Évaluation de la calibration d'un classifieur

Projet de thèse proposé par G. Russolillo et N. Niang

CNAM, Laboratoire CEDRIC, équipe MSDMA

April 15, 2022

Keywords : statistique, machine learning, classifieur probabiliste, calibration, recalibration

1 Introduction

On se place dans un cadre du classement de n observations, décrites par un ensemble de variables X, dans l'un des groupes définis par les Q modalités d'une variables qualitative Y. Les méthodes de classement (classifieurs) probabiliste fournissent une règle de prédiction qui permet d'affecter à chaque observation i telle que $X = x_i$ la probabilité $\hat{\pi}_{iq}$ d'observer l'évènement $Y = q$, c'est-à-dire que cette observation prenne la q-ème modalité de Y. C'est à partir de ces probabilités qu'une décision peut être prise pour chaque observation, par exemple en retenant la modalité la plus probable. Ce cadre correspond à de nombreuses situations très concrètes en médecine (choix de la thérapie sur un patient), marketing (choix de la publicité à afficher en ligne), banque (accord d'un crédit ou non), industrie (arrêt ou non d'une machine), fiscalité (détection de fraudes) par exemple.

La validation d'une méthode de classement probabiliste passe par l'évaluation de sa capacité discriminante et sa calibration. La discrimination fait référence à la capacité du modèle de bien classer les observations, autrement dit à séparer les observations prenant la modalité q de celles ne la prenant pas. La calibration quant à elle vise à vérifier si les probabilités des événements issues de la règle de prédiction coïncident avec les "vraies" probabilités, c'est-à-dire les taux des événements dans la population. Une méthode qui discrimine bien les observations n'est cependant pas nécessairement bien calibrée.

2 Contexte scientifique

La validation des classifieurs probabilistes passe parfois par la seule évaluation de leur capacité discriminante, alors que l'objectif peut être plutôt d'évaluer la calibration. C'est typiquement le cas lorsque on utilise de modèles de prédiction des risques pour appuyer la prise de décision médicale : une estimation fiable de la probabilité d'un

diagnostic (ou bien d'un pronostic) est essentielle pour le bon choix du traitement d'un patient.

Le problème de l'évaluation de la calibration des classifieurs suscite de plus en plus d'intérêt au sein de la communauté du machine learning [1]. Cela est d'abord dû à la nécessité d'estimer des probabilités fiables dans certaines applications. Par exemple, dans les soins de santé automatisés, le contrôle devrait être confié à des médecins humains lorsque la confiance d'une méthode de diagnostic de maladies est faible ; dans le contexte du marketing, une société de publicité en ligne doit prédire de manière exacte la probabilité qu'un utilisateur clique sur une annonce pour prendre une décision optimale quant à l'annonce à afficher. Au-delà de l'intérêt pour l'évaluation de la calibration suscité par ces contextes applicatifs, il existe également vif intérêt pour la recalibration a posteriori [2, 3] d'une règle de classement établie. C'est notamment le cas pour des classifieurs au pouvoir discriminant reconnu (réseaux de neurones, SVC, etc), pour lesquels la calibration peut être insuffisante. En particulier, les réseaux de neurones développés pendant les dernières années peuvent ne pas être bien calibrés [4].

Si l'évaluation de la capacité de discrimination d'un classifieur probabiliste est un problème bien résolu (via l'utilisation de l'indice de concordance ou autre [5]), celle de la calibration n'est pas évidente. En théorie, on devrait comparer la probabilité estimée d'un événement pour une observation avec la "vraie" probabilité de l'événement pour cette même observation. Or, on ne dispose que de la modalité observée issue de cette probabilité. Pour cette raison, on définit traditionnellement comme calibré au sens *fort* [6], un classifieur tel que $\forall i, q\ \hat{\pi}_{iq} = Pr(Y = 1|X = x_i)$; autrement dit, un classifieur qui fournit pour chaque profil x_i, une probabilité $\hat{\pi}_{iq}$ égale à la proportion des observations prenant la modalité q, parmi celles ayant le profil x_i.

En pratique, cette vérification est rendue impossible quand il existe un grand nombre de profils distincts, ce qui arrive notamment en présence de variables explicatives continues. Une définition dite *faible* de calibration est alors généralement utilisée. Celle-ci est basée sur l'idée que deux observations sont similaires lorsque le modèle leur assigne la même probabilité (ou bien des probabilités similaires) pour l'événement. Selon cette définition, une méthode est dite calibrée (au sens *faible*) si $\forall k, q\ \hat{\pi}_{kq} = Pr(Y = 1|\hat{\pi} = \hat{\pi}_{kq})$; autrement dit si, pour une probabilité estimée générique $\hat{\pi}_{kq}$, le taux d'observations prenant la modalité q parmi les observations auxquelles la méthode assigne la probabilité $\hat{\pi}_{kq}$ est effectivement égal à $\hat{\pi}_{kq}$.

Plusieurs critiques peuvent être soulevées vis-à-vis de cette procédure. Tout d'abord, elle s'appuie uniquement sur les probabilités estimées par le modèle. Or, des observations aux profils très différents, mais ayant la même probabilité d'événement peuvent alors être regroupées. Ensuite, ne disposant généralement pas de plusieurs observations avec une même probabilité estimée, on est amené à partitionner les observations selon des intervalles de probabilités disjoints. Cette approche nécessite de choisir le nombre et la largeur de ces intervalles. L'ensemble de ces choix n'est pas anodin car il influe directement sur le diagnostic de calibration et reste difficile en pratique.

3 Objectifs et approches proposées

Une calibration forte implique une calibration faible. Cependant, la réciproque est fausse et donc une calibration faible ne doit pas être interprétée comme une calibration forte [7]. L'objectif de cette thèse consistera à explorer de nouvelles stratégies de mesure de la calibration, qui permettent de s'approcher davantage de la définition de calibration au sens fort.

Un état de l'art sera réalisé sur les approches existantes dans les différentes disciplines où le problème de la calibration se pose (e.g. météorologie, métrologie, médecine). Plusieurs approches nouvelles pour la mesure de la calibration sont envisagées. Celles-ci s'appuieront sur de nouveaux critères pour la définition de groupes d'observations homogènes. L'utilisation de techniques de réechantillonnage et de génération de données artificielles seront également explorées. L'impact de ces nouvelles méthodes pour l'évaluation de la recalibration sera également étudié.

4 Calendrier prévisionnel

Le travail de thèse se déroulera selon les étapes suivantes

- Étape 1 (6 mois) : étude bibliographique détaillée sur la définition, la mesure et l'amélioration de la calibration d'un classifieur probabiliste.
- Étape 2 (9 mois) : développement des nouvelles méthodes d'évaluation de la calibration dans le cas d'une variable cible binaire
- Étape 3 (9 mois) : extension des méthodes proposées au cas d'une variable cible nominale polytomique et ordinale
- Étape 4 (6 mois) : évaluation de l'impact des méthodes proposées sur les procédures de recalibration

Les 6 derniers mois seront consacrés à la rédaction de la thèse.

References

[1] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3459–3467. PMLR, 16–18 Apr 2019.

[2] J Platt. Probabilities for sv machines. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74, Cambridge, MA, 2000. MIT Press.

[3] Meelis Kull, Telmo M. Silva Filho, and Peter Flach. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*, 11(2):5052 – 5080, 2017.

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017.

[5] Frank E. Harrell Jr., Kerry L. Lee, and Daniel B. Mark. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, 1996.

[6] Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J. Pencina, and Ewout W. Steyerberg. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*, 74:167–176, 2016.

[7] Margaret Pepe and Holly Janes. Methods for evaluating prediction performance of biomarkers and tests. In Mei-Ling Ting Lee, Mitchell Gail, Ruth Pfeiffer, Glen Satten, Tianxi Cai, and Axel Gandy, editors, *Risk Assessment and Evaluation of Predictions*, pages 107–142, New York, NY, 2013. Springer New York.