Calibration des modèles d'apprentissage pour l'amélioration des détecteurs automatiques d'exemples mal-étiquetés

Ilies Chibane*, Thomas George*,
Pierre Nodet*, Vincent Lemaire*

*Orange Innovation
prénom.nom@orange.com,
https://hellofuture.orange.com

Résumé. Les données mal-étiquetées sont un problème répandu qui dégrade la performance des modèles d'apprentissage automatique en contexte industriel. Les méthodes qui permettent de détecter les exemples mal-étiquetés reposent la plupart du temps sur l'introspection d'un modèle d'apprentissage, qui est entraîné puis sondé pour chaque exemple afin d'obtenir un score de confiance indiquant si l'étiquette fournie est bonne ou mauvaise. Dans cet article, nous étudions la calibration de ce modèle sous-jacent. Nous montrons empiriquement que l'emploi de méthodes de calibration améliore la précision et la robustesse de la détection d'exemples mal-étiquetés, ce qui permet d'obtenir une solution pratique et efficace pour des applications industrielles.

1 Introduction

La qualité d'un modèle d'apprentissage automatique est conditionnée par la qualité des données sur lesquelles il est entraîné. Or, la collection d'exemples d'apprentissage et leur annotation peuvent se révéler coûteuse pour les jeux de données les plus volumineux. Obtenir des étiquettes de qualité reste donc un problème majeur : on constate que même des jeux de données publics populaires tels que MNIST ou CIFAR-10/CIFAR-100, présentent des problèmes d'étiquetage (Northcutt et al., 2021a). Pour éviter un examen manuel de chaque étiquette, des détecteurs d'exemples mal-étiquetés visent à identifier ces exemples de manière automatique (Frénay et Verleysen, 2013). Ces détecteurs fournissent un *score de confiance* pour chaque exemple du jeu données sur lequel ils sont appliqués, indiquant si l'étiquette fournie peut être considérée comme bonne ou mauvaise. Parmi ceux-ci, les détecteurs basés sur l'*introspection* examinent s'il existe une différence de traitement entre les exemples bien et mal-étiquetés lors de l'apprentissage, mesurée à l'aide de *sondes*. Dans cet article, nous explorons une piste d'amélioration de ces détecteurs : nous proposons d'ajouter une étape intermédiaire de *calibration* du modèle d'apprentissage avant de le sonder (figure 1 et section 2.1).

Travaux connexes La détection automatique d'exemples mal étiquetées a fait l'objet de nombreux travaux de recherche. Ces travaux ont permis notamment de mettre en lumière une des limites de ces détecteurs, notamment qu'ils aient du mal à distinguer les exemples difficiles

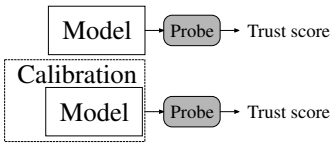

FIG. 1 – *Les méthodes de détection par introspection sondent directement le modèle appris sur le jeu de données d'apprentissage afin d'obtenir un score de confiance pour chaque exemple. Nous proposons d'ajouter une étape intermédiaire de calibration avant de sonder le modèle.*

(exemples rares ou proches d'une frontière de décision) des exemples mal-étiquetés. Cela se produit en particulier avec des classes déséquilibrées, où les exemples de la classe minoritaires ont plus tendance à être considérés comme mal-étiquetés que les exemples de la classe majoritaire (Northcutt et al., 2021b; Kuan et Mueller, 2022). Bien que le lien entre bruit d'étiquetage et calibration ait déjà été étudié (Joel Söderberg, 2023), savoir si calibrer un modèle avant de le sonder améliore la détection d'exemple mal-étiquetés reste une question ouverte.

Contributions

1. Nous proposons d'ajouter une étape de calibration des modèles d'apprentissage avant de les sonder dans les détecteurs par introspection ;
2. Nous évaluons la pertinence de notre solution sur une série de jeux de données avec un bruit d'étiquetage réaliste, en apprenant un classifieur sur une version filtrée du jeu de données initial débarrassée de ses exemples avec plus faible score de confiance, montrant une amélioration de l'équilibre des classes de la version filtrée du jeu de données par rapport à la version initiale ainsi qu'une meilleure performance prédictive des classifieurs appris sur le jeu de données filtré.

2 Contexte

En classification supervisée, à partir d'un jeu d'apprentissage d'exemples $(x, y_{\text{observée}})$, composés de variables d'entrée x (mots d'un texte, pixels d'une image, ...), et d'une étiquette $y_{\text{observée}}$ (bonne ou mauvaise review, chien ou chat, ...). L'objectif est d'apprendre une fonction f (modèle) de l'espace des variables d'entrées $\mathcal{X}$ à l'ensemble des classes $\mathcal{Y}$ qui généralise le mieux possible à des exemples non-vus à l'apprentissage : on cherche à ce que la décision prise par notre fonction pour un exemple dit test soit la même que l'étiquette qu'aurait obtenu cet exemple s'il avait été annoté. En particulier, la fonction f renvoie un vecteur de taille C contenant la confiance du modèle dans l'appartenance d'un exemple à chacune des classes possibles. Dans ce cas, la décision prise $\hat{y}$ par f est la classe avec la confiance la plus élevée, $\hat{y}(x) := \arg\max_c \hat{f}(x)_c$.

Dans cet article, nous faisons l'hypothèse que le processus d'étiquetage est déterministe : les exemples du jeu d'apprentissage sont échantillonnés à partir d'une distribution $\mathbb{P}(x)$, puis annotés avec une seule vraie étiquette possible y_{oracle}. Un exemple est dit mal-étiqueté si $y_{\text{observée}} \neq y_{\text{oracle}}(x)$.

La réussite de l'apprentissage de bonnes fonctions f dans le cadre de la classification supervisée tient, entre autres, dans la taille du jeu d'apprentissage. Or pour chacun de ces exemples, il est nécessaire de l'annoter, ce qui est un processus parfois long et coûteux. C'est pourquoi il est fréquent de vouloir automatiser ce processus, qui peut alors causer des erreurs d'étiquetage. Pour pallier ce problème, une approche possible est de détecter automatiquement ces erreurs d'étiquetage dans le jeu d'apprentissage (Frénay et Verleysen, 2013).

2.1 Méthodes d'introspection pour la détection d'exemples mal-étiquetés

Les détecteurs automatiques d'exemples mal-étiquetés sont des outils qui pour chaque exemple du jeu d'apprentissage renvoient un *score de confiance*. Les exemples avec les plus faibles scores sont alors susceptibles d'être analysés manuellement (ré-étiquetage actif) ou alors supprimés automatiquement (filtrage). Parmi ces détecteurs, nous nous focalisons sur les méthodes d'introspection, qui couvrent la plupart des méthodes de l'état de l'art.

Les méthodes d'introspection se reposent sur un modèle d'apprentissage déjà appris sur des exemples issus du jeu d'apprentissage. Pour savoir ce qu'il pense de l'étiquette donnée à un exemple, on interroge le modèle avec une *sonde*, telle que la confiance donnée à l'étiquette par le modèle. L'idée repose sur le fait que les modèles d'apprentissage sont un minimum robustes au bruit d'étiquetage en se concentrant principalement sur l'apprentissage du *signal* contenu dans les exemples bien étiquetés, plutôt que sur le *bruit* porté par les exemples mal-étiquetés.

La confiance n'est pas la seule manière de *sonder* un modèle. Pour les lecteurs intéressés par le fonctionnement de ces détecteurs et de leurs *sondes*, nous les renvoyons vers une étude systémique récente sur ce sujet (George et al., 2024).

2.2 Calibration des modèles d'apprentissage automatique

Un modèle d'apprentissage automatique est dit calibré si la confiance *prédite* par le modèle correspond à la fréquence *observée* que la classe prédite soit correcte. Il convient de noter qu'un modèle dont les performances sont médiocres peut être bien calibré si la confiance accordée à ses prédictions est faible.

Plus formellement, un classifeur est calibré sur le label le plus probable (top-label) [1] si :

$$\forall \tau \in [0,1],\ \underbrace{\mathbb{E}_{x \sim \mathbb{P}_{\leq \tau}(x)}\left[\delta_{y_{\text{oracle}}(x)=\hat{y}(x)}\right]}_{\text{fréquence de prédictions correctes}} = \underbrace{\mathbb{E}_{x \sim \mathbb{P}_{\leq \tau}(x)}\left[\hat{f}\left(x\right)_{\hat{y}(x)}\right]}_{\text{confiance du modèle}}$$

avec τ un seuil de confiance et $\mathbb{P}_{\leq \tau}\left(x\right)$ est $\mathbb{P}\left(x\right)$ restreinte aux valeurs de x pour lesquels la confiance du modèle dans la classe prédite $f\left(x\right)_{\hat{y}(x)}$ est inférieure ou égale à τ.

Par défaut, les classifieurs ne sont pas particulièrement bien calibrés. Un des méthodes de calibration la plus populaire est la régression isotonique. Elle repose sur un regroupement des exemples de manière croissante par la confiance prédite par le modèle qu'on souhaite calibrer. Dans chacun de ces groupes, on calcule la confiance moyenne ainsi que la précision moyenne sur un ensemble de calibration. Au moment de prédire on remplace la confiance du modèle par la précision moyenne du groupe dans lequel tombe l'exemple.

1. on utilise ici la calibration top-label car, en classification multi-classe, plusieurs définitions coexistent, voir section 3.1 de Perez-Lebel et al. (2023).

3 Notre proposition : calibrer pour mieux détecter

3.1 Intuition

Le modèle sous-jacent utilisé dans les méthodes de détection par introspection est primordial pour le fonctionnement de ces détecteurs, il doit être notamment bien adapté à la tâche d'apprentissage et un minimum robuste au bruit d'étiquetage (George et al., 2024). Dans cet article nous allons nous intéresser à une autre propriété potentiellement intéressante, sa calibration. Quitte à utiliser la confiance d'un modèle, autant qu'elle soit correctement calibrée ?

En effet, utiliser un modèle calibré semble être plein de promesses. En plus d'avoir de meilleur performances (Niculescu-Mizil et Caruana, 2005), calibrer un modèle améliore aussi sa robustesse (Zhang et al., 2023). Dans le cas où les classes à prédire sont déséquilibrées, la calibration permet d'éviter d'obtenir des modèles trop confiants dans des exemples étiquetés comme appartenant à la classe majoritaire, au détriment des exemples des autres classes (Huang et al., 2020). Enfin, cela permet d'obtenir un meilleur estimateur de la distribution a posteriori $\mathbb{P}(Y|X = x)$, ce qui peut améliorer la détection dans le cas où on utilise des *sondes* basées sur la confiance dans l'étiquette observée $\mathbb{P}(Y = y_{\text{observée}}|X = x)$.

3.2 Notre proposition

Nous proposons d'ajouter à chaque méthode de détection par introspection une étape intermédiaire de calibration avant de sonder le modèle, comme schématisé dans la Figure 1. Comme la calibration agit uniquement sur les confiances prédites par le modèle, cette nouvelle étape n'aura d'effet que sur des détecteurs qui utilisent des sondes basées sur la confiance. Dans ce cas, les *scores de confiance* sont calculés à partir des confiances calibrées modifiant ainsi la priorisation des exemples à analyser ou à filtrer.

4 Expérimentations avec un bruit d'étiquetage réaliste

4.1 Protocole expérimental

Pour évaluer l'efficacité de notre approche, nous utilisons un pipeline en 3 étapes :
détection ⇒ **filtrage** ⇒ **apprentissage**. Notre but est de savoir si *calibrer* le modèle d'apprentissage dans l'étape de **détection** permet d'obtenir le meilleur jeu de donnée **filtré** contenant les meilleurs exemples pour **apprendre** des modèles plus performants. La performance de notre modèle final sera évalué avec une fonction de perte d'entropie croisée sur un jeu de test indépendant. Cette méthodologie d'évaluation des détecteurs permet de pondérer les erreurs (faux positifs ou faux négatifs) en fonction de leur utilité pour entraîner le classifieur final.

Jeux de données : Nous utilisons les mêmes jeu de données et bruit d'étiquetage que dans George et al. (2024). Le benchmark contient 19 jeux de données de classification multi-classes sur des données tabulaires et textuelles. Les jeux de données sont variés, que ce soit en termes de nombre d'exemples, de nombre de variables, de nombre de classes ou bien d'équilibre de classes. Chacun de ces jeux de données est fourni avec une liste de règles d'annotations automatiques. Ces règles sont des fonctions s'appliquant à chacun des exemples d'apprentissage et

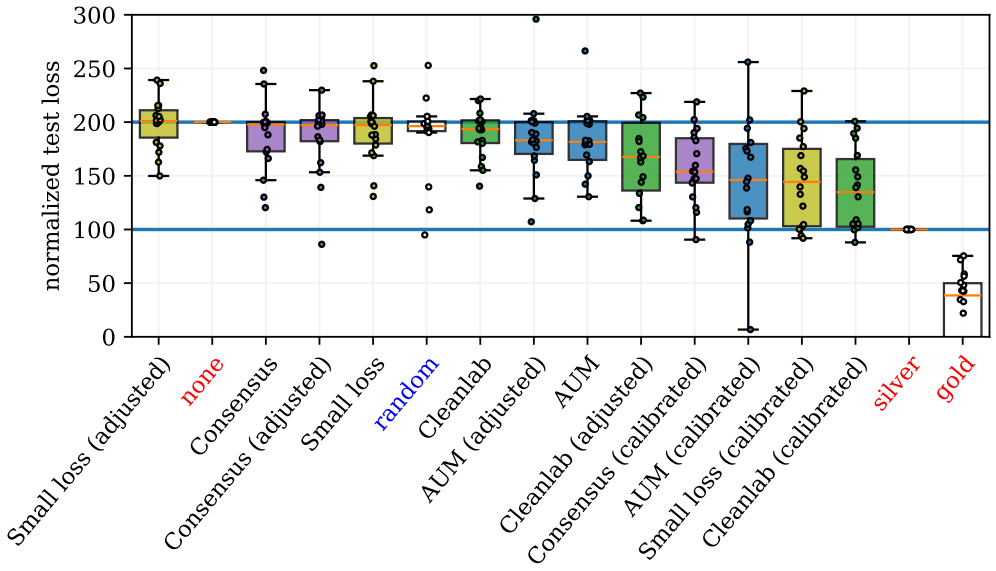

FIG. 2 – *Distribution (boite à moustache) de la fonction de perte normalisée (100 équivaut à la fonction de perte d'un modèle appris sur les exemples propres uniquement, 200 équivaut à la fonction de perte d'un modèle appris sur tous les exemples d'apprentissage, incluant les mal-étiquetés) en test du modèle appris après la pipeline en 3 étapes de différents détecteurs. La calibration est effectuée sur un ensemble de calibration sans bruit d'étiquetage. Le nom des méthodes inclut le nom du détecteur, et le nom de son amélioration (si existante).*

renvoyant soit une étiquette, soit une non-réponse. Comme plusieurs règles sont fournies par jeu de données, l'agrégation de ces règles donne une étiquette $\tilde{y}$ qui est une estimation bruitée de la vraie étiquette y_{oracle}, particulièrement dans les régions de l'espace des variables d'entrées où peu de règles donnent une réponse ou lorsque beaucoup de règles sont en désaccord.

Détecteurs testés : Dans ce benchmark, plusieurs détecteurs sont testés, notamment Area Under the Margin (AUM, Pleiss et al., 2020) qui somme les marges de prédiction au cours de l'apprentissage, Small Loss (Amiri et al., 2018; Jiang et al., 2018) qui utilise la valeur de la fonction de perte, Consensus Consistency (Jiang et al., 2021) qui regarde les désaccords entre plusieurs modèles appris sur des versions bootstrap du jeu d'apprentissage, et CleanLab (Northcutt et al., 2021b) qui évalue la confiance out-of-bag de modèles appris sur des sous-échantillons du jeu d'apprentissage.

Améliorations testées : Nous testons deux manières d'améliorer les détecteurs cités précédemment, en *calibrant* les modèles (notre proposition), et en *ajustant* les scores de confiances des modèles (Kuan et Mueller, 2022). Pour l'approche de Kuan et Mueller (2022), les scores de confiance sont *ajustés* de manière à ce que les scores de confiance moyens par classe soient normalisés. Pour notre approche, nous utilisons la régression isotonique pour calibrer les modèles. Sans amélioration, les détecteurs seront qualifiés d'*étalons*.

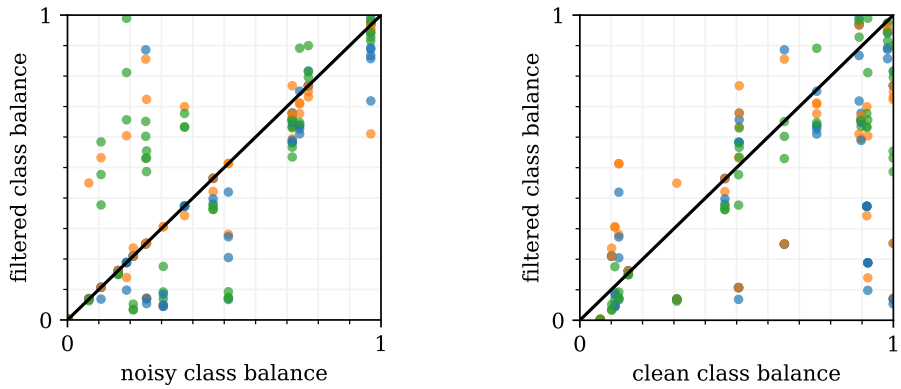

FIG. 3 – *Pour chaque paire détecteur/jeu de données (un point), on compare l'équilibre des classes du jeu de données avec bruit d'étiquetage (à gauche) et sans bruit d'étiquetage (à droite) sur l'axe des x, avec l'équilibre des classes du jeu de données filtré sur l'axe des y. Les points bleus • correspondent aux détecteurs sans calibration, les points oranges • correspondent aux détecteurs ajustés, et les points verts • correspondent aux détecteurs calibrés. Par rapport aux méthodes calibrées qui respectent plus le vrai équilibre des classes (les points verts sont autour de la droite $y = x$ dans la figure de droite), les méthodes ajustées tendent à plus respecter l'équilibre des classes bruitées (les points oranges sont autour de la droite $y = x$ dans la figure de gauche). Cependant, l'équilibre des classes bruitées peut être très différent du vrai équilibre des classes pour certaines formes de bruit.*

Pour choisir les hyperparamètres, nous exécutons une recherche aléatoire par détecteur et par jeu de données. On choisit alors le pipeline qui a donné le modèle obtenant la fonction de perte minimale sur un ensemble de validation propre. Ce modèle est alors évalué sur un ensemble de test et les résultats sont reportés dans la partie suivante.

4.2 Résultats

Est-ce que l'ordonnancement généré en calibrant permet d'apprendre de meilleurs modèles ? Pour chaque détecteur, on compare leur performances avec leur performances en ajoutant l'étape de *calibration* ou d'*ajustement*. Dans notre benchmark (figure 2), on observe que la fonction de perte la plus faible est obtenue en calibrant les modèles.

Comment change l'ordonnancement lorsqu'on calibre ? Nous faisons l'hypothèse que les détecteurs avec *calibration* sont meilleurs que les détecteurs *étalons* ou *ajustés* car ils ont tendance à prioriser des sous-populations d'exemples différentes. Dans le cas de déséquilibre de classe, les exemples des classes minoritaires sont clés pour la généralisation, c'est pourquoi dans la figure 3, nous observons comment l'équilibre des classes du jeu filtré se comporte par rapport au vrai équilibre des classes et à l'équilibre des classes du jeu bruité (ou observé). Pour généraliser au cas multi-classe, l'équilibre des classes est défini par le ratio du nombre d'exemples de la classe la plus petite par le nombre d'exemples de la classe la plus grande.

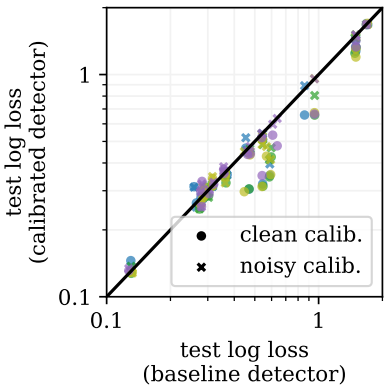

FIG. 4 – *Pour chaque pair détecteur/jeu de données (un point), on compare la perte en test entre un détecteur étalon sur l'axe des x, à la perte en test du détecteur calibré sur un ensemble de calibration propre (les cercles •) et du détecteur calibré sur un ensemble de calibration bruité (les croix ×) sur l'axe des y. Un ensemble de calibration propre est souvent le plus efficace (les cercles sont pour la plupart en dessous de la droite $y = x$). Un ensemble de calibration bruité, bien que moins efficace, ne dégrade jamais les performances par rapport à l'étalon (aucune croix au dessus de la droite $y = x$).*

On observe que les détecteurs *calibrés*, *ajustés* ou *étalons* se comportent différemment selon cet axe d'observation : les détecteurs *calibrés* produisent des jeux filtrés qui reproduisent la vraie distribution des classes de manière plus fidèle (figure 3 droite), alors que les détecteurs *ajustés* ont tendance à reproduire la distribution de classes des jeux bruités (figure 3 gauche). Les détecteurs calibrés semblent donc produire des scores de confiance indépendants de la classe de l'exemple, ce qui permet de mieux différencier les exemples difficiles (comme ceux des classes minoritaires) des exemples mal-étiquetés.

Peut-on calibrer sur un ensemble de calibration bruité ? Pour tester la robustesse de notre approche, nous regardons comment se comportent les détecteurs *calibrés* lorsque l'ensemble de calibration contient lui aussi des exemples mal-étiquetés, avec un processus d'étiquetage corrompu de la même manière que le processus d'étiquetage de l'ensemble d'apprentissage. La figure 4 nous montre que les performances des détecteurs *calibrés* avec un ensemble de calibration propre sont meilleurs que les détecteurs *calibrés* avec un ensemble de calibration bruité. En revanche, même en calibrant sur un ensemble bruité, on constate que les performances ne sont jamais dégradées par rapport à l'*étalon* : l'étape de calibration semble être toujours bénéfique, quelles que soient les conditions.

5 Conclusion

Dans cet article, nous avons proposé d'ajouter une étape intermédiaire de calibration aux détecteurs d'exemples mal-étiquetés par introspection. L'idée est d'avoir un meilleur modèle d'apprentissage, qui, lorsqu'on le sonde, permet d'obtenir de meilleurs scores de confiance. Nous avons montrés que dans des pipelines automatiques pour l'apprentissage avec du bruit d'étiquetage de la forme **détection** $\Rightarrow$ **filtrage** $\Rightarrow$ **apprentissage**, sur un benchmark exhaustif avec du bruits d'étiquetage réaliste, notre approche i) améliore la perte du modèle final en test ii) permet de maintenir l'équilibre des classes après filtrage par rapport au vrai équilibre des classes iii) est robuste au bruit d'étiquetage dans l'ensemble de calibration.

Références

Amiri, H., T. Miller, et G. Savova (2018). Spotting Spurious Data with Neural Networks. In *NAACL*.

Frénay, B. et M. Verleysen (2013). Classification in the presence of label noise : a survey. *IEEE Transactions on Neural Networks and Learning Systems*.

George, T., P. Nodet, A. Bondu, et V. Lemaire (2024). Mislabeled examples detection viewed as probing machine learning models : concepts, survey and extensive benchmark. *TMLR*.

Huang, L., J. Zhao, B. Zhu, H. Chen, et S. V. Broucke (2020). An experimental investigation of calibration techniques for imbalanced data. *IEEE Access*.

Jiang, L., Z. Zhou, T. Leung, L.-J. Li, et L. Fei-Fei (2018). Mentornet : Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*.

Jiang, Z., C. Zhang, K. Talwar, et M. C. Mozer (2021). Characterizing Structural Regularities of Labeled Data in Overparameterized Models. In *ICML*.

Joel Söderberg, M. (2023). The effect of model calibration on noisy label detection.

Kuan, J. et J. Mueller (2022). Model-agnostic label quality scoring to detect real-world label errors. In *ICML DataPerf Workshop*.

Niculescu-Mizil, A. et R. Caruana (2005). Predicting good probabilities with supervised learning. In *ICML*.

Northcutt, C., L. Jiang, et I. Chuang (2021b). Confident learning : Estimating uncertainty in dataset labels. *JAIR*.

Northcutt, C. G., A. Athalye, et J. Mueller (2021a). Pervasive label errors in test sets destabilize machine learning benchmarks. In *NeurIPS D&B*.

Perez-Lebel, A., M. L. Morvan, et G. Varoquaux (2023). Beyond calibration : estimating the grouping loss of modern neural networks. In *ICLR*.

Pleiss, G., T. Zhang, E. Elenberg, et K. Q. Weinberger (2020). Identifying Mislabeled Data using the Area Under the Margin Ranking. In *NeurIPS*.

Zhang, M., X. Zhao, J. Yao, C. Yuan, et W. Huang (2023). When noisy labels meet long tail dilemmas : A representation calibration method. In *ICCV*.

Summary

Mislabeled data is a pervasive issue that undermines the performance of machine-learning models across various industries. Methods for detecting mislabeled instances usually involve training a base machine learning model and then probing it for every instance in order to obtain a trust score that the provided label is genuine or incorrect. In this paper, we experiment with the calibration of this base model. Our empirical findings show that employing calibration methods improves the accuracy and robustness of mislabeled instance detection, providing a practical and effective solution for industry applications.