On the Combination of Relative Clustering Validity Criteria

Lucas Vendramin
Pablo A. Jaskowiak*
Ricardo J. G. B. Campello

Institute of Mathematics and Computer Sciences (ICMC)
University of São Paulo (USP) - Brazil

Funded by CNPq and FAPESP

Outline

- Background
- Combining Relative Validity Criteria
- Evaluating Relative Validity Criteria
- Experimental Evaluation
- Final Remarks and Future Perspectives

Clustering

Clustering *(Jain and Dubes, 1988)*

The process of organizing data objects in a convenient, valid and meaningful manner.

- No universal definition to the problem
  - Large number of clustering algorithms introduced
  - Active field of research with long history
- Given different algorithms and parameters
  - How to select the most appropriate ones?
  - We need a way to assess and evaluate the results

Cluster Validation

- Refers to procedures meant for evaluating clustering results in a quantitative and objective fashion

- Two different types of validation criteria

External
Compare clustering solution against expected structure (ground truth partition). Not appropriate for real world applications, in which there is no expected results.

Relative
Use same information as internal measures, i.e., only information from the data itself. These measures can be used to compare multiple solutions and select the "best" one.

Cluster Validation

- Plethora of relative validity criteria in the literature
  - Criteria performance depends on the application
    - Some studies provide guidance in restricted scenarios
  - It is difficult to select a criterion among such variety
    - It is prohibitive to conduct studies to identify the *best* relative validity criteria in each and every real world application scenario

- Are there any alternatives?
  - One alternative is to rely on results from multiple criteria
    - Combination of different relative validity criteria

Relative Criteria Combination

- Not much attention has been given to the subject
  - Although a few studies relied on criteria combination
    - Few datasets and criteria were considered
    - No systematical assessment has been conducted so far

- In this work we systematically evaluate
  - 4 different types of combinations of 28 relative criteria
  - Not interested in the performance of single criterion

Our Goal

Verify if combining relative validity criteria can be beneficial in practical applications, in which the user does not know which criterion is the best or the worst one

Combining Relative Criteria

- For a particular application scenario
  - Given different clustering algorithms, number of clusters
  - How to select the best partition?
  - More than 40 relative criteria in the literature!
    - Which one to choose?

- If the user has no clue on which criteria select, given a particular combination strategy, can he/she obtain
  - Results as good as the ones from the best criteria?
  - Better results than the worst criteria from the combination?

Combining Relative Criteria

- We consider a set of 28 different relative criteria

Calinski-Harabasz (VRC)

Dunn + 17 variants of Dunn

Silhouette Width Criterion

Point-Biserial

C-Index

PBM

$C/\sqrt{k}$

Alternative Simplified Silhouette

Davies-Bouldin

Alternative Silhouette

Simplified Silhouette

Combining Relative Criteria

- Given a set of partitions
  - Each partition is evaluated by different relative criteria
  - Criteria values are normalized between 0 and 1
  - Such values are then combined in four different ways

- We evaluate the following combination procedures
  - Mean
  - Harmonic Mean
  - Mean* (mean, removing the most discrepant value)
  - Median

- We consider combinations of 3 and 5 relative criteria

Evaluating Relative Validity Criteria

- Two different methodologies
  - Traditional Methodology (Milligan and Cooper, 1985)
  - Alternative Methodology (Vendramin et al., 2010)

Traditional Methodology

- Take N_D datasets with known ground truth solution
- For each dataset
  - Generate a collection of partitions of different quality and number of clusters, using different algorithms
  - Compute the values of relative criteria for all the partitions generated. Check whether the number of clusters of the best partition (as selected by each relative validity criteria) match the number of clusters for the ground truth partition
- For each relative criterion
  - Count the number of datasets for which it finds the *correct* number of clusters, as defined by the ground truth solution

Alternative Methodology

- Take N_D datasets with known ground truth solution
- For each dataset
  - Generate a collection of partitions of different quality and number of clusters, using different algorithms
  - Compute the values of relative criteria for all the partitions generated. For each criterion, compute the correlation between its values and external criterion values for all partitions generated for the particular dataset in hand
- For each relative criterion
  - Compute the mean and standard deviation of correlation values for each criterion, which can be used as a measure of their accuracy

Evaluating Relative Validity Criteria

- To analyze the effectiveness of combining different criteria we counted the number of combinations that
  - Outperformed all criteria involved in the combination
  - Outperformed at least one criteria in the combination

- A combination is an improvement (outperforms) the criteria that are involved in the combination if it
  - Gives the correct number of clusters in a larger number of datasets, considering the Traditional Methodology
  - Gives a better value of correlation with the external criterion, considering the Alternative Methodology

Experimental Setup

- Datasets
  - Synthetic data (Milligan and Cooper, 1985)
    - 972 datasets in total
  - ALOI datasets (Geusebroek et al., 2005)
    - Amsterdam Library of Object Images
    - 400 datasets in total
- Datasets have different number of
  - Clusters (2, 3, 4, 5, 6, 12, 14, and 16)
  - Objects (50, 75, 100, 125, and 500)
  - Dimensions (2, 3, 4, 7, 22, 23, and 24)

Experimental Setup

- Clustering algorithms
  - k-means
  - Hierarchical
    - Single-Linkage, Average-Linkage, Complete-Linkage, Ward's
- For each dataset, we consider as number of clusters
  - From 2 to $\sqrt{n}$, where n is the number of objects
- Given this setup we got a total of
  - 427,680 partitions for synthetic datasets
  - 14,000 partitions for ALOI datasets

Experimental Setup

- Traditional Methodology
  - Number of hits (correct number of clusters)

- Alternative Methodology
  - External Criteria
    - Adjusted Rand Index and Jaccard

  - Correlation Coefficient
    - Pearson and Weighted Goodman-Kurskal (Campello and Hruschka, 2009)

  - Statistical Tests
    - Friedman (mean) and Brown-Forsythe (variances)

Results and Discussion

- Synthetic Datasets
  - Improvements over *all* the criteria from the combination

Traditional Methodology 3 Criteria Combination

Combination Strategy	# Improvements (Percentage)
Mean	315 (9.62)
Harmonic	338 (10.32)
Mean-2	163 (4.98)
Median	174 (5.31)

Alternative Methodology 3 Criteria Combination

	# Improvements (Percentage)		
Combination	Mean	Variance	Both
Mean	22 (0.67)	10 (0.30)	4 (0.12)
Harmonic	52 (1.58)	239 (7.29)	43 (1.31)
Mean-2	3 (0.09)	4 (0.12)	0 (0)
Median	21 (0.64)	6 (0.18)	5 (0.15)

Results and Discussion

- Synthetic Datasets
  - Improvements over *at least one* criteria in the combination

Traditional Methodology 3 Criteria Combination

Combination Strategy	# Improvements (Percentage)
Mean	3274 (99.94)
Harmonic	3274 (99.94)
Mean-2	3264 (99.63)
Median	3264 (99.63)

Alternative Methodology 3 Criteria Combination

	# Improvements (Percentage)		
Combination	Mean	Variance	Both
Mean	3248 (99.14)	1777 (54.24)	1777 (54.24)
Harmonic	3100 (94.62)	2676 (81.68)	2587 (78.96)
Mean-2	2946 (89.92)	1685 (51.43)	1536 (46.88)
Median	3108 (94.87)	1475 (45.02)	1454 (44.38)

Results and Discussion

- ALOI Datasets
  - Improvements over *all* the criteria from the combination

Traditional Methodology 5 Criteria Combination

Combination	# Improvements (Percentage) Mean	Variance	Both
Mean	0 (0)	0 (0)	0 (0)
Harmonic	0 (0)	75 (16.23)	0 (0)
Mean-2	0 (0)	0 (0)	0 (0)
Median	0 (0)	0 (0)	0 (0)

Alternative Methodology 5 Criteria Combination

Combination	# Improvements (Percentage) Mean	Variance	Both
Mean	0 (0)	0 (0)	0 (0)
Harmonic	0 (0)	75 (16.23)	0 (0)
Mean-2	0 (0)	0 (0)	0 (0)
Median	0 (0)	0 (0)	0 (0)

Results and Discussion

- ALOI Datasets
  - Improvements over *at least one* criteria from the combination

Traditional Methodology 5 Criteria Combination

Combination Strategy	# Improvements (Percentage)
Mean	462 (100)
Harmonic	462 (100)
Mean-2	462 (100)
Median	462 (100)

Alternative Methodology 5 Criteria Combination

	# Improvements (Percentage)		
Combination	**Mean**	**Variance**	**Both**
Mean	462 (100.00)	456 (98.70)	456 (98.70)
Harmonic	462 (100.00)	462 (100.00)	462 (100.00)
Mean-2	462 (100.00)	435 (94.15)	435 (94.15)
Median	462 (100.00)	435 (94.15)	435 (94.15)

Results and Discussion

- When the user has no clue on which criteria select
  - Combining different criteria can bring improvements over the worst criterion, i.e., this one can be avoided

- We considered only “*blind*” combinations
  - Increasing number of criteria lead to
    - Decrease in accuracy considering all criteria from combination
    - Increase in accuracy only against the worst criteria
    - There is still no theory or guidelines on how, how many and which criteria select to compose relative criteria combinations

Results and Discussion

- The study opened venues for further considerations
  - How to select complimentary criteria?
  - How to guarantee minimum criterion accuracy?
  - How to normalize criteria results before combination?
  - Are there better ways for combining criteria then the quite simple and *naive* approaches considered here?

Future Work

- To answer such questions
  - Borrow concepts from Ensemble Theory
    - Minimum Complementarity and Accuracy
  - Consider carefully which criteria to select for combinations
    - Can we identify similar and dissimilar criteria?
    - Which k criteria are the best match, which combinations to avoid
  - Such concepts are well developed for other tasks, e.g.,
    - Classification
    - Clustering
    - Outlier Detection

Final Remarks

- We evaluated relative criteria combinations
  - 28 relative criteria
  - 4 different types of combinations
  - Real and synthetic datasets
  - 3 and 5 criteria combinations
  - Over 400.000 partitions
- Results were consistent for all scenarios under evaluation
- If the user knows which one is the best criterion, combinations do not provide any improvements
- However, if there is no evidence regarding which criterion to use, combination of relative criteria is a good choice for user

Acknowledgements

Brazilian Research Agencies

FAPESP CNPq

Any Questions?

pablo@icmc.usp.br

Thank You!