Université du Luxembourg
snt@uni.lu - snt.uni.lu

Data augmentation for classification tasks with imbalanced data

Votre rôle

Les tâches de classification sont parmi les plus populaires en analyse de données. L'apprentissage supervisé est le plus souvent utilisé comme méthode pour determiner si une donnée appartient à une classe particulière. L'idée principale de cette approche est de produire une fonction basée sur des données d'apprentissage, afin de prédire à quelle classe appartient une donnée étudiée. Cela signifie que le succès de l'utilisation d'un algorithme de classification par apprentissage dépends largement de la sélection des données de la base d'apprentissage.

La plupart de ces algorithmes nécessitent un nombre comparable d'exemples pour chacune des classes, mais il est souvent impossible de créer des ensembles de données équilibrés à partir de données réelles (problème dit "Imbalanced data" en langue anglophone). En effet, on retrouve cette problématique pour des applications telles que le diagnostic médical, la prédiction de maladies rares mais importantes; la détection de fraudes dans les opérations bancaires ; la détection des intrusions dans les réseaux, la gestion des risques, la prévision des pannes d'équipements techniques et la détection de défaillances dans le domaine manufacturier.

Un tel déséquilibre des classes dans les données d'apprentissage amène les algorithmes de classification à surestimer les classes associées aux groupes majoritaires en raison de leurs proportions accrues. Par conséquent, les données appartenant aux groupes minoritaires sont mal classées.

Diverses solutions ont été proposées dans la littérature pour résoudre les problèmes liés au déséquilibre des classe. Notamment l'ajout de données synthétiques dans les classes minoritaires dans le but de rééquilibrer la distribution des classes.

L'objectif de ce stage est d'étudier les méthodes de génération de données synthétiques pour le rééquilibrage de données. Plus précisément, il est demandé d'étudier l'impact du rééquilibrage de données sur les performances des algorithmes de classification de données. L'étude se portera dans un premier temps sur des données séquentielles (séries temporelles) puis si possible sur des images.

Ce stage pourra se poursuivre par une thèse, sous réserve de validation interne.

Université du Luxembourg
snt@uni.lu - snt.uni.lu

Références :

[1] Haibo He et E. Garcia, “Learning from Imbalanced Data”, IEEE Transactions on Knowledge and Data Engineering, t. 21, no 9, p. 1263-1284, sept. 2009
[2] J. M. Johnson et T. M. Khoshgoftaar, “Survey on deep learning with class imbalance”, Journal of Big Data, t. 6, no 1, p. 27, déc. 2019.
[3] B. K. Iwana et S. Uchida, “An Empirical Survey of Data Augmentation for Time Series Classification with Neural Networks”, 2020

Vous serez amené à effectuer les tâches suivantes :

- Passer en revue la littérature scientifique concernant la classification de données déséquilibrées et le rééquilibrage
- Implémenter divers algorithmes de classification et de rééquilibrage
- Mettre en œuvre une stratégie expérimentale
- Analyser les résultats de vos expériences
- Communiquer avec des doctorants et chercheurs de l’université pour rendre compte de vos résultats

L’équipe dans laquelle vous travaillerez

- Dorian Joubaud
- Dr. Sylvain Kubler: Superviseur
- Prof. Yves Le Traon: Directeur de l’équipe de recherche SerVal

Votre profil

- Étudiant.e Bac +5 en école d’informatique, université ou école d’ingénieur avec une formation en statistiques, machine learning et analyse de données
- Disposant d’une 1ère expérience sur des projets de machine learning et/ou deep learning
- Bonne maîtrise des langages de programmation (Python / R) ainsi que des frameworks en traitement de données (Pandas, Numpy), en visualisation de données (Matplotlib, Seaborn) et en Machine Learning/Deep Learning (Scikit-learn, Tensorflow, Keras, Torch)
- Curieux.euse, agile et possédant de bonnes capacités d’analyse et de synthèse
- Doté.e d’un bon relationnel et d’un fort esprit d’équipe
- Vous disposez d’un niveau de compréhension et d’expression en anglais vous permettant de communiquer avec des doctorants et chercheurs venu du monde entier.

Université du Luxembourg
snt@uni.lu - snt.uni.lu

Ce qui vous attend au SnT...

Des infrastructures passionnantes et des laboratoires uniques. Sur les deux campus du SnT, nos chercheurs peuvent se promener sur la lune au LunaLab, construire un nanosatellite ou contribuer à améliorer les véhicules autonomes. Les chercheurs du SnT s'engagent dans des projets axés sur la demande. Grâce à notre programme de partenariat, nous travaillons sur des projets avec plus de 45 partenaires industriels.

Faites partie d'une famille multiculturelle. Au SnT, nous comptons plus de 60 nationalités. Tout au long de l'année, nous organisons des événements de renforcement de l'esprit d'équipe, des activités de mise en réseau, etc.

En résumé

- Type de contrat : Stage 4 à 6 mois
- Début du stage : Selon disponibilité
- Temps de travail: Plein temps 40.0 heures par semaines
- Location: Luxembourg
- Gratification : ~ 1200€ mensuel

Comment postuler ?

Les candidatures doivent comprendre :

- CV
- Lettre de motivation

Merci d'envoyer ces documents à :

- sylvain.kubler@uni.lu
- dorian.joubaud@uni.lu

À propos de l'université du Luxembourg...

L'Iniversité du Luxembourg cherche à recruter des chercheurs au SnT (Interdisciplinary Centre for Security, Reliability and Trust).

Le SnT mène des recherches interdisciplinaires sur les systèmes et services ICT (Information and Communication Technologies) sûrs, fiables et dignes de confiance, souvent en collaboration avec des partenaires industriels, gouvernementaux ou internationaux. Le SnT est actif dans plusieurs projets de recherche internationaux financés par le programme Horizon2020 et l'Agence spatiale européenne. Pour plus d'informations, vous pouvez consulter : https://wwwfr.uni.lu/snt