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.

Notre travail utilise une extension de ce principe proposé par [Hechtlinger et al. (2018)]. Ils proposent d'utiliser pour produire la prédiction la densité $p(x|y)$ plutôt que $p(y|x)$. Cela permet notamment de différencier deux cas d'incertitudes différents : le premier prédit plus d'une étiquette compatible avec x en cas d'ambiguité et le deuxième prédit l'ensemble vide $\emptyset$ lorsque le modèle ne sait pas ou n'a pas vu un exemple similaire au cours de l'entraînement. Cette approche est rappelée dans la section 3. Cependant, les tests réalisés dans [Hechtlinger et al. (2018)] ne concernaient que les images et des CNN.

Dans la section 4, nous montrons expérimentalement que cette approche est bien générique, dans le sens où elle fonctionne pour des architectures (CNN, GRU et MLP) et des natures de données (images, textes, tabulaires) variées.

2 Prédiction conformelle

Soit $z_1 = (x_1, y_1), z_2 = (x_2, y_2), \ldots$ des paires successives constituant les exemples, avec $x_i \in X$ un objet et $y_i \in Y$ son étiquette. Nous notons $Z := X \times Y$ l'espace des exemples. La prédiction conformelle [Vovk et al. (2005)] est une méthode d'apprentissage en ligne transductive, qui fournira la $n^{ème}$ prédiction $\hat{y}_n$ en utilisant les caractéristiques observées x_n de l'exemple ainsi que les exemples précédents $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$. En général, l'exemple x_n avec sa vraie étiquette y_n est ensuite rajouté à l'ensemble d'apprentissage. Contrairement aux approches inductives, il n'y donc pas de modèle générique appris. Afin de définir un prédicteur conformel, il est nécessaire de définir avant d'autres notions.

On notera Z^* l'espace des séquences d'exemples. Dans la suite, nous considérerons toujours le cas $n > 0$, c'est-à-dire des séquences non-vide $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1}) \in Z^*$ quelconque. Un prédicteur de confiance a les caractéristiques suivantes : il permet de prédire un sous-ensemble de Y, et cette prédiction est assortie d'une probabilité d'erreur $\epsilon \in [0, 1]$ appelée niveau de signification (*significance level*) qui correspond à une garantie statistique de couverture de la vraie étiquette $1 - \epsilon$ appelée le niveau de confiance (*confidence level*). Formellement, un prédicteur de confiance Γ mesurable se définit comme suit :

$$\Gamma : Z^* \times X \times [0, 1] \to 2^Y, \tag{1}$$

où 2^Y dénote l'ensemble des parties de Y. On note Γ^ϵ le prédicteur Γ de confiance ϵ. Il doit être décroissant au sens de l'inclusion par rapport à ϵ, c'est à dire qu'on doit avoir :

$$\forall n > 0, \quad \forall \epsilon_1 \geq \epsilon_2 \quad , \Gamma^{\epsilon_1}(z^*, x_n) \subseteq \Gamma^{\epsilon_2}(z^*, x_n) \tag{2}$$

Pour un ϵ fixé, un prédicteur de confiance sera dit *valide* si le taux d'erreur moyen des prédictions ne dépasse pas ce taux, et *efficient* si la taille des ensembles prédits est aussi petite que possible. Pour construire un tel prédicteur, nous nous appuyons sur une mesure de non-conformité notée A_n. Cette mesure est un score qui nous indique à quel point un exemple z_i est différent des autres exemples $\wr z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n \int$. Le score de dissimilarité de z_i par rapport aux autres exemples est noté ainsi :

$$\alpha_i := A_n(\wr z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n \int, z_i), \tag{3}$$

En comparant α_i aux autres scores de non-conformité α_j avec $j \neq i$, nous calculons un « degré de signification » (*p-value*) π_i de z_i exprimant la proportion d'exemples moins

conformes que z_i, avec $\pi_i = {}^{|\{j=1,\ldots,n:\alpha_j \geq \alpha_i\}|}/_{n}$. Si π_i s'approche de la borne inférieure $1/n$ alors z_i est non conforme par rapport à la plupart des autres exemples (un exemple aberrant). Si, au contraire, il s'approche de la borne supérieure 1 alors z_i est très conforme.

À partir de ces π_i, il est possible de définir un prédicteur conformel en prédisant l'ensemble des étiquettes y telles que ${}^{|\{i=1,\ldots,n:\alpha_i \geq \alpha_n\}|}/_{n} > \epsilon$, avec pour $i = 1, \ldots, n-1$:

$$\alpha_i := A_n(\wr z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n \int, z_i) \text{ et } \alpha_n := A_n(\wr z_1, \ldots, z_{n-1} \int, z_n). \tag{4}$$

Construire un prédicteur conformel revient donc à définir une fonction de non-conformité. La section suivante présente une telle mesure s'appuyant sur des représentations issues de l'apprentissage profond ainsi que sur des modèles d'estimation de densité.

3 Présentation du modèle

[Hechtlinger et al. (2018)] utilise la méthode suivante : Soit $\hat{p}(x|y)$ une estimation de densité de $p(x|y)$ pour l'étiquette $Y = y$. On définit $\hat{t}_y$ pour qu'il soit le fractile empirique d'ordre ϵ des valeurs $\{\hat{p}(x_i|y)\}$; tel que :

$$\hat{t}_y = \sup \left\{ t : \frac{1}{n_y} \sum_i I(\hat{p}(x_i|y) \geq t) \geq 1 - \epsilon \right\} \tag{5}$$

où n_y est le nombre d'éléments appartenant à la classe y. On fixe $\Gamma^\epsilon(x) = \{y : \hat{p}(x|y) \geq \hat{t}_y\}$. Soit (X, Y) une nouvelle observation. Le papier [Hechtlinger et al. (2018)] montre que $|P(Y \in \Gamma^\epsilon(X)) - (1-\epsilon)| \xrightarrow{P} 0$ avec $\min_y n_y \to \infty$.

Ensuite, les données étiquetées sont divisées en deux parties : la première pour construire $\hat{p}(x|y)$, la seconde pour d'évaluer $\{\hat{p}(x_i|y)\}$ et définir $\hat{t}_y$. Puis, on fixe $C(x) = \{y : \hat{p}(x|y) \geq \hat{t}_y\}$. Cela garantit que les x avec une probabilité faible - c'est-à-dire les régions pauvre en exemples - soient classés comme $\emptyset$. Cette approche par division évite le grand coût des calculs de l'apprentissage profond dans le cas où l'approche en ligne est utilisée.

La qualité finale du prédicteur (son efficience, sa robustesse) dépendent en partie de l'estimateur de densité. [Lei et al. (2013)] tend à suggérer que l'utilisation d'estimateurs à noyaux donnent de bons résultats dans des conditions faibles.

4 Expériences effectuées

Afin d'examiner l'efficacité de la méthode conformelle sur différents types de données, trois jeux de données pour classification binaire ont été utilisés. Ils sont :

1. **CelebA** [Liu et al. (2015)] : jeu de données de visages avec plus de 200000 images de célébrités utilisées pour déterminer si une personne est un homme 1 ou une femme 0.
2. **IMDb** [Maas et al. (2011)] : contient plus de 50000 textes différents décrivant des critiques de films pour une analyse de sentiments (avec 1 représentant un avis positif et 0 indiquant un avis négatif).
3. **EGSS** [Arzamasov (2018)] : contient 10000 exemples pour l'étude de la stabilité du réseau électrique (1 représentant un réseau stable), avec 12 caractéristiques numériques.

4.1 Approche

L'approche globale utilisée respecte les conditions énumérées précédemment et utilise une estimation de la densité par noyaux gaussiens. Chaque jeu de données est divisé en ensembles d'entraînement, de validation et de test. Un modèle d'apprentissage profond dédié à chaque type de données est entraîné sur les données d'entraînement et de validation. L'avant-dernière couche dense sert de vecteur de caractéristiques de taille fixée pour chaque jeu de données et représentant l'objet (image, texte ou vecteur), qui sont utilisés pour la partie conformelle. L'architecture des modèles d'apprentissage profond est représentée dans la figure 1. Elle est construite suivant les étapes ci-dessous :

1. Utiliser un modèle d'apprentissage profond de base suivant le type de données. Dans le cas de CelebA, il s'agit d'un CNN avec un ResNet50 [He et al. (2016)] pré-entraîné sur ImageNet [Deng et al. (2009)] et ajusté sur CelebA. Pour IMDb, ce modèle est un GRU bidirectionnel prenant en entrée les données traitées avec un lexeur *(tokenizer)* et un remplissage *(padding)*. Pour EGSS, ce modèle est un perceptron multicouche (MLP).
2. Appliquer une couche dense intermédiaire et l'utiliser comme extracteur de caractéristiques avec un vecteur (de taille 50 pour CelebA et de taille 64 pour les autres) représentant l'objet, et qui sera utilisé ultérieurement pour la prédiction conformelle.
3. Ajouter une couche dense pour obtenir la classe prédite par le modèle (0 ou 1).

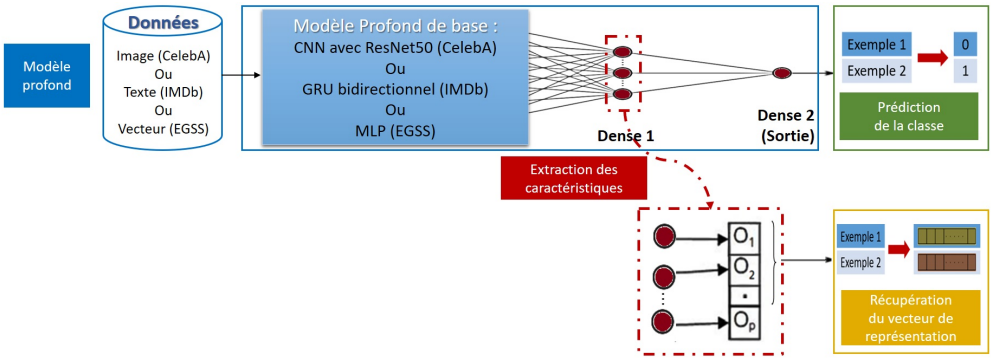

FIG. 1 – *Architecture des modèles d'apprentissage profond.*

En se basant sur les vecteurs récupérés, une estimation de densité par noyau gaussien est réalisée sur les données d'entraînement de chaque classe, afin d'obtenir les valeurs $P(x|y)$. Ensuite, l'ensemble de validation est utilisé pour calculer les scores de densité et les trier afin de déterminer le seuil à un ϵ donné de toutes les valeurs, ce qui permet de délimiter la région de densité de chaque classe. Enfin, l'ensemble de test est utilisé pour calculer les performances du modèle.

La visualisation des régions de densité (figure 2) se fait via le premier plan d'une Analyse en Composantes Principales. Les résultats montrent les régions distinctes des classes 0 (en rouge) et 1 (en bleu) avec une intersection non vide (en vert) qui représente une région d'incertitude aléatoire. Les points en dehors de ces trois régions appartiennent à la région d'incertitude épistémique, signifiant que le classifieur "ne sait pas".

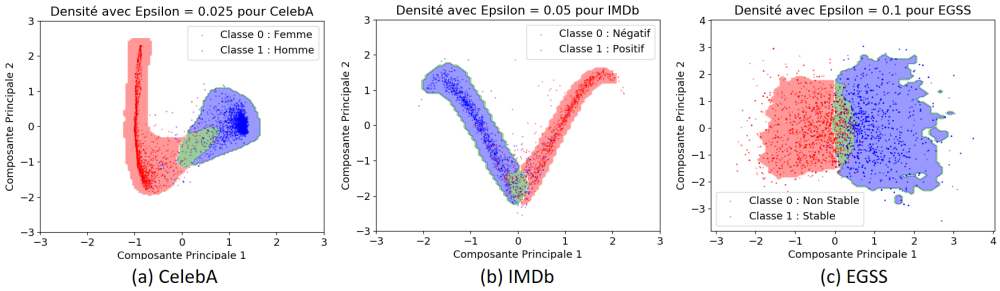

FIG. 2 – *Régions de densité de la prédiction conformelle.*

4.2 Résultats sur les exemples de test

Pour obtenir plus d'informations sur les résultats de cette expérience, la justesse *(Accuracy)* des modèles a été calculée avec différentes valeurs ϵ comprises entre 0,01 et 0,5 lors de la détermination du seuil de densité de prédiction conformelle comme suit :

— La justesse du DL : La justesse du modèle profond de base (CNN pour CelebA, GRU pour IMDb ou MLP pour EGSS) sur tous les exemples de test.

— La justesse du modèle conformel valide : La justesse du modèle conformel quand on considère seulement les prédictions singleton 0 ou 1 (sans prendre en considération les $\{0, 1\}$ et les ensembles vides).

— La justesse du DL valide : La justesse du modèle profond de base sur les exemples de test qui on été prédits en tant que 0 ou 1 par le modèle conformel.

Le pourcentage d'ensembles vides $\emptyset$ et $\{0, 1\}$ a également été calculé à partir de toutes les prédictions des exemples de test effectuées par le modèle de prédiction conformelle. Les résultats sont indiqués dans la figure 3.

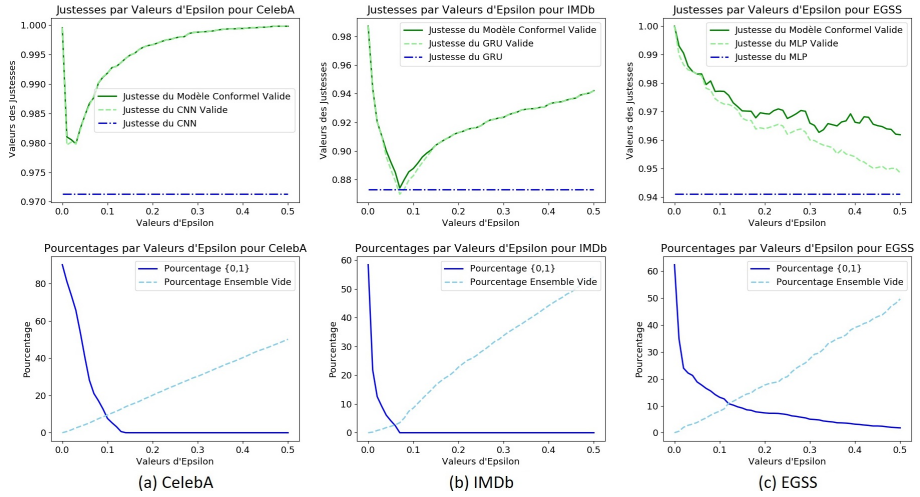

FIG. 3 – *La justesse et les pourcentages en fonction de* ϵ.

Les résultats montrent que la justesse du modèle conformel valide et la justesse du modèle d'apprentissage profond de base valide sont presque égales et sont meilleures que la justesse du modèle de base pour toutes les valeurs ϵ. Sur nos tests, l'ajout de la prédiction conformelle à un modèle profond ne le fait pas perdre en performance, et même l'améliore parfois (EGSS). Cela s'explique par le fait que le modèle de prédiction conformelle permet de s'abstenir de prédire (ensemble vide $\emptyset$) ou prédire les deux classes pour les exemples ambigus, permettant ainsi d'avoir une prédiction plus fiable de l'étiquette. Il est aussi remarqué qu'au fur et à mesure que ϵ croît, le pourcentage d'ensembles $\{0, 1\}$ prédits diminue jusqu'à ne plus en prédire (à $\epsilon = 0.15$ pour CelebA par exemple). A l'inverse, le contraire est observé avec le pourcentage d'ensembles vides $\emptyset$ qui augmente lorsque ϵ augmente.

4.3 Résultats sur les exemples bruités et étrangers

CelebA : Deux types de bruits ont été introduits : un bruit masquant des parties du visage et un autre gaussien sur l'ensemble des pixels. Ces perturbations et leurs prédictions sont illustrés dans la figure 4 avec "Pred" la prédiction du CNN et "C_Pred" celle du modèle conformel. Cet exemple montre que le CNN et le modèle de prédiction conformelle identifient correctement la femme dans l'image (a). Cependant, en masquant l'image (b), le CNN la prédit en tant qu'homme avec un score de 0.6 alors que le modèle de prédiction conformelle est plus prudent en indiquant qu'il ne sait pas ($\emptyset$). Lorsqu'on applique un bruit gaussien sur toute l'image (c), le CNN prédit que c'est un homme avec un score plus important de 0.91, alors que le modèle conformel prédit les deux classes. Pour les images aberrantes, les exemples (d), (e) et (f) illustrent la capacité du modèle conformel à identifier différentes images aberrantes en tant que tel ($\emptyset$) contrairement au modèle profond qui les prédit en tant qu'hommes avec un haut score.

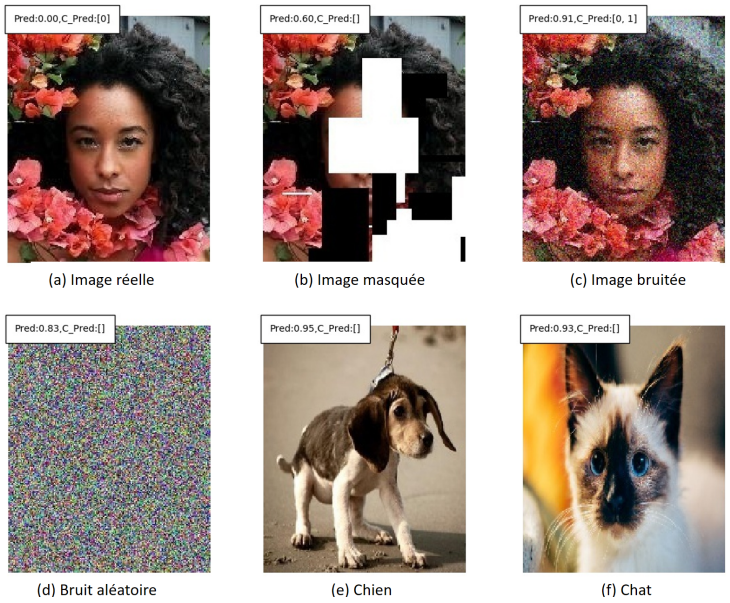

FIG. 4 – *Exemples d'images aberrantes et bruitées comparées à l'image réelle pour CelebA.*

IMDb : La figure 5 affiche une comparaison de deux textes avant et après le changement aléatoire de quelques mots (en gras) par d'autres mots du vocabulaire du modèle. Le texte réel prédit comme avis négatif par les deux modèles devient positif pour le GRU après perturbation. Néanmoins, le modèle conformel est plus prudent en indiquant qu'il peut s'agir des deux cas ($\{0, 1\}$). Pour l'exemple aberrant formé complètement de mots du vocabulaire, le modèle GRU prédit positif avec un score de 0.99, tandis que le modèle conformel dit qu'il ne sait pas ($\emptyset$).

Exemple réel	« **Every great** romantic comedy **needs** conflict between the romantic leads [...] **This story** falls completely flat in this **area** [...] suspense is flat, **there** is no anticipation, **and** there **really is no** allure [...] I was quite surprised. During the movie, I **expected them** more **to** play **a** game of checkers and chat **about** the weather than see **any** moving passion [...] While I'm a fan **of** both actors [...] The writing **was** very **weak, which** also **might** have **impacted** the **performances** [...] »	Etiquette : 0 Prédiction GRU : 0.0979318321 Prédiction Conformelle : 0
Exemple bruité	« **ambidexterous trentini** romantic comedy **dispassionately** conflict between the romantic leads [...] **fanout phoolan** falls completely flat in this **centerpiece** [...] suspense is flat, **binding** is no anticipation, **scroller** there **wiedzmin laudable thunderball** allure [...] I was quite surprised. During the movie, I **lives' 'are** more **subtextual** play **commemorations** game of checkers and chat **stratovarius** the weather than see **linking** moving passion [...] While I'm a fan **unti** both actors [...] The writing **ripoff's** very **nare releases** also **sharikov** have **maes** the **'sketching'** [...] »	Etiquette : 0, Prédiction GRU : 0.7479107976 Prédiction Conformelle : {0, 1}
Exemple aberrant	« wolverines 'sandwich' controversial posit homme subfunctions snowmobile symbiotic malamud challenge needle's personl witch's nonce wills' swooshes cobbled brash mcq wanky 'bought regenerated southstreet amazed ravenna 'mainly belyt hijixn shrugs deodorant mesquida anodynesprech romishness malice seldomely settling dispicable vocation [...] reduce macfarlane's disclosing officers' wiretapping balbao seagals mi3 dibnah romulan controls dolled maguire' [...] »	Prédiction GRU : 0.9996775389 Prédiction Conformelle : []

FIG. 5 – *Exemples d'un texte aberrant et d'un bruité comparé au texte réel pour IMDb.*

EGSS : La figure 6 affiche une comparaison des positions des exemples de tests sur les régions de densités avant (a) et après (b) l'ajout d'un bruit gaussien. Ceci montre que plusieurs exemples se positionnent en dehors des régions de densité après l'introduction des perturbations. Les exemples aberrants (c) créés en modifiant quelques caractéristiques de ces exemples de test avec des valeurs extrêmes (pour simuler une panne de capteur par exemple) sont encore plus éloignés des régions de densités, et reconnus comme tel par le modèle conformel ($\emptyset$).

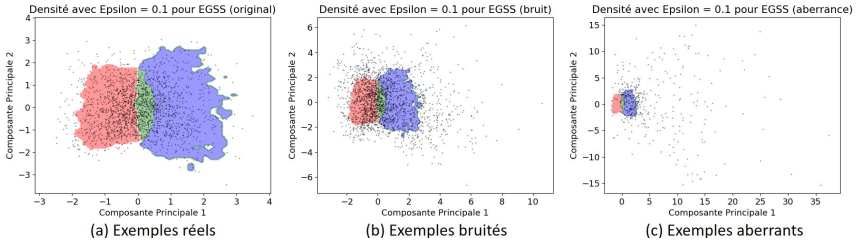

FIG. 6 – *Visualisation des exemples réels, bruités et aberrants pour EGSS.*

5 Conclusions et perspectives

Nous avons utilisé la prédiction conformelle et la technique de [Hechtlinger et al. (2018)] afin d'avoir un modèle d'apprentissage profond plus fiable et prudent. Les résultats montrent l'intérêt de cette méthode sur différentes données (image, texte, tabulaire) utilisées avec différentes architectures d'apprentissage profond (CNN, GRU et MLP). En effet, dans ces trois cas, le modèle conformel non seulement ajoute de la fiabilité et de la robustesse au modèle profond

en détectant les exemples ambigus mais aussi garde ou améliore la performance du modèle profond de base quand il ne prédit qu'une seule classe. Nous avons aussi illustré la capacité de la prédiction conformelle à gérer exemples bruités et aberrants pour les trois types de données. Ces expériences montrent que la méthode conformelle peut donner plus de robustesse et de fiabilité aux prédictions sur plusieurs types de données et d'architectures profondes de base.

Pour améliorer les expériences et résultats, les perspectives incluent l'optimisation de l'estimation de densité basée sur les réseaux de neurones. A ϵ fixé se pose par exemple le problème de trouver le modèle le plus efficient. Aussi, il serait utile de comparer la prédiction conformelle avec des méthodes de calibration, par exemple évidentielle [Denoeux (2019)].

Références

Arzamasov, V. (2018). UCI electrical grid stability simulated data data set.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, et L. Fei-Fei (2009). ImageNet : A Large-Scale Hierarchical Image Database. In *CVPR*.

Denoeux, T. (2019). Logistic regression, neural networks and dempster–shafer theory : A new perspective. *Knowledge-Based Systems 176*, 54–67.

He, K., X. Zhang, S. Ren, et J. Sun (2016). Deep residual learning for image recognition. In *CVPR*, pp. 770–778.

Hechtlinger, Y., B. Póczos, et L. Wasserman (2018). Cautious deep learning. *arXiv preprint arXiv :1805.09460*.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, Volume 14, pp. 1137–1145. Montreal, Canada.

Lei, J., J. Robins, et L. Wasserman (2013). Distribution-free prediction sets. *Journal of the American Statistical Association 108*(501), 278–287.

Liu, Z., P. Luo, X. Wang, et X. Tang (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, et C. Potts (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics : Human language technologies-vol. 1*, pp. 142–150.

Vovk, V., A. Gammerman, et G. Shafer (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Summary

Deep networks like other learning models can associate high trust to unreliable predictions. Making these models robust and reliable is therefore essential, especially for critical decisions. This experimental paper shows that the conformal prediction approach of [Hechtlinger et al. (2018)] brings a convincing solution to this challenge. Conformal prediction consists in predicting a set of classes covering the real class with a user-defined frequency. In the case of atypical examples, the conformal prediction will predict the empty set. Experiments show the good behavior of the conformal approach, especially when the data is noisy.