JCPC : Approche de Calibration des Probabilités des Classifieurs basée sur la règle de Jeffrey

Sara Kebir*, Karim Tabia*

* Univ. Artois, CNRS, CRIL, F-62300 Lens, France
{kebir,tabia}@cril.fr

Résumé. Dans de nombreuses applications critiques, les modèles d'apprentissage automatique doivent non seulement prédire l'étiquette de classe avec précision, mais aussi fournir la probabilité que la prédiction soit correcte. Cette probabilité détermine si l'on peut faire confiance ou non à la prédiction. Dans cet article, nous présentons une nouvelle approche pour calibrer les probabilités des modèles d'apprentissage automatique via une étape de post-traitement. Le point de départ de ce travail est l'observation que la calibration est plutôt meilleure sur un petit nombre de catégories ou de sous-ensembles de classes que sur un grand nombre de classes. L'approche de calibration que nous proposons, appelée *JCPC*, est basée sur la révision probabiliste des croyances et calibre les probabilités prédites sur les classes en utilisant les probabilités prédites sur les catégories. Notre étude expérimentale sur plusieurs jeux de données et modèles d'apprentissage automatique montre des résultats très prometteurs.

1 Introduction

La décennie actuelle est fortement marquée par l'omniprésence des systèmes intelligents basés sur l'IA dans un grand nombre de domaines. Cela a soulevé des questions sensibles et des défis pour la communauté de l'IA, en particulier lorsqu'il s'agit d'IA explicable et digne de confiance. Pour ce deuxième aspect, la calibration de la confiance est un facteur important pour une IA plus fiable. En effet, de nombreux systèmes et applications intelligents et innovants reposent en grande partie sur l'apprentissage machine (ML), ce qui donne lieu à de nouveaux problèmes et risques induits, pour la plupart, par la complexité des systèmes, leur opacité et la sensibilité de certaines applications critiques.

Une IA digne de confiance nécessite l'utilisation de modèles dont la confiance est bien calibrée. En effet, il faut non seulement faire des prédictions précises, mais aussi fournir des probabilités (interprétées comme la confiance du modèle) que les prédictions sont correctes. De telles probabilités permettent de savoir quand faire confiance au modèle, ce qui peut avoir des conséquences sur la manière dont les prédictions sont gérées dans le cas d'une prise de décision automatisée. Par exemple, des probabilités de confiance bien calibrées indiquent quand les prédictions sont susceptibles d'être incorrectes. Cela permet de gérer ces prédictions en conséquence. Certaines techniques d'apprentissage automatique, telles que les classifieurs de réseaux bayésiens, peuvent fournir directement des probabilités postérieures, tandis que de