CSAI: New Cluster Validation Index based on Stability Analysis

Adane Tarekegn, Bjørnar Tessem, Fazle Rabbi

Media Futures

Introduction

The validation of clustering solutions is a major challenge due to the absence of ground truth in many real-world applications and datasets. In this study, we introduce a novel Clustering Stability Assessment Index (CSAI) that offers a unified and quantitative approach to measuring the quality and consistency of clustering solutions. It employs a data resampling approach, where the training dataset is grouped into several partitions for cluster construction, from which the prediction of cluster memberships for test data points can be made. In this process, clustering stability is determined by assessing the correlation or similarity between the features of the training and testing data points within each cluster. While CSAI was initially motivated by the need to assess cluster quality in news event detection tasks, it is also applicable across various domains and data modalities.

CSAI is formally defined by the following equation. Let the dataset $\boldsymbol{X}$ consist of $\boldsymbol{p}$ data points, which are randomly divided into $\boldsymbol{m}$ subsets or partitions. We apply the same clustering method $\boldsymbol{A}$ to each partition, yielding $\boldsymbol{n}$ clusters for each.

$$\mathrm{CSAI}(\mathrm{X_i}, \mathrm{A}) = \frac{1}{\mathrm{m}} \sum_{i=1}^{m} \left(\frac{1}{w.n} \sum_{j=1}^{n} \sqrt{\frac{1}{s} \sum_{k=1}^{s} \left(\hat{x}_{ijk} - \hat{v}_{ijk} \right)^2} \right)$$

- $\hat{x}_{ijk}$ **and** $\hat{v}_{ijk}$: k_th feature obtained in the j_th cluster of i_th partition in the training and testing data respectively; m represent number of partitions in the dataset.
- **n** and **s** denote number of clusters and number of features associated to clusters, respectively.
- **w** is the weighting factor used to normalize the scores across different partitions. It is computed as:

$$\mathbf{w} = \max_{i=1}^{\mathrm{m}} \left(\max_{j=1}^{\mathrm{n}} \left(\max_{k=1}^{\mathrm{s}} \hat{x}_{ijk} - \min_{k=1}^{\mathrm{s}} \hat{x}_{ijk} \right) \right)$$

Main Contributions

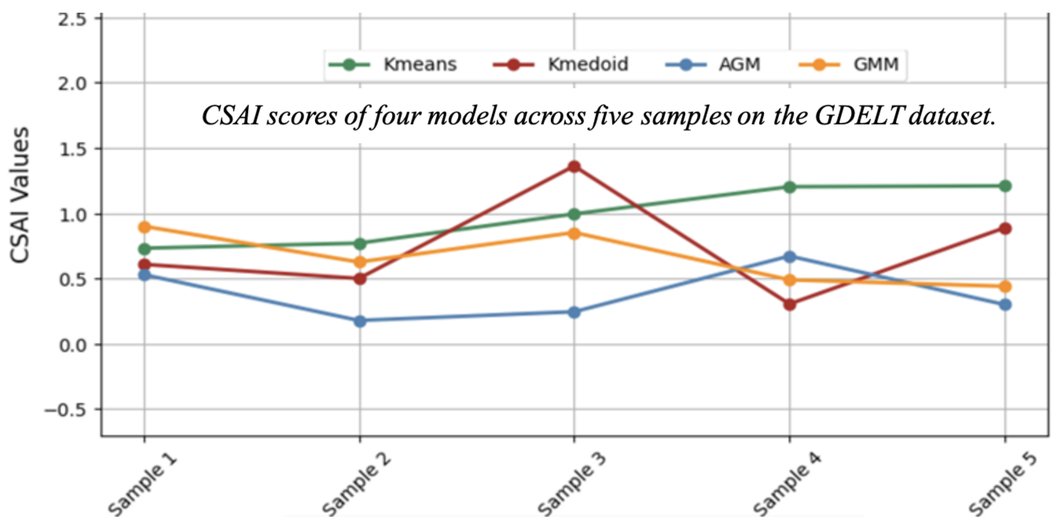
1. We propose CSAI as new approach to assess both validity and stability of the clustering solutions based on information from multiple features associated with clusters
2. We use publicly available benchmark textual datasets to evaluate the proposed method. Each dataset includes multiple partitions for cluster construction and a separate validation set to assess result quality.
3. The effectiveness of the CSAI index is compared against the widely used clustering validation indices using various clustering algorithms.
4. CSAI leverages the structure of features within clusters rather than centroids and uses the mean squared distance as a proximity measure.

Methods

Dataset: Three publicly available datasets: 20 newsgroups, MN-DS and arXiv datasets, and one curated news dataset from the Global Database of Events, Language, and Tone (GDELT), were used to evaluate the effectiveness of CSAI index.

Preprocessing: Basic pre-processing tasks were performed, such as normalization, and noise reduction. keywords were generated from the pre-processed dataset using KeyBERT, which leverages BERT embeddings to produce high-level descriptions or representations of the document.

Algorithms: Four different clustering algorithms were used, including, k-means and k-medoids (partitioning-based clustering), agglomerative clustering (hierarchical), and Gaussian mixture model (model-based clustering).



CSAI Scores of clustering algorithms with different embeddings

Algorithms	TF-IDF	GloVe	BERT	LLM
K-means clustering	0.8869	0.6970	0.4330	0.2034
K-medoid clustering	2.311	0.7402	0.7327	0.3607
Agglomerative clustering	0.3676	0.2714	0.2712	0.1137
Gaussian mixture model	1.0845	0.2268	0.6619	0.2922

Conclusion

We contributed CSAI, a new cluster evaluation index for dealing with the validity and robustness of clustering solutions. CSAI distinguishes itself by fully leveraging aggregated feature structures pertaining to clusters rather than cluster centroids. Our experimental results demonstrate that CSAI is a promising approach that can effectively evaluate the quality of clustering solutions and their stability. More importantly, the CSAI index exhibited the highest efficacy in agglomerative hierarchical clustering, highlighting its suitability for hierarchical clustering while also showcasing its effectiveness across a range of clustering algorithms.

PARTNERS

HOST

UNIVERSITY OF BERGEN

FUNDED BY

This research is funded by SFI MediaFutures partners and the Research Council of Norway (grant number 309339).

Forskningsrådet

Norwegian Centre for Research-based Innovation