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

nombreux autres classifieurs s'appuient sur certaines techniques pour fournir des probabilités de confiance. En pratique, de nombreux modèles donnent de mauvaises estimations des probabilités prédictives ; ils les surestiment souvent, comme dans le cas des forêts aléatoires et même des modèles modernes basés sur les réseaux profonds. Ensuite, des techniques de calibration sont souvent utilisées pour mieux calibrer les probabilités prédictives (Filho et al., 2021).

Nous proposons dans cet article une nouvelle approche pour calibrer les probabilités de prédiction d'un classifieur. Le point de départ 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 proposée, appelée *JCPC*, est basée sur la révision de croyances probabilistes avec des entrées incertaines et calibre les probabilités prédites sur les classes en utilisant les probabilités prédites sur des sous-ensembles de classes fournies par un modèle de calibration spécialement formé pour prédire les catégories. Ainsi, nous considérons la calibration comme une tâche de révision d'informations incertaines à la lumière de nouvelles entrées incertaines et plus fiables dans l'esprit de la règle de conditionnement de Jeffrey.

2 Préliminaires et notations

2.1 Calibration des probabilités prédictives

La classification est une tâche prédictive définie par deux ensembles de variables : Un ensemble de caractéristiques X= $\{X_1,..,X_n\}$ où $|X|$=n, et une variable cible discrète notée C prenant des valeurs dans son domaine D_C.

Un classifieur f est une fonction faisant correspondre chaque instance de données d'entrée x (vecteur instanciant chaque variable dans X) à une valeur du domaine D_C.

Intuitivement, un classifieur f est dit calibré (ou fournit des probabilités de prédiction calibrées) si, lorsqu'il prédit une étiquette $c_i \in D_C$ avec la probabilité $\hat{\mathrm{p}}_i$, cette prédiction sera correcte avec la probabilité $\hat{\mathrm{p}}_i$ (la probabilité $p_f(C = c_i|p = \hat{\mathrm{p}}_i)$ est calibrée si en moyenne la prédiction est correcte avec la probabilité $\hat{\mathrm{p}}_i$). Cela signifie que le classifieur quantifie avec précision son incertitude ou sa confiance lorsqu'il fait des prédictions.

2.2 Révision d'informations incertaines avec de nouvelles entrées incertaines : règle de Jeffrey

En se plaçant dans le cadre de la révision d'une information probabiliste codée par une distribution de probabilités antérieure p (sur un domaine discret) lorsqu'une nouvelle preuve ϕ devient disponible. Le conditionnement classique permet de mettre à jour l'information préalable p vers la distribution postérieure $p(.|\phi)$. Lorsque la nouvelle information n'est plus une évidence ou une observation, mais une évidence incertaine, alors nous avons affaire à une révision d'information incertaine avec des entrées incertaines. La règle de Jeffrey (Jeffrey, 1965) étend le conditionnement probabiliste classique au cas où la nouvelle information est incertaine. Elle permet de mettre à jour une distribution de probabilités initiale p en une distribution postérieure p' étant donné l'incertitude portant sur un ensemble d'événements mutuellement exclusifs et exhaustifs λ=$\{\lambda_1,...,\lambda_n\}$ (à savoir, λ est une partition de l'ensemble des états pos-

sibles Ω). Dans ce contexte, la nouvelle entrée est de la forme (λ_i, α_i), i=1..n où α_i désigne la nouvelle probabilité de λ_i.

Étant donné une distribution de probabilités p codant les croyances initiales et les nouvelles entrées sous la forme (λ_i, α_i) pour i=1..n, le degré de probabilité actualisé de tout événement $\phi \subseteq \Omega$ est obtenu comme suit :

$$p'(\phi) = \sum_{\lambda_i} \alpha_i \times \frac{p(\phi \cap \lambda_i)}{p(\lambda_i)} \quad (1)$$

La distribution postérieure p' obtenue à l'aide de la règle de Jeffrey existe toujours et elle est unique (Chan et Darwiche, 2005). Il est à noter que dans la règle de Jeffrey, les événements λ_i doivent être possibles dans la distribution antérieure (à savoir, $\forall \lambda_i \in \lambda$, $p(\lambda_i) > 0$).

3 Calibration des probabilités d'un classifieur basée sur la règle de Jeffrey

Le point de départ de ce travail est l'observation que sur les problèmes de classification impliquant un grand nombre de classes, il est souvent difficile de prédire correctement certaines classes, en particulier dans le cas de jeux de données non équilibrés. Ce qui est vrai pour la précision des prédictions l'est aussi pour les probabilités de confiance. Cette difficulté peut être essentiellement liée à la nature des données et aux spécificités des classifieurs utilisés.

Deux questions simples se posent alors : i) Si l'on regroupe les classes en catégories (sortes de super classes, afin d'être mieux représentées et de réduire le nombre de classes), peut-on améliorer la qualité des prédictions (en termes de précision et de calibration) ? ii) Si oui, serait-il possible d'exploiter les meilleures performances des prédictions sur les catégories (sous-ensembles des classes initiales) pour *rectifier* ou *calibrer* les prédictions d'un classifieur f ?

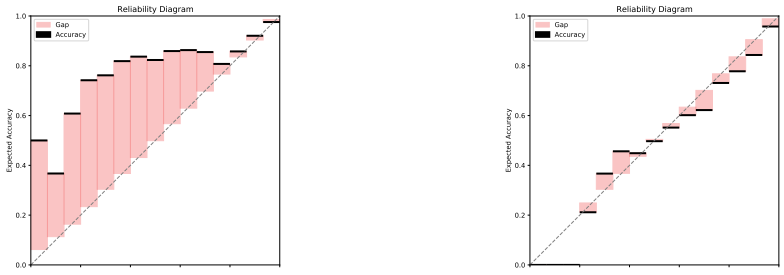
Pour la première question, la réponse est positive pour la plupart des jeux de données et des classifieurs testés, bien qu'avec des résultats différents selon la façon dont les classes $\{c_1, .., c_k\}$ sont regroupées en catégories $\{cat_1, .., cat_j\}$ (avec j<k). Dans la FIG. 1, on peut voir les diagrammes de fiabilité d'un classifieur SVM appris sur le jeu de données DBPedia.

Il est clair que le classifieur SVM construit sur 70 classes est mal calibré (voir l'écart avec la ligne de calibration parfaite) par rapport au SVM construit sur 9 catégories.

Pour la deuxième question, le fait d'avoir un classifieur plus calibré sur les catégories fournit des informations pertinentes sur les classes incluses dans chaque catégorie. Comment et dans quel cas utiliser ces probabilités sur les catégories afin de garantir une amélioration des probabilités de confiance du classifieur initial sera présenté dans ce qui suit.

Nous avons, d'un côté, des classes et une distribution de probabilités p fournie par le classifieur à calibrer f, et de l'autre côté, des catégories (une partition des classes) et une autre distribution de probabilités sur les catégories p' fournie par le classifieur f'. De plus, dans la plupart des cas, les probabilités des classifieurs f' sont plus calibrées comme illustré dans l'exemple de la FIG. 1 ce qui signifie que le classifieur f' est plus fiable en termes de calibration. Ceci place en quelque sorte notre problème dans le cadre de la révision d'informations incertaines par de nouvelles informations incertaines.

Il est tout à fait logique de mettre à jour la distribution p avec p' puisque cette dernière fournit des probabilités plus calibrées. Cela revient à donner la priorité à la nouvelle informa-



(a) SVM 70 (b) SVM 9

FIG. 1 – Diagrammes de fiabilité des classifieurs SVM sur deux niveaux de DBPedia

tion p' exactement en accord avec la règle de Jeffrey. Ainsi, les probabilités révisées p_c sont obtenues en suivant la règle de Jeffrey comme suit : $\forall c_i \in D_C$,

$$p_c(c_i) = p'(cat(c_i)) \times \frac{p(c_i)}{p(cat(c_i))}, \tag{2}$$

où $cat(c_i)$ désigne la catégorie de la classe c_i et $p(cat(c_i))$ la probabilité de toutes les classes de la catégorie $cat(c_i)$ calculée à partir de la distribution p, ($p(cat(c_i)) = \sum_{c_j \in cat(c_i)}(p(c_j))$). Il est à noter que la distribution postérieure p_c existe toujours et qu'elle est unique sauf si le classifieur f associe une probabilité nulle à $cat(c_i)$ [1].

Une autre idée pour améliorer les résultats de la calibration est de partir avec des probabilités pré-calibrées à la fois pour le classifieur à calibrer f et pour le classifieur de calibration f'. En effet, rien n'empêche dans ce cas d'utiliser les techniques de calibration existantes dans l'état-de-l'art pour pré-calibrer p et p' pour finalement réviser selon notre méthode *JCPC*.

Jusqu'à présent, nous avons brièvement présenté l'idée principale de notre approche de calibration *JCPC*. Afin de l'appliquer en pratique, il faut bien répondre à la question liée au regroupement des classes en catégories dans le cas où, pour le problème considéré, il n'existe pas de taxonomie ou de hiérarchie de classes qui puisse être utilisée directement. Cette question sera abordée dans la section suivante.

4 *JCPC* pour les problèmes sans hiérarchies de classes

Dans certains domaines, il existe des taxonomies et des hiérarchies de classes permettant de regrouper sémantiquement les classes en catégories mais cette option ne garantit pas nécessairement les meilleurs résultats. Le nombre et la composition des catégories est un des points clés pour avoir des probabilités bien calibrées. Pour les jeux de données sans taxonomie de classes, une solution serait de regrouper les classes en catégories en utilisant les techniques de l'état-de-l'art, mais cela donne souvent des clusters où les éléments de la même classe sont distribués dans différents clusters.

1. Si $p(cat(c_i))$=0, on peut soit appliquer *JCPC* en donnant plus de priorité au modèle des catégories $p_c(c_i)$=$\frac{p'(cat(c_i))}{|cat(c_i)|}$ où $|cat(c_i)|$ est le nombre de classes dans la catégorie $cat(c_i)$, soit conserver la probabilité initiale de la classe $p_c(c_i)$=$p(c_i)$.

Pour surmonter ce problème, nous suggérons de procéder différemment et proposons une procédure simple de *construction de la hiérarchie des classes*. Ainsi, au lieu d'appliquer des méthodes de clustering sur toutes les instances du jeu de données, nous calculons d'abord le centroïde (ou prototype) de chaque classe, puis nous utilisons des algorithmes de clustering agglomératifs (ascendants) (Müllner, 2011) pour les regrouper hiérarchiquement sous la forme d'un dendrogramme, le nombre optimal de catégories (ou clusters) est ensuite obtenu en utilisant une heuristique de l'état-de-l'art comme Elbow (Madhulatha, 2012).

5 Étude expérimentale

Les expériences évaluant notre approche peuvent être reproduites [2] et sont menées sur des jeux de données de classification de documents et d'images bien connus, à savoir : *DBPedia* [3], *Amazon products reviews (Amazon PR)* [4], *CIFAR-10* et *CIFAR-100* (Krizhevsky et Hinton, 2009).

Les expériences sont menées sur les jeux de données de classification de documents à l'aide de quatre classifieurs (ML), à savoir, classifieur Naïf Bayésien (NB), Forêts Aléatoires (RF), Regression Logistique (LR) et Machine à Vecteurs de Support (SVM). Pour les jeux de données de classification d'images, des réseaux neuronaux convolutifs (CNNs) très robustes ont été utilisés, à savoir ResNet-50 (He et al., 2016), DenseNet-121 (Huang et al., 2016), Inception-v3 (Szegedy et al., 2015), VGG-19 (Simonyan et Zisserman, 2014) et EfficientNet-B7 (Tan et Le, 2019). Ces derniers ont été pré-entraînés sur ImageNet (Deng et al., 2009).

Les mesures d'évaluation utilisées pour évaluer la précision et la calibration de la confiance prédite sont : La précision (ACC), la vraisemblance logarithmique négative (NLL), l'erreur de calibration attendue (ECE), et l'erreur de calibration maximale (MCE).

5.1 Résultats principaux

5.1.1 Impact de la hiérarchie des classes/catégorisation

Les résultats obtenus avec la méthode de construction de la hiérarchie que nous proposons sont compétitifs avec ceux obtenus en utilisant les catégories originales. Nous pouvons voir dans le tableau 1 que contrairement au clustering avec K-means, qui détériore la précision et l'ECE du modèle de classes, la construction de la hiérarchie en utilisant notre procédure améliore l'ECE. Bien qu'elle ne soit pas aussi bonne que la hiérarchie originale de CIFAR-100, l'amélioration est tout de même significative, surtout si l'on tient compte du fait que le modèle est très calibré au départ (avec une ECE de 1,75%). Ces résultats soulignent l'impact de la catégorisation des classes sur la qualité de la méthode de calibration proposée.

5.1.2 Résultats de l'application de *JCPC*

Le tableau 2 montre les résultats obtenus avec notre approche de calibration *JCPC* sur Amazon PR et DBPedia. Nous pouvons voir qu'à l'exception des RF sur Amazon PR, tous les

2. https://colab.research.google.com/drive/1nCn8WbtcWrYPmMXDF9zBZSSdU1XguO3V?usp=sharing
3. DBPedia dataset, https://www.kaggle.com/danofer/dbpedia-classes
4. Amazon PR, https://www.kaggle.com/kashnitsky/hierarchical-textclassification

Modèle		CIFAR-100			
		Acc%	NLL	ECE%	MCE%
ResNet50	Non cal	**74.83**	**0.85**	1.75	5.07
	K-means	68.07	1.09	2.45	8.81
	Construction de la hiérarchie	69.98	1.03	1.53	5.29
	Hiérarchie CIFAR-100	72.15	0.96	**1.28**	**4.21**

TAB. 1 – Comparaison des performances du ResNet50 avant et après la calibration *JCPC* sur CIFAR-100 en utilisant K-means, la construction de la hiérarchie et la hiérarchie CIFAR-100.

modèles testés montrent une amélioration de la qualité de la confiance, en maintenant ou en améliorant dans certains cas la précision initiale. Les résultats obtenus avec *JCPC* ont été améliorés davantage en utilisant *JCPC*-oracle avec des probabilités d'entrée p et $p^{'}$ pré-calibrées moyennant les techniques de l'état-de-l'art, à savoir la régression isotonique et la régression sigmoïde (Guo et al., 2017). Ce qui nous permet de surpasser leurs résultats.

Les résultats obtenus avec les CNNs sur CIFAR-10 et CIFAR-100 constituent un défi pour l'approche de calibration que nous proposons comme ils sont initialement assez bien calibrés. Les résultats présentés dans le tableau 3 montrent que la calibration en utilisant *JCPC* est un peu limitée avec les trois premiers CNNs appliqués à CIFAR-10. Les résultats obtenus avec les autres expériences sont très positifs sur les deux jeux de données, en particulier avec DenseNet-121 sur CIFAR-10. L'utilisation de l'oracle sur ces jeux de données a clairement surpassé la technique de l'état-de-l'art, température scaling (Guo et al., 2017), qui n'affecte toutefois pas la précision du modèle. Le VGG-19 sur CIFAR-10 est le seul modèle qui n'a été calibré ni par notre méthode ni par celle de l'état-de-l'art car il n'y a pratiquement aucune marge d'amélioration.

Modèle		Amazon PR				DBPedia			
		Acc%	NLL	ECE%	MCE%	Acc%	NLL	ECE%	MCE%
NB	Non cal	42.93	2.67	27.63	67.17	71.42	1.22	22.41	34.08
	JCPC	**53.01**	**2.05**	**26.49**	**53.33**	**71.71**	**1.14**	**20.51**	**32.09**
	Iso	68.19	1.50	13.03	24.46	**88.88**	**0.45**	12.13	**24.8**
	Sig	62.32	1.71	22.49	28.37	83.03	0.76	16.62	25.66
	JCPC-oracle	**70.12**	**1.38**	**9.40**	**20.67**	82.00	0.69	**8.53**	26.41
LR	Non cal	64.00	1.77	23.73	44.90	**92.30**	**0.36**	10.31	36.25
	JCPC	**67.44**	**1.42**	**18.25**	**32.79**	91.76	**0.35**	**09.23**	**29.74**
	Iso	67.32	1.47	11.51	19.33	91.76	0.41	15.99	33.26
	Sig	68.05	1.31	13.45	20.50	92.12	0.40	16.35	32.35
	JCPC-oracle	**70.27**	**1.13**	**9.10**	**18.76**	91.50	**0.36**	**8.38**	**24.76**
RF	Non cal	**67.19**	2.50	**05.16**	10.59	90.57	0.62	26.39	47.24
	JCPC	**67.39**	**2.40**	14.76	25.32	90.36	**0.60**	**25.66**	**43.03**
	Iso	64.39	5.16	21.14	46.88	**91.99**	0.39	2.71	13.30
	Sig	67.02	**1.73**	20.79	41.86	91.47	**0.32**	1.94	11.72
	JCPC-oracle	66.24	2.63	5.65	**9.12**	90.10	0.49	**1.41**	**8.46**
SVM	Non cal	62.40	1.65	16.85	40.86	83.77	0.87	21.16	54.12
	JCPC	**63.35**	**1.59**	**06.78**	**14.98**	81.01	0.90	**11.05**	**30.08**
	Iso	**65.89**	**1.52**	11.03	22.16	**85.49**	**0.64**	18.35	30.88
	Sig	24.59	2.65	**6.12**	33.11	33.67	2.55	**8.76**	40.95
	JCPC-oracle	63.35	1.59	6.78	**14.98**	81.01	0.90	11.05	**30.08**

TAB. 2 – Comparaison des performances des classifieurs avant et après calibration en utilisation *JCPC*, les méthodes de l'état-de-l'art et *JCPC*-oracle sur Amazon PR et DBPedia.

Modèle		CIFAR-10				CIFAR-100			
		Acc%	NLL	ECE%	MCE%	Acc%	NLL	ECE%	MCE%
ResNet-50	Non cal	92.07	0.24	1.50	8.64	**74.83**	**0.85**	1.75	5.07
	JCPC	**92.36**	0.24	1.94	12.78	72.15	0.96	**1.28**	**4.21**
	T-Scaling	92.07	0.23	1.21	10.23	74.83	0.85	2.21	5.91
	JCPC-oracle	**92.40**	**0.23**	**0.66**	**5.75**	72.15	0.96	**1.28**	**4.21**
Inception-v3	Non cal	**90.30**	0.32	2.78	9.21	**64.99**	1.3	6.75	13.10
	JCPC	90.09	0.33	2.79	10.12	63.21	**1.3**	**1.61**	**4.26**
	T-Scaling	90.30	0.32	0.74	9.19	64.99	1.3	2.07	7.40
	JCPC-oracle	90.02	**0.31**	**0.67**	**5.42**	63.21	**1.3**	**1.61**	**4.26**
VGG-19	Non cal	**88.86**	**0.32**	**0.56**	**4.11**	**67.66**	**1.14**	3.36	10.17
	JCPC	88.52	0.33	0.99	6.74	64.96	1.2	**2.59**	**7.79**
	T-Scaling	88.86	0.32	0.8	4.92	67.66	1.14	2.70	8.64
	JCPC-oracle	88.50	0.33	0.64	4.16	64.82	1.2	**2.00**	**7.28**
DenseNet-121	Non cal	93.37	0.2	1.61	11.22	**73.90**	**0.89**	3.10	8.68
	JCPC	**93.55**	**0.2**	**0.99**	**5.22**	69.21	1.0	**1.35**	**4.02**
	T-Scaling	93.37	0.2	0.51	**5.01**	73.90	0.89	1.55	5.26
	JCPC-oracle	**93.66**	**0.19**	**0.51**	9.77	71.52	0.98	**1.10**	**4.30**
EfficientNet-B7	Non cal	**95.86**	0.16	2.27	19.02	**77.15**	0.81	5.29	13.37
	JCPC	95.60	0.17	**2.09**	**14.09**	73.52	0.95	**4.34**	**7.41**
	T-Scaling	95.86	0.13	0.41	**8.67**	77.15	**0.79**	2.49	7.81
	JCPC-oracle	95.74	**0.13**	**0.33**	10.39	74.99	0.88	**1.36**	**7.05**

TAB. 3 – Comparaison des performances des classifieurs avant et après calibration en utilisation *JCPC*, les méthodes de l'état-de-l'art et *JCPC*-oracle sur CIFAR-10 et CIFAR-100.

6 Conclusions et discussions

Dans cet article, nous avons proposé une nouvelle approche pour calibrer les probabilités prédictives d'un classifieur par la révision d'informations incertaines et plus fiables, basée sur la règle de conditionnement de Jeffrey. En plus d'être simple, notre approche garantit le plus souvent une meilleure calibration que les méthodes de l'état-de-l'art.

L'un des problèmes les plus importants de l'approche proposée est la qualité du modèle de catégories qui dépend fortement de la qualité de la hiérarchie construite à partir du jeu de données et de l'utilisation de la pré-calibration. Pour les travaux futurs, l'une des pistes et questions ouvertes est de travailler sur la meilleure façon de construire la hiérarchie des classes spécifiquement pour améliorer la qualité de la calibration et d'effectuer une analyse approfondie des cas où nous sommes sûrs d'avoir une amélioration grâce à notre approche.

Remerciements. Ce travail a été soutenu par le projet Vivah 'Vers une Intelligence artificielle à VisAge Humain' soutenu par l'ANR, et par le projet ANR CROQUIS (Collecting, Representing, cOmpleting, merging, and Querying heterogeneous and UncertaIn waStewater and stormwater network data), subvention ANR-21-CE23-0004 de l'Agence Nationale de la Recherche (ANR).

Références

Chan, H. et A. Darwiche (2005). On the revision of probabilistic beliefs using uncertain evidence. *Artif. Intell. 163*(1), 67–90.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, et L. Fei-Fei (2009). ImageNet : A Large-Scale Hierarchical Image Database. In *CVPR09*.

Filho, T. S., H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, et P. Flach (2021). Classifier calibration : How to assess and improve predicted class probabilities : a survey. *CoRR abs/2112.10327*.

Guo, C., G. Pleiss, Y. Sun, et K. Q. Weinberger (2017). On calibration of modern neural networks. In D. Precup et Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning*, Volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR.

He, K., X. Zhang, S. Ren, et J. Sun (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

Huang, G., Z. Liu, L. van der Maaten, et K. Q. Weinberger (2016). Densely connected convolutional networks.

Jeffrey, R. C. (1965). *The Logic of Decision*. New York, NY, USA : University of Chicago Press.

Krizhevsky, A. et G. Hinton (2009). Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario.

Madhulatha, T. S. (2012). An overview on clustering methods.

Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms.

Simonyan, K. et A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, et Z. Wojna (2015). Rethinking the inception architecture for computer vision.

Tan, M. et Q. V. Le (2019). Efficientnet : Rethinking model scaling for convolutional neural networks.

Summary

In many critical applications, machine learning models must not only predict the class label accurately, but also provide the probability that the prediction is correct. This probability determines whether or not to trust the prediction. In this paper, we present a novel approach for calibrating the probabilities of machine learning models via a post-processing step. The starting point of this work is the observation that calibration is rather better on a small number of categories or subsets of classes than on a large number of classes. Our proposed calibration approach, named *JCPC*, is based on probabilistic belief update and calibrates the predicted probabilities on classes using the predicted probabilities on categories. Our experimental study on many datasets and machine learning models show very promising results.