LIMOS - Axe SIC

Thesis supervisor: Violaine Antoine (MCF), violaine.antoine@uca.fr
Co-supervisor : Sébastien Salva (PR), sebastien.salva@uca.fr

Towards a better evaluation of Machine Learning models

Summary :
There exist many machine learning algorithms, whose advantages and disadvantages are generally well known. However, it remains difficult to determine which algorithm to use for a specific dataset, as algorithm guidelines only include the most basic dataset characteristics (such as the number of attributes, number of classes, etc.) [A]. Yet, dataset characteristics are far more complex and have a significant impact on algorithm performance. This thesis aims to better identify the characteristics for which a machine learning algorithm would perform well or poorly. To achieve this, a study of data characteristics must be conducted. This study includes defining characterization metrics, computing numerical indicators, comparing these indicators with model performance, and selecting the most relevant ones. We will also explore the generation of data with specified characteristics and propose an analysis of the robustness of these characteristics. Our focus will be first on unsupervised classification algorithms.

Subject :
Numerous classification algorithms (supervised or unsupervised) are proposed each year to address one or more specific needs, such as handling complex and heterogeneous data [1], ensuring data quality [2], and incorporating output uncertainties [3,4], among others. However, few studies focus on the effectiveness of these algorithms based on the intrinsic properties of datasets, such as completeness, robustness, and class overlap. This is mainly due to the lack of consensus on the methods or properties used to characterize datasets. Consequently, only a few metrics exist, and no comprehensive metric allows for the comparison of datasets of the same type.

The thesis focuses on this problematic and proposes the following objectives:

1) propose formulas and define metrics to characterize datasets. Measuring these characteristics offers numerous advantages. For example, they can help researchers better present the strengths and weaknesses of new unsupervised classification algorithms they develop. This study can be approached by considering various performance metrics (such as the Silhouette score, Rand index, etc.).

2) Associate a type of dataset, defined by its characteristics, with a set of classification algorithms. The goal is to create guidelines that, in the context of an applied problem with a specific dataset, help to select the most suitable algorithms.

3) Propose methods to generate new datasets based on the properties that a user wishes to analyze. Since some datasets may be insufficient for studying certain characteristics, we also propose to develop an application that generates synthetic data to address the lack of datasets with specific properties.

4) Study the robustness of these characteristics. Specifically, analyze whether the dataset characteristics and model performance are affected when a small percentage of errors is introduced or when potentially incorrect data (e.g., outliers) are removed from the dataset.

The thesis subject includes five steps:

1. A comprehensive study of key characteristics for numerical data classification. This includes listing and categorizing these characteristics, selecting appropriate measures to quantify them, and considering different notions of similarity to compare the proximity between datasets.
2. The definition of metrics to quantify these characteristics independently of the dataset. These metrics will help formalize similarity concepts and compare dataset proximity. They may incorporate performance

measures of well-known algorithms or models (such as k-means, DBSCAN, and hierarchical clustering). Additionally, a global metric integrating all previous metrics will be defined to facilitate dataset comparison.

3. The implementation of an application that generates multiple synthetic datasets based on user-specified characteristics (or combinations of characteristics). While many synthetic datasets already exist, some properties are challenging to generate quickly and efficiently. For instance, increasing the number of attributes in a dataset while maintaining a specific cluster overlap rate and ensuring a balanced distribution of attribute importance is a complex task.

4. The analysis of the performance of various classification algorithms using characterizations from multiple datasets. This will allow classification algorithms to be categorized based on dataset characteristics.

5. Finally, the goal is to identify and match the most effective algorithms to one or more datasets associated with a specific problem.

Publications /references associated to the subject:

Webography

[A] https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Bibliography

[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