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 22 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 metrics: accuracy, area under curve,

F_1-measure, precision, recall, sensitivity, specificity, and many others[1]. Only vague guidelines are usually given to face this choice.

Regarding such metrics, we can also ask: are they all well-founded and self-consistent? is it possible that the use of any of them lead to contradictions? Several studies show that using area under curve, for example, may imply hidden contradictions with the data or the assumptions used for our inference[2].

If we have many more data for one of the two classes – a common predicament in medical applications – we must face the "class-imbalance problem"

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.

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.

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.

Swets, J. A. (1988): *Measuring the accuracy of diagnostic systems*. Science **240**[4857], 1285–1293. doi:10.1126/science.3287615.

[1] Sammut & Webb 2017. [2] Swets 1988; Baker & Pinsky 2001; Lobo et al. 2008.