RJ 10219 October 5, 2001
Computer Science

Research Report

AN INFORMATION-THEORETIC EXTERNAL CLUSTER-VALIDITY MEASURE

Byron E. Dom

LIMITED DISTRIBUTION NOTICE

IBM Research Division
Yorktown Heights, New York • San Jose, California • Zurich, Switzerland

AN INFORMATION-THEORETIC EXTERNAL CLUSTER-VALIDITY MEASURE

Byron E. Dom

October 5, 2001

ABSTRACT:

In this paper we propose a measure of similarity/association between two partitions of a set of objects. Our motivation is the desire to use the measure to characterize the quality or accuracy of clustering algorithms by somehow comparing the clusters they produce with "ground truth" consisting of classes assigned to the patterns by manual means or some other means in whose veracity there is confidence. Such measures are referred to as "external". Our measure also allows clusterings with different numbers of clusters to be compared in a quantitative and principled way. Our evaluation scheme quantitatively measures how useful the cluster labels of the patterns are as predictors of their class labels. When all clusterings to be compared have the same number of clusters, the measure is equivalent to the mutual information between the cluster labels and the class labels. In cases where the numbers of clusters are different, however, it computes the reduction in the number of bits that would be required to encode (compress) the class labels if both the encoder and decoder have free access to the cluster labels. To achieve this encoding the estimated conditional probabilities of the class labels given the cluster labels must also be encoded. These estimated probabilities can be seen as a "model" for the class labels and their associated code length as a "model cost". In addition to defining the measure we compare it to other commonly used external measures and demonstrate its superiority as judged by certain criteria.

Keywords: clustering, clustering accuracy, clustering validity

2. The Evaluation Problem

Evaluation of the *validity*[3] (quality or accuracy) of the output of clustering algorithms is a difficult problem in general. Measures or indices of cluster validity can be divided into two types: *external* and *internal*[9]. External validity criteria measure how well the clustering results match some prior knowledge about the data. It is assumed that this information is not, in general, computable from $\mathbf{X}$. Perhaps the most common form of external information is a set of classes (categories) and class labels for the objects corresponding to $\mathbf{X}$. These are usually obtained via manual classification.

The use of some measure based solely on the feature data $\mathbf{X}$ (an *internal* measure), begs the question: Why not just use this measure itself as an objective function for clustering? This may in fact be possible in some cases where the objective function used does exactly capture what is desirable in a particular application and there is a feasible algorithm for finding the optimal clustering. In such cases the evaluation problem is moot. In other cases, of course, the answer to this question may have to do with computational feasibility - it may not be possible to devise an algorithm to efficiently find the associated optimal clustering.

In many (if not most) applications, however, clustering algorithms attempt to do what humans can do quite well, albeit slowly relative to the speed of a computer. This human sufficiency is especially true in the case of document clustering, for example, where natural language understanding and vast amounts of world knowledge (or specialized domain knowledge) are used by humans. In such applications the best accuracy/quality measure will therefore be based on human subjective judgments. One way to obtain this is to ask humans to judge the quality of the results directly. This is an expensive and time consuming process however and every algorithm (or variation of a single algorithm) tried will require a new set of subjective judgments.

An alternative to this is to ask humans to cluster the data set into what they consider to be an appropriate set of clusters. This is done *once* to obtain a set C of what we will refer to as *class* labels: $C = \{c_i | i = 1, 2, \ldots, n\}$, $c_i \in \mathcal{C}$. The idea is that the intended users of the algorithm would be quite happy if the algorithm had produced these classes as clusters. They are thus treated as the ideal clustering and quality is judged based on some measure of how well the cluster labels produced by the algorithm(s) agree with the class labels. Any accuracy assessment based on this notion is thus measuring the quality of the clustering *relative* to the particular classification represented by C. Another classification will obviously result in a different measure in general. Despite this weakness, *external* measures tend to be the most reliable and are therefore usually preferable when class labels are available[4]. In this paper we propose an external validity measure appropriate for flat (non-hierarchical) clustering where a ground-truth classification is available for evaluation purposes.

[3] This term is used by different workers in slightly different ways. We use it only to refer to a measure applied to the results after a clustering algorithm has done its job; not something that can be used in the process of clustering.

[4] See Appendix C for further discussion of external versus internal validity measures.

As in the case of the cluster labels, we can think of C as a sample drawn from a population described by a probability distribution $\{p(c)\}$. Also, we can think of the set of pairs $\{(c_i, k_i)\}$ associated with C and K as a sample drawn from a population described by the distribution $\mathcal{P} \equiv \{p(c,k)\}$.

3. Summaries of the Class-Cluster Relationship

A *complete* characterization of the behavior of a particular algorithm when used to cluster a given data set is, of course, contained in the individual objects (e.g. documents) themselves i.e. which objects were assigned to which clusters. Some amount of anecdotal evidence of this type is invaluable in diagnosing the behavior of the clustering algorithm. For large numbers of objects, however, the objects in aggregate are more than can be dealt with in this manner. Some reduced information is essential. For a *partitional* clustering the usual first level of reduction is expressed by the two-dimensional contingency table $\mathcal{H} \equiv \{h(c,k)\}$, where $h(c,k)$ is the number of objects labeled class c that are assigned to cluster k by f. In a perfect (from an external measure's point of view) clustering $\mathcal{H}$ is a square matrix (i.e. $|\mathcal{C}| = |\mathcal{K}|$) and only one non-zero element per row/column. This will be diagonal if associated classes and clusters are given the same label. Two associated definitions are the one-dimensional *marginal* tables $h(c) \equiv \sum_k h(c,k)$ and $h(k) \equiv \sum_c h(c,k)$.

A further reduction is embodied in the 2×2 contingency table $\mathcal{A} = \{a_{ij} | i,j \in \{0,1\}\}$, where the elements a_{ij} are counts of pairs of vectors $\{x_p, x_q\}$. The row index value i indicates the state of the pairs with respect to the classes. A value of 0 indicates pairs that were assigned to the same class, whereas a value of 1 corresponds to pairs occuring in different classes. Similarly for the column index j except that it corresponds to clusters rather than classes. The following are the associated definitions and formulas in terms of $\mathcal{H}$.

1. a_{00} is the number of pairs of vectors that are found in both the same class and the same cluster:

$$a_{00} = \sum_c \sum_k \binom{h(c,k)}{2} = \sum_{c,k} \frac{h(c,k)[h(c,k)-1]}{2}. \quad (2)$$

2. a_{01} is the number of pairs occuring in the same class, but not the same cluster:

$$\begin{aligned} a_{01} &= \sum_c \sum_k \sum_{k'<k} h(c,k)h(c,k') \\ &= \sum_c \binom{h(c)}{2} - a_{00}. \end{aligned} \quad (3)$$

3. a_{10} is the number of pairs occuring in the same cluster, but not the same class;

$$\begin{aligned} a_{10} &= \sum_k \sum_c \sum_{c'<c} h(c,k)h(c',k) \\ &= \sum_k \binom{h(k)}{2} - a_{00}. \end{aligned} \quad (4)$$

4. a_{11} is the number of pairs of vectors that are found in neither the same class nor the same cluster;

$$\begin{aligned} a_{11} &= \sum_{c}\sum_{c'<c}\sum_{k}\sum_{k'<k} h(c,k)h(c',k') \\ &= \binom{n}{2} - (a_{00}+a_{01}+a_{10}). \end{aligned} \tag{5}$$

5. Column and row sums of $\mathcal{A}$ eg. $a_{0\bullet}$ and $a_{\bullet 0}$: The symbol "$\bullet$" in place of an index indicates that the index is summed over. For example:

$$a_{0\bullet} = \sum_{i=0}^{1} a_{0i} = \sum_{c}\binom{h(c)}{2}. \tag{6}$$

$$a_{\bullet 0} = \sum_{i=0}^{1} a_{i0} = \sum_{k}\binom{h(k)}{2}. \tag{7}$$

6. $\sum_i \sum_j a_{ij} = M \equiv \binom{n}{2}$ the total number of pairs.

4. An Information-Theoretic External Validity Measure

In this section we describe our proposed measure.

4.1. Comparing Clusterings Where the Number of Clusters is Fixed

We first treat the special case of clustering problems where the number of clusters is known (or at least fixed) in advance. For such problems our measure is essentially equivalent[5] to the *conditional entropy*[6] $H(C|K)$, which is defined as

$$H(C|K) = -\sum_{c=1}^{|\mathcal{C}|}\sum_{k=1}^{|\mathcal{K}|} p(c,k)\log p(c|k).$$

We don't know the distribution $p(C,K)$ however and we are thus forced to estimate it. For this we use $\mathcal{H}$ and we refer to the associated estimate of $H(C|K)$ as the *empirical* conditional entropy and denote it by $\tilde{H}(C|K)$:

$$\tilde{H}(C|K) = -\sum_{c=1}^{|\mathcal{C}|}\sum_{k=1}^{|\mathcal{K}|}\frac{h(c,k)}{n}\log\frac{h(c,k)}{h(k)} = \tilde{H}(C,K) - \tilde{H}(K).$$

[5] Our general measure (described in the following section) includes an additional term that depends primarily on $|\mathcal{C}|$, $|\mathcal{K}|$ and n (See Appendix B) and is relatively insensitive to other details of $\mathcal{H}$. We omit this additional term here for pedagogical purposes.

[6] See [2] for a discussion of conditional entropy and mutual information.

Alternately and equivalently we might define and use the *empirical mutual information*, $\tilde{I}(C;K)$

$$\tilde{I}(C;K) = \tilde{H}(C) - \tilde{H}(C|K), \tag{8}$$

where

$$\tilde{H}(C) = -\sum_{c=1}^{|C|} \frac{h(c)}{n} \log \frac{h(c)}{n}.$$

These ($\tilde{H}(C|K)$ and $\tilde{I}(C;K)$) are equivalent because $\tilde{H}(C)$ is fixed for a given ground-truth set.

Why should we expect $\tilde{H}(C|K)$ to be a good measure of clustering quality? Given our stated general approach, this is equivalent to asking why we should expect this to be a good measure of the usefulness of the elements K as predictors of the corresponding elements of C. $\tilde{H}(C|K)$ is the number of bits[7] required to encode (compress) the class labels of all the objects using $\mathcal{H}$ as a model and assuming that both the encoder and the decoder "know" it and $\{k_i\}$. The worst-case value of $\tilde{H}(C|K)$ is just $\tilde{H}(C)$, the number of bits it takes to encode $\{c_i\}$ with no help from $\{k_i\}$. This case corresponds to $\tilde{I}(C;K) = 0$. The best-case value of $\tilde{H}(C|K)$ is 0, which corresponds to perfect correspondence between c and k. This corresponds to $\tilde{I}(C;K) = \tilde{H}(C)$.

The conditional entropy was used as an external validity measure in [1] and mutual information was used as one in [20, 19]. Also workers in the area have discussed mutual information as a measure of association between categorical attributes[10, 18]. We assume that it has not been used more often as an external validity measure for clustering because it is not viable for comparing clusterings with different numbers of clusters. This is discussed in some detail in the following section.

4.2. Clustering for a Variable Number of Clusters

Here we propose a measure for comparing the output of two clustering algorithms designed to determine the number of clusters as well as clustering the feature vectors. While $\tilde{H}(C|K)$ and $\tilde{I}(C;K)$ are excellent measures of clustering quality when the number of clusters is fixed, they have one deficiency: they don't take the number of clusters into consideration directly. If two different clusterings to be compared have the same $\tilde{H}(C|K)$, but different numbers of clusters, the one with fewer clusters will usually be preferable. An extreme case that demonstrates this is a clustering that assigns a different value of k to every object. This will have the smallest possible code length $\tilde{H}(C|K) = 0$, but in reality provides no useful information. What is missing in the coding scenario used in justifying our use of $\tilde{H}(C|K)$ is that, for the cluster labels to be useful as predictors, the distribution ($\mathcal{P}$ or $\mathcal{H}$) must be available to the decoder. In the coding analogy this must therefore be transmitted from the encoder to the decoder. In the extreme cluster-per-object example this distribution is equivalent to a list of the class labels of all the objects.

[7] Assuming base-2 logarithms.

4.2.1. The Cost of Encoding $\mathcal{H}$ In Section 4.1 we explained the use of conditional entropy as a validity measure when the number of clusters is fixed. In what follows, to handle the case of a variable number of clusters, we add the cost of encoding our estimation ($\mathcal{H}$ in this case) of the distribution $\mathcal{P}$. The decoder knows $\{k_i\}$ and therefore $\{h(k)\}$. What the encoder needs to transmit are $\mathcal{H}$ using some predetermined encoding scheme and then C using a code constructed using $\mathcal{H}$. The code length for C is equal to $n\tilde{H}(C|K)$. Obtaining a code-length for $\mathcal{H}$ is somewhat more involved, however. Think of $\mathcal{H}$ as a matrix with rows indexed by c and columns by k. The quantity $h(k)$ is equal to the sum of the elements in the k^{th} column $\mathcal{H}$. The number of possible columns corresponding to $h(k)$ is the number of $|\mathcal{C}|$-component vectors with non-negative integer components summing to $h(k)$. This is given by the following combinatorial formula:

$$\binom{h(k)+|\mathcal{C}|-1}{|\mathcal{C}|-1}. \tag{9}$$

Thus we can encode the k^{th} column by specifying an integer index from 1 through this number. If we assume that all columns consistent with $h(k)$ are equally likely (uniform-prior assumption), the code length for this integer index is the log of (9). Thus the number of bits required to encode the entire matrix $\mathcal{H}$ when the $\{h(k)\}$ are already known is the log of the number of $\mathcal{H}$'s consistent with $\{h(k)\}$, which is given by:

$$\log \prod_{k=1}^{|\mathcal{K}|} \binom{h(k)+|\mathcal{C}|-1}{|\mathcal{C}|-1} \;=\; \sum_{k=1}^{|\mathcal{K}|} \log \binom{h(k)+|\mathcal{C}|-1}{|\mathcal{C}|-1}. \tag{10}$$

Our clustering quality measure (smaller values are better) is the entire encoding cost (data plus model) per object, which is given by:

$$\mathcal{Q}_0(C,K) \;=\; \tilde{H}(C|K) \;+\; \frac{1}{n}\sum_{k=1}^{|\mathcal{K}|} \log \binom{h(k)+|\mathcal{C}|-1}{|\mathcal{C}|-1}. \tag{11}$$

To actually use the encoding scheme implicit in this we would need an additional term of $\log n$ bits to encode $|\mathcal{C}|$, but we omit it because it is fixed for a constant for a given ground truth set. This (11) can be seen as an application of the *minimum description length* principle (MDL)[15, 16].

Note that this measure is not symmetric with respect to C and K, which is reasonable. The status of these two sets of labels is not equivalent in this context.

4.2.2. Extreme Cases The following list enumerates several extreme cases of our measure.

1. Perfect clustering: The minimum possible value of (11) occurs when there are exactly as many clusters as classes and there is a one-to-one correspondence between classes

and clusters. In this case $\tilde{H}(C|K)$ is equal to zero (the cluster labels give perfect knowledge of the class labels) and we are left only with (10), the code length associated with $\mathcal{H}$, which is equal (in this case) to:

$$\frac{1}{n}\sum_{c=1}^{|\mathcal{C}|}\log\binom{h(c)+|\mathcal{C}|-1}{|\mathcal{C}|-1}. \tag{12}$$

2. $n \to \infty$ for fixed $|\mathcal{C}|, |\mathcal{K}|$ and $\mathcal{P}$: This case is worked out in detail in Appendix B. We repeat the result here.

   $$\mathcal{Q}_0(C,K) \;\sim\; H(C|K) \;+\; |\mathcal{K}|(|\mathcal{C}|-1)\frac{\log n}{n}. \tag{13}$$

   For very large n the $\frac{\log n}{n}$ term will become insignificant leaving

   $$\mathcal{Q}_0(C,K) \;\sim\; H(C|K). \tag{14}$$

   For two clusterings with the same conditional entropy $H(C|K)$, but different numbers of clusters, the $\log n$ term will always be important, however, and will be the deciding factor that causes the clustering with the smallest number of clusters to be rated best (smallest Q).

3. Uncorrelated $\mathcal{C}$ and $\mathcal{K}$: The worst-case (maximum) value of $\mathcal{Q}_0$ will occur where the cluster labels provide no useful information about the class labels. We consider three sub cases of this uncorrelated case.

   (a) $|\mathcal{K}| = 1$: Considering the 1st term in (11), we have $\tilde{H}(C|K) = \tilde{H}(C)$. The 2nd term, the model-encoding cost reduces to $(1/n)\log\binom{n+|\mathcal{C}|-1}{|\mathcal{C}|-1}$, the cost of encoding $\{h(c)\}$.

   $$\mathcal{Q}_0(C,K;|\mathcal{K}|=1) \;=\; \tilde{H}(C) \;+\; \frac{1}{n}\log\binom{n+|\mathcal{C}|-1}{|\mathcal{C}|-1}. \tag{15}$$

   (b) $1 < |\mathcal{K}| \ll n$: As long as the cluster labels have no useful information about the class labels[8] (i.e. $p(c|k) = p(c)$), $\mathcal{Q}_0$ will increase with $|\mathcal{K}|$. For complete lack of information to strictly hold, however, we must have $\forall c, k\; h(c,k) = h(c)/|\mathcal{K}|$, which can only strictly hold when the $\{h(c)\}$ are all exactly divisible by $|\mathcal{K}|$. For $n \gg |\mathcal{C}|$ and $n \gg |\mathcal{K}|$, this can be made to hold approximately for any n. As $|\mathcal{K}| \to n$, however, it will be impossible to have the no-information case.

   (c) $|\mathcal{K}| = n$: In this case $\tilde{H}(C|K) = 0$ because the class labels are completely determined by the cluster labels, when $\mathcal{H}$ is known. The model-encoding cost,

[8] That is, information that can be utilized when one knows $\mathcal{H} = \{h(c,k)\}$.

on the other hand, reduces to $n \log |\mathcal{C}|$, the cost of simply directly encoding the class labels assuming they are a priori equally likely.

$$\mathcal{Q}_0(C, K; |\mathcal{K}| = n) \quad = \quad \log |\mathcal{C}| \tag{16}$$

Either of these may be the maximum value, depending on $\tilde{H}(C)$. For example, clearly, if the class labels are equally likely *a priori*, the 1-cluster cost will be the largest. On the other hand, if all the classes are empty except one, the n-cluster cost is the largest.

4.2.3. Other Possible Forms of the Measure We can modify this definition to make it analogous to mutual information. That is, we can use the difference between this code length (11) and the number of bits that would have been required to encode the c values using $-\log p(c)$ bits each. This code-length difference is equal to:

$$\begin{aligned} \mathcal{Q}_1(C,K) \; = \; \hat{\mathcal{I}}(C;K) = \; & \tilde{I}(C;K) \; + \; \tfrac{1}{n}\Big[\log\tbinom{n+|\mathcal{C}|-1}{|\mathcal{C}|-1} \\ & - \textstyle\sum_{k=1}^{|\mathcal{K}|} \log\tbinom{h(k)+|\mathcal{C}|-1}{|\mathcal{C}|-1}\Big], \end{aligned} \tag{17}$$

There is an important qualitative difference between this and $\tilde{I}(C;K)$, which is symmetric with respect to C and K whereas $\hat{\mathcal{I}}(C;K)$ is not.

As we discuss in the following section, most other external measures have the property that the best possible value is equal to one, whereas the worst is zero. Here we transform ours to have this $(0, 1]$ property. Our basic measure is given by the code length given in (11).

$$\mathcal{Q}_2(C,K) = \frac{\frac{1}{n}\sum_{c=1}^{|\mathcal{C}|} \log\binom{h(c)+|\mathcal{C}|-1}{|\mathcal{C}|-1}}{\mathcal{Q}_0(C,K)} \tag{18}$$

Clearly $\mathcal{Q}_2 = 0$ can only be approached as $n \to \infty$.

5. A Survey of Other External Validity Measures

An extensive review of related association measures prior to 1959 including work in the late 19th century can be found in two papers by Goodman and Kruskal[5, 6].

5.1. Classification Error

An external measure that is sometimes used in cases where the number of clusters is equal to the number of classes is *classification error*. If the rows and columns of $\mathcal{H}$ are made to correspond by associating the majority class in each cluster with the cluster itself, then $\mathcal{H}$ can be viewed as the *confusion matrix* of pattern recognition and the sum of its off-diagonal elements divided by the total number of objects is the *total classification error*. The problem with this is that it ignores how incorrect classifications are distributed among the other clusters. Being distributed uniformly randomly among the other clusters is arguably much worse than all going to a single cluster.

Applying this measure to the case where the number of clusters is different than the number of classes is also problematic. This problem is addressed in [11][9] by defining a "normalized Hamming distance" in which the following associations are used:

$$\hat{c}(k) \equiv \arg\max_c h(c,k), \tag{19}$$
$$\hat{k}(c) \equiv \arg\max_k h(c,k). \tag{20}$$

A "directional Hamming distance" $D_H(\cdot;\cdot)$ is defined as

$$D_H(C;K) = \textstyle\sum_k \sum_{c \neq \hat{c}(k)} h(c,k) \tag{21}$$
$$D_H(K;C) = \textstyle\sum_c \sum_{k \neq \hat{k}(c)} h(c,k) \tag{22}$$

The *normalized Hamming distance* is then defined in terms of these as follows.

$$\tilde{D}_H(C,K) = 1 - \frac{D_H(C;K) + D_H(K;C)}{2n} \tag{23}$$

5.2. Measures Based on $\mathcal{A}$

Jain and Dubes[9] list the following four commonly used external measures of partitional validity. All are functions of the matrix $\mathcal{A}$, defined in Section 3.

- Rand[14]: $\frac{a_{00}+a_{11}}{\binom{n}{2}}$
- Jacard[13]: $\frac{a_{00}}{a_{00}+a_{01}+a_{10}}$
- Fowlkes and Mallows[4]: $\frac{a_{00}}{\sqrt{\tilde{a}}}$, where $\tilde{a} \equiv a_{0\bullet}a_{\bullet 0}$
- Γ statistic(Hubert and Schultz)[8]: $\frac{Ma_{00}-\tilde{a}}{\sqrt{\tilde{a}(M-a_{0\bullet})(M-a_{\bullet 0})}}$

6. Comparison Methodology

How does one compare clustering accuracy measures? Judging the accuracy of clustering algorithms is a difficult enough problem, but now we present ourselves with the even more difficult meta-question of how one compares accuracy measures. First, as we stated in the introduction, we only consider so-called *external* measures which judge accuracy by seeing how closely, in some appropriate sense, the partition of objects produced by a clustering algorithm agrees with what is deemed an ideal (ground truth) partition. By this constraint we have obviously reduced our question to one of what constitutes the best measure of agreement between the two clusterings.

[9] This is also discussed in [7].

Our argument will proceed as follows. While acknowledging that what constitutes a desirable accuracy measure depends on the particular application, we assert that having a general measure is desirable for assessing general clustering algorithms designed to be used in many applications and, even in those cases where a particular application is targeted, measuring the most relevant quantity (e.g. time saved at some particular task) may be infeasible. We believe that our measure is superior to others as a general measure, though we must acknowledge that the choice of which measure to use is, to a certain extent, a matter of taste. We make a concession to this fact in our arguments. First we argue that our measure is superior and should therefore be used in all appropriate cases i.e. comparing a partition to a ground-truth partition. Second, we argue that, even if one doesn't accept that our measure is intrinsically superior, it must at least be acknowledged that it has all the desirable qualitative properties and produces different results from all other measures in some cases and should therefore be accepted as, at least another measure in the set to choose from. In support of these assertions

1. We show that our measure has all the desirable properties.
2. We argue on philosophical grounds that our measure is superior because of its information-theoretic basis.
3. We show that other measures give counter-intuitive (if not simply incorrect) results in certain cases, but that our measure always give the desired behavior.
4. We show that our measure gives results different from those produced by other commonly used measures.

6.1. A Parametric Form for $p(c, k)$

To identify desirable properties of a clustering accuracy measure we first examine how one might characterize the statistical relationship between the class labels and cluster labels corresponding to a set of objects. A data set for which there are known class labels (ground truth) may itself be characterized by the number of objects n, the number of classes $|\mathcal{C}|$ and the distribution of class labels $\{h(c)\}$, which can be thought of as a sample drawn from a distribution characterized by a probability function $\{p(c)\}$. In addition to these data-set properties the output of a clustering algorithm is further characterized by the number of clusters and joint distribution of class and cluster labels: $\mathcal{H} \equiv \{h(c, k)\}$. We can also consider this to be a sample drawn from a population characterized by $\mathcal{P} \equiv \{p(c, k)\}$.

We next define a family of distributions over (c, k) with the hope that this family captures most of the essential characteristics of such distributions from the perspective of characterizing the accuracy of clustering algorithms. Members of this family are identified by values of certain parameters as follows:

- $|\mathcal{C}|$: number of classes

- $(\forall c \in \mathcal{C})\ p(c) \;=\; 1/|\mathcal{C}|$
- $|\mathcal{K}|$: number of clusters
- Decomposition of $\mathcal{K}$ into two disjoint subsets $\mathcal{K}_u$ ("useful") and $\mathcal{K}_n$ ("noise"). The cardinalities of these subsets are given by $|\mathcal{K}_u|$ and $|\mathcal{K}_n|$ respectively and clearly $|\mathcal{K}| = |\mathcal{K}_u| + |\mathcal{K}_n|$. The roles of these two cluster subsets are as follows.
  - $\mathcal{K}_n$: These clusters are completely noise in the sense that there is no correlation between these cluster labels and class labels. Also:
  $$(\forall k \in \mathcal{K}_n)\ (\forall c \in \mathcal{C})\ \ p(c|k) \;=\; 1/|\mathcal{C}|.$$
  - $\mathcal{K}_u$: The clusters in the *useful* set $\mathcal{K}_u$ are correlated with the classes and $\mathcal{K}_u$ is further decomposed into $|\mathcal{C}|$ subsets $\{\mathcal{K}(c)|c \in \mathcal{C}\}$ and correspondingly $\mathcal{C}$ is decomposed into subsets $\{\mathcal{C}(k)|k \in \mathcal{K}_u\}$. The role of these subsets is as follows. For a given c the probabilities $p(k|c)$ are equal for all $k \in \mathcal{K}(c)$. While the sizes of the $\{\mathcal{K}(c)\}$ could be left as parameters also, we determine them automatically as follows.
    - * If $|\mathcal{C}| = |\mathcal{K}_u|$, then $|\mathcal{K}(c)| = 1$ and $\mathcal{K}(c)$ consists of cluster $k = c$.
    - * If $|\mathcal{C}| < |\mathcal{K}_u|$, then $(\forall k \in \mathcal{K}_u)\,|\mathcal{C}(k)| = 1$ and at least some of the $\{\mathcal{C}(k)\}$ will overlap, corresponding to more than one cluster. Cluster-to-class assignments proceed as follows:
      - · The first $\lceil |\mathcal{K}_u|/|\mathcal{C}| \rceil$ clusters are assigned to class 1.
      - · The next $\lceil (|\mathcal{K}_u| - |\mathcal{C}(1)|)/(|\mathcal{C}| - 1) \rceil$ clusters are assigned to class 2.
      - · and so on .... The number of clusters to be assigned to the next class being given by the *ceiling* of the ratio of the number of unassigned clusters remaining to the number of unassigned classes remaining.
    - * If $|\mathcal{C}| > |\mathcal{K}_u|$, then the class-to-cluster assignments proceed in a manner exactly analogous to the cluster-to-class assignments in the $|\mathcal{C}| < |\mathcal{K}_u|$ case.
- ϵ: total error probability:
$$\sum_{c \in \mathcal{C}} \sum_{k \notin \mathcal{K}(c)} p(c,k) \;=\; \epsilon$$
$$(\forall c \in \mathcal{C})\ (\forall k \in \mathcal{K}(c))\ \ p(k|c) \;=\; \frac{1-\epsilon}{|\mathcal{K}(c)|}$$
- ϵ_1, ϵ_2: error components: $\epsilon = \epsilon_1 + \epsilon_2$
  - ϵ_1:
  $$(\forall c \in \mathcal{C})\ [\forall (k \in \mathcal{K}_u) \wedge (k \notin \mathcal{K}(c))]\ \ p(k|c) \;=\; \frac{\epsilon_1}{|\mathcal{K}_u| - |\mathcal{K}(c)|}$$

- ϵ_2:

$$(\forall c \in \mathcal{C})\ (\forall k \in \mathcal{K}_n)\ \ p(k|c) \ = \ \frac{\epsilon_2}{|\mathcal{K}_n|}$$

$$\sum_{c \in \mathcal{C}} \sum_{k \in \mathcal{K}_n} p(c,k) \ = \ \epsilon_2$$

Three examples of $\{p(c,k)\}$ are presented in Tables 1, 2 and 3.

Table 1: Class-cluster joint probability distribution $\mathcal{P} \equiv \{p(c,k)\}$ with $|\mathcal{K}_u| = |\mathcal{C}| = 5$ and $\epsilon_1 = \epsilon_2 = 0$

	← cluster →				
class ↓	1	2	3	4	5
1	0.2	0.0	0.0	0.0	0.0
2	0.0	0.2	0.0	0.0	0.0
3	0.0	0.0	0.2	0.0	0.0
4	0.0	0.0	0.0	0.2	0.0
5	0.0	0.0	0.0	0.0	0.2

Table 2: Class-cluster joint probability distribution $\mathcal{P}$ with $|\mathcal{K}_u| = |\mathcal{C}| = 5$, $|\mathcal{K}_n| = 0$, $\epsilon_1 = 0.2$ and $\epsilon_2 = 0$

	← cluster →				
class ↓	1	2	3	4	5
1	0.16	0.01	0.01	0.01	0.01
2	0.01	0.16	0.01	0.01	0.01
3	0.01	0.01	0.16	0.01	0.01
4	0.01	0.01	0.01	0.16	0.01
5	0.01	0.01	0.01	0.01	0.16

6.2. Desirable Characteristics of Clustering Accuracy Measures

Assume that some accuracy measure $\mathcal{M}(\mathcal{H})$ increases monotonically from 0 to 1 as accuracy improves, with $\mathcal{M} = 1$ corresponding to the perfect clustering where $|\mathcal{K}| = |\mathcal{C}|$ and each cluster corresponds to one and only one class. Assume further that we intend to apply it to the expected $\mathcal{H}$ or $\mathcal{A}$ corresponding to the family of distributions just defined above. For a fixed n and $|\mathcal{C}|$ what behavior do we desire of $\mathcal{M}$ with respect to the parameters

Table 3: Class-cluster joint probability distribution $\mathcal{P}$ with $|\mathcal{K}_u| = |\mathcal{C}| = 5$, $|\mathcal{K}_n| = 3$, $\epsilon_1 = 0.2$ and $\epsilon_2 = 0.3$. clusters $\{6, 7, 8\}$ are *noise*, while $\{1, 2, 3, 4, 5\}$ are *useful.*

	← cluster →							
class ↓	1	2	3	4	5	6	7	8
1	0.10	0.01	0.01	0.01	0.01	0.02	0.02	0.02
2	0.01	0.10	0.01	0.01	0.01	0.02	0.02	0.02
3	0.01	0.01	0.10	0.01	0.01	0.02	0.02	0.02
4	0.01	0.01	0.01	0.10	0.01	0.02	0.02	0.02
5	0.01	0.01	0.01	0.01	0.10	0.02	0.02	0.02

of this family[10]: $|\mathcal{K}_u|, |\mathcal{K}_n|, \epsilon_1$ and ϵ_2? Expressing the answer in terms of differences and derivatives with respect to single parameters while all others are held fixed, we have:

Table 4: **Desirable Properties of $\mathcal{M}$**

P1: For $|\mathcal{K}_u| < |\mathcal{C}|$ and $\Delta|\mathcal{K}_u| \leq (|\mathcal{C}| - |\mathcal{K}_u|)$, $\frac{\Delta\mathcal{M}}{\Delta|\mathcal{K}_u|} > 0$.

P2: For $|\mathcal{K}_u| \geq |\mathcal{C}|$, $\frac{\Delta\mathcal{M}}{\Delta|\mathcal{K}_u|} < 0$.

P3: $\frac{\Delta\mathcal{M}}{\Delta|\mathcal{K}_n|} < 0$.

P4: $\frac{\partial\mathcal{M}}{\partial\epsilon_1} \leq 0$ with equality holding only when $|\mathcal{K}_u| = 1$.

P5: $\frac{\partial\mathcal{M}}{\partial\epsilon_2} \leq 0$ with equality holding only when $|\mathcal{K}_n| = 0$.

Summarizing, $\mathcal{M}$ should indicate that a clustering is worse whenever:

- the number of *useful* clusters varies away from $|\mathcal{K}_u| = |\mathcal{C}|$
- the number of *noise* clusters $|\mathcal{K}_n|$ increases
- either of the error parameters ϵ_1 and ϵ_2 increases

6.3. Analysis of $\mathcal{Q}_{0,2}$ Vis a Vis the Desiderata of Section 6.2

For simplicity we will base our analysis/discussion on the asymptotic forms of our measure given in (13) and (14). Here we consider the variation with respect to the various model parameters.

[10] All others are determined by these four.

- **P1** and **P2** ($|\mathcal{K}_u|$): Certainly, increasing $|\mathcal{K}_u|$ beyond $|\mathcal{C}|$ will increase both $H(C|K)$ and the model cost. Decreasing it from $|\mathcal{C}|$ will increase $H(C|K)$, while decreasing the model cost, however. We see from (14) that the increase in $H(C|K)$ will dominate, asymptotically.

- **P3** ($|\mathcal{K}_n|$): Increasing $|\mathcal{K}_n|$ while holding ϵ_2 fixed will increase the model cost while having no effect on $H(C|K)$. The former is obvious, while the latter is due to the fact that only the fraction ϵ_2 of objects that are assigned to noise clusters (not the number of noise clusters) affects $H(C|K)$.

- **P4** and **P5** (ϵ_1 and ϵ_2): Increasing either ϵ_1 or ϵ_2 will clearly increase $H(C|K)$, but have no effect on the asymptotic model cost $|\mathcal{K}|(|\mathcal{C}|-1)\log n$. Thus $\frac{\partial \mathcal{Q}_0}{\partial \epsilon_{1,2}} \geq 0$, which implies that $\frac{\partial \mathcal{Q}_2}{\partial \epsilon_{1,2}} \leq 0$.

Thus we have shown that our measure satisfies all of our desired characteristics asymptotically. In the following section we will do an exploration of the ability of all of the measures discussed here to satisfy these characteristics for a certain set of test cases.

7. Comparison Results

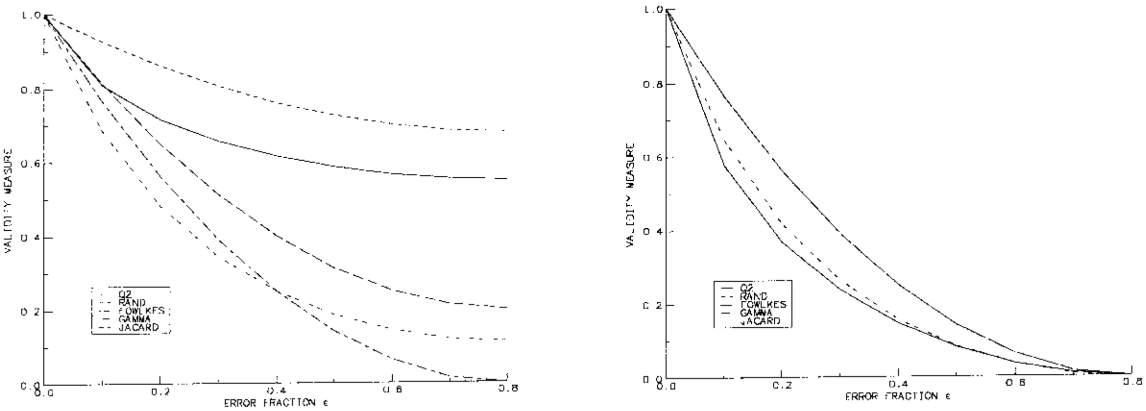
Our comparison of the measures proceeds as follows. All results are obtained by first computing the expected values of $\mathcal{H}$ and $\mathcal{A}$ for the given $\mathcal{P}$ (i.e. corresponding to specific values of the parameters $|\mathcal{C}|, |\mathcal{K}_u|, |\mathcal{K}_n|, \epsilon_1$ and ϵ_2) and then using them to compute values of the various validity measures[11]. In most of the results we present the $\mathcal{Q}_2$ form of our measure because it has the same general behavior as the other measures in that $\mathcal{Q}_2 = 1$ corresponds to a perfect clustering and values decrease toward 0 as the clustering accuracy gets worse. Both forms of the measure ($\mathcal{Q}_0$ and $\mathcal{Q}_2$) obviously produce the same ordering of a given group of clusterings all compared against the same ground truth.

7.1. Simple Cases

We begin by examining two simple cases. The first of these (See Figure 1) has $|\mathcal{C}| = |\mathcal{K}| = |\mathcal{K}_u| = 5$ and $|\mathcal{K}_n| = 0$. The ϵ_1 parameter is then varied over a range of $[0, 0.8]$. All measures show the desired (and expected) behavior with their values decreasing with increasing ϵ_1.

In the second case we examine, we hold $|\mathcal{K}_n|, \epsilon_1$ and ϵ_2 all fixed at 0 and $|\mathcal{C}| = 6$, while increasing $|\mathcal{K}_u|$ from 6 to 60. These results are shown in Figure 2. All measures show the desired/expected behavior decreasing monotonicly (from one) with $|\mathcal{K}_u|$.

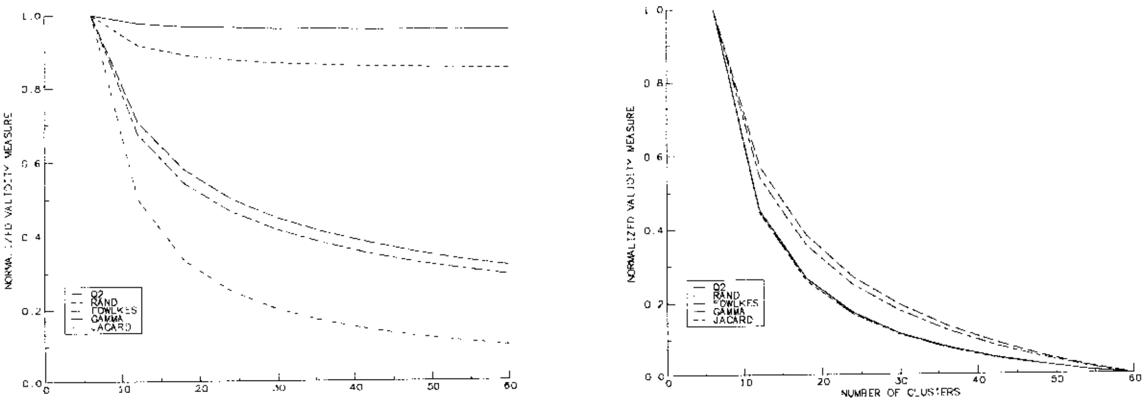
[11] Our measures $\mathcal{Q}_0$ and $\mathcal{Q}_2$ require computing factorials of certain h values. These are computed using the Γ function so that meaningful results are obtained when the expected h values are non-integer.



(a)

(b)

Figure 1: The result of computing several external validity measures including $\mathcal{Q}_2$ while varying the error parameter ϵ_1 and holding $|\mathcal{C}| = |\mathcal{K}| = |\mathcal{K}_u| = 5$ ($|\mathcal{K}_n| = 0$). (a) not renormalized; (b) normalized to all vary over the range [0,1] over the range of ϵ investigated.



(a)

(b)

Figure 2: The result of computing several external validity measures including $\mathcal{Q}_2$ while varying $|\mathcal{K}_u|$ the number of useful clusters and fixing the other model parameters at the following values: $|\mathcal{C}| = 6, |\mathcal{K}_n| = 0, \epsilon_1 = 0, \epsilon_2 = 0$. (a) not renormalized; (b) normalized to all vary over the range [0,1] over the range of ϵ investigated.

7.2. A More Detailed Comparison Over a Broad Range of Model Parameters

7.2.1. Comparison Results The calculated accuracy-measure values reported in this section correspond to fixed values of $|\mathcal{C}|(|\mathcal{C}| = 5)$ and $n(n = 500)$. All other model parameters were varied over the following values.

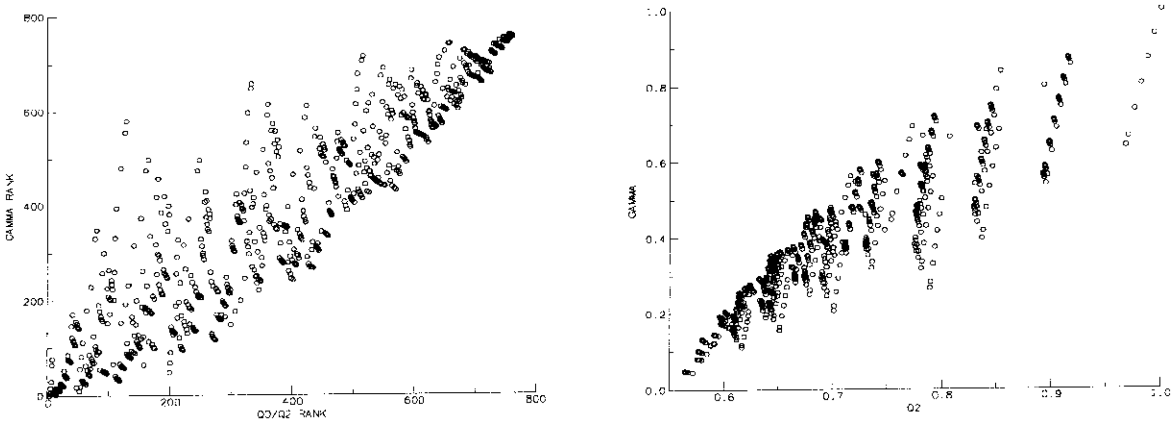
- $|\mathcal{K}_u|$: 10 values: $2, 3, \ldots, 11$
- $|\mathcal{K}_n|$: 7 values: $0, 1, \ldots, 6$
- ϵ_1: 4 values: $0, 0.066..., 0.133..., 0.2$
- ϵ_2: 4 values: $0, 0.1, 0.2, 0.3$

Running through the complete range of these values would result in $1120 = 10 \times 7 \times 4 \times 4$ instances of $\{p(c, k)\}$, but when meaningless combinations (e.g. $|\mathcal{K}_n| = 0$ with non-zero ϵ_2 values) are eliminated, there are 760 valid combinations. For each valid parameter combination values of all 7 measures ($\mathcal{Q}_0$, $\mathcal{Q}_2$, Rand, Fowlkes, Gamma, Jacard and Hamming) were calculated.

While this is clearly not a complete characterization of the behavior of these measures, it has the following objectives.

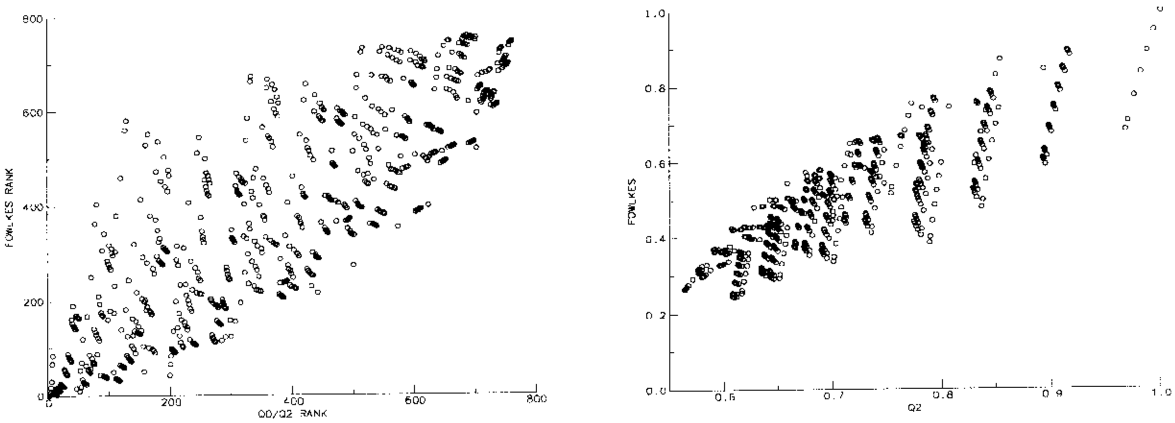
- To confirm the analysis of our measure in Section 6.3. The absence of any violations of our desired characteristics, while not proving that they hold in all cases, is evidence in support of that contention.
- To show that our measure produces a unique ranking of the various cases, a ranking different from that produced by other measures.
- To see if any of the other measures violate our desirability criteria in any of these cases, which would suggest that our measure is superior as judged by our criteria.

The results of these calculations and and their associated ranks according to all the measures are displayed graphically in Figures 3 through 7. Each figure consists of two parts: (a) the ranks (of the various cases) produced by one of the other measures plotted versus the ranks produced by our measure; (b) the values of the other measure plotted versus the corresponding $\mathcal{Q}_2$ values. As can be seen from these figures, the values of the measures and their rankings over the range of model-parameter values explored, the other measures, while clearly correlated with our measure ($\mathcal{Q}_2$), produce different rankings. Thus our measure is a clear alternative with its own unique ranking.



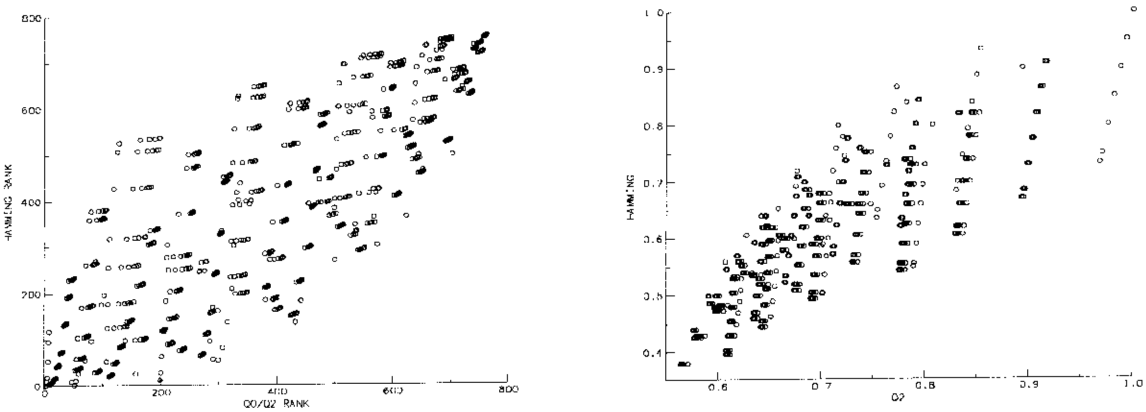
(a) (b)

Figure 3: A comparison of the Gamma measure and Q_2 (a) ranks & (b) values



(a) (b)

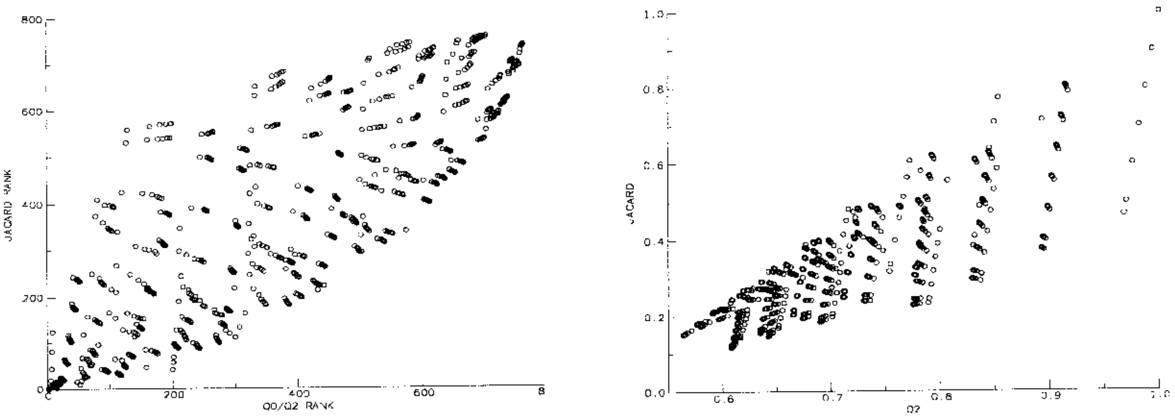
Figure 4: A comparison of the Fowlkes-and-Mallows measure and Q_2 (a) ranks & (b) values



(a)

(b)

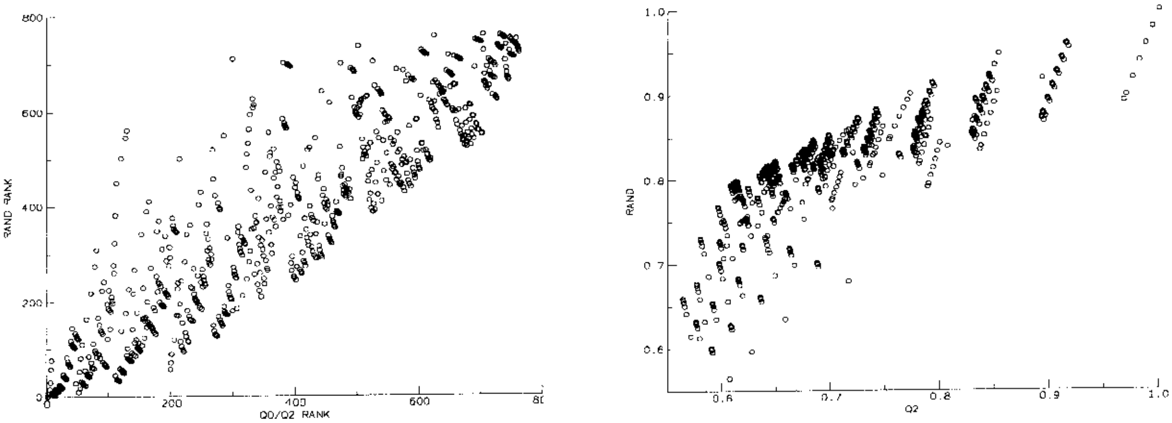
Figure 5: A comparison of the Hamming measure and $\mathcal{Q}_2$ (a) ranks & (b) values



(a)

(b)

Figure 6: A comparison of the Jacard measure and $\mathcal{Q}_2$ (a) ranks & (b) values



(a)

(b)

Figure 7: A comparison of the Rand measure and $\mathcal{Q}_2$ (a) ranks & (b) values

7.2.2. Examination of Whether or not the Measures Satisfy the Desirable Properties of Section 6.2 We analyzed the results of our calculations to see if the desirable properties listed in Table 4 of Section 6.2 were satisfied. Checking all instances of the associated differences, we found the following.

- Our measure ($\mathcal{Q}_2$) satisfied all tests.
- **P4:** All measures satisfied all tests of the $\frac{\partial \mathcal{M}}{\partial \epsilon_1} < 0$ criterion.
- **P5:** All measures except Rand satisfied all tests of the $\frac{\partial \mathcal{M}}{\partial \epsilon_2} < 0$ criterion. We observed 29 cases where Rand failed this test. All occured for cases where $|\mathcal{K}_u| < |\mathcal{C}|$; most for $|\mathcal{K}_u| = 2$ and only 4 for $|\mathcal{K}_u| = 3$ with none for $|\mathcal{K}_u| \geq 4$.
- **P1 and P2:** We observed 14 instances of measures failing the $\frac{\Delta \mathcal{M}}{\Delta |\mathcal{K}_u|}$ tests. Of these 12 were Rand and 2 were Hamming. The two Hamming cases were minor infractions where $\frac{\Delta \mathcal{M}}{\Delta |\mathcal{K}_u|} = 0$ was observed between $|\mathcal{K}_u| = 10$ and 11 for particular instances of the other model parameters. All the Rand errors involved detecting the peak at either $|\mathcal{K}_u| = 6$ or 7. All were for large values of the error parameters ϵ_1 and ϵ_2.
- **P3:** The test that created the most problems for the measures other than ours was $\frac{\Delta \mathcal{M}}{\Delta |\mathcal{K}_n|} < 0$. This difference ratio was measured for each of the measures at each of 6 $|\mathcal{K}_n|$ values $1, 2, \ldots, 6$ for every combination of the following values of the other three model parameters.
  - $|\mathcal{K}_u|$: 10 values: $2, 3, \ldots, 11$
  - ϵ_1: 4 values: $0, 0.066..., 0.133..., 0.2$
  - ϵ_2: 3 values: $0.1, 0.2, 0.3$

  Any time one or more errors was detected in a sequence of measure values corresponding to the sequence of $|\mathcal{K}_n|$ values $1, 2, \ldots, 6$, it was counted as one error. The distribution of these errors over the measures is given in Table 5. Note that three of the measures (Rand, Gamma and Hamming) failed on every one of the 120 instances of the parameter triple $(|\mathcal{K}_u|, \epsilon_1, \epsilon_2)$. The Hamming measure failed because it showed no sensitivity to this parameter i.e. $\frac{\Delta \mathcal{M}}{\Delta |\mathcal{K}_n|} = 0$.

Table 5: The number of cases (out of 120) where the various measures failed the $\frac{\Delta \mathcal{M}}{\Delta |\mathcal{K}_n|} < 0$ test.

Rand	Fowlkes	Gamma	Jacard	Hamming
120	103	120	80	120

8. Discussion and Conclusion

We have proposed and evaluated a new external cluster-validity measure based on information-theoretic considerations. We have also examined the behavior of this measure and its ability to satisfy certain desirability criteria. We have also compared it with other commonly used external validity measures.

In general, the answer to the question of which clustering accuracy measure is *best* will depend on the particular application and it is certainly impossible to anticipate every possible application. In those cases where it is judged that accuracy is best measured by comparing the results of a clustering algorithm to some ideal ("ground truth"), however, our measure is appropriate. At the very least, we can say that the measure we propose offers one more choice to the list of similarity measures appropriate for such comparisons and this measure may give a relative ranking (among various clustering results) that is different from that produced by other measures. We would like to say something stronger, however. We believe that the measure proposed here is superior to other measures in a certain fundamental sense. *Information theory* has, since its inception in 1948[17], clearly demonstrated the viability of code length as a measure of information content. The subsequent development of the theory of *algorithmic complexity*[12] extended these ideas and ultimately led to the *minimum description length principle*[15, 16], which distilled the essence of these and extended them further. For this reason we feel that our measure, which embodies these principles, is superior to the other measures discussed here, which we consider to be more heuristic in nature. This is, of course, a philosophical argument. In support of it we have shown that, when compared with other measures, our measure is the only one that satisfies all of a set of desiderata related to how measure values should vary with certain features of the class-cluster distribution.

Acknowledgment

We would like to thank Alex Cozzi for many helpful and detailed comments after a careful reading of an earlier draft of this document.

References

[1] P. S. Bradley, Usama Fayyad, and Cory Reina. Scaling clustering algorithms to large databases. In Rakesh Agrawal and Paul Stolorz, editors, *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining.* AAAI Press, August 1998.

[2] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley, New York, 1991.

[3] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society*, B39:1–38, 1977.

[4] E. Fowlkes and C. Mallows. A method for comparing two hierarchical clusterings. *Journal of American Statistical Association*, 78:553–569, 1983.

[5] L. A. Goodman and W. H. Kruskal. Measures of association for cross classification. *Journal of the American Statistical Association*, 49:732–764, December 1954.

[6] L. A. Goodman and W. H. Kruskal. Measures of association for cross classification II: Further discussion and references. *Journal of the American Statistical Association*, 54:123–163, March 1959.

[7] Qian Huang and Byron Dom. Quantitative methods of evaluating image segmentation. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. IEEE, October 1995.

[8] L. J. Hubert and J. Schultz. Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology*, 29:190–241, 1976.

[9] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988. ISBN 0-13-022278-X.

[10] Nicholas Jardine and Robin Sibson. *Mathematical Taxonomy*. Wiley Interscience, London, 1971. ISBN 0-471-44050-7.

[11] T. Kanungo, B. Dom, W. Niblack, and D. Steele. MDL-based multi-band image segmentation using a fast region merging scheme. Technical Report 9960, IBM Research Division, May 1995.

[12] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:3–18, 1965.

[13] Glenn W. Milligan, Lisa M. Sokol, and S.C. Soon. The effect of cluster size, dimensionality and the number of clusters on recovery of true cluster structure. *IEEE Trans PAMI*, 5(1):40–47, 1983.

[14] W. M. Rand. Objective criterion for evaluation of clustering methods. *Journal of American Statistical Association*, 66:846–851, 1971.

[15] J. Rissanen. Modelling by shortest data description. *Automatica*, 14:465–471, 1978.

[16] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15. World Scientific Series in Computer Science, 1989.

[17] C. E. Shannon. A mathematical theory of communication. *Bell Syst Tech J.*, (3):379–423, 1948.

[18] Peter H. A. Sneath and Robert R. Sokal. *Numerical Taxonomy*. W. H. Freeman and Co., San Francisco, 1973. ISBN 0-7167-0697-0.

[19] Shivakumar Vaithyanathan and Byron Dom. Generalized model selection for unsupervised learning in high dimensions. In S. A. Solla, T. K. Leen, and K. R. Müller, editors, *Proceedings of Neural Information Processing Systems.* MIT Press, November 1999.

[20] Shivakumar Vaithyanathan and Byron Dom. Model selection in unsupervised learning with applications to document clustering. In I. Brakto and S. Dzeroski, editors, *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 423–433, San Fancisco, June 1999. Morgan Kaufman.

[21] Shivakumar Vaithyanathan and Byron Dom. Hierarchical unsupervised learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, Stanford University; Stanford, CA, June 2000.

[22] C. T. Zahn. Graph-theoretical methods for detecting and describing Gestalt clusters. *IEEE Transactions on Computers C*, C-20:68–86, 1971.

A. Computing the Expectation $E\{\mathcal{A}\}$

Here we discuss the general problem of computing expectations of functions of the contingency table $\mathcal{H}$ (e.g. the elements of $\mathcal{A}$). Following that we address specific problem of obtaining expressions for $E\{\mathcal{A}\}$.

In formal statistical clustering analyses we assume that the set of objects being clustered is a sample drawn from some very large *population* (a.k.a. "process" or "source"). The characteristics of this population, when combined with some clustering procedure, such as the f of (1), yield a probability distribution $\mathcal{P} \equiv \{p(c,k)\}$, where $p(c,k)$ is the probability that an object drawn randomly from that population will belong to *class* c and *cluster* k and $p(c,k)$ is the Bernoulli parameter for that outcome. Using $\mathcal{P}$, the distribution describing the random variable $\mathcal{H}$ will be the following *multinomial* distribution.

$$p(\mathcal{H}) \quad = \quad \binom{n}{\{h(c,k)\}} \prod_{c,k} [p(c,k)]^{h(c,k)}, \tag{24}$$

where $\binom{n}{\{h(c,k)\}}$ represents the multinomial coefficient:

$$\binom{n}{\{h(c,k)\}} = \frac{n!}{\prod_{c,k} h(c,k)!}.$$

We can also write the distribution for the number of times out of n that a randomly drawn object will belong to class c and cluster k. That is the *binomial distribution*:

$$p(h(c,k)) \quad = \quad \binom{n}{h(c,k)} [p(c,k)]^{h(c,k)} [1 - p(c,k)]^{n-h(c,k)}.$$

Using this it is easy to show that

$$E\left\{\binom{h(c,k)}{2}\right\} \;=\; \binom{n}{2}[p(c,k)]^2,$$

and therefore:

$$\begin{aligned} E\{a_{00}\} &= \sum_c \sum_k E\left\{\binom{h(c,k)}{2}\right\} \\ &= \binom{n}{2} \sum_c \sum_k [p(c,k)]^2 \end{aligned} \qquad (25)$$

Using these results and (2) through (5) we obtain:

$$\begin{aligned} E\{a_{01}\} &= \sum_c E\left\{\binom{h(c)}{2}\right\} \;-\; E\{a_{00}\} \\ &= \binom{n}{2}\left\{\sum_c [p(c)]^2 \;-\; \sum_c \sum_k [p(c,k)]^2\right\} \\ &= \binom{n}{2} \sum_c \sum_k \sum_{k'<k} p(c,k)p(c,k') \end{aligned} \qquad (26)$$

$$\begin{aligned} E\{a_{10}\} &= \sum_k E\left\{\binom{h(k)}{2}\right\} \;-\; E\{a_{00}\} \\ &= \binom{n}{2}\left\{\sum_k [p(k)]^2 \;-\; \sum_c \sum_k [p(c,k)]^2\right\} \\ &= \binom{n}{2} \sum_k \sum_c \sum_{c'<c} p(c,k)p(c',k) \end{aligned} \qquad (27)$$

$$\begin{aligned} E\{a_{11}\} &= \binom{n}{2} \;-\; (E\{a_{00}\} + E\{a_{01}\} + E\{a_{10}\}) \\ &= \binom{n}{2}\left\{1 \;+\; \sum_c \sum_k [p(c,k)]^2 \right. \\ &\qquad \left. - \sum_c [p(c)]^2 - \;\sum_k [p(k)]^2\right\} \\ &= \binom{n}{2} \sum_c \sum_{c'<c} \sum_k \sum_{k'<k} p(c,k)p(c',k') \end{aligned} \qquad (28)$$

$$E\{a_{0\bullet}\} = \sum_c E\left\{\binom{h(c)}{2}\right\} = \binom{n}{2}\sum_c [p(c)]^2, \tag{29}$$

$$E\{a_{\bullet 0}\} = \sum_k E\left\{\binom{h(k)}{2}\right\} = \binom{n}{2}\sum_c [p(k)]^2. \tag{30}$$

B. Asymptotic Form of $\mathcal{Q}_0(C, K)$

Here we examine the asymptotic ($n \to \infty$) behavior of our quality measure (11), which we repeat here:

$$\mathcal{Q}_0(C, K) \;=\; \tilde{H}(C|K) \;+\; \frac{1}{n}\sum_{k=1}^{|\mathcal{K}|} \log\binom{h(k) + |\mathcal{C}| - 1}{|\mathcal{C}| - 1}.$$

For the first term in this equation we have:

$$\tilde{H}(C|K) \;\sim\; H(C|K).$$

The second term (the model cost) takes a little more work to analyze. First consider the behavior of $\binom{n}{k}$ as $n \to \infty$ for fixed k.

$$\binom{n}{k} \;\equiv\; \frac{n!}{(n-k)!\,k!} \;=\; \frac{\prod_{i=n-k+1}^{n} i}{k!}.$$

The numerator in this expression can be written:

$$n(n-1)(n-2)\,\ldots\,(n-k+1),$$

which has k terms and thus is a k^{th}-degree polynomial in n for which the coefficient of the highest order term (n^k) is 1. Thus, as $n \to \infty$

$$\binom{n}{k} \;\sim\; \frac{n^k}{k!}$$

and therefore

$$\log\binom{n}{k} \;\sim\; k\log n.$$

Next we apply this result to the individual terms:

$$\log\binom{h(k) + |\mathcal{C}| - 1}{|\mathcal{C}| - 1}$$

As $n \to \infty$, $h(k) \sim n\,p(k)$ and thus

$$\log\binom{h(k) + |\mathcal{C}| - 1}{|\mathcal{C}| - 1} \;\sim\; (|\mathcal{C}| - 1)\log(n\,p(k) + |\mathcal{C}| - 1) \;\sim\; (|\mathcal{C}| - 1)\log n.$$

Using this, we have

$$\sum_{k=1}^{|\mathcal{K}|} \log \binom{h(k) + |\mathcal{C}| - 1}{|\mathcal{C}| - 1} \sim |\mathcal{K}|(|\mathcal{C}| - 1) \log n,$$

which gives

$$\mathcal{Q}_0(C, K) \sim H(C|K) + |\mathcal{K}|(|\mathcal{C}| - 1)\frac{\log n}{n}.$$

For very large n the $\frac{\log n}{n}$ term will become insignificant leaving

$$\mathcal{Q}_0(C, K) \sim H(C|K).$$

For two clusterings with the same conditional entropy $H(C|K)$, but different numbers of clusters, the $(\log n)/n$ term will always be important, however, and will be the deciding factor.

C. Discussion of External vs. Internal Validity Measures

Here we present a supplementary discussion of issues related to the question of which of the two types of validity measures (internal and external) is most appropriate in a given situation. First we discuss a case where an internal measure may be appropriate - doing clustering using the process of density estimation. Following that we expand on the case for using external measures and then we discuss a crucial issue in the use of external measures - obtaining the appropriate ground-truth set.

C.1. The Case for Internal Measures: Validated Likelihood

An internal validation measure of sorts may be appropriate in certain cases, for example, in the common approach of doing clustering by doing *density estimation.* This process involves *estimating* a probability model $p(x)$ that characterizes the data generating mechanism. With the usual assumption that the data are *i.i.d.*, the probability of the complete data set $\mathbf{X}$ is given by:

$$p(\mathbf{X}) = \prod_{i=1}^{n} p(x_i)$$

The model $p(x)$ usually has both a *structure* and associated parameters that must be estimated. A specification of the structure includes the number of clusters. As an example we might fit $\mathbf{X}$ using a mixture density of the form

$$p(x; \alpha, \theta) = \sum_{j=1}^{M} \alpha_j f(x; \theta_j), \tag{31}$$

where the $\{f(x;\theta_j)\}$ are the *component densities* and the $\{\alpha_j\}$ are the *mixing coefficients* ($\sum_{j=1}^{N}\alpha_j = 1$). For example a mixture of Gaussians is frequently used and the fitting is usually performed using the *Expectation Maximization* (EM) algorithm [3].

When this mixture fitting is used as part of a clustering procedure, the mixing coefficients are interpreted as marginal cluster probabilities $p(j)$ and the component densities as conditional densities $p(x|j)$. For these we can compute $p(j|x)$:

$$p(j|x) \quad = \quad \frac{p(x|j)\,p(j)}{\sum_i p(x|i)\,p(i)}.$$

To do partitional (a.k.a. "hard") clustering, we assign the object corresponding to x to the cluster j for which $p(j|x)$ is maximum. The values $\{p(j|x)\}$ can also be interpreted as *degrees of membership* in a "soft" clustering scheme.

A good (and standard) way to evaluate the density-estimation process is to take a *validation* set $\mathbf{X}_v$ drawn from the same population as $\mathbf{X}$ and use the model estimated from $\mathbf{X}$ to compute the probability (or *density*) of $\mathbf{X}_v$ (also referred to as the "validated likelihood" of the model). The reason for using an external validation set rather than the original data $\mathbf{X}$ is to preclude an overly optimistic estimate due to *over fitting.*

A potential problem with this validation-set strategy is that the amount of data available may be limited and the relationship between the number of objects/vectors n and the number of adjustable parameters in the optimum model may be such that all of the data is needed to get reliable estimates of the model structure and parameters. In this case the technique of *cross validation* can be used. In *k-fold cross validation* $\mathbf{X}$ is divided into k non-overlapping subsets of n/k vectors. For each of these subsets, the model-estimation process is performed on its complement and the likelihood is then measured on the withheld subset. This is performed for each of the subsets and the k likelihoods are averaged to get the final estimate. In the extreme form of this (known as "leave-one-out") $k = n$.

This measure is *internal* in the sense the the score is based on the feature values x, but it is *external* in the sense that the feature vectors on which it is based are not from the set used to do the mixture fitting i.e. they are from a separate *validation set.*

When density estimation is used to do clustering, evaluation using validated likelihood may be reasonable. A problem is that it may give weight to the ability of the model to describe the distribution of features that have very little ability to discriminate among the underlying clusters. The extreme form of these features are the so-called "noise features" identified (for example) in the model-selection/clustering algorithms of [20, 21]. These are features that follow the same distribution in all clusters. Another case where validated likelihood may not be appropriate is where the clustering algorithm determines which features to use in clustering. Thus two different density-estimation-based clustering algorithms may be operating in different feature spaces and their associated likelihood values can therefore not be compared directly.

C.2. The Case for External Measures

C.2.1. Cases where Internal Validation is Impossible In addition to the cases described above where density estimation is used to do clustering and yet the use of validated likelihood may still be problematic, there are cases where density estimation isn't used. In fact, in some cases, there are no feature values to be described; only pairwise similarity values. Many clustering algorithms that can handle either of these cases are based on methods of graph analysis. An early example is based on the *minimum spanning tree* (MST)[22]. In such cases external methods would seem to provide the best choice (if not the only choice) as validation measures.

C.2.2. The Ground Truth The goodness of the quality assessment performed using an external measure is obviously limited by the quality of the validation set and its ground-truth labels. The set of objects should be representative of those the algorithm will encounter in practice and their labeling should obviously reflect the task it is desired that the clustering algorithm perform.

In certain cases constructing such a validation set may be problematic. For example in some cases there may be more than one meaningful way to cluster the set of objects. One aspect of this is a granularity issue - if the objects fit in a hierarchy of categories (i.e. a taxonomy), what level in that hierarchy should the ground-truth labels reflect?[12] A more difficult issue is the following. Suppose one wishes to evaluate a general purpose clustering algorithm but there are multiple unrelated ways that one might cluster the objects. How should the validation-set objects be labeled in these cases? We would propose the following answers to these.

1. In the first case (the hierarchy) the labels should reflect the lowest level in the hierarchy. If a clustering algorithm chooses to cluster at a higher level in the hierarchy, it will still receive a reasonable accuracy score due to the power of the labels of the higher level nodes in predicting the bottom-level labels of associated objects, which would reduce the number of bits required to encode those bottom-level labels.

2. In the second case (multiple possible organizations) a set of classes corresponding to the Cartesian product of all the sets (organizations) should be constructed and each class in the combined set treated as a different class. Thus if there were three possible organizations corresponding to the three sets of classes $A = \{a_i\}$, $B = \{b_j\}$ and $C = \{c_k\}$, the new set of classes will be $A \times B \times C = \{a_i b_j c_k\}$, where the class $a_i b_j c_k$ corresponds to those objects in the intersection $a_i \cap b_j \cap c_k$. In this case, the best accuracy score will obviously be obtained by discovering the complete $A \times B \times C$ organization, but discovering any of the single organizations (A, B or C) or associated pairs ($A \times B$ and so on) will be rewarded because knowing the associated labels will allow the full $A \times B \times C$ labels to be encoded with fewer bits.

[12] In this work we have limited ourselves to flat partitional clustering. A variation of our measure for hierarchical clustering is planned for future work.

It is obviously also important that the number of objects per class in the validation set be sufficient to allow the underlying structure to be captured. The Cartesian-product classification scheme just outlined may result in a large number of labeled objects being required to have a sufficient validation set. If practical considerations preclude these large numbers of objects, however, then one must proceed carefully, attempting to observe, for example, when a clustering algorithm has discovered a viable structure that is not reflected in the validation set.