2013 IEEE 13th International Conference on Data Mining

Classification-Based Clustering Evaluation

John S. Whissell, Charles L.A. Clarke
David R. Cheriton School of Computer Science
University of Waterloo, Waterloo, Ontario, N2L3G1
Email: {jswhisse,claclark}@uwaterloo.ca

***Abstract*—The evaluation of clustering quality has proven to be a difficult task. While it is generally agreed that application-specific human assessment can provide a reasonable gold standard for clustering evaluation, the use of human assessors is not practical in many real situations. As a result, machine computable internal clustering quality measures (*CQMs*) are often used in the evaluation process. However, CQMs have their own drawbacks. Despite their extensive use in clustering research and applications, many CQMs have been shown to lack generality. In this paper we present a new CQM with general applicability. The basis of our CQM is a pattern recognition view of clustering's purpose: *the unsupervised prediction of behavior from populations*. This purpose translates naturally into our new classifier based CQM which we refer to as *informativeness*. We show that informativeness can satisfy core CQM axioms defined in prior research. Additionally, we provide experimental support, showing that informativeness can outperform many established CQMs by detecting a larger variety of meaningful structures across a range of synthetic datasets, while at the same time exhibiting good performance on each individual dataset. Our results indicate that informativeness provides a highly general and effective CQM.**

***Keywords*-clustering methods;**

I. Introduction

Despite the popularity of clustering in data mining research and applications, it is an ill-posed problem [1]. In particular, evaluating clustering quality has been shown to be a complicated and confusing task. For a given data set, it is often unclear what it means for one clustering to be better than another.

Issues like these, combined with the recent demand for big data analytics, have driven an increasing amount of research towards clustering evaluation. Even though there is a large volume of research on clustering evaluation [2]–[7], it is widely accepted that the best way to evaluate a clustering is through application-specific human assessment. If a clustering helps with whatever task an individual has at hand, then it is a good clustering for them.

Judging clusterings situationally with human assessors would be sensible as it focuses on a clustering's intended use(s), but it is often not practical for various reasons. To account for this issue, internal clustering quality measures (*CQMs*), requiring no external input about what is expected from the clustering, are often used to aid people in selecting good clusterings. CQMs have their own issues. They have been found lacking in terms of generality, and recent theoretical works have highlighted other issues with them.

It seems that the design of a universally applicable CQM may not be possible. However, that does not mean that we cannot design a CQM that is more generally applicable than those currently used, while still being effective in individual situations. In this paper we present and evaluate a CQM that meets this goal.

The basis of our new CQM is a pattern recognition view of clustering's purpose: *the unsupervised prediction of behavior from populations* [1]. We present a classifier-based CQM, which we refer to as *informativeness*, which is a natural translation of this purpose to the domain of CQMs. We show that informativeness can satisfy CQM axioms defined in previous research [3]. Additionally, we analyze experimental clustering of synthetic datasets with various structures, which together suggest that informativeness is highly general, while at the same time being effective on individual datasets.

To the best of our knowledge, the use of classifiers in the evaluation of clusterings has scarcely been directly examined. Our previous work is one of the very few on the subject [8]. Distinct from our simplistic use of classifiers there, here we present a formal CQM that uses multiple classifiers in evaluating clusterings.

The rest of this paper proceeds as follows. Section II defines the notation used in this paper. In Section III, we present *informativeness*, our new CQM that uses classifiers. Section IV shows that informativeness can satisfy CQM axioms from previous research. In Section V we compare informativeness against several well-known CQMs using clusterings on synthetic datasets of varying structures. The results of this experiment indicate that informativeness is more general than the competitors we considered, while at the same time being effective on individual datasets. Section VI provides a conclusion and discussion of future work.

II. Preliminaries

This section defines the notation used in this paper. Let X be a dataset of n objects, and x_i be the ith object of X.

A *clustering* C over X is a partitioning of X into k disjoint sets referred to as *clusters*, where c_i is the ith cluster of C. The number of objects in c_i is denoted as $|c_i|$, and

1550-4786/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDM.2013.28

$p(c_i) = \frac{|c_i|}{n}$. The cluster x_i belongs to in C is denoted as $c(x_i)$. A clustering is *trivial* if it has only one cluster, or if every cluster contains one object.

For a pair of objects $x_i, x_j \in X$, $x_i \sim_C x_j$ denotes that x_i and x_j are members of the same cluster in C, and $x_i \nsim_C x_j$ denotes that they are in different clusters in C. A distance function is denoted as d, where $d(x_i, x_j)$ is that function's distance measurement between x_i and x_j. A C over (X, d) refers to C being a clustering of X, where X's object pair distances are defined using d.

A classifier type is denoted as f, where $f_{\text{C,X}}$ is f, trained on all of C and X, and $f_{\text{C,X,v}}(x_i)$ is the predicted label for an $x_i \in X$, obtained through v-fold cross validation training of f on C and X.

Finally, a CQM is a function of the form $M(C, X, *) \rightarrow \mathbb{R}$, where $*$ is the additional parameters taken by M.

III. Informativeness

In this section we present our new CQM. Recall from the introduction that the basis of our new CQM is the notion that clustering's purpose is the unsupervised prediction of behavior from populations. While prediction suggests some form of classification, there are typically no labels during clustering, preventing classification's use. However, in the context of a CQM, which is computed after a clustering is complete, we do in-fact have a labeling that a classifier can use—the clustering itself. Given this, and the prediction oriented basis of our CQM, we argue that it is natural to use classifiers in it. The formal definition of our classifier based CQM, which refer to as *informativeness*, is given in Algorithm 1.

Algorithm 1 Informativeness

Input: int v, clustering C, dataset X,
  classifier types f^*
$\text{A}_{C,X,f^*,v} \leftarrow 0$
for all $f \in f^*$ **do**
  for $i \leftarrow 1$ to k **do**
    $r_{f_{\text{C,X,v}}}(c_i) \leftarrow \frac{|\{x_j \in X : x_j \in c_i \wedge c_i = f_{\text{C,X,v}}(x_j)\}|}{n}$
  end for
  $\text{A}_{C,X,f,v} \leftarrow -\sum_{i=1}^{k} r_{f_{\text{C,X,v}}}(c_i) \log(p(c_i))$
  $\text{A}_{C,X,f^*,v} \leftarrow \max(\text{A}_{C,X,f^*,v}, \text{A}_{C,X,f,v})$
end for
$\text{H}(C) \leftarrow -\sum_{i=1}^{k} p(c_i) \log(p(c_i))$
$\text{I}(C, X, f^*, v) \leftarrow \frac{\text{A}_{C,X,f^*,v} - \frac{\text{H}(C)}{k}}{\frac{(k-1)\text{H}(C)}{k}}$
return $\text{I}(C, X, f^*, v)$

The value $\text{I}(C, X, f^*, v)$ is the informativeness of C. The general process of informativeness is: *measure how well each classifier type can predict population behavior using the clustering, and take the quality of the best prediction from this as representing the clustering's quality.*

The $r_{f_{\text{C,X,v}}}(c_i)$ values are the fraction of objects in X that are correctly assigned to each cluster when using v-fold cross validation labeling. They are used in computing $\text{A}_{C,X,f,v}$ values, where each $\text{A}_{C,X,f,v}$ is informativeness' estimation of how well $f_{\text{C,X}}$ will predict population behavior. Note that it is principled to use our crossfold validation labels for this prediction measure because clusterings are typically considered to be independent and identically distributed samples of their populations.

The definition of $\text{A}_{C,X,f,v}$ has an encoding interpretation to it. Consider the process of repeatedly classifying unseen members of the population X was drawn from using $f_{\text{C,X}}$. Imagine the output of this process as a stream of cluster ids from C. Then to minimize the stream size over infinitely many classifications, each cluster id should be assigned a code of length $-\log(p(c_i))$. Now, consider that only some of the cluster ids $f_{\text{C,X}}$ produces will be correct. The expected number of correct bits that we will receive from reading a single cluster id from the stream produced by $f_{\text{C,X}}$ on the population X was drawn from is exactly $\text{A}_{C,X,f,v}$. In essence, it is a direct measure of the amount of population information C gives us. We argue it is an intuitive way to measure C's prediction quality, as we aimed to do.

Another aspect of note in informativeness is the use of multiple classifiers, with only the best performing (the one with the highest $\text{A}_{C,X,f,v}$) being used in $\text{A}_{C,X,f^*,v}$. Our motivation for this feature was keeping informativeness as general as possible. Classifier types are suited to specific structures, so using many of them, and taking only the best result, allows informativeness to detect more types of clusterings as high quality, increasing its generality. This is in contrast to typical CQMs, which will fail to detect clusterings as high quality that do not match their singular notion of what makes a clustering good. Admittedly, this generality comes at a cost—using more classifier also means that informativeness' effectiveness may diminish with respect to specific structures. However, our experiment suggests that judicious selection of classifiers can mitigate this problem.

Line 12 is a correction of $\text{A}_{C,X,f^*,v}$ for chance. The general form of the correction that we used is that of the widely adopted Hubert and Arabie formula [4]. For $\text{A}_{C,X,f^*,v}$, the formula translates to:

$$\frac{\text{A}_{C,X,f^*,v} - \text{E}[\text{A}_{C,X,f^*,v}]}{\max(\text{A}_{C,X,f^*,v}) - \text{E}[\text{A}_{C,X,f^*,v}]}, \tag{1}$$

where $\text{E}[\text{A}_{C,X,f^*,v}]$ is the expected $\text{A}_{C,X,f^*,v}$ by chance. Defining a random classifier as one that places objects in each cluster with equal likelihood, and using that as our measure of $\text{E}[\text{A}_{C,X,f^*,v}]$, we have:

$$\text{E}[\text{A}_{C,X,f^*,v}] = -\sum_{i=1}^{k} \frac{p(c_i)}{k} \log(p(c_i)) = \frac{\text{H}(C)}{k}, \tag{2}$$

where:

$$\text{H}(C) = -\sum_{i=1}^{k} p(c_i) \log(p(c_i)), \tag{3}$$

and:

$$\max(\mathrm{A}_{C,X,f^*,v}) = \mathrm{H}(C). \tag{4}$$

Substituting Eq. 2 and 4 into Eq. 1, we obtain:

$$\frac{\mathrm{A}_{C,X,f^*,v} - \frac{\mathrm{H}(C)}{k}}{\frac{(k-1)\mathrm{H}(C)}{k}} = \mathrm{I}(C, X, f^*, v), \tag{5}$$

As a final note with respect to this correction, when multiple clusterings have the same $\mathrm{I}(C, X, f^*, v)$, the one(s) with the most clusters should be considered better. This is due to them scoring the same despite having to deal with a more complicated classification problem (i.e., more clusters/classes), which in turn suggests a clearer clustering structure. We apply this notion in our experiment in Section V.

The final aspect of our implementation that we discuss is its time complexity. In that respect, we note that it is based on the classifier types and the v parameter. For our experiment in Section V, we were able to compute all the informativeness results in less than one day on a single computer. This suggests that informativeness is fast enough to run in practical situations.

Proving general superiority of a CQM is likely not possible for the same reason that designing a universal CQM is not. However, one can provide evidence that may suggest it. In that regard, we have already discussed informativeness' basis, how it can account for quality by chance, and its generality. In the following sections, we will provide more evidence. We will show that informativeness can satisfy CQM axioms suggested in previous research. Additionally, we will provide an experiment on synthetic datasets that will show that it can behave in a highly general yet effective manner. Together, these results provide a strong motivation for informativeness' use in practice.

IV. Clustering Quality Measure Axioms

Ackerman and Ben-David [3] suggest four axioms that all CQMs should satisfy. While they are careful to note that satisfying their axioms does not prove that a CQM is reasonable, they provide a strong case for the converse, i.e., if a CQM fails to satisfy them it is unreasonable. It is therefore useful to show that informativeness can satisfy them. In this section we will provide a theorem for each of the four axioms that proves that informativeness can satisfy it.

For our proofs, we consider informativeness when it uses only an r-nearest neighbor classifier (*RNN*) and leave-one-out crossfold validation. We denote the application of informativeness with these settings to a C over X, where the RNN classifies using d, as $\mathrm{I}_{\text{RNN}}(C, X, d)$. We assume that there are no tied object pair distances in X, and further that every object has a unique $c \in C$ which has the most members in the object's r-nearest neighborhood. Note that these restrictions/parameters for informativeness are used only to yield clearer proofs in this work. We have similar proofs for many different parameterized versions of informativeness.

Before presenting our axiom proofs, we define the crossfold validation labeling behavior of an RNN.

Definition 1 (Crossfold validation labels from an RNN)**.** *Let f be an RNN, trained on C and X with leave-one-out crossfold validation and some d. Then its labeling function for all $x_i \in X$ is:*

$$f_{C,X,n}(x_i) \leftarrow \arg\max_{c_j \in C} \sum_{l=1}^{r} \alpha(c_j, nn(x_i)_l), \tag{6}$$

where $nn(x_i)$ is a list of all $x_j \in X$ other than x_i, ordered from least to greatest $d(x_i, x_j)$, $nn(x_i)_l$ is the lth object in this list, and α is an indicator function of the form:

$$\alpha(c_i, x_j) = \begin{cases} 1, & \text{if } c(x_j) = c_i \\ 0, & \text{otherwise.} \end{cases}$$

We now give our proofs.

A. Scale Invariance

Definition 2 (Scale Invariance [3])**.**
CQM M satisfies scale invariance *if, for any C over (X, d), and every positive number λ, we have $M(C, X, d) = M(C, X, \lambda d)$.*

Lemma 1 (Informativeness' score for a C over X, when using f^*, is based only on the cluster sizes ($|c_i|$s), and the frequency of correct label predictions for each c_i by each $f \in f^*$ ($r_{f_{C,X,v}}(c_i)$s))**.**
Proof. *This follows directly from the definition of $I(C, X, f^*, v)$ in Algorithm 1.* □

Theorem 1 (I_{RNN} satisfies scale invariance)**.**
Proof. *The $|c_i|$ values remain unchanged in a test for scale invariance. Further, multiplying all distances by a uniform amount does not change ordering of nearest neighbors for any object. By Definition 1, this means predicted labels do not change either. Given this, and Lemma 1, it follows that $I_{RNN}(C, X, d) = I_{RNN}(C, X, \lambda d)$.* □

B. Weak Local Consistency

Definition 3 (C-Weakly Locally Consistent Variant [3])**.** *Distance function d' is a C-*weakly locally consistent variant *for a C over (X,d), if the following properties hold:*

1. *For all $c_i \in C$ there exists a constant $\lambda \leq 1$ such that for all $x_j, x_k \in c_i$ we have $d(x_j, x_k) \geq \lambda d'(x_j, x_k)$.*
2. *For all x_i, x_j in different clusters, we have $d(x_i, x_j) \leq d'(x_i, x_j)$.*
3. *There exists some set R, where R contains exactly one object from every cluster in C, such that for some constant $\lambda \geq 1$, for all $x_i, x_j \in R$ we have $d(x_i, x_j) \geq \lambda d'(x_i, x_j)$.*

Definition 4 (Weak Local Consistency [3])**.** *CQM M satisfies* weak local consistency *if, for any C over (X, d), and*

C-weakly locally consistent variant *of* d *denoted as* d'*, we have* $M(C, X, d) \leq M(C, X, d')$.

The proof that $\mathrm{I}_{\mathrm{RNN}}$ satisfies weak local consistency requires first showing that adding a classifier f to informativeness' computation never increases its score more than when adding a classifier f' such that when trained, f' makes at least as many correction label predictions as f does for every cluster in C.

Lemma 2 (For a C over X, let f and f' be classifier such that $\forall_{c_i \in C} r_{f_{C,X,v}}(c_i) \leq r_{f'_{C,X,v}}(c_i)$. Then $\forall_{f^*} \mathrm{I}(C, X, f^* \cup f, v) \leq \mathrm{I}(C, X, f^* \cup f', v)$)**.**
Proof. *Given Lemma 1, and the use of the* max *function in selecting which* $A_{C,X,f,v}$ *to use in* $I(C, X, f^*, v)$*, it suffices to show that* $A_{C,X,f,v} \leq A_{C,X,f',v}$*. Let* $\triangle_{c_i} = r_{f'_{C,X,v}}(c_i) - r_{f_{C,X,v}}(c_i)$*. Then:*

$$\begin{aligned} A_{C,X,f',v} &= -\sum_{i=1}^{k} r_{f'_{C,X,v}}(c_i) \log(p(c_i)) \\ &= -\sum_{i=1}^{k} (r_{f_{C,X,v}}(c_i) + \triangle_{c_i}) \log(p(c_i)) \\ &= A_{C,X,f,v} - \sum_{i=1}^{k} \triangle_{c_i} \log(p(c_i)). \end{aligned}$$

As $-\sum_{i=1}^{k} \triangle_{c_i} \log(p(c_i)) \geq 0$*, we have* $A_{C,X,f,v} \leq A_{C,X,f',v}$. □

Theorem 2 ($\mathrm{I}_{\mathrm{RNN}}$ satisfies weak local consistency)**.**
Proof. *The* $|c_i|$ *values remain unchanged in a test for weak local consistency. Given this, and Lemma 1 and 2, it suffices to show that for any* C *over* (X, d)*, and* d' *that is a* C*-weakly local consistent variant of* d*, all correct classifications made using* d *are correct when using* d'*. If* C *consists of a single cluster, this is trivially true. Otherwise, for some correctly classified object* x_i *let* x_j *and* x_l *be elements of* X *such that* $x_i \sim_C x_j$*,* $x_i \nsim_C x_l$*, and:*

$$d(x_i, x_j) < d(x_i, x_l). \tag{7}$$

From the definition of a C*-weakly locally consistent variant we have:*

$$d(x_i, x_j) \geq \lambda d'(x_i, x_j)$$

and:

$$d(x_i, x_l) \leq d'(x_i, x_l).$$

Setting $\lambda = 1.0$*, and merging these two inequalities with Eq. 7, we obtain:*

$$d'(x_i, x_j) \leq d(x_i, x_j) \leq d(x_i, x_l) < d'(x_i, x_l). \tag{8}$$

In order for it to be possible for some $x_i \in X$ *to become incorrectly classified when using* d' *we must find an* $x_j \sim_C x_i$*, where* x_j *is the* u*th element of* $nn(x_i)$*, and an* $x_l \nsim_C x_i$*, where* x_l *is* u'*th element* $nn(x_i)$*,* $u' \geq u$*, such that* $d(x_i, x_j) < d(x_i, x_l)$ *and* $d'(x_i, x_j) > d'(x_i, x_l)$*. However, Eq. 8 shows such a pair of objects cannot exist. Given this, a correctly classified object using* d *is correctly classified when using* d'*.* □

C. Co-final Richness

Definition 5 (C-Consistent Variant [3])**.** *Distance function* d' *is a* C-consistent variant *for a* C *over* (X,d) *if for all* $x_i, x_j \in X$*, when* $x_i \sim_C x_j$ *we have* $d(x_i, x_j) \geq d'(x_i, x_j)$*, and when* $x_i \nsim_C x_j$ *we have* $d(x_i, x_j) \leq d'(x_i, x_j)$*.*

Definition 6 (Co-final Richness [3])**.** *CQM* M *satisfies co-final richness if, for any non-trivial pair of clusterings* C *over* (X, d) *and* C' *over* (X, d')*, there exists a* C*-consistent variant of* d*, denoted as* d''*, such that* $M(C, X, d'') \geq M(C', X, d')$*.*

Theorem 3 ($\mathrm{I}_{\mathrm{RNN}}$ satisfies co-final richness if the size of each cluster in the clusterings it is used on are always at least $r + 1$)**.**
Proof. *For any* C *over* (X, d)*, a* C*-consistent variant of* d*, denoted as* d''*, can be can defined in the following manner:*

$$d''(x_i, x_j) = \begin{cases} 0, & \text{if } x_i = x_j \\ \max\limits_{x_i, x_j \in X} 2d(x_i, x_j), & \text{if } x_i \nsim_C x_j \\ \min\limits_{x_i, x_j \in X} \frac{1}{2} d(x_i, x_j), & \text{otherwise.} \end{cases} \tag{9}$$

Based on this definition of d''*, when computing* $I_{RNN}(C, X, d'')$ *the* r *nearest neighbors of every* $x_i \in X$ *will share the same label as* x_i*, ensuring correct classification. This gives* $I_{RNN}(C, X, d'') = 1$*, the maximum possible. Therefore we have* $I_{RNN}(C, X, d'') \geq I_{RNN}(C', X, d')$*.* □

D. Isomorphism Invariance

Definition 7 (Clustering Isomorphism [3])**.** *Clusterings* C *and* C' *over* (X, d) *are isomorphic, denoted as* $C \approx_d C'$*, if there exists a distance preserving isomorphism* $\phi : X \to X$ *such that* $x_i \sim_C x_j$ *if and only if* $\phi(x_i) \sim_{C'} \phi(x_j)$*.*

Definition 8 (Isomorphism Invariance [3])**.** *CQM* M *satisfies isomorphism invariance if for any* C *and* C' *over* (X, d)*, where* $C \approx_d C'$*, we have* $M(C, X, d) = M(C', X, d)$*.*

Theorem 4 ($\mathrm{I}_{\mathrm{RNN}}$ satisfies isomorphism invariance)**.**
Proof. *For any* C *and* C' *over* (X, d)*,* $C \approx_d C'$ *implies a mapping* ϕ' *from clusters in* C *to* C' *exists, such that* $|\phi'(c_i)| = |c_i|$*, and the labeling behavior of* $f_{C,X,v}$ *is:*

$$f_{C,X,n}(\phi(x_i)) \leftarrow \arg \max_{\phi'(c_j) \in C'} \sum_{l=1}^{r} \alpha(\phi'(c_j), nn(\phi(x_i))_l).$$

This gives:

$$\forall_{\phi'(c_i) \in C'} r_{f_{C,X,v}}(c_i) = r_{f_{C,X,v}}(\phi'(c_i)).$$

It then follows from Lemma 1 that we have $I(C, X, d) = I(C', X, d)$ *when* $C \approx_d C'$*.* □

In the context of the entire paper, our axiom proofs some additional motivation for informativeness' use in practice.

V. Synthetic Dataset Experiment

In our synthetic dataset experiment we compared informativeness to other CQMs on clusterings of synthetic datasets with a variety of structures. The synthetic datasets we used are detailed in Section V-A, the clustering algorithms in Section V-B, the CQMs we compared informativeness against in Section V-C, and the classifiers used by informativeness in Section V-D.

For the experiment, we generated 50 instances of each dataset and clustered each instance with each clustering algorithm using from two to 20 clusters, giving 4750 (50*5*19) clusterings of each dataset. We then computed each CQM for each clustering.

In Section V-E we analyze the results of our experiment using a classical number of clusters estimation approach [9].

A. Datasets

We used five synthetic datasets in our experiment, each of which is detailed below.

6GAUSS consisted of six Gaussian clusters with identity covariance, each with 500 points in five dimensions. Their means were randomly assigned a value from zero to 10 in each dimension. Cluster means were required to be at least four Euclidean distance apart, and points were required to within two Euclidean distance of their cluster mean.

PAIRED consisted of three pairs of Gaussian clusters with identity covariance, each with 500 points in five dimensions. Each pair of Gaussians was placed around a mean with a randomly assigned value in each dimension from zero to 20 such that the Euclidean distance between paired Gaussian clusters was between four and eight, and the Euclidean distance between non-paired Gaussians was at least 12. Additionally, points were required to be within two Euclidean distance of their cluster mean.

ELONG consisted of five Gaussian clusters with identity covariance, each with 300 points in five dimensions. Their means were randomly assigned a value from zero to 50 in each dimension. To create elongated clusters in different dimensions, we multiplied the values of a single, distinct dimension for each cluster by 15. Cluster means were required to be at least five Euclidean distance apart.

UNIFORM consisted of eight clusters, each with 300 points in three dimensions. Each cluster had its points uniformly distributed in a 3x3x3 box around a randomly assigned center in a 10x10x10 cube. Cluster centers were required to be five Euclidean distance apart.

RINGS consisted of 2 ring clusters centered around (0,0), a larger outer ring with radius 2 and a smaller inner ring of radius 1. 400 points were evenly spaced by degrees on the inner ring. A random noise component between 0 and 0.1 was then added to the x and y coordinates of all the points. The outer ring was created in a similar fashion, except 1200 points were used.

B. Clustering Algorithms

We used five well-known clustering algorithms in our experiment: k-means [10], repeated bisecting k-means, UPGMA [11], complete linkage [12], and single linkage [12]. Our k-means algorithm used Lloyd's method [10], with the initial centroids being selected randomly from objects in the dataset. Our implementation of repeated bisecting k-means split the largest remaining cluster in two using our k-means algorithm, until the desired number of clusters was reached. Finally, we used Euclidean distance for UPGMA, complete linkage, and single linkage.

C. Competing Evaluation Measures

We compared informativeness against four well-known and studied CQMs; *silhouette width* [13] (*SW*), the *Davies-Bouldin index* [14] (*DB*), the *Calinski-Harabasz index* [15] (*CH*), and the *Dunn index* [16] (*DN*).

D. Classification Algorithms

For efficiency reasons we restricted ourselves to three classifier types: a five nearest neighbor classifier, a C4.5 decision tree, and a Rocchio classifier. For the latter two, we used Weka[1] implementations. We implemented the Rocchio classifier ourselves. For parameters, our Rocchio classifier used Euclidean distance. For the other two, we used their default parameter settings in Weka. Ten-fold cross validation was when computing informativeness.

E. Results and Discussion

To analyze each CQM's behavior with respect to picking the optimal number of clusters, we grouped our clusterings of each dataset by sample. Then, for each sample, we recorded the number of clusters in the optimal scoring clustering for each CQM. We also recorded the adjusted mutual information [7] of those clusterings with their true labelings. This provided us with a measure how good the clusterings actually were that was more robust than simply counting the number of clusters. Table I gives number of clusters estimations, AMI from those estimations, and a Tukey's honestly significantly different test on the AMI values.

It is clear from Table I that no CQM had significantly higher AMI than informativeness for any dataset we tested. On the other hand, all the other CQMs had significantly worse average AMI than informativeness for two or more of the datasets. The actual magnitude of the differences was often substantial as well. Based on these results, we concluded that informativeness was the best at selecting the single optimal clustering for datasets in our experiment. This suggests that informativeness might be of use in real clustering applications.

[1] http://www.cs.waikato.ac.nz/ml/weka/

Table I

FREQUENCY OF NUMBER OF CLUSTER ESTIMATIONS BY EACH CQM FOR EACH DATASET, ALONG WITH THE AVERAGE AMI FROM THE CLUSTERINGS USED IN THOSE ESTIMATIONS. BOLDED VALUES DENOTE THE COLUMN FOR THE TRUE NUMBER OF CLUSTERS IN EACH DATASET. WE INCLUDE THE RESULTS OF A TUKEY'S HONESTLY SIGNIFICANTLY DIFFERENT (*HSD*) TEST ON THE AMI VALUES WITH $p = 0.01$. AN O INDICATES THAT THE ROW IS SIGNIFICANTLY BETTER THAN THE COLUMN, AN X INDICATES THE OPPOSITE, AND A $-$ INDICATES NO STATISTICALLY SIGNIFICANT DIFFERENCE.

	Number of Clusters Estimations									Tukey's Test			
6GAUSS	2	3	4	5	**6**	7	8	9+	AMI	**SW**	**DB**	**CH**	**DN**
Inf.	0	0	0	0	**49**	1	0	0	.999	-	-	-	O
SW	0	1	2	5	**42**	0	0	0	.983		-	-	-
DB	1	2	2	12	**33**	0	0	0	.958			-	-
CH	0	0	0	0	**50**	0	0	0	1				O
DN	3	3	2	0	**42**	0	0	0	.950				
PAIRED	2	3	4	5	**6**	7	8	9+	AMI	**SW**	**DB**	**CH**	**DN**
Inf.	0	1	10	20	**19**	0	0	0	.911	O	O	O	O
SW	0	50	0	0	**0**	0	0	0	.783		-	-	-
DB	0	50	0	0	**0**	0	0	0	.783			-	-
CH	0	48	0	0	**2**	0	0	0	.791				-
DN	1	49	0	0	**0**	0	0	0	.779				
ELONG	2	3	4	**5**	6	7	8	9+	AMI	**SW**	**DB**	**CH**	**DN**
Inf.	0	1	4	**9**	13	12	4	7	.910	-	O	O	-
SW	0	1	11	**25**	1	1	2	9	.916		O	O	-
DB	0	0	1	**2**	1	2	2	42	.808			O	-
CH	0	0	0	**0**	0	0	0	50	.738				X
DN	8	5	5	**22**	6	1	0	3	.851				
UNIFORM	2	3	4	5	6	7	**8**	9+	AMI	**SW**	**DB**	**CH**	**DN**
Inf.	0	0	0	0	0	0	**47**	3	.997	-	-	-	O
SW	0	0	0	0	0	0	**50**	0	1		-	-	O
DB	0	0	0	0	0	0	**50**	0	1			-	O
CH	0	0	0	0	0	0	**50**	0	1				O
DN	27	5	5	1	2	1	**2**	7	.590				
RINGS	**2**	3	4	5	6	7	8	9+	AMI	**SW**	**DB**	**CH**	**DN**
Inf.	**22**	7	3	10	2	3	2	1	.670	O	O	O	O
SW	**0**	0	0	0	0	6	34	10	.527		-	-	O
DB	**0**	0	0	0	0	0	0	50	.494			-	O
CH	**0**	0	0	0	0	0	0	50	.431				O
DN	**21**	12	0	0	0	11	4	2	.181				

VI. CONCLUSION

In this paper we presented *informativeness*, a novel CQM based on the notion that clustering's purpose is the prediction of behavior from populations. We adapted classifiers to estimate the quality of this prediction for an individual clustering in informativeness, accounting for chance as well as generality in our implementation. Additionally, we showed that informativeness can satisfy CQM axioms defined in previous research, and that it performs better overall than a number of well-known CQMs on synthetic datasets of varied structures. Together, our implementation, results, and analysis suggest that informativeness can be of genuine use in practical clustering situations.

As future work on informativeness, we aim to study its behavior when using a variety of classifiers. In particular, we would like to identify a set of classifiers that makes it highly general and effective, but also fast to compute. We also plan to investigate if certain classifiers are particularly suited to use with informativeness in particular domains. Finally, we are currently investigating the use of informativeness in real applications.

REFERENCES

[1] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, pp. 651–666, 2010.

[2] M. Ackerman and S. Ben-David, "Clusterability: A theoretical study," in *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, 2009, pp. 1–8.

[3] ——, "Measures of clustering quality: A working set of axioms for clustering," in *Neural Information Processing Systems*, 2008, pp. 121–128.

[4] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, 1985.

[5] J. Kleinberg, "An impossibility theorem for clustering," in *Neural Information Processing Systems*, 2002, pp. 446–453.

[6] T. Lange, M. L. Braun, V. Roth, and J. M. Buhmann, "Stability-based model selection," in *Neural Information Processing Systems*, 2003.

[7] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Is a correction for chance necessary?" in *Proceedings of the 26th International Conference on Machine Learning*, 2009.

[8] J. S. Whissell, C. L. A. Clarke, and A. Ashkan, "Clustering web queries," in *Proceedings of the 18th ACM International Conference on Information and Knowledge Management*, 2009, pp. 899–908.

[9] G. W. Milligan, "A monte carlo study of thirty internal criterion measures for cluster analysis," *Psychometrika*, vol. 46, pp. 187–199, 1981.

[10] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, pp. 129–137, 1982.

[11] L. Kafuman and P. Rousseeuw, *Finding groups in data: An introduction to cluster analysis*. Wiley (New York), 1990.

[12] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Computing Reviews*, vol. 31, pp. 264–323, 1999.

[13] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[14] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, pp. 224–227, 1979.

[15] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications and Statistics*, vol. 1, pp. 1–27, 1974.

[16] J. C. Dunn, "Well separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, pp. 95–104, 1974.