COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université (2)

THÈSE DE DOCTORAT DE

NANTES UNIVERSITE (1)

ECOLE DOCTORALE N° 641
Mathématiques et Sciences et Technologies du numérique, de l'Information et de la Communication
Spécialité : *Informatique* (3)

Par

« Xihui WANG » (4)

« Classification Multi-Labels en flux » (5)

« Comparaisons d'approches et nouvelles propositions »

Thèse présentée et soutenue à « l'amphithéâtre du LS2N », le « 28-02-2023 » (6)
Unité de recherche : Laboratoire des Sciences du Numérique de Nantes (LS2N – équipe Duke) (7)

Rapporteurs avant soutenance :

Julien Velcin	Professeur des universités, Université de Lyon 2
Jean-Charles Lamirel	Maître de conférences HDR, Université de Strasbourg

Composition du Jury :
Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition du jury doit être revue pour s'assurer qu'elle est conforme et devra être répercutée sur la couverture de thèse

Président :	Prénom Nom	Fonction et établissement d'exercice (8)*(à préciser après la soutenance)*
Examinateurs :	Armelle Brun	Professeure des universités, Université de Lorraine–Nancy
	Mustapha Lebbah	Professeur des universités, Université Paris Saclay – Versailles
Dir. de thèse :	Pascale Kuntz	Professeure des universités, Université de Nantes

Invité(s)

Frank Meyer	Docteur Ingénieur R&D, Orange Labs Lannion

Titre : Classification Multi-Labels en flux : Comparaisons d'approches et nouvelles propositions.

Mots clés : classification multi-labels, flux de données, dérive conceptuelle

Résumé : Avec l'évolution conjointe des volumes de données à traiter et de la nature même de ces données, les algorithmes de classification multi-labels sont confrontés à un défi majeur : leur capacité à apprendre des modèles à partir de données en flux et à s'adapter aux changements de leurs distributions statistiques au fil du temps en prenant en compte des ressources matérielles limitées en stockage et en calcul. Dans cette thèse, nous abordons ce défi pour deux types de données : des flux stationnaires et non stationnaires. Pour la classification multi-labels de flux stationnaires nous avons développé un nouvel algorithme (MLT-ML) qui, avec une faible complexité temporelle, permet d'obtenir des performances en prédiction compétitives en exploitant les corrélations entre labels pour partitionner l'espace de recherche à chaque instant et réduire ainsi la complexité de l'apprentissage. Pour la classification de flux non-stationnaires nous avons développé successivement deux nouveaux algorithmes (ODM et A2ML) qui combinent une mémoire à court terme et une mémoire à long terme. Cette combinaison permet une adaptation efficace des modèles d'apprentissage aux dérives de concepts. En particulier, nous avons montré expérimentalement l'apport dans A2ML de l'introduction d'une règle d'échantillonnage biaisée pour la gestion de la mémoire à long terme ainsi que l'efficacité de la création de nouveaux clusters associés à l'apparition de nouveaux labels dans le flux. Pour combler l'absence de protocoles d'évaluation consensuels pour la classification multi-labels sur des données en flux, nous avons développé un nouveau cadre de simulation qui permet d'introduire explicitement des dérives de différents types et donc de mieux comprendre les changements de comportements des différentes stratégies de classification. Les comparaisons avec les meilleurs algorithmes de l'état de l'art menées sur des flux non stationnaires de plus de 50 000 exemples confirment le niveau élevé de performances de notre nouvel algorithme A2ML qui a une complexité temporelle significativement plus réduite que tous les autres.

Title : Multi-Labels Stream Classification : Comparisons of approaches and new proposals.

Keywords : multi-Labels classification, data stream, concept drift

Abstract : Due to the ever-increasing number of current applications, multi-label classification algorithms are facing a major challenge: their capacity for learning models from streaming data that include changes in distribution over time, while constantly coming up against limited computational and storage resources. In this thesis, we first study the multi-label classification problem on stationary streams and propose a new algorithm MLT-ML. This algorithm not only has a very low time complexity, but also has a high prediction performance by using the labels' correlation to partition the label space at each time. Then, we provide two new algorithms, ODM and A2ML, for non-stationary streams, which both combine a short-term memory with a long-term one. This combination ensures an efficient adaptation to the various types of concept drift. In particular, by using the biased reservoir sampling strategy and creating new clusters for new labels, A2ML can adapt to drift more effectively than ODM and its efficiency will not decrease over time. In addition, in order to further understand the behavior of the algorithm on the non-stationary stream, we also propose a new evaluation protocol to generate various types of concept drift. The experimentation confirmes A2ML's high levels of performance, and reveal computation times that are lower than those of the state of the art.

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université