Human Cluster Evaluation and Formal Quality Measures: A Comparative Study

Joshua M. Lewis
josh@cogsci.ucsd.edu
Dept. of Cognitive Science
University of California, San Diego

Margareta Ackerman
mackerma@uwaterloo.ca
Cheriton School of Computer Science
University of Waterloo

Virginia R. de Sa
desa@cogsci.ucsd.edu
Dept. of Cognitive Science
University of California, San Diego

Abstract

Clustering quality evaluation is an essential component of cluster analysis. Given the plethora of clustering techniques and their possible parameter settings, data analysts require sound means of comparing alternate partitions of the same data. When proposing a novel technique, researchers commonly apply two means of clustering quality evaluation. First, they apply formal Clustering Quality Measures (CQMs) to compare the results of the novel technique with those of previous algorithms. Second, they visually present the resultant partitions of the novel method and invite readers to see for themselves that it uncovers the correct partition. These two approaches are viewed as disjoint and complementary.

Our study compares formal CQMs with human evaluations using a diverse set of measures based on a novel theoretical taxonomy. We find that some highly natural CQMs are in sharp contrast with human evaluations while others correlate well. Through a comparison of clustering experts and novices, as well as a consistency analysis, we support the hypothesis that clustering evaluation skill is present in the general population.

Keywords: clustering; validity indices; psychophysics; visual perception; machine learning

Introduction

Clustering is a fundamental data analysis tool that aims to group similar objects. It has been applied to a wide range of disciplines such as astronomy, bioinformatics, psychology, and marketing. Successful clustering often requires using a number of different clustering techniques and then comparing their output. The evaluation of clusterings is an integral part of the clustering process, needed not only to compare partitions to each other, but also to determine whether *any* of them are sufficiently good.[1]

As there is no universal clustering objective, there is no consensus on a formal definition of clustering. As a result, there are a wide variety of Clustering Quality Measures (CQMs), also known as internal validity indices, that aim to evaluate the quality of clusterings. To compare clusterings, researchers often select a CQM, which assigns a numerical value to a partition representing its quality.

Researchers rarely rely on CQMs alone. There is a deep implicit assumption running through the clustering literature that human judgment of clustering quality is quite good. Authors visually present the resultant partitions and invite readers to see for themselves that the new method performs well. To take one example, in their influential paper on spectral clustering Ng, Jordan and Weiss write, "The results are surprisingly good... the algorithm reliably finds clusterings consistent with what a human would have chosen." (Ng, Jordan, & Weiss, 2002) Up until now, clustering quality measures and human judgment were considered complementary approaches to clustering evaluation. Most papers that present novel clustering algorithms include these two types of evaluations separately.

Our study compares formal CQMs with human evaluations to determine how often they agree, and whether certain CQMs correlate better with human judgments than others. We also evaluate the consistency of human responses–if humans are very inconsistent, then it is unlikely that they are good judges of cluster quality (an ideal measure is stable on the same partition). Further, we separate our human subjects into expert and non-expert groups to determine whether clustering evaluation requires experience, and identify divergent strategies between the groups.

To sharpen our focus on a small set of CQMs, we construct a property-based taxonomy of CQMs that distinguishes them on grounds beyond their particular mathematical formulations. The CQMs selected for the study are diverse in that they each satisfy a distinct set of these properties.

Previous studies have investigated how humans choose the number of groups (Lewis, 2009) and partition data (Santos & Sá, 2005) in a clustering setting, but these approaches only show what humans think are the optimal partitions rather than how they judge partition quality in general. Our study uses a set of non-optimal partitions that humans partially order by quality, giving us more detailed quality judgments than in past work. Intuitively, in (Lewis, 2009) and (Santos & Sá, 2005) subjects took on the role of a k-choosing algorithm and a clustering algorithm (respectively), whereas in this study subjects are in the role of clustering evaluators.

Our main findings are as follows. Many CQMs with natural mathematical formalizations disagree with human evaluations. On the other hand, we identify CQMs whose evaluations are well correlated with those of humans. In particular, we find that Silhouette (Rousseeuw, 1987) and Dunn's Index (Dunn, 1974) are highly correlated with human evaluations. Our findings also indicate that there is sufficient similarity between the evaluations of novices and experts to suggest that clustering evaluation is a task that does not require specific training (though it may benefit from training). This opens the door for using human computation resources such as Amazon's Mechanical Turk to quickly solicit a large number of clustering quality judgments from novices as part of the data analysis process. Nevertheless, experts show much less sensitivity to the number of clusters and relate more closely to a greater range of clustering quality measures than novices, indicating a nuanced approach to the evaluation problem. Re-

[1] If no good clusterings have been found the underlying dataset may have no good clustering (the data is not "clusterable", see (Ackerman & Ben-David, 2009) for more on clusterability).

garding consistency, we find that even novices are more consistent in their evaluations than our set of CQMs.

Clustering quality measures

In this section we introduce the formal machinery describing the CQMs selected for our study.

Let X be a finite domain set. A *distance function* is a symmetric function $d : X \times X \to R^+$, such that $d(x,x) = 0$ for all $x \in X$. A *k-clustering* $C = \{C_1, C_2, \ldots, C_k\}$ of dataset X is a partition of X into k disjoint subsets (so, $\cup_i C_i = X$). A *clustering* of X is a k-clustering of X for some $1 \leq k \leq |X|$. Let $|C|$ denote the number of clusters in clustering C. For $x, y \in X$ and clustering C of X, we write $x \sim_C y$ if x and y belong to the same cluster in C and $x \not\sim_C y$, otherwise. Finally, a CQM is a function that maps clusterings to real numbers.

Gamma: This measure was proposed as a CQM by (Baker & Hubert, 1975) and it is the best performing measure in (Milligan, 1981). Let d^+ denote the number of times that a pair of points that was clustered together has distance smaller than two points that belong to different cluster, whereas d^- denotes the opposite result.

Formally, let $d^+(C) = |\{\{x,y,x',y'\} \mid x \sim_C y, x' \not\sim_C y', d(x,y) \leq d(x',y')\}|$, and $d^-(C) = |\{\{x,y,x',y'\} \mid x \sim_C y, x' \not\sim_C y', d(x,y) \geq d(x',y')\}|$. The *Gamma* measure of C is $\frac{d^+(C) - d^-(C)}{d^+(C) + d^-(C)}$.

Silhouette: The Silhouette measure was defined by (Rousseeuw, 1987). Silhouette is the default clustering quality measure in MATLAB.

Let $dist(x, C_i) = avg_{y \in C_i} d(x,y)$. The *silhouette of a point* x with respect to clustering C is $S(x,C) = \frac{\min_{j \neq i} dist(x,C_j) - dist(x,C_i)}{\max(\min_{j \neq i} dist(x,C_j), dist(x,C_i))}$ where $x \in C_i$. The *silhouette of a clustering* C is $sum_{x \in X} S(x,C)$.

Dunn's Index: Dunn's Index (Dunn, 1974) compares the maximum within-cluster distance to the minimum between-cluster distances. *Dunn's Index* of C is $\frac{\min_{x \not\sim_C y} d(x,y)}{\max_{x \sim_C y} d(x,y)}$.

Average Between and Average Within: The Average Between and Average Within measures evaluate the between-cluster separation and within-cluster homogeneity, respectively. The *average between* of C is $avg_{x \not\sim_C y} d(x,y)$. The *average within* of C is $avg_{x \sim_C y} d(x,y)$.

Calinski-Harabasz: The Calinski-Harabasz measure (Caliński & Harabasz, 1974) makes use of cluster centers. Let $c_i = \frac{1}{|C_i|} \sum_{x \in C_i} |x|$ denote the center-of-mass of cluster C_i, and $\bar{x}$ the center-of-mass of X. Let $B(C) = \sum_{C_i} |C_i| |c_i - \bar{x}|^2$ and $W(C) = \sum_{C_i} \sum_{x \in C_i} |x - c_i|^2$. The *Calinski-Harabasz* of C is $\frac{n-k}{k-1} \cdot \frac{B(C)}{W(C)}$.

Weighted inter-intra: The weighted inter-intra measure is proposed by (Strehl, 2002). It compares the homogeneity of the data to its separation. Let $intra(C_i) = avg_{x,y \in C_i} d(x,y)$ and $inter(C_i, C_j) = avg_{x \in C_i, y \in C_i} d(x,y)$. The *Weighted inter-intra* of a clustering C is $(1 - \frac{2k}{n}) \cdot (1 - \frac{\sum_i \frac{1}{n-|C_i|} \sum_{j \neq i} inter(C_i,C_j)}{\sum_i \frac{2}{|C_i|-1} intra(C_i)})$, where n is the number of points in the dataset.

Methods

We ran two groups of human subjects and a group of clustering quality measures on a partition evaluation task. Our human subjects were divided into a novice group with little or no knowledge of clustering methods and an expert group with detailed knowledge of clustering methods.

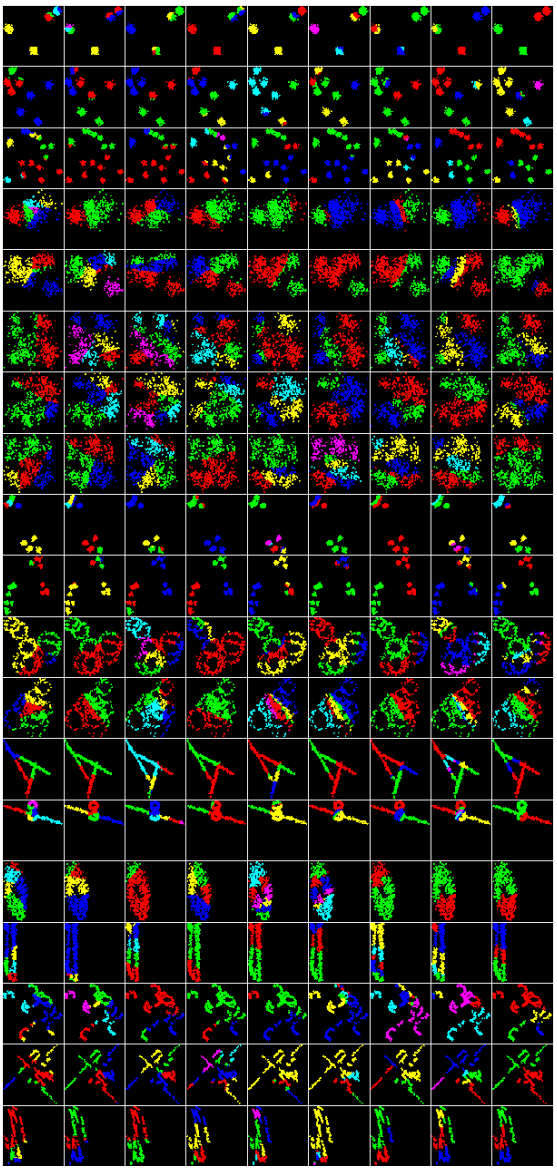

Figure 1: All stimuli. Datasets are in rows; partitions are in columns.

Human subjects and stimuli

Twelve human subjects were recruited for this project as the novice group, 9 female and 3 male, with an average age of 20.3 years. The novice subjects have no previous exposure to clustering. The expert group consists of 5 people and includes the authors of this paper. All experts have studied clustering in an academic setting, and 4 have done research on the subject.

We used 19 different two dimensional datasets to generate our clustering stimuli, drawn from (Lewis, 2009), and chosen to represent a range of dataset types including mixtures of Gaussians and datasets with hierarchical structure. In order to maintain responsiveness of the stimulus presentation interface, we subsampled 500 points randomly from each dataset. We use synthetic datasets in order to better generate a wide range of stimuli, and our datasets are 2D to facilitate visualization.

Each dataset is randomly clustered nine times in the following manner. For each of the nine clusterings, we first draw the number of partitions, k, from a uniform distribution over the integers 2 to 6. Second we choose cluster centroids using two strategies: for four of the clusterings we randomly select k centroids from the original dataset, and for five of the clusterings we select k centroids from a Laplacian Eigenmap embedding of the data. Finally we color points based on the identity of their nearest centroid in the appropriate space. The goal of this approach is to create stimuli with varied clustering quality.

Each trial consisted of all nine different partitions of the same dataset randomly arranged per trial in a 3 by 3 grid (see Figure 1 for a visualization of all the stimuli). The datasets were shown as scatter plots with colored points on a black background to reduce brightness-related eye strain. For novice subjects, trials were organized into three blocks of 19, where each dataset appeared once per block and the order of the datasets within each block was randomized. Expert subjects were tested on one block of non-randomized datasets. We instructed subjects to choose the two best partitioned displays and the one worst partitioned display from the nine available on every trial.

Analysis

We analyzed our novice subjects for internal consistency of their positive and negative ratings across blocks and found that even our least consistent subject performed well above chance. We did not exclude any subjects due to inconsistency and we did not analyze internal consistency for experts as they were only tested on one block.

To analyze consistency across subjects we use Fleiss' κ (Fleiss, 1971) and include neutral responses. Fleiss' κ measures the deviation between observed agreement and the agreement attributable to chance given the relative frequency of ratings and normalized for the number of raters. Neutral ratings are twice as frequent as non-neutral, and positive ratings are twice as frequent as negative ratings, so the compensation for relative frequency in Fleiss' κ makes it well-suited to our data. In addition, we perform a consistency analysis on the clustering quality measures by discretizing their ratings in a manner similar to the human data.

We analyze the relationship between novice ratings, expert ratings and clustering quality measures by calculating the Pearson's correlation coefficient, ρ, between ratings. To make the responses as comparable as possible we normalize response vectors to a length of one within each dataset. Human subjects have to rank two positive and one negative partition per dataset, even if every partition is quite bad, so by normalizing within dataset we make the CQM responses similar in structure—partitions are judged only relative to other partitions within a dataset.

Because cluster centroids are chosen randomly, increasing k is likely to increase the chance of getting an undesirable partition (e.g. a partition with very few data points). Additionally, partitions with higher k require more effort to interpret, and therefore we might expect novice subjects to be biased towards a lower k. For these reasons our correlations control for k by partialing out a vector of k values for each partition. Geometrically this is equivalent to projecting each response vector onto the hyperplane orthogonal to the vector of k values.

Results

Correlation

Table 1 shows correlation coefficients between all measures for both expert and novice responses, with k factored out. The correlation between expert and novice human positive ratings is higher than the correlation between any CQM and either human positive rating. The negative human ratings have a similarly high correlation. The absolute values of the correlation coefficients between CQMs and expert ratings are strictly greater than or equal to those between CQMs and novice ratings, indicating a closer relationship between expert strategies and the dataset characteristics summarized by the CQMs when k is factored out. k itself correlates very strongly with the novices and less so with the experts. Silhouette provides the best overall correlation with expert ratings, and Avg Within provides the best overall correlation with novice ratings (save k).

Consistency

The most undesirable form of inconsistency across subjects or CQMs is both positive and negative responses to the same stimulus. For experts, stimuli with a number of positive responses 3 or higher never receive a negative rating, and only once does this occur for stimuli with 2 positive responses. In contrast the CQMs exhibit much more disagreement and novices seem to fall somewhere in between. The quantitative measure κ bears this out: CQMs score 0.128, novices score 0.183 and experts score 0.213. κ ranges from -1 to 1, with -1 representing complete disagreement, 1 representing complete agreement and 0 representing the amount of agreement expected by chance. While there is no standard significance test for differences in κ, the rating scale suggested by Landis and Koch (Landis & Koch, 1977) would classify the CQM and novice rater groups each as in slight agreement, and the expert raters as in fair agreement. To test whether any one measure was significantly harming CQM consistency we left each out in turn from the analysis and found values ranging from 0.098 to 0.172, which is in line with the CQM consistency with no measure left out, and in every case less con-

Table 1: Correlation coefficients between human responses and CQMs with k factored out (except for the k column). Text in bold (excluding k column) if $p < .0025$ after Bonferroni correction for $n = 20$ comparisons per subject group and $\alpha = .05$.

ρ	Expert Positive	Expert Negative	Novice Positive	Novice Negative	Gamma	Silhouette	Dunn	Avg Within	Avg Btw	CH	W-Inter/Intra	k
Expert Pos	1	**-.35**	**.56**	-.19	-.15	**.46**	**.40**	**-.39**	**.34**	**.44**	.19	-.43
Expert Neg		1	-.13	**.44**	.09	**-.27**	-.12	**.44**	-.18	**-.36**	**-.30**	.32
Novice Pos			1	-.04	-.13	**.39**	**.40**	-.20	.23	**.30**	.04	-.73
Novice Neg				1	.08	**-.27**	.01	**.30**	-.07	**-.25**	**-.27**	.71

Table 2: A summary of the number of partitions for which a high degree of agreement was achieved by the raters. If a partition is classified as negative or positive by 90% - 100% of raters, it would be added to the top row, and similarly for the other buckets. The total possible number of agreed upon partitions is 51 (19 stimuli * 3 possible negative/positive responses per stimulus).

% Majority	Experts	Novices	CQMs
90% - 100%	4	0	0
80% - 90%	15	3	1
70% - 80%	0	2	7
60% - 70%	20	9	0
50% - 60%	0	20	19
Sum >= 50%	39	34	27

sistent than the novice subjects. Finally, we left out both Avg Within and Between, since they measure quality on intentionally simple and distinct dimensions, and found a κ of 0.110.

In Table 2 we summarize the consistency of experts, novices and cluster quality measures. It shows how often certain percentages of raters are able to agree on negative or positive ratings for particular stimuli. Experts agree over 50% of the time on more samples (39), than do novices (34) or CQMs (27).

Discussion

Comparing human evaluations with CQMs

Some natural quality measures have low correlation with human evaluations. Most notably, Gamma has low correlation with both positive and negative human ratings for both novices and experts. W-Inter/Intra has low correlation with the positive ratings of both subject groups. This shows that a natural mathematical formalization does not suffice to guarantee that the evaluations of clusterings produced using the CQM will seem natural to humans.

There are also CQMs that correlate well with human evaluations. Of these the most notable are CH and Silhouette. These two popular measures correlate well with both expert and novice evaluations, on both the positive and negative ratings.

Comparing experts with novices

Evaluations of experts and novices have a correlation score of 0.56, higher than the correlation of any CQM with any of the two subject groups. This suggests that a cluster evaluation skill is present in the general population.

On the other hand, we observe some interesting differences between the two groups of subjects. One of the most notable differences between experts and novices is that, while both groups prefer clusterings with fewer clusters, novices rely much more heavily on this heuristic.

Experts seem to use more, and more complex strategies than novices. Positive expert ratings correlate well with two more measures than positive novice ratings. No measure considered correlates better with novice ratings than with expert ratings, and in the great majority of cases the correlation is higher with expert ratings.

With a cover of at most six domain elements on any input dataset (see Definition 5 below), Dunn's measure is (according to this measure of complexity) the simplest measure that we explore. While positive expert evaluations correlate well with five distinct measures, Dunn's measure is one of three measures that correlate well with novice evaluations. This further illustrates that novices rely on fewer simpler strategies, which indicates that expert evaluations may be more sophisticated and reliable.

Consistency

Given the difficulty of knowing whether humans or CQMs do a reasonable job of evaluating clustering quality, one might hope that at least they are consistent across individuals (or measures). Consistency indicates that some repeatable process is at work and that its repeatability is minimally affected by changes in input. Of course CQMs are perfectly consistent on a within measure basis—given the same partition they will always report the same quality–and one is tempted to suggest that between measure consistency is an unfair point of comparison; aren't all the measures using quite different evaluative procedures, and didn't we select them to be distinct?

We did, but CQMs purport to evaluate clustering quality in general. Insofar as they evaluate this more nebulous property they should be consistent, even if their methods differ. As it turns out, they are somewhat consistent with each other, just not as consistent as humans. Further, the consistency story did not vary when we tested all the leave-one-out subsets of CQMs, indicating that CQM consistency is not being skewed by just one divergent measure.

Human experts are the most consistent group in this study. This lends empirical support to the common practice of seeking human visual evaluations of partition quality. Novices are less consistent, and as discussed above there is evidence that the evaluations they provide are less sophisticated. Despite the unfavorable comparison to experts, it is notable that subjects with no formal knowledge of cluster analysis are able to respond more consistently than a set of CQMs. This lends credence to the notion that our ability to evaluate partitions is acquired in the natural course of visual development.

A Property-Based Taxonomy of CQMs

In the absence of formal guidelines for CQM selection[2], in particular for selecting a versatile set of CQMs, we develop a property-based framework for distinguishing CQMs based on such a framework for clustering algorithms discussed in (Ackerman, Ben-David, & Loker, 2010b) (also see (Bosagh-Zadeh & Ben-David, 2009) and (Ackerman, Ben-David, & Loker, 2010a)). The framework consists of identifying natural properties of CQMs and classifying measures based on the properties that they satisfy. For the purposes of our study we use this framework to select meaningfully versatile CQMs. This taxonomy may have independent interest for choosing CQMs in other settings. Note that these properties are descriptive only, and not necessarily desirable.

Our taxonomy of CQMs follows a line of work on theoretical foundations of clustering beginning with the famous impossibility result by (Kleinberg, 2003), which showed that no clustering function can simultaneously satisfy three specific properties. (Ackerman & Ben-David, 2008) reformulate these properties in the setting of CQMs, and show that these properties are consistent and satisfied by many CQMs. We follow up on (Ackerman & Ben-David, 2008) by studying natural properties that can be used to distinguish between CQMs.

In Table 3, we present a taxonomy of our seven clustering quality measures. Each property, defined below, aims to capture some fundamental feature that is satisfied by some measures.

Normed clustering quality measures

A *clustering quality measure m* takes a domain set X, a distance function d over X, and a clustering C of X, and outputs a non-negative real number. Some quality measures are defined over normed vector spaces. *Normed CQMs* take a quadruple of the form $(V, X, C, \|\cdot\|)$, where V is a vector space, X a finite subset of V, and $\|\cdot\|$ is a norm over V. Normed CQMs can rely on centers-of-mass of clusters that are not necessarily in X, but are part of the vector-space V. Observe that the centers-of-mass are not defined for un-normed CQMs. We define the properties for CQMs in general, but one can apply any property to a normed CQM by using the norm to define the distance function. That is, set $d(x,y) = \|x-y\|$ for all $x, y \in X$.

[2]Although there are no formal guidelines for CQM selection, some interesting heuristics haven been proposed, see, for example, (Vendramin, Campello, & Hruschka, 2009).

Table 3: A taxonomy of the seven quality measures used in the study.

	Gamma	Silhouette	Dunn	Avg Within	Avg Btw	CH	W-Inter/Intra
Order-consist.	✓	*X*	*X*	*X*	*X*	*X*	*X*
Sep-invariant	*X*	*X*	*X*	✓	*X*	*X*	*X*
Hom-invariant	*X*	*X*	*X*	*X*	✓	*X*	*X*
Bounded	✓	✓	*X*	*X*	*X*	*X*	*X*
Constant Cover	*X*	*X*	✓	*X*	*X*	*X*	*X*
Norm-based	*X*	*X*	*X*	*X*	*X*	✓	*X*

Invariance and consistency properties

Invariance properties describe changes to the underlying data that do not affect the quality of a clustering. Consistency properties describe similarity conditions under which clusterings have similar quality. We propose two new invariance properties.

Definition 1 (Separation Invariance)**.** *A CQM m is* separation-invariant *if for all X and distance functions d and d' over X where $d(x,y) = d'(x,y)$ for all $x \sim_C y$, $m(C,X,d) = m(C,X,d')$.*

A separation invariant CQM is not affected by changes to between-cluster distances. Conversely, homogeneity invariant CQMs depend only on between-cluster distances, and are invariant to changes to within-cluster distances.

Definition 2 (Homogeneity Invariance)**.** *A CQM m is* homogeneity-invariant *if for all X and distance functions d and d' over X where $d(x,y) = d'(x,y)$ for all $x \not\sim_C y$, $m(C,X,d) = m(C,X,d')$.*

Observe that separation-invariance and homogeneity-invariance can also be viewed as consistency properties. An additional consistency property, order consistency, is an adaptation of an analogous property of clustering functions presented in (Jardine & Sibson, 1971). Order consistency describes CQMs that depend only on the order of pairwise distances.

Definition 3. *A CQM m is* order consistent *if for all d and d' over X such that for all $p,q,r,s \in X$, $d(p,q) < d(r,s)$ if and only if $d'(p,q) < d'(r,s)$, $m(C,X,d) = m(C,X,d')$.*

Domain and range properties

A bounded range can aid in interpreting the results of a CQM, in particular if the bounds are attainable by some clusterings.

Definition 4 (Bounded)**.** *A CQM m is* bounded *if there exist datasets X_1 over d_1 and X_2 over d_2, and clusterings C_1 of X_1 and C_2 of X_2, so that $m(C_1,X_1,d_1) \leq m(C,X,d) \leq m(C_2,X_2,d_2)$ for all C,X, and d.*

Our next property describes the quantity of domain elements that effect the CQM. First, we introduce the notion of an m-cover of a clustering, a subset of the domain which has the same quality as the entire set. For clustering C of X, and $X' \subseteq X$, let $C|X'$ denote the clustering C' of X' where for all $x,y \in X'$, $x \sim_{C'} y$ if and only if $x \sim_C y$.

An *m-cover* of clustering C of X is any set $R \subseteq X$, so that $m(X,k) = m(R,C|R)$. We define clustering quality measures that have a constant size cover for all clusterings.

Definition 5 (Bounded Cover)**.** *A CQM m has* bounded cover *if there exists a constant r so that for every data set X and clustering C of X, there exists an m-cover of C of cardinality at most r.*

CQMs that have a bounded cover search the domain space for some local features, ignoring most of the information in the dataset.

Conclusions

We perform an empirical study comparing human evaluations of clustering with formal clustering quality measures. To select a versatile set of CQMs, we develop a theoretical property-based taxonomy of CQMs. Our study shows that some CQMs with seemingly natural mathematical formulations yield evaluations that disagree with human perception. On the other hand, we identify CQMs (CH and Silhouette) that have significant correlation with human evaluations.

Our consistency analysis reveals that even novices are at least as consistent a broad set of CQMs, and perhaps more consistent. We also find significant correlations between the evaluations of expert and novice subjects. This lends support to the common practice of seeking human visual evaluations of partition quality. If one needs to evaluate a very large number of partitions it may be reasonable to use human computation via a service such as Mechanical Turk to rank partitions efficiently (or at least throw out the really bad ones). Finally, experts appear to use more sophisticated strategies than novices, indicating that training can improve human clustering evaluation performance.

Acknowledgments

This work is funded by NSF Grant #SES-0963071, Divvy: Robust and Interactive Cluster Analysis (PI Virginia de Sa). Thanks to Cindy Zhang for valuable code contributions.

References

Ackerman, M., & Ben-David, S. (2008). Measures of clustering quality: A working set of axioms for clustering. In *Advances in neural information processing systems.*

Ackerman, M., & Ben-David, S. (2009). Clusterability: A theoretical study. *Proceedings of AISTATS-09, JMLR: W&CP*, *5*, 1–8.

Ackerman, M., Ben-David, S., & Loker, D. (2010a). Characterization of Linkage-based Clustering. In *Proceedings of colt.*

Ackerman, M., Ben-David, S., & Loker, D. (2010b). Differentiating clustering paradigms: a property-based approach. In *Advances in neural information processing systems.*

Baker, F., & Hubert, L. (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, *70*(349), 31–38.

Bosagh-Zadeh, B., & Ben-David, S. (2009). A uniqueness theorem for clustering. In *Proceedings of the 25th conference on uncertainty in artificial intelligence, auai press.*

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation*, *3*(1), 1–27.

Dunn, J. (1974). Well-separated clusters and optimal fuzzy partitions. *Cybernetics and Systems*, *4*(1), 95–104.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382.

Jardine, N., & Sibson, R. (1971). *Mathematical taxonomy*. John Wiley and Sons, Inc., New York.

Kleinberg, J. (2003). An impossibility theorem for clustering. In *Advances in neural information processing systems 15: Proceedings of the 2002 conference* (p. 463).

Landis, J. R., & Koch, G. G. (1977, March). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lewis, J. M. (2009). Finding a better k: A psychophysical investigation of clustering. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (p. 315-320).

Milligan, G. (1981). A Monte-Carlo study of 30 internal criterion measures for cluster-analysis. *Psychometrika*, *46*, 187–195.

Ng, A. Y., Jordan, M., & Weiss, Y. (2002). On spectral clustering: analysis and an algorithm. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14* (pp. 849–856). Cambridge, MA: MIT Press.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Santos, J., & Sá, J. M. de. (2005). *Human clustering on bi-dimensional data: An assessment* (Tech. Rep. No. 1). INEB Instituto de Engenharia Biomédica, Porto, Portugal. Available from `http://www.di.ubi.pt/~lfbaa/entnetsPubs/JMS_TechReport2005_1.pdf`

Strehl, A. (2002). Relationship-based clustering and cluster ensembles for high-dimensional data mining.

Vendramin, L., Campello, R., & Hruschka, E. (2009). *On the comparison of relative clustering validity criteria.* Sparks.