March 2022

abstract

[Luca] I find it very difficult to structure the paper: there seems to be issues at several levels in the development and use of binary classifiers (and classifiers in general) within machine-learning.
Here are some relevant points:

- There should be a distinction between "inference" (or forecast, prediction, guess) and "decision" (or action, choice). In particular, the possible situations we may be uncertain about and the possible decisions available may be completely different things. A clinician, for example, may be uncertain about "cancer" vs "non-cancer", while the choices are about "drug treatment 1" vs "drug treatment 2" vs "surgery".
- Probability theory & decision theory say that in order to make self-consistent decision we need two things: (a) the probabilities for the possible situations, (b) the utilities of the decisions given each possible situation.
- A useful machine-learning algorithm should therefore give us one of two things:
  - either the *probabilities* of the uncertain situations ("cancer" vs "non-cancer" in the example above),
  - or the final decision ("drug treatment 1" vs "drug treatment 2" vs "surgery" in the example above).

  Current machine-learning classifiers do not give us either: the output in the example above would be "cancer" vs "non-cancer", often without probabilities.
- So there are two possible solutions to the problem above:
  - We must build a classifier that outputs probabilities. The 0–1 outputs of current classifiers cannot properly interpreted as probabilities, for various reasons.
  - We must build a classifier that output *decisions*: so not "cancer" vs "non-cancer", but "drug treatment 1" vs etc..

1 Valuation metrics, amounts of data, inferences, and decisions

Let's consider the simple example of a binary classifier and several dilemmas that appear in its development, choice, and use.

At the moment of evaluating different classifier algorithms, or different hyperparameter settings for one algorithm, we are avalanched by a choice of possible evaluation scales: accuracy, area under curve,

F_1-measure, mean square contingency[1] also known as Matthews correlation coefficient[2], precision, recall, sensitivity, specificity, and many others[3]. Only vague guidelines are usually given to face this choice. A thorough analysis and discussion of several such scales was given by Goodman & Kruskal (1954; 1959; 1963; 1972).

We can also ask: are all these scales well-founded and self-consistent? is it possible that the use of any of them leads to contradictions? The literature abounds with studies showing that some scale X may imply hidden contradictions with the data or the assumptions used for our inference, and is therefore worse than some other scale Y. See for example Baker & Pinsky (2001), Lobo et al. (2008), Hand & Christen (2018), Zhu (2020) and Goodman & Kruskal's papers cited above for instances of criticisms of area under the curve, F_1-measure, Matthews correlation coefficient, and other scales.

If we have many more data for one class than for the other – a common predicament in medical applications – we must face the "class-imbalance problem": the classifier ends up classifying all data as belonging to the more numerous class[4], which may be an undesirable action if the misclassification of cases belonging to the less numerous class entails high costs.

The three points above turn out to be tightly related and to have a common solution. We show that

1. the admissible valuation scales for a binary classifier form a two-dimensional family; that is, the choice of a specific scale corresponds to the choice of two numbers. Such choice is problem-dependent and cannot be given a priori.
2. scales that are not

2 Overview of decision theory

Decision theory makes a distinction between

a the possible situations we are uncertain about

b the possible choices we can make.

This distinction is important, in fact in some cases the numbers of possible uncertain situations

[1] Yule 1912 denoted "r" there. [2] Matthews 1975. [3] Sammut & Webb 2017. [4] Sammut & Webb 2017; Provost 2000.

Bibliography

("de X" is listed under D, "van X" under V, and so on, regardless of national conventions.)

Baker, S. G., Pinsky, P. F. (2001): *A proposed design and analysis for comparing digital and analog mammography special receiver operating characteristic methods for cancer screening*. J. Am. Stat. Assoc. **96**[454], 421–428. doi:10.1198/016214501753168136.

Goodman, L. A., Kruskal, W. H. (1954): *Measures of association for cross classifications*. J. Am. Stat. Assoc. **49**[268], 732–764. doi:10.1080/01621459.1954.10501231. See corrections Goodman, Kruskal (1957; 1958) and also Goodman, Kruskal (1959; 1963; 1972).

— (1957): *Corrigenda: Measures of association for cross classifications*. J. Am. Stat. Assoc. **52**[280], 578. doi:10.1080/01621459.1957.10501415. See Goodman, Kruskal (1954).

— (1958): *Corrigenda: Measures of association for cross classifications*. J. Am. Stat. Assoc. **53**[284], 1031. doi:10.1080/01621459.1958.10501492. See Goodman, Kruskal (1954).

— (1959): *Measures of association for cross classifications. II: Further discussion and references*. J. Am. Stat. Assoc. **54**[285], 123–163. doi:10.1080/01621459.1959.10501503. See also Goodman, Kruskal (1954; 1963; 1972).

— (1963): *Measures of association for cross classifications. III: Approximate sampling theory*. J. Am. Stat. Assoc. **58**[302], 310–364. doi:10.1080/01621459.1963.10500850. See correction Goodman, Kruskal (1970) and also Goodman, Kruskal (1954; 1959; 1972).

— (1970): *Corrigenda: Measures of association for cross classifications. III: Approximate sampling theory*. J. Am. Stat. Assoc. **65**[330], 1011. doi:10.1080/01621459.1970.10481142. See Goodman, Kruskal (1963).

— (1972): *Measures of association for cross classifications, IV: Simplification of asymptotic variances*. J. Am. Stat. Assoc. **67**[338], 415–421. doi:10.1080/01621459.1972.10482401. See also Goodman, Kruskal (1954; 1959; 1963).

Hand, D., Christen, P. (2018): *A note on using the F-measure for evaluating record linkage algorithms*. Stat. Comput. **28**[3], 539–547. doi:10.1007/s11222-017-9746-6.

Lobo, J. M., Jiménez-Valverde, A., Real, R. (2008): *AUC: a misleading measure of the performance of predictive distribution models*. Glob. Ecol. Biogeogr. **17**[2], 145–151. doi:10.1111/j.1466-8238.2007.00358.x, https://www2.unil.ch/biomapper/Download/Lobo-GloEcoBioGeo-2007.pdf.

Matthews, B. W. (1975): *Comparison of the predicted and observed secondary structure of T4 phage lysozyme*. Biochim. Biophys. Acta **405**[2], 442–451. doi:10.1016/0005-2795(75)90109-9.

Provost, F. (2000): *Machine learning from imbalanced data sets 101*. Tech. rep. WS-00-05-001. (AAAI, Menlo Park, USA). https://aaai.org/Library/Workshops/2000/ws00-05-001.php.

Sammut, C., Webb, G. I., eds. (2017): *Encyclopedia of Machine Learning and Data Mining*, 2nd ed. (Springer, Boston). doi:10.1007/978-1-4899-7687-1. First publ. 2011.

Yule, G. U. (1912): *On the methods of measuring association between two attributes*. J. R. Stat. Soc. **75**[6], 579–652. doi:10.1111/j.2397-2335.1912.tb00463.x.

Zhu, Q. (2020): *On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset*. Pattern Recognit. Lett. **136**, 71–80. doi:10.1016/j.patrec.2020.03.030.