Adapting the Right Measures for K-means Clustering

Junjie Wu
Sch'l of Eco. & Mgt
Beihang University
wujj@buaa.edu.cn

Hui Xiong
Rutgers Business Sch'l
Rutgers University
hxiong@rutgers.edu

Jian Chen
Sch'l of Eco. & Mgt
Tsinghua University
jchen@mail.tsinghua.edu.cn

ABSTRACT

Clustering validation is a long standing challenge in the clustering literature. While many validation measures have been developed for evaluating the performance of clustering algorithms, these measures often provide inconsistent information about the clustering performance and the best suitable measures to use in practice remain unknown. This paper thus fills this crucial void by giving an organized study of 16 external validation measures for K-means clustering. Specifically, we first introduce the importance of measure normalization in the evaluation of the clustering performance on data with imbalanced class distributions. We also provide normalization solutions for several measures. In addition, we summarize the major properties of these external measures. These properties can serve as the guidance for the selection of validation measures in different application scenarios. Finally, we reveal the interrelationships among these external measures. By mathematical transformation, we show that some validation measures are equivalent. Also, some measures have consistent validation performances. Most importantly, we provide a guide line to select the most suitable validation measures for K-means clustering.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; I.5.3 [**Pattern Recognition**]: Clustering

General Terms

Measurement, Experimentation

Keywords

Cluster Validation, External Criteria, K-means

1. INTRODUCTION

Clustering validation has long been recognized as one of the vital issues essential to the success of clustering applications [10]. Despite the vast amount of expert endeavor spent on this problem [7], there is no consistent and conclusive solution to cluster validation. The best suitable measures to use in practice remain unknown. Indeed, there are many challenging validation issues which have not been fully addressed in the clustering literature. For instance, the importance of normalizing validation measures has not been fully established. Also, the relationship between different validation measures is not clear. Moreover, there are important properties associated with validation measures which are important to the selection of the use of these measures but have not been well characterized. Finally, given the fact that different validation measures may be appropriate for different clustering algorithms, it is necessary to have a focused study of cluster validation measures on a specified clustering algorithm at one time.

To that end, in this paper, we limit our scope to provide an organized study of external validation measures for K-means clustering [14]. The rationale of this pilot study is as follows. K-means is a well-known, widely used, and successful clustering method. Also, external validation measures evaluate the extent to which the clustering structure discovered by a clustering algorithm matches some external structure, e.g., the one specified by the given class labels. From a practical point view, external clustering validation measures are suitable for many application scenarios. For instance, if external validation measures show that a document clustering algorithm can lead to the clustering results which can match the categorization performance by human experts, we have a good reason to believe this clustering algorithm has a practical impact on document clustering.

Along the line of adapting validation measures for K-means, we present a detailed analysis of 16 external validation measures, as shown in Table 1. Specifically, we first establish the importance of measure normalization by highlighting some unnormalized measures which have issues in the evaluation of the clustering performance on data with imbalanced class distributions. In addition, to show the importance of measure normalization, we also provide normalization solutions for several measures. The key challenge here is to identify the lower and upper bounds of validation measures. Furthermore, we reveal some major properties of these external measures, such as consistency, sensitivity, and symmetry properties. These properties can serve as the guidance for the selection of validation measures in different application scenarios. Finally, we also show the interrelationships among these external measures. We show that some validation measures are equivalent and some measures have consistent validation performances.

KDD'09, June 28–July 1, 2009, Paris, France.
Copyright 2009 ACM 978-1-60558-495-9/09/06 ...$5.00.

Table 1: External Cluster Validation Measures.

	Measure	Notation	Definition	Range
1	Entropy	E	$-\sum_i p_i(\sum_j \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i})$	$[0, \log K']$
2	Purity	P	$\sum_i p_i(\max_j \frac{p_{ij}}{p_i})$	(0,1]
3	F-measure	F	$\sum_j p_j \max_i [2 \frac{p_{ij}}{p_i} \frac{p_{ij}}{p_j} / (\frac{p_{ij}}{p_i} + \frac{p_{ij}}{p_j})]$	(0,1]
4	Variation of Information	VI	$-\sum_i p_i \log p_i - \sum_j p_j \log p_j - 2 \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j}$	$[0, 2 \log \max(K, K')]$
5	Mutual Information	MI	$\sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j}$	$(0, \log K']$
6	Rand statistic	R	$[\binom{n}{2} - \sum_i \binom{n_{i\cdot}}{2} - \sum_j \binom{n_{\cdot j}}{2} + 2 \sum_{ij} \binom{n_{ij}}{2}] / \binom{n}{2}$	(0,1]
7	Jaccard coefficient	J	$\sum_{ij} \binom{n_{ij}}{2} / [\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2} - \sum_{ij} \binom{n_{ij}}{2}]$	[0,1]
8	Fowlkes and Mallows index	FM	$\sum_{ij} \binom{n_{ij}}{2} / \sqrt{\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}}$	[0,1]
9	Hubert Γ statistic I	Γ	$\frac{\binom{n}{2} \sum_{ij} \binom{n_{ij}}{2} - \sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}}{\sqrt{\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2} [\binom{n}{2} - \sum_i \binom{n_{i\cdot}}{2}] [\binom{n}{2} - \sum_j \binom{n_{\cdot j}}{2}]}}$	(-1,1]
10	Hubert Γ statistic II	Γ'	$[\binom{n}{2} - 2 \sum_i \binom{n_{i\cdot}}{2} - 2 \sum_j \binom{n_{\cdot j}}{2} + 4 \sum_{ij} \binom{n_{ij}}{2}] / \binom{n}{2}$	[0,1]
11	Minkowski score	MS	$\sqrt{\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2} - 2 \sum_{ij} \binom{n_{ij}}{2}} / \sqrt{\sum_j \binom{n_{\cdot j}}{2}}$	$[0, +\infty)$
12	classification error	ε	$1 - \frac{1}{n} \max_\sigma \sum_j n_{\sigma(j),j}$	[0,1)
13	van Dongen criterion	VD	$(2n - \sum_i \max_j n_{ij} - \sum_j \max_i n_{ij}) / 2n$	$[0, 1)$
14	micro-average precision	MAP	$\sum_i p_i(\max_j \frac{p_{ij}}{p_i})$	(0,1]
15	Goodman-Kruskal coefficient	GK	$\sum_i p_i(1 - \max_j \frac{p_{ij}}{p_i})$	[0,1)
16	Mirkin metric	M	$\sum_i n_{i\cdot}^2 + \sum_j n_{\cdot j}^2 - 2 \sum_i \sum_j n_{ij}^2$	$[0, 2\binom{n}{2})$

Note: $p_{ij} = n_{ij}/n,\ p_i = n_{i\cdot}/n,\ p_j = n_{\cdot j}/n$.

Most importantly, we provide a guide line to select the most suitable validation measures for K-means clustering. After carefully profiling these validation measures, we believe it is most suitable to use the normalized van Dongen criterion (VD_n) which has a simple computation form, satisfies mathematically sound properties, and can measure well on the data with imbalanced class distributions. However, for the case that the clustering performance is hard to distinguish, we may want to use the normalized Variation of Information (VI_n) instead, since the measure VI_n has high sensitivity on detecting the clustering changes.

2. EXTERNAL VALIDATION MEASURES

In this section, we introduce a suite of 16 widely used external clustering validation measures. To the best of our knowledge, these measures represent a good coverage of the validation measures available in different fields, such as data mining, information retrieval, machine learning, and statistics. A common ground of these measures is that they can be computed by the contingency matrix as follows.

Table 2: The Contingency Matrix.

Partition C

		C_1	C_2	$\cdots$	$C_{K'}$	$\sum$
Partition P	P_1	n_{11}	n_{12}	$\cdots$	$n_{1K'}$	$n_{1\cdot}$
	P_2	n_{21}	n_{22}	$\cdots$	$n_{2K'}$	$n_{2\cdot}$
	$\cdot$	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdot$
	P_K	n_{K1}	n_{K2}	$\cdots$	$n_{KK'}$	$n_{K\cdot}$
	$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot K'}$	n

The Contingency Matrix. Given a data set D with n objects, assume that we have a partition $P = \{P_1, \cdots, P_K\}$ of D, where $\bigcup_{i=1}^{K} P_i = D$ and $P_i \bigcap P_j = \phi$ for $1 \leq i \neq j \leq K$, and K is the number of clusters. If we have "true" class labels for the data, we can have another partition on D: $C = \{C_1, \cdots, C_{K'}\}$, where $\bigcup_{i=1}^{K'} C_i = D$ and $C_i \bigcap C_j = \phi$ for $1 \leq i \neq j \leq K'$, where K' is the number of classes. Let n_{ij} denote the number of objects in cluster P_i from class C_j, then the information on the overlap between the two partitions can be written in the form of a contingency matrix, as shown in Table 2. Throughout this paper, we will use the notations in this contingency matrix.

The Measures. Table 1 shows the list of measures to be studied. The "Definition" column gives the computation forms of the measures by using the notations in the contingency matrix. Next, we briefly introduce these measures.

The entropy and purity are frequently used external measures for K-means [20, 26]. They measure the "purity" of the clusters with respect to the given class labels.

F-measure was originally designed for the evaluation of hierarchical clustering [19, 13], but has also been employed for partitional clustering. It combines the precision and recall concepts from the information retrieval community.

The Mutual Information (MI) and Variation of Information (VI) were developed in the field of information theory [3]. MI measures how much information one random variable can tell about another one [21]. VI measures the amount of information that is lost or gained in changing from the class set to the cluster set [16].

The Rand statistic [18], Jaccard coefficient, Fowlkes and Mallows index [5], and Hubert's two statistics [8, 9] evaluate the clustering quality by the agreements and/or disagreements of the pairs of data objects in different partitions.

The Minkowski score [1] measures the difference between the clustering results and a reference clustering (true clusters). And the difference is computed by counting the disagreements of the pairs of data objects in two partitions.

The classification error takes a classification view on clustering [2]. It tries to map each class to a different cluster so as to minimize the total misclassification rate. The "σ" in Table 1 is the mapping of class j to cluster $\sigma(j)$.

The van Dongen criterion [23] was originally proposed for evaluating graph clustering. It measures the representativeness of the majority objects in each class and each cluster.

Finally, the micro-average precision, Goodman-Kruskal coefficient [6] and Mirkin metric [17] are also popular measures. However, the former two are equivalent to the purity measure and the Mirkin metric is equivalent to the Rand statistic ($M/2\binom{n}{2} + R = 1$). As a result, we will not discuss these three measures in the future sections.

In summary, we have 13 (out of 16) candidate measures. Among them, P, F, MI, R, J, FM, Γ, and Γ' are positive measures — a higher value indicates a better clustering performance. The remainder, however, consists of measures based on the distance notion. Throughout this paper, we will use the acronyms of these measures.

3. DEFECTIVE VALIDATION MEASURES

In this section, we present some validation measures which will produce misleading validation results for K-means on data with skewed class distributions.

3.1 K-means: The Uniform Effect

One of the unique characteristic of K-means clustering is the so-called uniform effect; that is, K-means tends to produce clusters with relatively uniform sizes [25]. To quantify the uniform effect, we use the coefficient of variation (CV) [4], a statistic which measures the dispersion degree of a random distribution. CV is defined as the ratio of the standard deviation to the mean. Given a sample data objects $X = \{x_1, x_2, \ldots, x_n\}$, we have $CV = s/\bar{x}$, where $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $s = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}$. CV is a dimensionless number that allows the comparison of the variations of populations that have significantly different mean values. In general, the larger the CV value is, the greater the variability in the data.

Example. Let CV_0 denote the CV value of the "true" class sizes and CV_1 denote the CV value of the resultant cluster sizes. We use the `sports` data set [22] to illustrate the uniform effect by K-means. The "true" class sizes of `sports` have $CV_0 = 1.02$. We then use the CLUTO implementation of K-means [11] with default settings to cluster `sports` into seven clusters. We also compute the CV value of the resultant cluster sizes and get $CV_1 = 0.42$. Therefore, the CV difference is $DCV = CV_1 - CV_0 = -0.6$, which indicates a significant uniform effect in the clustering result.

Indeed, it has been empirically validated that the 95% confidence interval of CV_1 values produced by K-means is in [0.09, 0.85] [24]. In other words, for data sets with CV_0 values greater than 0.85, the uniform effect of K-means can distort the cluster distribution significantly.

Now the question is: Can these widely used validation measures capture the negative uniform effect by K-means clustering? Next, we provides a necessary but not sufficient criterion to testify whether a validation measure can be effectively used to evaluate K-means clustering.

3.2 A Necessary Selection Criterion

Assume that we have a sample document data containing 50 documents from 5 classes. The class sizes are 30, 2, 6, 10 and 2, respectively. Thus, we have $CV_0 = 1.166$, which implies a skewed class distribution.

For this sample data set, we assume there are two clustering results as shown in Table 3. In the table, the first result consists of five clusters with extremely balanced sizes. This is also indicated by $CV_1 = 0$. In contrast, for the second result, the five clusters have varied cluster sizes with $CV_1 = 1.125$, much closer to the CV value of the "true" class sizes. Therefore, from a data distribution point of view, the second result should be better than the first one.

Indeed, if we take a closer look on contingency Matrix I in Table 3, we can find that the first clustering partitions the objects of the largest class C_1 into three balanced sub-clusters. Meanwhile, the two small classes C_2 and C_5 are totally "disappeared" — they are overwhelmed in cluster P_5 by the objects from class C_3. In contrast, we can easily identify all the classes in the second clustering result, since they have the majority objects in the corresponding clusters. Therefore, we can draw the conclusion that the first clustering is indeed much worse than the second one.

Table 3: Two Clustering Results.

I	C_1	C_2	C_3	C_4	C_5
P_1	10	0	0	0	0
P_2	10	0	0	0	0
P_3	10	0	0	0	0
P_4	0	0	0	10	0
P_5	0	2	6	0	2

II	C_1	C_2	C_3	C_4	C_5
P_1	27	0	0	2	0
P_2	0	2	0	0	0
P_3	0	0	6	0	0
P_4	3	0	0	8	0
P_5	0	0	0	0	2

As shown in Section 3.1, K-means tends to produce clusters with relatively uniform sizes. Thus the first clustering in Table 3 can be regarded as the negative result of the uniform effect. So we establish the first necessary but not sufficient criterion for selecting the measures for K-means as follows.

Criterion 1. *If an external validation measure cannot capture the uniform effect by K-means on data with skewed class distributions, this measure is not suitable for validating the results of K-means clustering.*

Next, we proceed to see which existing external cluster validation measures can satisfy this criterion.

3.3 The Cluster Validation Results

Table 4 shows the validation results for the two clusterings in Table 3 by all 13 external validation measures. We highlighted the better evaluation of each validation measure.

As shown in Table 4, only three measures, E, P and MI, cannot capture the uniform effect by K-means and their validation results can be misleading. In other words, these measures are not suitable for evaluating the K-means clustering. These three measures are defective validation measures.

3.4 The Issues with the Defective Measures

Here, we explore the issues with the defective measures. First, the problem of the **entropy** measure lies in the fact that it cannot evaluate the integrity of the classes.

We know $E = -\sum_i p_i \sum_j \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i}$. If we take a random variable view on cluster P and class C, then $p_{ij} = n_{ij}/n$ is the joint probability of the event: $\{P = P_i \bigwedge C = C_j\}$, and $p_i = n_{i\cdot}/n$ is the marginal probability. Therefore, $E = \sum_i p_i \sum_j -p(C_j|P_i) \log p(C_j|P_i) = \sum_i p_i H(C|P_i)$ $= H(C|P)$, where $H(\cdot)$ is the Shannon entropy [3]. The above implies that the entropy measure is nothing but the conditional entropy of C on P. In other words, if the objects in each large partition are mostly from the same class, the entropy value tends to be small (indicating a better clustering quality). This is usually the case for K-means clustering on highly imbalanced data sets, since K-means tends to partition a large class into several pure sub-clusters. This leads to the problem that the integrity of the objects from the same class has been damaged. The entropy measure cannot capture this information and penalize it.

The **mutual information** is strongly related to the entropy measure. We illustrate this by the following Lemma.

Lemma 1. *The mutual information measure is equivalent to the entropy measure for cluster validation.*

Proof. By information theory, $MI = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_i p_j} = H(C) - H(C|P) = H(C) - E$. Since $H(C)$ is a constant for any given data set, MI is essentially equivalent to E. □

The **purity** measure works in a similar fashion as the entropy measure. That is, it measures the "purity" of each cluster by the ratio of the objects from the majority class. Thus, it has the same problem as the entropy measure for evaluating K-means clustering.

In summary, entropy, purity and mutual information are defective measures for validating K-means clustering.

Table 4: The Cluster Validation Results

	E	P	F	MI	VI	R	J	FM	Γ	Γ'	MS	ε	VD
I	**0.274**	**0.920**	0.617	**1.371**	1.225	0.732	0.375	0.589	0.454	0.464	0.812	0.480	0.240
II	0.396	0.9	**0.902**	1.249	**0.822**	**0.857**	**0.696**	**0.821**	**0.702**	**0.714**	**0.593**	**0.100**	**0.100**

3.5 Improving the Defective Measures

Here, we give the improved versions of the above three defective measures: entropy, mutual information, and purity.

Lemma 2. *The Variation of Information measure is an improved version of the entropy measure.*

Proof. If we view cluster P and class C as two random variables, it has been shown that $VI = H(C) + H(P) - 2MI = H(C|P) + H(P|C)$ [16]. The component $H(C|P)$ is nothing but the entropy measure, and the component $H(P|C)$ is a valuable supplement to $H(C|P)$. That is, $H(P|C)$ evaluates the integrity of each class along different clusters. Thus, we complete the proof. □

By Lemma 1, we know MI is equivalent to E. Therefore, VI is also an improved version of MI.

Lemma 3. *The van Dongen criterion is an improved version of the purity measure.*

Proof. $VD = \frac{2n-\sum_i \max_j n_{ij} - \sum_j \max_i n_{ij}}{2n} = 1 - \frac{1}{2}P - \frac{\sum_j \max_i n_{ij}}{2n}$. Apparently, $\sum_j \max_i n_{ij}/n$ reflects the integrity of the classes and is a supplement to the purity measure. □

4. MEASURE NORMALIZATION

In this section, we show the importance of measure normalization and provide normalization solutions to some measures whose normalized forms are not available.

4.1 Normalizing the Measures

Generally speaking, normalizing techniques can be divided into two categories. One is based on a statistical view, which formulates a baseline distribution to correct the measure for randomness. A clustering can then be termed "valid" if it has an unusually high or low value, as measured with respect to the baseline distribution. The other technique uses the minimum and maximum values to normalize the measure into the [0,1] range. We can also take a statistical view on this technique with the assumption that each measure takes a uniform distribution over the value interval.

The Normalizations of R, FM, Γ, Γ', J and MS. The normalization scheme can take the form as

$$S_n = \frac{S - E(S)}{\max(S) - E(S)}, \tag{1}$$

where $\max(S)$ is the maximum value of the measure S, and $E(S)$ is the expected value of S based on the baseline distribution. Some measures derived from the statistics community, such as R, FM, Γ and Γ', usually take this scheme.

Specifically, Hubert and Arabie (1985) [9] suggested to use the multivariate hypergeometric distribution as the baseline distribution in which the row and column sums are fixed in Table 2, but the partitions are randomly selected. This determines the expected value as follows.

$$E(\sum_i \sum_j \binom{n_{ij}}{2}) = \frac{\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}}{\binom{n}{2}}. \tag{2}$$

Based on this value, we can easily compute the expected values of R, FM, Γ and Γ' respectively, since they are the linear functions of $\sum_i \sum_j \binom{n_{ij}}{2}$ under the hypergeometric distribution assumption. Furthermore, although the exact maximum values of the measures are computationally prohibited under the hypergeometric distribution assumption, we can still reasonably approximate them by 1. Then, according to Equation (1) and (2), we can finally have the normalized R, FM, Γ and Γ' measures, as shown in Table 5.

Table 5: The Normalized Measures.

Measure		Normalization
1	R_n, Γ'_n	$(m - m_1 m_2/M)/(m_1/2 + m_2/2 - m_1 m_2/M)$
2	J'_n, MS'_n	$(m_1 + m_2 - 2m)/(m_1 + m_2 - 2m_1 m_2/M)$
3	FM_n	$(m - m_1 m_2/M)/(\sqrt{m_1 m_2} - m_1 m_2/M)$
4	Γ_n	$(mM - m_1 m_2)/\sqrt{m_1 m_2 (M - m_1)(M - m_2)}$
5	VI_n	$1 + 2\frac{\sum_i \sum_j p_{ij} \log(p_{ij}/p_i p_j)}{(\sum_i p_i \log p_i + \sum_j p_j \log p_j)}$
6	VD_n	$\frac{(2n - \sum_i \max_j n_{ij} - \sum_j \max_i n_{ij})}{(2n - \max_i n_{i\cdot} - \max_j n_{\cdot j})}$
7	F_n	$(F - F_-)/(1 - F_-)$
8	ε_n	$(1 - \frac{1}{n} \max_\sigma \sum_j n_{\sigma(j),j})/(1 - 1/\max(K, K'))$

Note: (1) $m = \sum_{i,j} \binom{n_{ij}}{2}$, $m_1 = \sum_i \binom{n_{i\cdot}}{2}$, $m_2 = \sum_j \binom{n_{\cdot j}}{2}$, $M = \binom{n}{2}$.
(2) $p_i = n_{i\cdot}/n$, $p_j = n_{\cdot j}/n$, $p_{ij} = n_{ij}/n$.
(3) Refer to Table 1 for F, and Procedure 1 for F_-.

The normalization of J and MS is a little bit complex, since they are not linear to $\sum_i \sum_j \binom{n_{ij}}{2}$. Nevertheless, we can still normalize the equivalent measures converted from them. Let $J' = \frac{1-J}{1+J} = \frac{2}{1+J} - 1$ and $MS' = MS^2$.

It is easy to show $J' \Leftrightarrow J$ and $MS' \Leftrightarrow MS$. Then based on the hypergeometric distribution assumption, we have the normalized J' and MS' as shown in Table 5. Since J' and MS' are negative measures — a lower value implies a better clustering, we normalize them by modifying Equation (1) as $S_n = (S - \min(S))/(E(S) - \min(S))$.

Finally, we would like to point out some interrelationships between these measures as follows.

Proposition 1.

$$(1) \quad (R_n \equiv \Gamma'_n) \Leftrightarrow (J'_n \equiv MS'_n).$$
$$(2) \quad \Gamma_n \equiv \Gamma.$$

The above proposition indicates that the normalized Hubert Γ statistic I (Γ_n) is the same as Γ. Also, the normalized Rand statistic (R_n) is the same as the normalized Hubert Γ statistic II (Γ'_n). In addition, the normalized Rand statistic (R_n) is equivalent to J'_n, which is the same as MS'_n. Therefore, we have three independent normalized measures including R_n, FM_n and Γ_n for further study. Note that this proposition can be easily proved by mathematical transformation. Due to the space limitation, we omit the proof.

The Normalizations of VI and VD. Another normalization scheme is formalized as $S_n = \frac{S - \min(S)}{\max(S) - \min(S)}$.

Some measures, such as VI and VD, often take this scheme. However, to know the exact maximum and minimum values is often impossible. So we usually turn to a reasonable approximation, e.g., the upper bound for the maximum, or the lower bound for the minimum.

When the cluster structure matches the class structure perfectly, $VI = 0$. So, we have $\min(VI) = 0$. However, finding the exact value of $\max(VI)$ is computationally infeasible. Meila suggested to use $2 \log \max(K, K')$ to approximate $\max(VI)$ [16], so the normalized VI is $\frac{VI}{2 \log \max(K,K')}$.

The VD in Table 1 can be regarded as a normalized measure. In this measure, $2n$ has been taken as the upper bound [23], and $\min(VD) = 0$.

However, we found that the above normalized VI and VD cannot well capture the uniform effect of K-means, because the proposed upper bound for VI or VD is not tight enough. Therefore, we propose new upper bounds as follows.

Lemma 4. *Let random variables C and P denote the class and cluster sizes respectively, $H(\cdot)$ be the entropy function, then $VI \leq H(C) + H(P) \leq 2\log\max(K', K)$.*

Lemma 4 gives a tighter upper bound $H(C) + H(P)$ than $2\log\max(K', K)$ which was provided by Meila [16]. With this new upper bound, we can have the normalized VI_n as shown in Table 5. Also, we would like to point out that, if we use $H(P)/2 + H(C)/2$ as the upper bound to normalize mutual information, the VI_n can be equivalent to the normalized mutual information MI_n ($VI_n + MI_n = 1$).

Lemma 5. *Let $n_{i\cdot}$, $n_{\cdot j}$ and n be the values in Table 2, then $VD \leq (2n - \max_i n_{i\cdot} - \max_j n_{\cdot j})/2n \leq 1$.*

Due to the page limit, we omit some proofs. The above two lemmas imply that the tighter upper bounds of VI and VD are the functions of the class and cluster sizes. Using these two new upper bounds, we can derive the normalized VI_n and VD_n in Table 5.

The Normalization of F and ε have been seldom discussed in the literature. As we know, $\max(F) = 1$. Now the goal is to find a tight lower bound. In the following, we propose a procedure to find the lower bound of F.

Procedure 1: The computation of F_-.

```
Let $n^* = \max_i n_{i\cdot}$.
Sort the class sizes so that $n_{\cdot[1]} \leq n_{\cdot[2]} \leq \cdots \leq n_{\cdot[K']}$.
Let $a_j = 0$, for $j = 1, 2, \cdots, K'$.
for $j = 1 : K'$
    if $n^* \leq n_{\cdot[j]}$,    $a_j = n^*$, break.
    else    $a_j = n_{\cdot[j]}$, $n^* \leftarrow n^* - n_{\cdot[j]}$.
$F_- = (2/n)\sum_{j=1}^{K'} a_j/(1 + \max_i n_{i\cdot}/n_{\cdot[j]})$.
```

With the above procedure, we can have the following lemma, which finds a lower bound for F.

Lemma 6. *Given F_- computed by Procedure 1, $F \geq F_-$.*

Proof. It is easy to show:

$$F = \sum_j \frac{n_{\cdot j}}{n} \max_i \frac{2n_{ij}}{n_{i\cdot} + n_{\cdot j}} \geq \frac{2}{n} \max_i \sum_j \frac{n_{ij}}{n_{i\cdot}/n_{\cdot j} + 1} \tag{3}$$

Let us consider an optimization problem as follows.

$$\min_{x_{ij}} \sum_j \frac{x_{ij}}{n_{i\cdot}/n_{\cdot j} + 1}$$

$$s.t.\ \sum_j x_{ij} = n_{i\cdot};\ \forall j,\ x_{ij} \leq n_{\cdot j};\ \forall j,\ x_{ij} \in \mathbb{Z}_+$$

For this optimization problem, to have the minimum objective value, we need to assign as many objects as possible to the cluster with highest $n_{i\cdot}/n_{\cdot j} + 1$, or equivalently, with smallest $n_{\cdot j}$. Let $n_{\cdot[0]} \leq n_{\cdot[1]} \leq \cdots \leq n_{\cdot[K']}$ where the virtual $n_{\cdot[0]} = 0$, and assume $\sum_{j=0}^{l} n_{\cdot[j]} < n_{i\cdot} \leq \sum_{j=0}^{l+1} n_{\cdot[j]}$, $l \in \{0, 1, \cdots, K'-1\}$, we have the optimal solution:

$$x_{i[j]} = \begin{cases} n_{\cdot[j]},\ 1 \leq j \leq l; \\ n_{i\cdot} - \sum_{k=1}^{l} n_{\cdot[k]},\ j = l+1; \\ 0,\ l+1 < j \leq K'. \end{cases}$$

Therefore, according to (3), $F \geq \frac{2}{n}\max_i \sum_{j=1}^{K'} \frac{x_{i[j]}}{n_{i\cdot}/n_{\cdot[j]}+1}$.

Let $F_i = \frac{2}{n}\sum_{j=1}^{K'} \frac{x_{i[j]}}{n_{i\cdot}/n_{\cdot[j]}+1} = \frac{2}{n}\sum_{j=1}^{K'} \frac{x_{i[j]}/n_{i\cdot}}{1/n_{\cdot[j]}+1/n_{i\cdot}}$. Denote "$x_{i[j]}/n_{i\cdot}$" by "$y_{i[j]}$", and "$\frac{1}{1/n_{\cdot[j]}+1/n_{i\cdot}}$" by "$p_{i[j]}$", we have $F_i = \frac{2}{n}\sum_{j=1}^{K'} p_{i[j]} y_{i[j]}$. Next, we remain to show

$$\arg\max_i F_i = \arg\max_i n_{i\cdot}.$$

Assume $n_{i\cdot} \leq n_{i'\cdot}$, and for some l, $\sum_{j=0}^{l} n_{\cdot[j]} < n_{i\cdot} \leq \sum_{j=0}^{l+1} n_{\cdot[j]}$, $l \in \{0, 1, \cdots, K'-1\}$. This implies that

$$y_{i[j]} \begin{cases} \geq y_{i'[j]},\ 1 \leq j \leq l; \\ \leq y_{i'[j]},\ l+1 < j \leq K'. \end{cases}$$

Since $\sum_{j=1}^{K'} y_{i[j]} = \sum_{j=1}^{K'} y_{i'[j]} = 1$ and $j \uparrow \ \Rightarrow\ p_{i[j]} \uparrow$, we have $\sum_{j=1}^{K'} p_{i[j]} y_{i[j]} \leq \sum_{j=1}^{K'} p_{i[j]} y_{i'[j]}$.

Furthermore, according to the definition of $p_{i[j]}$, we have $p_{i[j]} \leq p_{i'[j]}$, $\forall\ j \in \{1, \cdots, K'\}$. Therefore,

$$F_i = \frac{2}{n}\sum_{j=1}^{K'} p_{i[j]} y_{i[j]} \leq \frac{2}{n}\sum_{j=1}^{K'} p_{i[j]} y_{i'[j]} \leq \frac{2}{n}\sum_{j=1}^{K'} p_{i'[j]} y_{i'[j]} = F_i',$$

which implies that "$n_{i\cdot} \leq n_{i'\cdot}$" is the sufficient condition for "$F_i \leq F_i'$". Therefore, by Procedure 1, we have $F_- = \max_i F_i$, which finally leads to $F \geq F_-$. Thus we complete the proof. □

Therefore, $F_n = (F - F_-)/(1 - F_-)$, as listed in Table 5. Finally, as to ε, we have the following lemma.

Lemma 7. *Given $K' \leq K$, $\varepsilon \leq 1 - 1/K$.*

Proof. Assume $\sigma_1 : \{1, \cdots, K'\} \to \{1, \cdots, K\}$ is the optimal mapping of the classes to different clusters, i.e.,

$$\varepsilon = 1 - \frac{\sum_{j=1}^{K'} n_{\sigma_1(j),j}}{n}.$$

Then we construct a series of mappings $\sigma_s : \{1, \cdots, K'\} \mapsto \{1, \cdots, K\}$ ($s = 2, \cdots, K$) which satisfy

$$\sigma_{s+1}(j) = \mathrm{mod}(\sigma_s(j), K) + 1,\ \forall j \in \{1, \cdots, K'\},$$

where "$\mathrm{mod}(x, y)$" returns the remainder of positive integer x divided by positive integer y. By definition, σ_s ($s = 2, \cdots, K$) can also map $\{1, \cdots, K'\}$ to K' different indices in $\{1, \cdots, K\}$ as σ_1. More importantly we have $\sum_{j=1}^{K'} n_{\sigma_1(j),j} \geq \sum_{j=1}^{K'} n_{\sigma_s(j),j}$, $\forall s = 2, \cdots, K$, and $\sum_{s=1}^{K}\sum_{j=1}^{K'} n_{\sigma_s(j),j} = n$.

Accordingly, we have $\sum_{j=1}^{K'} n_{\sigma_1(j),j} \geq \frac{n}{K}$, which implies $\varepsilon \leq 1 - 1/K$. The proof is completed. □

Therefore, we can use $1 - 1/K$ as the upper bound of ε, and the normalized ε_n is shown in Table 5.

4.2 The *DCV* Criterion

Here, we present some experiments to show the importance of DCV ($CV_1 - CV_0$) for selecting validation measures.

Experimental Data Sets. Some synthetic data sets were generated as follows. Assume we have a two-dimensional mixture of two Gaussian distributions. The means of the two distributions are [-2,0] and [2,0], respectively. And their covariance matrices are exactly the same as $[\sigma^2\ 0;\ 0\ \sigma^2]$.

Therefore, given any specific value of σ^2, we can generate a simulated data set with 6000 instances, n_1 instances from the first distribution, and n_2 instances from the second one, where $n_1 + n_2 = 6000$. To produce simulated data sets with imbalanced class sizes, we set a series of n_1 values: {3000, 2600, 2200, 1800, 1400, 1000, 600, 200}. If $n_1 = 200$, $n_2 = 5800$, we have a highly imbalanced data set with $CV_0 = 1.320$. For each mixture model, we generated

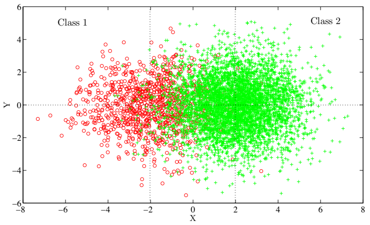
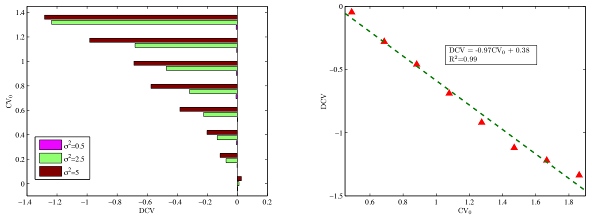


Figure 1: A Simulated Data Set ($n_1 = 1000$, $\sigma^2 = 2.5$).



(a) Simulated Data Sets. (b) Sampled Data Sets.

Figure 2: Relationship of CV_0 and DCV.

8 simulated data sets with CV_0 ranging from 0 to 1.320. Further, to produce data sets with different clustering tendencies, we set a series of σ^2 values: $\{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$. As σ^2 increases, the mixture model tends to be more unidentifiable. Finally, for each pair of σ^2 and n_1, we repeated the sampling 10 times, thus we can have the average performance evaluation. In summary, we produced $8 \times 10 \times 10 = 800$ data sets. Figure 1 shows a sample data set with $n_1 = 1000$ and $\sigma^2 = 2.5$.

We also did sampling on a real-world data set `hitech` to get some sample data sets with imbalanced class distributions. This data set was derived from the San Jose Mercury newspaper articles [22], which contains 2301 documents about computers, electronics, health, medical, research and technology. Each document is characterized by 126373 terms, and the class sizes are 485, 116, 429, 603, 481 and 187, respectively. We carefully set the sampling ratio for each class, and get 8 sample data sets with the class-size distributions (CV_0) ranging from 0.490 to 1.862, as shown in Table 6. For each data set, we repeated sampling 10 times, so we can observe the averaged clustering performance.

Experimental Tools. We used the MATLAB 7.1 [15] and CLUTO 2.1.1 [11] implementations of K-means. The MATLAB version with the squared Euclidean distance is suitable for low-dimensional and dense data sets, while CLUTO with the cosine similarity is used to handle high-dimensional and sparse data sets. Note that the number of clusters, i.e., K, was set to match the number of "true" classes.

The Application of Criterion 1. Here, we show how we can apply Criterion 1 for selecting measures. As pointed out in Section 3.1, K-means tends to have the uniform effect on imbalanced data sets. This implies that for data sets with skewed class distributions, the clustering results by K-means tend to be away from "true" class distributions.

Table 6: The Sizes of the Sampled Data Sets.

Data Set	1	2	3	4	5	6	7	8
Class 1	100	90	80	70	60	50	40	30
Class 2	100	90	80	70	60	50	40	30
Class 3	100	90	80	70	60	50	40	30
Class 4	250	300	350	400	450	500	550	600
Class 5	100	90	80	70	60	50	40	30
Class 6	100	90	80	70	60	50	40	30
CV_0	0.49	0.686	0.88	1.078	1.27	1.47	1.666	1.86

Table 8: The Benchmark Data Sets.

Data Set	Source	#Class	#Case	#Feature	CV_0
cacmcisi	CA/CI	2	4663	41681	0.53
classic	CA/CI	4	7094	41681	0.55
cranmed	CR/ME	2	2431	41681	0.21
fbis	TREC	17	2463	2000	0.96
hitech	TREC	6	2301	126373	0.50
k1a	WebACE	20	2340	21839	1.00
k1b	WebACE	6	2340	21839	1.32
la1	TREC	6	3204	31472	0.49
la2	TREC	6	3075	31472	0.52
la12	TREC	6	6279	31472	0.50
mm	TREC	2	2521	126373	0.14
ohscal	OHSUMED	10	11162	11465	0.27
re0	Reuters	13	1504	2886	1.50
re1	Reuters	25	1657	3758	1.39
sports	TREC	7	8580	126373	1.02
tr11	TREC	9	414	6429	0.88
tr12	TREC	8	313	5804	0.64
tr23	TREC	6	204	5832	0.93
tr31	TREC	7	927	10128	0.94
tr41	TREC	10	878	7454	0.91
tr45	TREC	10	690	8261	0.67
wap	WebACE	20	1560	8460	1.04
DLBCL	KRBDSR	3	77	7129	0.25
Leukemia	KRBDSR	7	325	12558	0.58
LungCancer	KRBDSR	5	203	12600	1.36
ecoli	UCI	8	336	7	1.16
pageblocks	UCI	5	5473	10	1.95
letter	UCI	26	20000	16	0.03
pendigits	UCI	10	10992	16	0.04
MIN	-	2	77	7	0.03
MAX	-	26	20000	126373	1.95

Note: CA-CACM, CI-CISI, CR-CRANFIELD, ME-MEDLINE.

To further illustrate this, let us take a look at Figure 2(a) of the simulated data sets. As can be seen, for the extreme case of $\sigma^2 = 5$, the DCV values decrease as the CV_0 values increase. Note that DCV values are usually negative since K-means tends to produce clustering results with relative uniform cluster sizes ($CV_1 < CV_0$). This means that, when data become more skewed, the clustering results by K-means tend to be worse. From the above, we know that we can select measures by observing the relationship between the measures and the DCV values. As the DCV values go down, the good measures are expected to show worse clustering performances. Note that, in this experiment, we applied the MATLAB version of K-means.

A similar trend can be found in Figure 2(b) of the sampled data sets. That is, as the CV_0 values go up, the DCV values decrease, which implies worse clustering performances. Indeed, DCV is a good indicator for finding the measures which cannot capture the uniform effect by K-means clustering. Note that, in this experiment, we applied the CLUTO version of K-means clustering.

In the next section, we use the Kendall's rank correlation (κ) [12] to measure the relationships between external validation measures and DCV. Note that, $\kappa \in [-1, 1]$. $\kappa = 1$ indicates a perfect positive rank correlation, whereas $\kappa = -1$ indicates an extremely negative rank correlation.

4.3 The Effect of Normalization

In this subsection, we show the importance of measure normalization. Along this line, we first apply K-means clustering on the simulated data sets with $\sigma^2 = 5$ and the sampled data sets from `hitech`. Then, both unnormalized and normalized measures are used for cluster validation. Finally, the rank correlation between DCV and the measures are computed and the results are shown in Table 7.

As can be seen in the table, if we use the unnormalized measures to do cluster validation, only three measures, namely R, Γ, Γ', have strong consistency with DCV on both groups of data sets. VI, VD and MS even show strong con-

Table 7: The Correlation between DCV and the Validation Measures.

$\boldsymbol{\kappa}$	VI	VD	MS	ε	F	R	J	FM	Γ	Γ'
Simulated Data	***-0.71***	***0.79***	***-0.79***	1.00	1.00	1.00	0.91	***0.71***	1.00	1.00
Sampled Data	***-0.93***	***-1.00***	***-1.00***	***0.50***	***0.21***	1.00	***0.50***	***-0.43***	0.93	1.00

$\boldsymbol{\kappa}$	VI_n	VD_n	MS'_n	ε_n	F_n	R_n	J'_n	FM_n	Γ_n	Γ'_n
Simulated Data	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Sampled Data	1.00	1.00	1.00	***0.50***	***0.79***	1.00	1.00	1.00	0.93	1.00

Note: Poor or even negative correlations have been highlighted by the bold and italic fonts.

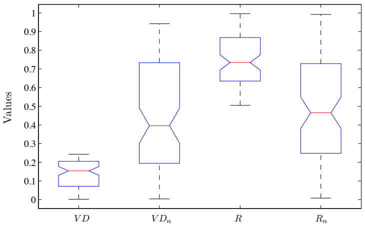
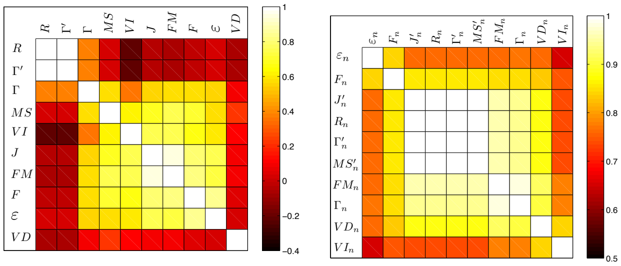


Figure 3: Un-normalized and Normalized Measures.



(a) Unnormalized Measures. (b) Normalized Measures.

Figure 4: Correlations of the Measures.

flict with DCV on the sampled data sets, since their κ values are all close to -1 on sampled data. In addition, we notice that F, ε, J and FM show weak correlation with DCV.

Table 7 shows the rank correlations between DCV and the normalized measures. As can be seen, all the normalized measures show perfect consistency with DCV except for F_n and ε_n. This indicates that the normalization is crucial for evaluating K-means clustering. The proposed bounds for the measures are tight enough to capture the uniform effect in the clustering results.

In Table 7, we can observe that both F_n and ε_n are not consistent with DCV. This indicates that normalization does not help F and ε too much. The reason is that the proposed lower bound for F and upper bound for ε are not very tight. Indeed, the normalizations of F and ε are very challenging. This is due to the fact that they both exploit relatively complex optimization schemes in the computations. As a result, we cannot easily compute the expected values from a multivariate hypergeometric distribution perspective, and it is also difficult to find tighter bounds.

Nevertheless, the above experiments show that the normalization is very valuable. In addition, Figure 3 shows the cluster validation results of the measures on all the simulated data sets with σ^2 ranging from 0.5 to 5. It is clear that the normalized measures have much wider value range than the unnormalized ones along [0,1]. This indicates that the values of normalized measures are more spread in [0, 1].

In summary, to compare cluster validation results across different data sets, we should use normalized measures.

5. MEASURE PROPERTIES

In this section, we investigate measure properties, which can serve as the guidance for the selection of measures.

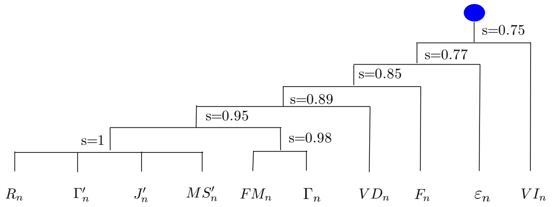


Figure 5: The Measure Similarity Hierarchy.

Table 9: $M(\mathfrak{R}_1) - M(\mathfrak{R}_2)$.

	R_n	FM_n	Γ_n	VD_n	F_n	ε_n	VI_n
R_n	0.00	0.09	0.13	0.08	0.10	0.26	-0.01
FM_n	0.09	0.00	0.04	0.00	0.10	0.22	-0.10
Γ_n	0.13	0.04	0.00	0.04	0.14	0.22	-0.06
VD_n	0.08	0.00	0.04	0.00	0.05	0.20	-0.18
F_n	0.10	0.10	0.14	0.05	0.00	0.08	-0.08
ε_n	0.26	0.22	0.22	0.20	0.08	0.00	0.04
VI_n	-0.01	-0.10	-0.06	-0.18	-0.08	0.04	0.00

5.1 The Consistency between Measures

Here, we define the consistency between a pair of measures in terms of the similarity between their rankings on a series of clustering results. The similarity is measured by the Kendall's rank correlation. And the clustering results are produced by the CLUTO version of K-means clustering on 29 benchmark real-world data sets listed in Table 8. In the experiment, for each data set, the cluster number is set to be the same as the "true" class number.

Figure 4(a) and 4(b) show the correlations between the unnormalized and normalized measures, respectively. One interesting observation is that the normalized measures have much stronger consistency than the unnormalized measures. For instance, the correlation between VI and R is merely -0.21, but it reaches 0.74 for the corresponding normalized measures. This observation indeed implies that the normalized measures tend to give more robust validation results, which also agrees with our previous analysis.

Let us take a closer look on the normalized measures in Figure 4(b). According to the colors, we can roughly find that R_n, Γ'_n, J'_n, MS'_n, FM_n and Γ_n are more similar to one another, while VD_n, F_n, VI_n and ε_n show inconsistency with others in varying degrees. To gain the precise understanding, we do hierarchical clustering on the measures by using their correlation matrix. The resultant hierarchy can be found in Figure 5 ("s" means the similarity). As we know before, R_n, Γ'_n, J'_n and MS'_n are equivalent, so they have perfect correlation to one another, and form the first group. The second group contains FM_n and Γ_n. These two measures behave similarly, and have just slightly weaker consistency with the measures in the first group. Finally, VD_n, F_n, ε_n and VI_n have obviously weaker consistency with other measures in a descending order.

Furthermore, we explore the source of the inconsistency among the measures. To this end, we divide the data sets in Table 8 into two repositories, where $\mathfrak{R}_1$ contains data sets with $CV_0 < 0.8$, and $\mathfrak{R}_2$ contains the rest. Then we compute the correlation matrices of the measures on the two repositories respectively (denoted by $M(\mathfrak{R}_1)$ and $M(\mathfrak{R}_2)$), and observe their difference $(M(\mathfrak{R}_1) - M(\mathfrak{R}_2))$ in Table 9. As can be seen, roughly speaking, all the measures except VI_n show weaker consistency with one another on data sets in $\mathfrak{R}_2$. In other words, while VI_n acts in the opposite way, most measures tend to disagree with one another on data sets with highly imbalanced classes.

5.2 Properties of Measures

In this subsection, we investigate some key properties of external clustering validation measures.

Table 10: Two Clustering Results.

I	C_1	C_2	C_3	$\sum$	II	C_1	C_2	C_3	$\sum$
P_1	**3**	**4**	12	19	P_1	**0**	**7**	12	19
P_2	**8**	**3**	12	23	P_2	**11**	**0**	12	23
P_3	12	12	0	24	P_3	12	12	0	24
$\sum$	23	19	24	66	$\sum$	23	19	24	66

Table 11: The Cluster Validation Results.

	R_n	FM_n	Γ_n	VD_n	F_n	ε_n	VI_n
I	0.16	0.16	0.16	0.71	0.32	0.77	0.78
II	0.24	0.24	0.24	0.71	0.32	0.70	0.62

The Sensitivity. The measures have different sensitivity to the clustering results. Let us illustrate this by an example. For two clustering results in Table 10, the differences between them are the numbers in bold. Then we employ the measures on these two clusterings. Validation results are shown in Table 11. As can be seen, all the measures show different validation results for the two clusterings except for VD_n and F_n. This implies that VD_n and F_n are less sensitive than other measures. This is due to the fact that both VD_n and F_n use maximum functions, which may loose some information in the contingency matrix. Furthermore, VI_n is the most sensitive measure, since the difference of VI_n values for the two clusterings is the largest.

Impact of the Number of Clusters. We use the data set `la2` in Table 8 to show the impact of the number of clusters on the validation measures. Here, we change the cluster numbers from 2 to 15. As shown in Figure 6, the measurement values for all the measures will change as the increase of the cluster numbers. However, the normalized measures including VI_n, VD_n and R_n can capture the same optimal cluster number 5. Similar results can also be observed for other normalized measures, such as F_n, FM_n and Γ_n.

A Summary of Math Properties. We summarize five math properties of measures as follows. Due to the space limit, we omit the proofs here.

Property 1 (Symmetry). *A measure O is symmetric, if $O(M^T) = O(M)$ for any contingence matrix M.*

The *symmetry* property treats the pre-defined class structure as one of the partitions. Therefore, the task of cluster validation is the same as the comparison of partitions. This means transposing two partitions in the contingency matrix should not bring any difference to the measure value. This property is not true for F_n which is a typical measure in asymmetry. Also, ε_n is symmetric if and only if $K = K'$.

Property 2 (N-invariance). *For a contingence matrix M and a positive integer λ, a measure O is n-invariant, if $O(\lambda M) = O(M)$, where n is the number of objects.*

Intuitively, a mathematically sound validation measure should satisfy the *n-invariance* property. However, three measures, namely R_n, FM_n and Γ_n cannot fulfill this requirement. Nevertheless, we can still treat them as the asymptotically n-invariant measures, since they tend to be n-invariant as the increase of n.

Table 12: Math Properties of Measures.

	F_n	VI_n	VD_n	ε_n	R_n	FM_n	Γ_n
P1	No	Yes	Yes	Yes**	Yes	Yes	Yes
P2	Yes	Yes	Yes	Yes	No	No	No
P3	Yes*	Yes*	Yes*	Yes*	No	No	No
P4	No	Yes	Yes	No	No	No	No
P5	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: Yes* — Yes for the un-normalized measures.
Yes** — Yes for $K = K'$.

Property 3 (Convex additivity). *Let $P = \{P_1, \cdots, P_K\}$ be a clustering, P' be a refinement of P[1], and P'_l be the partitioning induced by P' on P_l. Then a measure O is convex additive, if $O(M(P,P')) = \sum_{l=1}^{K} \frac{n_l}{n} O(M(I_{P_l}, P'_l))$, where n_l is the number of data points in P_l, I_{P_l} represents the partitioning on P_l into one cluster, and $M(X,Y)$ is the contingency matrix of X and Y.*

The *convex additivity* property was introduced by Meila [16]. It requires the measures to show additivity along the lattice of partitions. Unnormalized measures including F, VD, VI and ε hold this property. However, none of the normalized measures studied in this paper holds this property.

Property 4 (Left-domain-completeness). *A measure O is left-domain-complete, if, for any contingence matrix M with statistically independent rows and columns,*

$$O(M) = \begin{cases} 0, & O \text{ is a positive measure;} \\ 1, & O \text{ is a negative measure.} \end{cases}$$

When the rows and columns in the contingency matrix are statistically independent, we should expect to see the poorest values of the measures, i.e., 0 for positive measures and 1 for negative measures. Among all the measures, however, only VI_n and VD_n can meet this requirement.

Property 5 (Right-domain-completeness). *A measure O is right-domain-complete, if, for any contingence matrix M with perfectly matched rows and columns,*

$$O(M) = \begin{cases} 1, & O \text{ is a positive measure;} \\ 0, & O \text{ is a negative measure.} \end{cases}$$

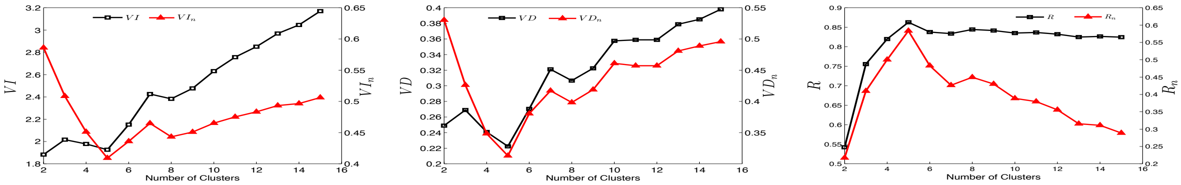
This property requires measures to show optimal values when the class structure matches the cluster structure perfectly. The above normalized measures hold this property.

5.3 Discussions

In a nutshell, among 16 external validation measures shown in Table 1, we first know that Mirkin metric (M) is equivalent to Rand statistic (R), and micro-average precision (MAP) and Goodman-Kruskal coefficient (GK) are equivalent to the purity measure (P) by observing their computational forms. Therefore, the scope of our measure selection is reduced from 16 measures to 13 measures. In Section 3, our analysis shows that purity, mutual information (MI), and entropy (E) are defective measures for evaluating K-means clustering. Also, we know that variation of information (VI) is an improved version of MI and E, and van Dongen criterion (VD) is an improved version of P. As a result, our selection pool is further reduced to 10 measures.

In addition, as shown in Section 4, it is necessary to use the normalized measures for evaluating K-means clustering, since the normalized measures can capture the uniform effect by K-means and allow to evaluate different clustering results on different data sets. By Proposition 1, we know

[1] “P' be a refinement of P” means P' is the descendant node of node P in the lattice of partitions. See [16] for details.



(a) (b) (c)

Figure 6: Impact of the Number of Clusters.

that the normalized Rand statistic (R_n) is the same as the normalized Hubert Γ statistic II (Γ'_n). Also, the normalized Rand statistic is equivalent to J'_n, which is the same as MS'_n. Therefore, we only need to further consider R_n and can exclude J'_n, Γ'_n as well as MS'_n. The results in Section 4 show that the normalized F-measure (F_n) and classification error (ε_n) cannot well capture the uniform effect by K-means. Also, these two measures do not satisfy some math properties in Table 12. As a result, we can exclude them. Now, we have five normalized measures: VI_n, VD_n, R_n, FM_n, and Γ_n. In Figure 5, we know that the validation performances of R_n, FM_n, and Γ_n are very similar to each other. Therefore, we only need to consider to use R_n.

From the above study, we believe it is most suitable to use the normalized van Dongen criterion (VD_n), since VD_n has a simple computation form, satisfies all mathematically sound properties as shown in Table 12, and can measure well on the data with imbalanced class distributions. However, for the case that the clustering performances are hard to distinguish, we may want to use the normalized variation of information (VI_n) instead[2], since VI_n has high sensitivity on detecting the clustering changes. Finally, R_n can also be used as a complementary to the above two measures.

6. CONCLUDING REMARKS

In this paper, we compared and contrasted external validation measures for K-means clustering. As our results revealed, it is necessary to normalize validation measures before they can be employed for clustering validation, since unnormalized measures may lead to inconsistent or even misleading results. This is particularly true for data with imbalanced class distributions. Along this line, we also provide normalization solutions for the measures whose normalized solutions are not available. Furthermore, we summarized the key properties of these measures. These properties should be considered before deciding what is the right measure to use in practice. Finally, we investigated the relationships among these validation measures. The results showed that some validation measures are mathematically equivalent and some measures have very similar validation performances.

7. ACKNOWLEDGEMENTS

This research was partially supported by the National Natural Science Foundation of China (NSFC) (No. 70621061, 70890082, and 70521001), the Rutgers Seed Funding for Collaborative Computing Research, and the National Science Foundation (NSF) of USA via grant number CNS 0831186.

8. REFERENCES

[1] A. Ben-Hur and I. Guyon. Detecting stable clusters using principal component analysis. In *Methods in Molecular Biology*. Humana press, 2003.

[2] M. Brun, C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh, and E.R. Dougherty. Model-based evaluation of clustering validation measures. *Pattern Recognition*, 40:807–824, 2007.

[3] T.M. Cover and J.A. Thomas. *Elements of Information Theory (2nd Edition)*. Wiley-Interscience, 2006.

[4] M. DeGroot and M. Schervish. *Probability and Statistics (3rd Edition)*. Addison Wesley, 2001.

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

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

[7] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Cluster validity methods: Part i. *SIGMOD Rec.*, 31(2):40–45, 2002.

[8] L. Hubert. Nominal scale response agreement as a generalized correlation. *British Journal of Mathematical and Statistical Psychology*, 30:98–103, 1977.

[9] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[10] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[11] G. Karypis. Cluto — software for clustering high-dimensional datasets, version 2.1.1. Oct. 2007.

[12] M.G. Kendall. *Rank Correlation Methods*. New York: Hafner Publishing Co., 1955.

[13] B. Larsen and C. Aone. Fast and effective text mining using linear-time document clustering. In *KDD*, 1999.

[14] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *BSMSP, Vol. I, Statistics*. University of California Press, 1967.

[15] MathWorks. K-means clustering in statistics toolbox.

[16] M. Meila. Comparing clusterings—an axiomatic view. In *ICML*, pages 577–584, 2005.

[17] B. Mirkin. *Mathematical Classification and Clustering*. Kluwer Academic Press, 1996.

[18] W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850, 1971.

[19] C.J.V. Rijsbergen. *Information Retrieval (2nd Edition)*. Butterworths, London, 1979.

[20] Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. In *Workshop on Text Mining, KDD*, 2000.

[21] A. Strehl, J. Ghosh, and R.J. Mooney. Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search, AAAI*, pages 58–64, 2000.

[22] TREC. Text retrieval conference. Oct. 2007.

[23] S. van Dongen. Performance criteria for graph clustering and markov cluster experiments. *TRINS= R0012*, Centrum voor Wiskunde en Informatica. 2000.

[24] J. Wu, H. Xiong, J. Chen, and W. Zhou. A generalization of proximity functions for k-means. In *ICDM*, 2007.

[25] H. Xiong, J. Wu, and J. Chen. K-means clustering versus validation measures: A data distribution perspective. In *KDD*, 2006.

[26] Y. Zhao and G. Karypis. Criterion functions for document clustering: Experiments and analysis. *Machine Learning*, 55(3):311–331, 2004.

[2] Note that the normalized variation of information is equivalent to the normalized mutual information.