Classification multi-label en flux adaptée à des ressources limitées

Hugo Peuzet*, Frank Meyer**, Pascale Kuntz ***

* Orange SA, LS2N
hugo.peuzet@ls2n.fr
** Orange SA
franck.meyer@orange.com
*** LS2N
pascale.kuntz@univ-nantes.fr

Résumé. La classification multi-label consiste à prédire pour une instance donnée un vecteur de labels à partir du vecteur d'attributs en entrée. De nouvelles applications suscitent aujourd'hui le développement de la classification multi-label en flux où les données ne sont plus connues dans leur intégralité au moment de la phase d'apprentissage mais sont découvertes au fil du temps. Cette communication présente une comparaison expérimentale de quatre stratégies basées sur des paradigmes différents (arbre de Hoeffding, forêt aléatoire adaptative, régression multi-cibles, réseau de neurones à une couche cachée avec mémoire) adaptées au cadre nouveau de la classification multi-label continue de données en flux sous contraintes de ressources.

1 Résumé étendu

La classification multi-label, qui consiste a prédire pour une instance donnée un vecteur de labels à partir du vecteur d'attributs en entrée, a connu un grand développement cette dernière décennie (Zheng et al. (2019)). L'explosion du volume des données disponibles dans différents domaines a conduit au développement de la classification multi-label extrême capable de traiter des données d'ordre jusqu'à $X \times y$ où $X = 10^6$ est l'ordre de grandeur du nombre d'attributs et $y = 10^6$ est l'ordre de grandeur du nombre de labels (Prabhu et Varma (2014)). Aujourd'hui, de nouvelles questions conduisent au développement de la classification multi-label en flux où les données ne sont plus connues dans leur intégralité au moment de la phase d'apprentissage mais sont découvertes au fil du temps. Citons par exemple la modération sur des réseaux sociaux produisant des flux de données qui peut entraîner des décisions rapides de gestion d'informations interdites et dangereuses et qui nécessite des algorithmes capables de s'adapter en temps réel aux changements dans des flux de données continus connaissant des évolutions rapides telles que l'apparition de nouveaux labels ou des changements de distributions statistiques.

Associé à ces données évolutives, une nouvelle problématique est en train d'émerger : l'apprentissage multi-label continu. Dans son cadre général, l'apprentissage continu (Van de Ven et Tolias (2019)) consiste à apprendre séquentiellement différentes tâches au cours du temps.

Une tâche est définie comme un ensemble de données appartenant à un domaine, un espace de sortie ou une distribution spécifique (Hu et al. (2022)). L'enjeu principal de l'apprentissage est de ne pas oublier complètement les connaissances liées aux tâches précédemment apprises lors de la prise en considération d'une nouvelle tâche.

De façon générale, les scénarii les plus courants de l'apprentissage continu restent basés sur un apprentissage par lot hors ligne : l'algorithme apprend séquentiellement des lots de données correspondant chacun à une tâche différente en effectuant de nombreuses itérations. Cependant, cette approche est de plus en plus critiquée par son manque de réalisme vis-à-vis des situations réelles (Prabhu et al. (2020), Verwimp et al. (2023)). Elle repose en particulier sur des hypthèses simplificatrices telles que notamment : (i) le partitionnement du flux de données en lot indépendants et identiquement distribués correspondant chacun à une tâche spécifique ; (ii) les intersections vides des sous-ensembles de labels associés à chaque tâche ; (iii) l'absence de restriction du budget de calcul ou de temps impactant la volumétrie des calculs effectués par le modèle. Pour palier ces limites, la stratégie d'apprentissage continu en ligne est en plein essor (OCL) (Cai et al. (2021)). Elle consiste à apprendre sur chaque lot de données courant en une seule passe. Une extension plus radicale qualifiée d'apprentissage en ligne en flux et en continu (OSCL) consiste à apprendre un seul exemple du lot à la fois (Gunasekara et al. (2023)).

Dans cette communication, nous étendons ce paradigme à la classification multi-label en flux en lui ajoutant un nouvel objectif de plus en plus considéré en intelligence artificielle : la capacité des modèles d'apprentissage à fonctionner avec des ressources de calcul et de stockage "raisonnables". Pour atteindre leur niveau de performances actuel les modèles n'ont cessé d'avoir recours à des ressources matérielles de plus en plus conséquentes. Cependant, des aspirations éthiques et écologiques commencent à plaider pour le développement d'une IA plus frugale (Evchenko et al. (2021)). La frugalité peut être intégrée à plusieurs niveaux de la chaîne de traitement : en entrée pour minimiser les coûts associés à l'acquisition des données et dans les phases d'apprentissage et d'inférence pour minimiser l'utilisation de ressources en mémoire et en calcul. Dans cette communication, nous nous focalisons sur ces deux phases pour élaborer des modèles qui puissent fonctionner dans un environnement matériel restreint à une machine "standard" sans recours à des accès aux données et ressources de calcul distribuées sur d'autres serveurs et calculateurs. Nous présentons ici un nouveau protocole expérimental permettant d'éprouver différentes stratégies de classification multi-label en flux sous contrainte volontaire de ressources. Nous l'éprouvons sur une comparaison de quatre stratégies basées sur des paradigmes différents : une approche basée sur un arbre de Hoeffding (Hulten et al. (2001)), une approche basée sur une forêt aléatoire adaptative (Gomes et al. (2017)), une régression multi-cibles (iSOUPtree) (Osojnik et al. (2018)), et une approche basée sur un réseau de neurones à une couche cachée couplé à une mémoire.

Références

Cai, Z., O. Sener, et V. Koltun (2021). Online continual learning with natural distribution shifts : An empirical study with visual data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8281–8290.

Evchenko, M., J. Vanschoren, H. H. Hoos, M. Schoenauer, et M. Sebag (2021). Frugal machine learning. *arXiv preprint arXiv :2111.03731*.

Gomes, H. M., A. Bifet, J. Read, J. P. Barddal, F. Enembreck, B. Pfharinger, G. Holmes, et T. Abdessalem (2017). Adaptive random forests for evolving data stream classification. *Machine Learning 106*, 1469–1495.

Gunasekara, N., B. Pfahringer, H. M. Gomes, et A. Bifet (2023). Survey on online streaming continual learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pp. 6628–6637. ijcai. org.

Hu, H., O. Sener, F. Sha, et V. Koltun (2022). Drinking from a firehose : Continual learning with web-scale natural language. *IEEE Transactions on Pattern Analysis and Machine Intelligence 45*(5), 5684–5696.

Hulten, G., L. Spencer, et P. Domingos (2001). Mining time-changing data streams. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 97–106.

Osojnik, A., P. Panov, et S. Džeroski (2018). Tree-based methods for online multi-target regression. *Journal of Intelligent Information Systems 50*, 315–339.

Prabhu, A., P. H. Torr, et P. K. Dokania (2020). Gdumb : A simple approach that questions our progress in continual learning. In *Computer Vision–ECCV 2020 : 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 524–540. Springer.

Prabhu, Y. et M. Varma (2014). Fastxml : A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 263–272.

Van de Ven, G. M. et A. S. Tolias (2019). Three scenarios for continual learning. *arXiv preprint arXiv :1904.07734*.

Verwimp, E., S. Ben-David, M. Bethge, A. Cossu, A. Gepperth, T. L. Hayes, E. Hüllermeier, C. Kanan, D. Kudithipudi, C. H. Lampert, et al. (2023). Continual learning : Applications and the road forward. *arXiv preprint arXiv :2311.11908*.

Zheng, X., P. Li, Z. Chu, et X. Hu (2019). A survey on multi-label data stream classification. *IEEE Access 8*, 1249–1275.

Summary

Multi-label classification aims to predict a label vector for a given instance from a feature vector as input. New applications now stimulate the development of multi-label data stream classification where the data are not known in their entirety during learning phase, but are discovered over time. This communication presents an experimental comparison between four strategies based on different paradigms (Hoeffding tree, adaptive random forest, multi-target regression, a neural networks with one hidden layer and a memory) adapted to the new framework of the continual multi-label data stream classification with resource constraints.