Prédiction conformelle profonde pour des modèles robustes

Soundouss Messoudi*, Sylvain Rousseau*, Sébastien Destercke*

*HEUDIASYC - UMR CNRS 7253, Université de Technologie de Compiègne
57 avenue de Landshut, 60203 COMPIEGNE CEDEX - FRANCE
https://www.hds.utc.fr/, prenom.nom@hds.utc.fr

Résumé. Les réseaux profonds, comme d'autres modèles, peuvent associer une confiance élevée à des prédictions peu fiables. Rendre ces modèles robustes et fiables est donc essentiel, surtout pour les décisions critiques. Ce papier montre expérimentalement que la prédiction conformelle, et plus particulièrement l'approche de [Hechtlinger et al. (2018)], apporte une solution convaincante à ce défi. La prédiction conformelle fournit un ensemble de classes couvrant la vraie classe avec avec une fréquence choisie au préalable par l'utilisateur. Dans le cas où l'exemple à prédire est atypique, la prédiction conformelle prédira l'ensemble vide. Les expériences menées montrent le bon comportement de l'approche conformelle, en particulier lorsque les données sont bruitées.

1 Introduction

L'apprentissage automatique et les modèles profonds sont aujourd'hui partout. Il a cependant été démontré que ces modèles peuvent fournir des scores et donc une confiance élevée dans une prédiction manifestement erronée. Ainsi, une image de chien peut être reconnue de manière presque certaine comme un panda, suite à un bruitage invisible à l'oeil nu. De plus, les réseaux profonds se prêtant peu à l'explication et à l'interprétabilité de par leur nature, il est d'autant plus important de rendre leurs décisions robustes et fiables.

Il existe plusieurs méthodes pour estimer la confiance à avoir dans les prédictions d'algorithmes d'apprentissage automatique. L'estimation « hold-out » est une des plus anciennes. Cette méthode calcule le niveau de confiance en utilisant un jeu de test pour estimer la confiance, qui correspond au taux d'erreur observé sur ce jeu de test. Si cette méthode est simple à mettre en oeuvre, l'évaluation de ses performances est sujette à une variance plus importante quand la taille des données de test est petite ou lorsqu'on a des données aberrantes [Kohavi et al. (1995)]. Elle ne fournit pas non plus les mêmes garanties statistiques que la prédiction conformelle.

Initialement, la prédiction conformelle [Vovk et al. (2005)] est une méthode d'apprentissage en ligne transductive qui effectue entraînement, apprentissage et prédiction simultanément, et fournit des prédictions parfois sous forme d'ensemble de classes dont la fiabilité statistique (le pourcentage moyen de recouvrement de la vraie classe par l'ensemble prédit) est garantie sous des hypothèses faibles. En pratique, la méthode utilise les prédictions et observations passées pour fournir une prédiction à la probabilité d'erreur ϵ préalablement définie. La méthode est générique, puisqu'elle ne requiert qu'une mesure de non-conformité pour être appliquée. Le principe de la prédiction conformelle sera rappelé en section 2.