LIMOS - Axe SIC

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

Title of PhD subject
Towards Better Evaluation of Machine Learning Models.

Summary :
There are many machine learning algorithms whose advantages and disadvantages are generally well known. However, it remains difficult to determine which algorithm should be used for a specific dataset, since algorithm usage guidelines typically include only the simplest dataset characteristics (number of attributes, number of classes, etc.) [A]. In reality, the characteristics of a dataset are far more complex and have a non-negligible influence on algorithm performance.
This thesis aims to better identify the dataset characteristics under which a machine learning algorithm would achieve good or poor performance. To achieve this goal, a study of data characteristics must be conducted. This study includes the definition of characterisation metrics, the computation of numerical indicators, the comparison of these indicators with model performance, and the selection of the most relevant ones.
We will also study the generation of datasets with specified characteristics and propose to analyse the robustness of these characteristics. We will focus on unsupervised classification algorithms, at least in the initial phase of the work.

Topic of the thesis:
Many classification algorithms (supervised or unsupervised) are proposed each year in order to address one or several specific needs such as the management of complex and heterogeneous data [1], data quality issues [2], or the incorporation of uncertainty in the output [3,4], among others. However, few studies focus on the effectiveness of these algorithms with respect to the intrinsic properties of datasets, such as completeness, robustness, or class overlap. This is mainly due to the fact that there is no consensus on the methods or properties that can be used to characterize datasets. Consequently, few metrics exist—and no global metric currently allows datasets of the same type to be compared.

This PhD research addresses this issue and proposes the following objectives:

1. Propose formulas and define metrics to characterise datasets. Measuring these characteristics offers several advantages. For instance, they could help researchers better present the strengths and weaknesses of the new unsupervised classification algorithms they propose. This study could be approached by considering various performance metrics (Silhouette score, Rand index, etc.).
2. Associate a type of dataset, defined by its characteristics, with a set of classification algorithms. The ultimate goal would be to create usage guidelines that would allow practitioners, when dealing with a specific dataset in an applied problem, to select the most appropriate algorithm(s).
3. Propose methods for generating new datasets according to the properties that a user wishes to analyse. Since some datasets may not be sufficient to study certain characteristics, we also propose to create an application capable of generating synthetic data in order to compensate for the lack of datasets exhibiting specific properties.
4. Study the robustness of these characteristics.

We will examine whether dataset characteristics and model performance are affected when a small percentage of noise or errors is introduced, or when potentially incorrect data (e.g., outliers) are removed from the dataset.

The PhD project will include five stages

1. A comprehensive study of the characteristics that are important for the classification of numerical data. This includes identifying and categorising these characteristics, selecting measures that allow them to be quantified, and reflecting on the notion(s) of similarity to be considered when comparing the proximity between two datasets.
2. The definition of metrics allowing these characteristics to be quantified independently of the dataset.

These metrics will formalise notions of similarity and allow the comparison of proximity between two datasets. They may incorporate performance measures of algorithms or models with well-known behavior (such as k-means, DBSCAN, and hierarchical clustering). A global metric combining all previously defined metrics to compare various datasets may then be proposed.

3. The implementation of an application that allows users to generate multiple synthetic datasets based on requested characteristics (or combinations of characteristics). Although many synthetic datasets already exist, some properties remain difficult to generate quickly and efficiently. For instance, increasing the number of attributes in a dataset while maintaining a specific level of cluster overlap and an equal distribution of attribute importance is challenging.
4. The study of the performance of several classification algorithms using the characterization of multiple datasets. Classification algorithms could thus be categorized according to dataset characteristics.
5. Finally, the ultimate objective is to match the most effective algorithm to one or several datasets associated with a specific application problem.

Publications / References Related to the Topic

[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