The Journal of Open Source Software

ClusterValidityIndices.jl: Batch and Incremental Metrics for Unsupervised Learning

Sasha Petrenko[1] **and Donald C. Wunsch II**[1]

1 Missouri University of Science and Technology

DOI: 10.21105/joss.03527

Software

- Review
- Repository
- Archive

Editor: Adi Singh

Reviewers:

- @rMassimiliano
- @malmaud
- @malmaud
- @rMassimiliano

Submitted: 28 May 2021
Published: 21 November 2022

Summary

ClusterValidityIndices.jl is a Julia package for evaluating the performance of clustering algorithms without the aid of supervised labels. Cluster Validity Indices (CVI) provide a metric of the over- or under-partitioning of an arbitrary clustering algorithm with only the original data and labels assigned by the clustering algorithm. Furthermore, there exist formulations of every CVI such that they may run incrementally (i.e. Incremental CVIs, or ICVI), streaming alongside the clustering algorithm and producing the same results as in their batch implementations. Using a standard interface, each CVI in this package can be run with any clustering algorithm to produce a metric of that algorithm's performance in scenarios where explicit supervised labels do not exist, which is extremely useful in real-world applications where that is often the case.

Statement of need

CVIs are useful as one of the only methods of determining the performance of a clustering algorithm in the absence of explicit labels (Arbelaitz et al., 2013). Furthermore, ICVIs can measure the performance of clustering algorithms as they are running in a computationally tractable manner, which is incredibly useful in a variety of streaming clustering applications (Brito Da Silva et al., 2020).

There exist many CVIs in the literature, and their algorithmic and programmatic requirements are often very similar. Despite their utility in machine learning applications, however, there does not exist to date a unified repository of their implementations in Julia. Furthermore, new incremental variations of these algorithms are regularly developed in the literature without the ability to update the original implementations. The purpose of this package is to create a unified framework and repository of CVIs so as to fill the gap left by most metrics in this machine learning problem subset.

Target Audience

This package is principally intended as a resource for researchers in machine learning, clustering, and the research fields that utilize these tools to other ends, such as the statistical analysis of time series. However, implementing these algorithms in the Julia programming language brings all of the benefits of Julia itself, such as the computational speed comparable to that of being implemented in a low-level language such as C while having the syntactic transparency of a high-level language such as MATLAB or Python. Being implemented in Julia allows this package to be understood and expanded upon by research scientists while also being able to be used in resource-demanding production environments.

Comparison to Existing Implementations

Many implementations of various CVIs exist as a result of reproducibility efforts in the literature for use by the machine learning research community. However, each of these implementations vary in degrees of programming accessibility, language, number of implemented CVIs, and incremental variant implementations. Due to gaps left by the incomplete feature overlap of existing CVI packages, the objective of this package is to make many batch and incremental CVIs available in a fast, free, open-source, and documented repository.

Researchers associated with the CRIStAL Laboratory of the University of Lille, France have implemented a CVI package in the MATLAB programming language for the reproducibility of a survey of CVIs (José-García, 2021; José-García & Gómez-Flores, 2021). This repository contains a large variety of CVIs from the literature, though they are lacking their incremental variants.

A variety of CVIs are also implemented in a package for time series clustering in the R statistical computing language (Sarda-Espinosa, 2022). However, these CVIs are implemented as a submodule of the R package rather than being standalone, impeding their visibility, and incremental variants are also not implemented.

Although each of these CVI projects and the very many and disparate implementations of individual CVIs in the literature combined may implement the majority of CVI algorithms relevant to modern research and engineering, together they lack cohesion in programming language and usage. Furthermore, though software implementations in the MATLAB programming language are syntactically accessible, this is to the detriment of those in research and industry without a private MATLAB license. In contrast, the R programming language is open and extremely powerful for research scientists and statisticians yet generally less suited for production environments.

The Julia programming langauge is selected for this open-source CVI package implementation due to its high-level syntactic ease of use and speed of development without comprimising computational efficiency for production environments. Furthermore, the objective of this package is to contain both batch and incremental implementations of many CVIs with clear documentation and simple usage for research scientists and engineers alike.

Cluster Validity Indices

Cluster Validity Indices (CVIs) are designed to tackle the problem of creating a metric of performance for unsupervised algorithms where the true answer is unknown. Clustering is a ubiquitous unsupervised learning paradigm, so the terminology and development of CVIs principally target clustering algorithms.

Because the clustering problem statement means that one does not have true labels to measure how well or poorly these algorithms perform, the most that can be done is to create a metric of the validity of the solution. This translates to how much an algorithm over- or under-partitions the data (i.e., how eager or reticent it is to create new categories) and quantifies how the algorithm structures its solution in its compactness (i.e., the density of the prescribed cluster regions) and its connectedness (i.e., how much disparate points in a cluster can be said to still belong to the same category).

In general, CVIs take a set of samples and the labels prescribed to them by a clustering algorithm, and they return a criterion value that is generally a positive real number. This criterion value often does not have an upper bound, varies greatly in behavior between CVIs, and changes as new samples are labeled and the CVI is reprocessed. In fact, it is often the trendlines of these values that provide the most information about the clustering process rather than the values themselves.

CVIs are originally derived to work on batches of samples and labels. However, there exist incremental variants that are proven to be mathematically equivalent to their batch counterparts (Brito Da Silva et al., 2020). These incremental CVIs (ICVIs) mitigate the computational overhead of computing these metrics online, such as in a streaming clustering scenarios.

Acknowledgements

This package is developed and maintained with sponsorship by the Applied Computational Intelligence Laboratory (ACIL) of the Missouri University of Science and Technology. This project is supported by grants from the Army Research Labs Night Vision Electronic Sensors Directorate (NVESD), the DARPA Lifelong Learning Machines (L2M) program under Contract No. FA8650-18-C-7831, Teledyne Technologies, and the National Science Foundation. The material, findings, and conclusions here do not necessarily reflect the views of these entities.

This research was also sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-22-2-0209. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, *46*(1), 243–256. https://doi.org/10.1016/j.patcog.2012.07.021

Brito Da Silva, L. E., Melton, N. M., & Wunsch, D. C. (2020). Incremental cluster validity indices for online learning of hard partitions: Extensions and comparative study. *IEEE Access*, *8*(i), 22025–22047. https://doi.org/10.1109/ACCESS.2020.2969849

José-García, A. (2021). *Cluster validity index toolbox*. https://github.com/adanjoga/cvik-toolbox

José-García, A., & Gómez-Flores, W. (2021). A survey of cluster validity indices for automatic data clustering using differential evolution. *Proceedings of the Genetic and Evolutionary Computation Conference*, 314–322. https://doi.org/10.1145/3449639.3459341

Sarda-Espinosa, A. (2022). *Time series clustering along with optimizations for the dynamic time warping distance*. https://cran.r-project.org/package=dtwclust