

Parallel Evaluation of Large-Scale Hierarchical Results

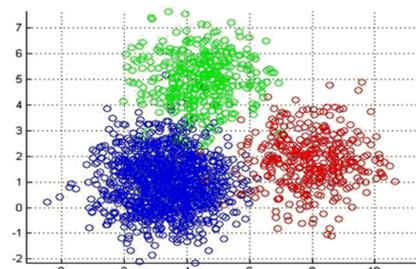
<https://sites.google.com/site/ramscruzdavid>

David Cruz-Rodriguez

Research Alliance in Math and Science
Polytechnic University of Puerto Rico
Mentor: Carlos Rojas, PhD

Introduction

Data analysis can be divided in two major areas (exploratory or confirmatory), but the key elements in both types of procedures are the natural grouping (clustering) or the goodness of the fit (validity). Cluster is the organization of a collection of distributions, usually represented as a vector or a point in multidimensional space, into cluster based on similarity.



Flat Clustering
A set of clusters without any explicit structure that would relate clusters to each other. Two of the well-known flat clustering algorithms are the k-means and Expectation-maximization (EM).

Figure 1. K-means: a nonhierarchical clustering and the most important flat clustering algorithm. Also, k-means is represented by assigning each pattern to the closest cluster center. The main purpose of k-means is to classify the data.

The points belonging to the same cluster are given the same label. The variety of techniques for representing data, measuring proximity (similarity) between data-sets has produced the necessity of better acceptance and evaluation measures of the obtained data clustering. It is very important to understand the difference between clustering (unsupervised learning) and discriminant analysis (supervised learning). In supervised learning a human imposes in the data; the problem is to label a newly encountered. The clustering technique is useful for several exploratory pattern analysis, grouping decision making, and machine learning situation.

Hierarchical Clustering

Hierarchical clustering is based on being informative than the unstructured set of clusters by flat clustering. Hierarchical clustering methods can be grouped in two general classes (agglomerative or divisive).

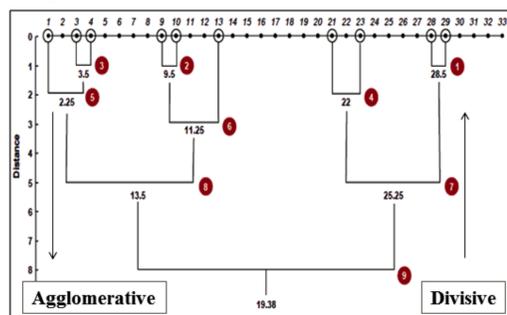
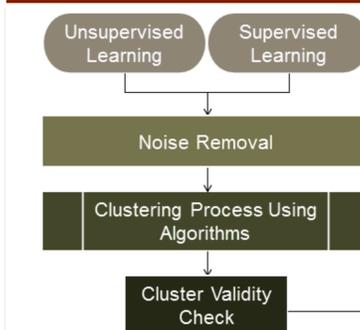


Figure 2. Dendrogram: Depicts seven patterns labeled 1,3,4,9,10,13,21,23,28 and 29 in three clusters. A hierarchical algorithm yields a dendrogram representing the nested grouping of patterns and similarity levels at which groupings change.

The purpose of this research is to survey the core concepts and techniques of the large variety of cluster validation. Where appropriate, references will be made to key tenet from clustering. Generally speaking, there are three types of clustering validation techniques: external, internal, and relative criteria. This research addresses only external and internal criteria.

Exploration and Analysis



Counting pairs of points are based on which two clustering's may agree or disagree. The confusion matrix is the overlapping between the pairs of points that can only fall one of the four constraints:

- SS is the number of pairs of items belonging to the same cluster and partition;
- SD is the number of pairs belonging to the same cluster and different partition;
- DS is the number of pairs belonging to a different cluster and the same partition;
- DD is the number of pairs belonging to a different cluster and partition.

Figure 3. Data Exploration: Finding hidden knowledge from the data sets. Data clustering is a tool which aids the data exploration process, and a technique of partitioning the data set without known priori information.

Cluster Validation
A process or technique of finding a set of data that best falls into natural categories without any prior information is called cluster validation.

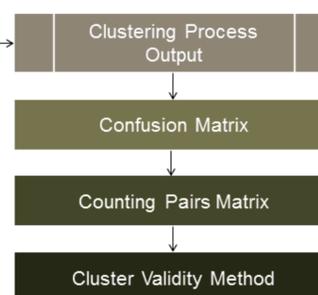


Figure 4. Cluster Validity Check: The process that describes the cluster validity check is used for the class criteria (external) of comparing clustering by counting pairs.

Validation Indexes

Validation Indexes	Notation	Indexes Restriction		Clustering Method	O(*)
		H C RB	S vs. Q		
1 Rand	$\frac{SS + DD}{(SS + SD + DS + DD)}$	X X		Flat	O(n)
2 Jaccard	$\frac{SS}{(SS + SD + DS)}$	X X		Flat	O(n)
3 Entropy	$\sum_{j=1}^m \frac{c_j}{N} E_j$	X X	X	Flat	O(n)
4 Purity	$\sum_{j=1}^m \frac{c_j}{n} P_j$	X	X	Flat	O(n)
5 CPCC	$\frac{(\frac{1}{M}) \sum_{i=1}^{N-1} \sum_{j+1}^N a_{ij} c_{ij} - \mu_P \mu_C}{\sqrt{[(\frac{1}{M}) \sum_{i=1}^{N-1} \sum_{j+1}^N a_{ij}^2 - \mu_P^2][(\frac{1}{M}) \sum_{i=1}^{N-1} \sum_{j+1}^N c_{ij}^2 - \mu_C^2]}}$			Hierarchical	O(n^2)

- Legend 1.** Validation indexes
1. H: Homogeneity
 2. C: Completeness
 3. RB: Rag bag
 4. S vs. Q: Cluster size vs. quantity
 5. O(*): Big O notation (Complexity)
 6. CPCC: Cophenetic correlation coefficient index

Table 1. Validation Indexes table: This table depicts two types of validation indexes: internal and external criterion. Also, the majority of the validation indexes are flat clustering with a high big O notation.

Results and Conclusions

This research analyzes and compares two clustering validity index approaches the external and internal criterion; carrying out analyses of six external indexes and four internal indexes. A large scale of data-set was used, which replicate the k-means output (partition and cluster). The data set was evaluated using the metric based on counting pairs "in parallel" finding a better complexity O(log n).

Intuitively, by observing the table 1 you can discern that the clustering nonhierarchical is the most popular. Thus, nonhierarchical algorithms possess a major drawback. It is very common to find them with a slower function in the running time of the algorithm. Yet, the hierarchical algorithms contains a better complexity in forms of a quadratic.

In spite of, the research was concentrated on how to create novel clustering measures to assess the data. Moreover, to try to make possible flat indexes into hierarchical evaluations.

Figure 5. Parallel Counting Pairs Matrix Product: The figure depicts the 2-D Mesh.

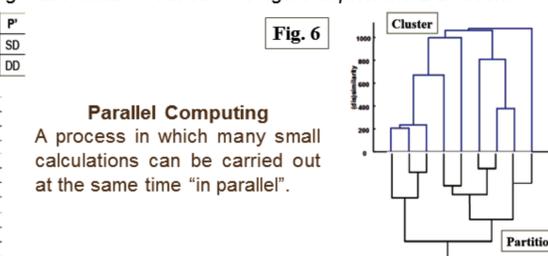
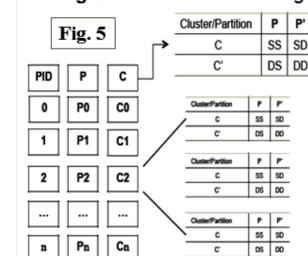


Figure 6. Hierarchical Counting Pair: The figure depicts the hierarchical counting pair evaluation in order to compare the measure of agreement.

In conclusion, some of the external validity indexes can be computed as soon as you compute the measure of agreement between the partition and the cluster from a hierarchical clustering technique. Empowering, from the prevalent external, more novel ideas to evaluate this very important hierarchical algorithms technique. By employing parallel computing into the flat index to evaluate the hierarchical algorithm, the complexity prevails the same but it is important to step-up the calculation running time.

There is a need for developing quality measure that assesses the quality of the partitioning quality and algorithm complexity. This will contribute to the better usage of data mining techniques for the extraction of valid, interesting, and patterns knowledge.

References

Amigo, E., Gonzalo, J., Ariles, J., & Verdejo, F. (2008). A Comparison of Extrinsic Clustering Evaluation Metrics Based on Formal Constraints. *Springer Standard Collection*, 12(4), 461 - 486. doi: 10.1007/s10791-008-9066-8.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). Cluster Validity Methods: Part 1. *ACM Special Interest Group on Management of Data Record*, 31(3), 42 - 45.

Meila, M. (2005). Comparing Clustering. *Association for Computing Machinery*, 577 - 584. doi: 10.1007/BF01908075.