Comprehensive Cross-Hierarchy Cluster Agreement Evaluation

David M. Johnson, Caiming Xiong, Jing Gao and **Jason J. Corso**
State University of New York at Buffalo
Department of Computer Science and Engineering
338 Davis Hall
Buffalo, NY, 14260-2500
{davidjoh,cxiong,jing,jcorso}@buffalo.edu

Abstract

Hierarchical clustering represents a family of widely used clustering approaches that can organize objects into a hierarchy based on the similarity in objects' feature values. One significant obstacle facing hierarchical clustering research today is the lack of general and robust evaluation methods. Existing works rely on a range of evaluation techniques including both internal (no ground-truth is considered in evaluation) and external measures (results are compared to ground-truth semantic structure). The existing internal techniques may have strong hierarchical validity, but the available external measures were not developed specifically for hierarchies. This lack of specificity prevents them from comparing hierarchy structures in a holistic, principled way. To address this problem, we propose the Hierarchy Agreement Index, a novel hierarchy comparison algorithm that holistically compares two hierarchical clustering results (e.g. ground-truth and automatically learned hierarchies) and rates their structural agreement on a 0-1 scale. We compare the proposed evaluation method with a baseline approach (based on computing F-Score results between individual nodes in the two hierarchies) on a set of unsupervised and semi-supervised hierarchical clustering results, and observe that the proposed Hierarchy Agreement Index provides more intuitive and reasonable evaluation of the learned hierarchies.

Introduction and Related Work

Research into flat clustering methods benefits greatly from the availability of powerful cluster evaluation tools, ranging from the original Rand Index (Rand 1971) to more modern methods such as V-Measure (Rosenberg and Hirschberg 2007), that allow researches to effectively judge their clustering results against some ground-truth objective, and thus compare the relative performance of different methods.

Unfortunately, *hierarchical* clustering does not enjoy such a wealth of viable, established evaluation techniques. While there do exist well-grounded internal (i.e. with no ground truth information) measures of hierarchy quality (**?**), researchers interested on measuring the external, semantic meaning of a hierarchy are forced to resort to a wide array of questionable and limited methods. Some, for instance, simply test their methods on non-hierarchical data, and cut their hierarchies to produce a flat segmentation for evaluation (**?**; **?**).

Currently, the most common external evaluation technique is a method we refer to as Cluster F-Measure (CFM). It was originally proposed in (**?**) and works (on either flat or hierarchical clusterings) by matching each ground-truth cluster/hierarchy node to the "best" test cluster/hierarchy node (where match quality is determined by $F(c_1, c_2) = \frac{2 \cdot P \cdot R}{P+R}$: the F-score for cluster c_2 on cluster c_1). The overall score, then, is just the cluster-size-weighted average of the best scores. This method is applicable to hierarchies, but not specific to them, and when applied to a hierarchical clustering solution completely ignores the actual structure of the hierarchy, treating each node as an unrelated flat cluster. We would argue that a strong tool for evaluating hierarchical clustering solutions must take account of the hierarchical relationships between nodes and elements, and that the CFM approach is thus inadequate.

We instead propose a novel evaluation technique that directly encodes and compares the entire structure of two hierarchies. We refer to our method, which can be thought of as an extension of the classical Rand Index to the hierarchical case, as the Hierarchy Agreement Index (HAI).

Hierarchy Agreement Index

In both the Rand Index and our proposed HAI, the total comparison score is computed via:

$$\mathcal{S}(\mathcal{C}_1, \mathcal{C}_2) = 1 - \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} |f(x_i, x_j, \mathcal{C}_1) - f(x_i, x_j, \mathcal{C}_2)| \ , \tag{1}$$

where $\mathcal{C}_1$ and $\mathcal{C}_2$ are the two clustering solutions being compared, N is the number of elements being clustered and f is a function that extracts some value in the range [0,1] describing the relatedness of two points under the given clustering scheme.

For the Rand Index, $f(x_i, x_j, \mathcal{C})$ is an indicator function, returning 1 if x_i and x_j are are in the same cluster under $\mathcal{C}$ and 0 otherwise. For hierarchy evaluation, we instead need an f that can return a range of values representing an arbitrary number of different degrees of relatedness. Moreover, f must be general to a wide range of different hierarchy structures. Most notably, it must yield meaningful values for

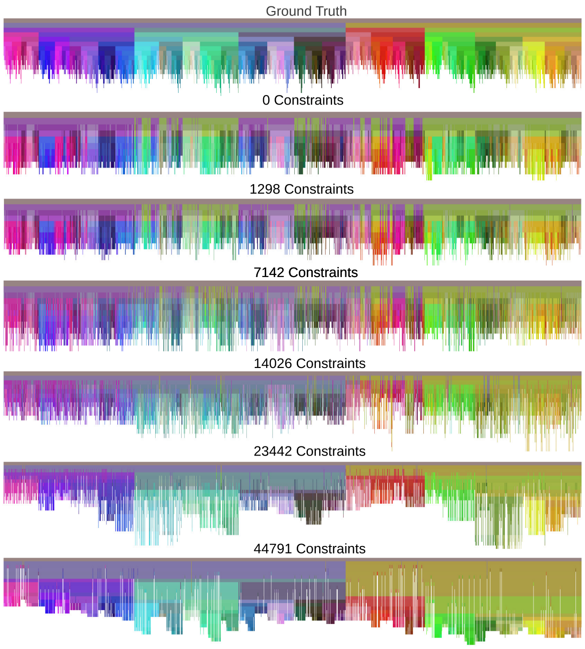

Figure 1: Results of hierarchical clustering with varying numbers of constraints on an example dataset we created. The data consists of 2500 randomly generated RGB color values, with a "ground truth" hierarchy constructed by running UPGMA (**?**) on the data in Lab space. We then ran a semi-supervised hierarchical clustering algorithm on the data (in HSV space), with varying numbers of triplet constraints drawn from the ground truth. The resulting hierarchies are visualized above. *(view in color)*

both binary and non-binary hierarchy trees, as well as both deep hierarchies (such as those produced by agglomerative clustering algorithms) and shallow hierarchies (such as most human-generated semantic hierarchies).

To satisfy these requirements, we define the *hierarchy distance* $d_{\mathcal{H}}(a, b)$ between two elements a and b in cluster hierarchy $\mathcal{H}$. Let $n^{a,b}$ be the smallest node in the hierarchy containing both a and b, and let $size(n) = \frac{|n_{\mathcal{D}}|}{N}$ (i.e. the proportion of the total element set found in node n). Now let $d_{\mathcal{H}}(a, b) = size(n^{a,b})$, or 0 if $n^{a,b}$ is a leaf node (because two elements that lie in the same leaf node are maximally close under that hierarchy).

Using hierarchy distance, our cluster agreement measure becomes:

$$\mathrm{HAI}(\mathcal{H}_1, \mathcal{H}_2) = 1 - \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} |d_{\mathcal{H}_1}(x_i, x_j) - d_{\mathcal{H}_2}(x_i, x_j)| \tag{2}$$

The distance $d_{\mathcal{H}}$ is independent of the specific structure of the hierarchy because it is unaffected by the number of intervening nodes between x_i, x_j and n^{x_i, x_j}. Additionally, by defining the hierarchy distance between two elements within the same leaf node as 0, we can meaningfully represent distance even in shallow hierarchies. While shallow hierarchy trees may contain very large leaf nodes, distance measures within these trees will not be distorted because node-pairs within those leaf nodes will still be considered maximally close.

Constraints	HAI	CFM
0	0.701	0.489
1298	0.705	0.472
7142	0.749	0.442
14026	0.856	0.452
23442	0.923	0.512
44791	0.991	0.784

Table 1: HAI (this paper) and CFM evaluation against ground truth on synthetic color data.

Thus, using equation 2, we can compare the relatedness of each point pair in the dataset under two input hierarchies and aggregate the results to achieve a single unified measure of hierarchy correspondence. Like the Rand Index, this measure will yield a correspondence score of 1 only for cases where the two cluster solutions are functionally identical ("functionally" because, again, two hierarchy trees, such as a binary and non-binary tree, may encode the same information despite having somewhat different tree structures).

Experiments

To validate our measure, we performed semi-supervised hierarchical clustering, with varying numbers of constraints, on a set of easily-visualized synthetic color data (see Figure 1). We then used both CFM and HAI to measure the agreement between each resulting hierarchy and the ground truth. The details of the particular semi-supervised hierarchical clustering method are beyond the scope of this short paper, but in essence it uses triplet constraints drawn from the ground truth hierarchy to iteratively perform flat semi-supervised clustering on the data, ultimately yielding a divisive hierarchy.

Figure 1 shows that as we add constraints to the semi-supervised method the resulting cluster hierarchy increasingly appears to resemble the ground truth hierarchy. The HAI results in Table 1 coincide well with this intuition, showing first minor gains, then a significant improvement by the 14000 constraint level. By contrast, the CFM measure actually worsens initially, and fails to detect any improvement in the hierarchy until the 23000 constraint level.

Conclusion and Future Work

Effective cluster hierarchy analysis is an important problem, and we believe HAI can fill a needed niche as an evaluation tool for hierarchical clustering algorithms. In addition to exploring variants on the HAI algorithm—an adjusted-for-chance version, for instance—we plan to perform a more rigorous evaluation of the measure, with more theoretically-grounded test data/hierarchies. However, we believe this initial experiment offers some useful validation for the method.

References

Rand, W. 1971. Objective criteria for the evaluation of clustering methods. *JASA*.

Rosenberg, A., and Hirschberg, J. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*.