Formalisation de la classification multi-Labels en flux et son application en cybersécurité : une étude approfondie

X. WANG[1], F. Meyer[2], P. Kuntz[3]

[1] Centre de recherche et d'innovation de Talan
[2] Orange Labs
[3] Laboratoire des sciences du numérique de Nantes

5 mai 2023

Résumé

Un nombre croissant d'applications actuelles confronte les algorithmes de classification multi-labels à un défi majeur : celui d'apprendre des modèles avec des ressources de calcul et de stockage limitées et à partir de données en flux intégrant des changements de distribution de données au fil du temps. Dans cet article, nous proposons un tour d'horizon de cette problématique avec un attention particulière pour des applications dans le domaine de la cybersécurité.

Mots-clés

Classification Multi-Labels, Flux de données, Dérive Conceptuelle, Cybersécurité.

Abstract

Due to the ever-increasing number of current applications, multi-labels classification algorithms are facing a major challenge : their capacity for learning models from streaming data including changes in distributions over time, while constantly coming up against limited computational and storage resources. In this article, we offer an in-depth analysis of this problem which will serve as a basis for a proposal of algorithms for applications in the cybersecurity domain.

Keywords

Classification Multi-Labels, Data Stream, Concept Drift.

1 Introduction

La classification multi-labels a connu un fort développement cette dernière décennie stimulé par l'essor des besoins applicatifs qui nécessitent de prendre en compte des relations de dépendance entre les labels et qui ne peuvent donc plus s'appuyer sur les algorithmes éprouvés de la classification mono-label sous l'hypothèse d'indépendance. Par exemple, pour l'annotation de textes, de nombreuses labels coexistant peuvent être liés entre eux - un article sur l'actualité concernant une conférence sur le changement climatique peut être étiqueté à la fois par les labels « politique » et « environnement »[1]. En exploitant la corrélation entre ces labels, des travaux ont montré que la classification multi-labels améliore grandement la précision de la prédiction pour la tâche d'annotation des textes. Au cours des dernières décennies, de nombreuses applications (e.g. annotation de textes [2], d'images [3] ou de menaces d'intrusion dans les réseaux [4]) ont donc stimulé la recherche sur les techniques de classification multi-labels.

Cependant, les modèles classiques sont aujourd'hui confrontés à deux défis. Leur complexité computationnelle se heurte à l'augmentation du nombre de labels pouvant être pris en compte dans certaines applications. Les jeux de données peuvent en contenir des dizaines de milliers [9]. De plus, la majorité des modèles s'appuie sur une disponibilité de l'ensemble des données pour leur traitement. Or, ces données peuvent être produites de manière continu et en grande quantité. Les ressources de calcul et de stockage pouvant être limitées par des contraintes techniques et des coûts énergétiques [6], leur traitement en grand nombre et leur stockage permanent ne sont plus toujours possible. S'ajoute à ces évolutions la nécessité de prendre en compte dans les modèles les changements de distribution des données qui peuvent apparaître au cours du temps, ce qu'on appelle la dérive conceptuelle [8].

Ces nouveaux besoins motivent la recherche sur des techniques de classification multi-labels en flux (en anglais Multi-Label Stream Classification ou MLSC) comme en témoigne les études et les travaux récents menées dans ce domaine [12]. Dans cette revue de la littérature, nous avons constaté que la recherche sur ce problème n'a fait qu'effleurer le sujet. Dans cet article, nous proposons un aperçu via une formalisation détaillée du problème de classification multi-labels en flux.

La classification multi-labels en flux revêt une importance considérable dans le domaine de la cybersécurité, notamment dans la détection d'URL malveillantes. Les URL malveillantes concernent des contenus non sollicités tels que le spam et le phishing, attirant ainsi des utilisateurs peu méfiants qui deviennent victimes d'escroqueries, subissant des pertes financières, des vols d'informations privées et des installations de logiciels malveillants. Ces activités malveillantes engendrent chaque année des pertes financières s'élevant à plusieurs milliards de euros. Un système capable de détecter rapidement ces URL malveillantes et d'appliquer les modifications nécessaires peut jouer un rôle cru-

cial dans la lutte contre un large éventail de menaces en matière de cybersécurité. Cependant, actuellement, peu de modèles se concentrent spécifiquement sur ce problèmes. C'est pourquoi, nous proposons de porter une attention particulière sur ce sujet en fin d'article.

Dans la section suivante, nous définissons le problème en détaillant les caractéristiques et les contraintes liées à la MLSC. La dernière section clôture l'article en détaillant les travaux en cours traitant du problème de détection d'URLs malveillantes.

2 Définition du problème

Nous considérons la classification supervisée multi-label comme un problème de prédiction consistant, pour un exemple donné et une collection de labels considérés, à lui associer un sous-ensemble de labels le caractérisant. Chaque exemple est associé à au moins un label. Dans le contexte de l'apprentissage en flux, nous considérons un flux de données infini où chaque exemple arrive à une grande vitesse et selon une distribution de probabilité inconnue. La tâche de l'apprentissage en flux de données consiste à générer un nouveau modèle à chaque instant à partir du modèle précédent et du nouvel exemple avant l'arrivée du prochain exemple. Le nouveau modèle doit pouvoir donner une prédiction associée aux labels pour le prochain exemple.

2.1 Caractéristiques du problème

Dans ce qui suit, nous nous concentrerons d'abord sur les caractéristiques des données multi-labels, puis nous décrirons les caractéristiques des données en flux.

Les caractéristiques de données multi-labels. Trois caractéristiques sont présentes dans les données multi-labels : la volumétrie de l'espace des labels, la présence de corrélations entre labels et le déséquilibre dans la distribution des labels [10].

- **Espace de sortie à grande dimension.** Le nombre de labels prédéterminés est noté l. Un jeu de données multi-labels peut avoir 2^l combinaisons de labels possibles. Plus le nombre de labels augmente, plus le volume de l'espace de sortie croît exponentiellement et des données peuvent se retrouver « isolées » dans l'espace de recherche. Cela est problématique pour les méthodes qui nécessitent un nombre significatif de données pour l'apprentissage.
- **La présence de corrélations entre labels.** Les labels ne sont généralement pas indépendants les uns des autres ; ils sont corrélés et peuvent apparaître conjointement à des fréquences différentes. Par exemple, les articles de journaux sont plus susceptibles d'être associés à la fois aux catégories "science" et "environnement" qu'aux catégories "environnement" et "sport". Dans une base de données de films, les labels "famille" et "guerre" n'apparaîtront sans doute que très rarement ensemble.
- **Le déséquilibre dans la distribution des labels.** Des déséquilibres peuvent apparaître à deux niveaux : dans la distribution marginale des labels ou dans leurs distributions jointes. Par exemple pour un jeu d'images d'animaux domestiques, la combinaison de labels "chat" et "chien" apparaît plus souvent que la combinaison de labels, "chat" et "serpent" ; de même le label "chat" peut apparaître beaucoup plus souvent que le label "serpent".

Les caractéristiques de données en flux. Dans le contexte du problème en flux, outre les caractères soutenu et sans fin de la génération des données à traiter, une autre caractéristique majeure est **la dérive conceptuelle** : la distribution de données peut évoluer d'une manière imprévue sur le flux [11]. Ce phénomène peut se produire à la suite de changements dans l'environnement. Un exemple typique est la manière dont les périodes de croissance des cultures sont actuellement modifiées en réponse au changement climatique. Un autre exemple est celui en cybersécurité, où les contenus des URLs malveillantes changent constamment et deviennent plus difficiles à identifier à mesure que les techniques évoluent.

Les données étant multi-labels, il convient également de noter que la distribution des labels change en permanence. Par exemple, les principaux sujets d'actualité que les gens suivent changent d'un jour à l'autre : la fréquence du label "économie" augmente considérablement pendant la période d'inflation et la fréquence du label "féminisme" augmente considérablement autour du 8 mars. Non seulement la distribution marginale des labels peut changer, mais la distribution jointe des labels peut également changer. Par exemple, lorsque le climat change, la relation entre la politique et l'environnement est plus étroite que la relation entre la politique et l'économie.

2.2 Contraintes du problème

En réalité, la quantité de mémoire disponible pour le stockage des données et la capacité de calcul des algorithmes sont limitées par le matériel physique. Dans le même temps, les données qu'ils doivent traiter sont continuellement générées dans de nombreuses applications. A titre illustratif, chaque seconde, des milliers de requêtes URLs sont réalisées sur internet et 300 heures de vidéo par minute sont mises en ligne sur Youtube.

Par conséquent, afin de répondre aux exigences posées par les contraintes imposées par le MLSC, un algorithme doit non seulement être capable d'apprendre chaque donnée entrante en utilisant une quantité limitée de mémoire, mais en même temps son temps d'apprentissage doit être contrôlé dans un intervalle de temps très limité. De plus, les modèles doivent non seulement accumuler les informations au cours du temps lorsque la distribution de donnée est stable afin de prédire plus précisément, mais aussi s'adapter rapidement lorsque une dérive conceptuelle se produit.

3 Travaux antérieurs

La formalisation du problème MLSC permet de cerner les caractéristiques et les contraintes du problèmme de détection des URLs malveillantes dans le domaine de la cybersécurité.
Les URLs (*Uniform Resource Locator*) sont utilisées pour référencer des ressources sur Internet. Chaque URL possède une structure spécifique (e.g, le protocole, le domaine, le port, le chemin, etc). Les attaquants tentent souvent de modifier un ou plusieurs éléments de la structure de l'URL pour inciter les utilisateurs à accéder à des URL malveillantes. Ces URLs malveillantes redirigent les utilisateurs vers des ressources ou des pages où le pirate peut exécuter du code sur l'ordinateur de l'utilisateur, rediriger les utilisateurs vers des sites web indésirables, des sites web malveillants ou d'autres sites d'hameçonnage, ou télécharger des logiciels malveillants [13]. Ces URLs menant vers des sites web malveillants sont une menace courante et sérieuse pour la cybersécurité.
Ces attaques informatiques sont parfois de grande ampleur et se propagent très rapidement. Si un type d'attaque se prépare, plusieurs URLs et plusieurs sites peuvent être concernés. Des alertes peuvent être générées par des organismes de lutte contre les cyberattaques, mais ces alertes ne concernent qu'une portion des URLs malveillantes. L'enjeu est donc alors d'apprendre très rapidement les motifs discriminants associés à ces URLs, pour être capable de reconnaître et d'alerter dans les instants qui suivent toute URL potentiellement liée à une attaque en cours.
Pour ce type d'application, un mécanisme d'apprentissage en flux peut s'avérer être pertinent. En effet la capacité de réactivité et de généralisation en temps réel, à partir de quelques exemples d'URLs dangereuses, aux autres URLs dynamiquement générées correspondant à la même attaque est cruciale. Un système qui apprend d'une manière traditionnelle, chaque semaine ou même chaque nuit n'est pas assez réactif et peut passer à côté d'URLs malveillantes qui ont plusieurs heures ou plusieurs jours pour se propager puis être consultées par les clients. Par ailleurs la qualification des types de menaces, non exclusive, est importante ; certaines menaces sont corrélées et pourraient être mieux identifiées via une analyse multi-labels. D'autres menaces peuvent être moins importantes, ou correspondre à une « zone grise » où le blocage systématique doit être remplacé par un message indiquant aux utilisateurs de prendre des précautions.

4 Conclusion

Dans cet article, nous avons présenté une formalisation des caractéristiques et des contraintes de la classification multi-labels en flux. Nous avons mis en évidence la nécessité pour les modèles de prendre en compte la grande dimension de l'espace de sortie, la présence de corrélations entre les labels, le déséquilibre dans la distribution des labels, ainsi que l'adaptation des modèles aux changements de la distribution de données avec des ressources limitées. Cette analyse permet d'avoir une vision plus globale de l'enjeu du problème et contribue à orienter le développement de nouvelles approches pour aborder ce problème complexe. Enfin, nous avons complété l'article par une ouverture sur les applications spécifiques au domaine de la cybersécurité, et plus spécifiquement dans la détection d'URL malveillantes. En perspective, nous nous intéresserons plus spécifiquement à la représentation à utiliser pour structurer des URLs et aux modèles de classification multi-labels en flux pour répondre à cette problèmatique.

Références

[1] Lang, Ken. "Newsweeder : Learning to filter netnews." Machine learning proceedings 1995. Morgan Kaufmann, 1995. 331-339.

[2] Chalkidis, Ilias, et al. Large-scale multi-label text classification on EU legislation. arXiv preprint arXiv :1906.02192 (2019).

[3] Liu, Yang, et al. "SVM based multi-label learning with missing labels for image annotation." *Pattern Recognition* 78 (2018) : 307-317.

[4] Almusawi, A. and Amintoosi, H. (2018). DNS tunneling detection method based on multilabel support vector machine. *Security and Communication Networks*, 2018, 1-9.

[5] Tarekegn, A. N., et al. (2021). A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118, 107965.

[6] Domingos, Pedro, and Geoff Hulten. Mining high-speed data streams. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2000.

[7] Zheng, Xiulin, et al. A survey on multi-label data stream classification. *IEEE Access* 8 (2019) : 1249-1275.

[8] Gama, João, et al. A survey on concept drift adaptation. *ACM computing surveys (CSUR)* 46.4 (2014) : 1-37.

[9] Weston, Jason, Samy Bengio, and Nicolas Usunier. Wsabie : Scaling up to large vocabulary image annotation. (2011).

[10] Tsoumakas, Grigorios, and Ioannis Katakis. Multi-label classification : An overview. *International Journal of Data Warehousing and Mining (IJDWM)* 3.3 (2007) : 1-13.

[11] Lu, Jie, et al. Learning under concept drift : A review. *IEEE transactions on knowledge and data engineering* 31.12 (2018) : 2346-2363.

[12] Zheng, Xiulin, et al. "A survey on multi-label data stream classification." IEEE Access 8 (2019) : 1249-1275.

[13] Do Xuan, Cho, Hoa Dinh Nguyen, and Victor Nikolaevich Tisenko. "Malicious URL detection based on machine learning." International Journal of Advanced Computer Science and Applications 11.1 (2020).