LIMOS - Axe SIC

Directeur de thèse : Violaine Antoine (MCF HDR), violaine.antoine@uca.fr
Co-encadrant : Sébastien SALVA (PU), sebastien.salva@uca.fr

Titre du sujet de thèse
Vers une meilleure évaluation des modèles de machine learning

Résumé du sujet de thèse :
Il existe de nombreux algorithmes de machine learning pour lesquelles leurs avantages et inconvénients sont globalement connus. Néanmoins, il reste difficile de savoir quel algorithme utiliser pour un jeu de données spécifique, car les guides d'utilisation des algorithmes incluent seulement les caractéristiques les plus simples des jeux de données (nombre d'attributs, nombre de classes, etc.) [A]. Or les caractéristiques d'un jeu de données sont bien plus complexes et ont une influence non négligeable sur la performance des algorithmes. Cette thèse vise à mieux renseigner les caractéristiques pour lesquels un algorithme de machine learning aurait de bonnes/mauvaise performances. Pour cela, une étude des caractéristiques des données doit être réalisée. Cette étude comprend la définition de métriques de caractérisation, le calcul d'indicateurs numériques, la comparaison de ces indicateurs avec les performances de modèles et la sélection des plus pertinents. Nous étudierons également la génération de données dont les caractéristiques seront précisées. Et nous proposerons d'étudier la robustesse de ces caractéristiques. Nous nous focaliserons sur les algorithmes de classification non supervisés, au moins dans un premier temps.

Sujet :
De nombreux algorithmes de classification (supervisés ou non) sont proposés chaque année afin de répondre à un ou plusieurs besoins particuliers tels que la gestion de données complexes et hétérogènes [1], la qualité des données [2], la prise en compte en sortie d'incertitudes [3,4], etc. Néanmoins, peu de travaux s'attardent sur l'efficacité de ces algorithmes en fonction des propriétés intrinsèques des jeux de données, e.g. la complétude, la robustesse, le chevauchement de classes. Ceci est surtout dû au fait qu'il n'existe pas de consensus sur les moyens ou propriétés permettant de caractériser des jeux de données. Par conséquent, peu de métriques et aucune métrique globale ne permet de comparer des jeux de données de même type.
La thèse se penche sur cette problématique et propose les objectifs suivants:
1) proposer des formules et définir des métriques pour caractériser des jeux de données. Les avantages à mesurer ces caractéristiques sont nombreux. Par exemple, elles peuvent aider un chercheur à mieux présenter les avantages et les inconvénients des nouveaux algorithmes de classification non supervisée qu'il propose. Cette étude pourra être abordée en considérant les diverses métriques de performance (Silhouette score, Rand index, etc.);
2) associer un type de jeu de données défini par ses caractéristiques à un ensemble d'algorithmes de classification. La finalité serait de créer des guides d'utilisation qui permettent, dans le cadre d'un problème appliqué avec un jeu de données spécifique, de choisir le ou les algorithmes les plus intéressants ;
3) proposer des méthodes pour générer de nouveaux jeux de selon les propriétés qu'un utilisateur cherche à vérifier. Comme certains jeux de données peuvent ne pas être suffisants pour analyser certaines caractéristiques, nous proposons également de créer une application qui génère des données synthétiques afin de combler le manque de jeux avec certaines caractéristiques;
4) étudier la robustesse de ces caractéristiques. Les caractéristiques du jeu de données et la performance des modèles sont-elles impactées si un petit pourcentage d'erreur est ajouté ou bien si des données potentiellement incorrectes (ex : outliers) sont supprimées du jeu.

Le sujet de thèse comprend cinq étapes :
1. l'étude exhaustive des caractéristiques d'importance pour la classification de données numériques. Cela comprend leur listing et leur catégorisation, le choix de mesures permettant de quantifier ces caractéristiques, et une réflexion sur la ou les notions de similarité à considérer pour comparer la proximité entre deux jeux de données.

2. La définition de métriques permettant de quantifier ces caractéristiques indépendamment du jeu de données. Ces métriques permettront de formaliser les notions de similarité, de comparer la proximité entre deux jeux de données. Ces métriques pourront incorporer des mesures de performance d'algorithmes ou modèles dont le comportement est connu (tels que k-means, dbscan et la classification hiérarchique). Puis la définition d'une métrique globale qui fédère l'ensemble des métriques précédentes et permet de comparer divers jeux de données pourra être donnée ;
3. L'implémentation d'une application permettant, selon une demande de caractéristique (ou de combinaison de caractéristiques) d'un utilisateur, de créer plusieurs jeux de données synthétiques). Bien que de nombreux jeux de données synthétiques existent déjà, certaines propriétés sont difficiles à générer rapidement et/ou efficacement. Par exemple, il est complexe d'augmenter le nombre d'attributs d'un jeu de données tout en conservant un taux spécifique de chevauchement entre clusters et une répartition équitable de l'importance des attributs.
4. L'étude de la performance de plusieurs algorithmes de classification à l'aide des caractérisations de plusieurs jeux de données. Les algorithmes de classification pourraient ainsi être classifiés selon des caractéristiques.
5. Enfin, la finalité est de mettre en correspondance l'algorithme le plus efficace pour un ou des jeux de données lié à une problématique spécifique.

Publications /références associées au sujet :
Webographie
[A] https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html
Bibliographie
[1] N. Wagner, V. Antoine, M-M. Mialon, R. Lardy, M. Silberberg, J. Koko, I. Veissier. Machine Learning to detect behavioural anomalies in dairy cows under subacute ruminal acidosis. Computers and Electronics in Agriculture. 170(105233), 2020.
[2] Roxane Jouseau, Sébastien Salva, and Chafik Samir, On Studying the Effect of Data Quality on Classification Performance, 23rd International Conference on Intelligent Data Engineering and Automated Learning (IDEAL), 24-26 November 2022, Manchester, UK
[3] V. Antoine, J. Guerrero, G. Romero. Possibilistic fuzzy c-means with partial supervision. Fuzzy Sets and Systems, Vol 449, pp 162-186, 2022.
[4] V. Antoine, J. Guerrero, J. Xie. Fast semi-supervised evidential clustering. International Journal of Approximate Reasonning, Vol. 133, pp 116-132, 2021.
[5] N. Kerckhove, N. Delage, S. Cambier, N. Cantagrel, E. Serra, F. Marcaillou, C. Maindet-Dominici, P. Picard, G. Martiné, R. Deleens, A-P Trouvin, L. Fourel, G. Espagne-Dubreuilh, L. Douay, S. Foulon, B. Dufraisse, C. Gov, E. Viel, F. Jedryka, S. Pouplin, C. Lestrade, E. Combe, S. Perrot, D. Perocheau, V. De Brisson, P. Vergne-Salle, P. Mertens, B. Pereira, A. Djiberou Mahamadou, V. Antoine, A. Corteval, A. Eschalier, C. Dualé, N. Attal, N. Authier. eDOL, a new mHealth application and web platform for the self-monitoring and the medical follow-up of chronic pain patients: feasibility study. Journal of Medical Internet Research, 2022