Information-Theoretic Distance Measures for Clustering Validation: Generalization and Normalization

Ping Luo, Hui Xiong, *Senior Member*, *IEEE*, Guoxing Zhan, Junjie Wu, and Zhongzhi Shi, *Senior Member*, *IEEE*

Abstract—This paper studies the generalization and normalization issues of information-theoretic distance measures for clustering validation. Along this line, we first introduce a uniform representation of distance measures, defined as quasi-distance, which is induced based on a general form of *conditional entropy*. The quasi-distance possesses three properties: symmetry, the triangle law, and the minimum reachable. These properties ensure that the quasi-distance naturally lends itself as the external measure for clustering validation. In addition, we observe that the ranges of the distance measures are different when they apply for clustering validation on different data sets. Therefore, when comparing the performances of clustering algorithms on different data sets, distance normalization is required to equalize ranges of the distance measures. A critical challenge for distance normalization is to obtain the ranges of a distance measure when a data set is provided. To that end, we theoretically analyze the computation of the maximum value of a distance measure for a data set. Finally, we compare the performances of the partition clustering algorithm K-means on various real-world data sets. The experiments show that the normalized distance measures have better performance than the original distance measures when comparing clusterings of different data sets. Also, the normalized Shannon distance has the best performance among four distance measures under study.

Index Terms—Clustering validation, entropy, information-theoretic distance measures, K-means clustering.

1 Introduction

Clustering analysis [9] provides insight into the data by partitioning the objects into groups (clusters) of objects, such that objects in a cluster are more similar to each other than to objects in other clusters. A longstanding challenge of clustering research is about how to validate clustering results [2], [3], [5], [7], [8], [11], [12], [14], [15]. A promising direction is the use of information-theoretic distance measures, such as Shannon Entropy [22] and Goodman-Kruskal coefficient [24], [10], as external criteria for clustering validation. In other words, these information-theoretic distance measures are used to compare the clustering output with the "true" partition[1] determined by the class label information. In this case, these *external measures* are viewed as the measurement of distances between two partitions of the data.

However, the lack of understanding of the characteristic of these information-theoretic distance measures hinders the use of these measures for clustering validation substantially. To this end, Meila [19] provided some basic requirements of information-theoretic distance measures for clustering validation, such as *refinement additivity*, *join additivity*, and *convex additivity*. As a further step, in this paper, we introduce a uniform representation of quasi-distance, for information-theoretic distance measures. The quasi-distance possesses three properties: symmetry, the triangle law, and the minimum reachable. These properties ensure that a quasi-distance measure naturally lends itself as the external criteria for clustering validation.

In general, there are two application scenarios of information-theoretic distance measures for clustering validation. First, these distance measures can be used to compare clusterings of a given data set by different clustering algorithms. Second, these measures can also be used to compare clusterings of different data sets by a specific clustering algorithm. For instance, in order to find the characteristic of data (high dimensionality, the size of the data, the sparseness of the data, and scales of attributes) that may strongly affect the performance of a clustering algorithm [25], multiple data sets with different

- *P. Luo is with the Institute of Computing Technology, Chinese Academy of Sciences, and also with the Hewlett-Packard Labs China, SP Tower A505, Tshinghua Science Park, HaiDian District, Beijing 100084, P.R. China. E-mail: ping.luo@hp.com.*
- *H. Xiong is with the Department of Management Science and Information Systems, Rutgers Business School, Rutgers University, Ackerson Hall 200K, 180 University Avenue, Newark, NJ 07102. E-mail: hxiong@rutgers.edu.*
- *G. Zhan is with the Department of Computer Science, Wayne State University, 420 State Hall, 5143 Cass Ave., Detroit, MI 48202. E-mail: gxzhan@wayne.edu.*
- *J. Wu is with the Department of Information Systems, School of Economics and Management, Beihang University, A1004, New Main Building, No. 37, Xue Yuan Road, Hai Dian District, Beijing 100191, P.R. China. E-mail: wujj@buaa.edu.cn.*
- *Z. Shi is with the Institute of Computing Technology, Chinese Academy of Sciences, Room 534, Building of Institute of Computing Technology, No. 6 Kexueyuan Nanlu, Zhongguan Cun Hai Dian District, Beijing 100080, P.R. China. E-mail: shizz@ics.ict.ac.cn.*

Manuscript received 24 July 2007; revised 7 July 2008; accepted 8 Sept. 2008; published online 16 Sept. 2008.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2007-07-0383. Digital Object Identifier no. 10.1109/TKDE.2008.200.

1. We will use the terms *partition* and *clustering* of a data set interchangeably in this paper.

characteristics are required to be clustered, and their results are then analyzed.

For the above second scenario, we have observed that the ranges of distance measures are different for different data sets. In other words, to do a fair comparison, distance normalization is required to equalize ranges of the distance measures. A critical challenge for distance normalization is to obtain the ranges of a distance measure. To that end, we theoretically analyze the computation of the maximum value of a distance measure for a data set. Our study reveals that the exact computation of the maximum value of a distance measure is usually difficult to find. As a result, we provide an approximate computation form of the maximum values for these distance measures. We also show that there are some cases in which the maximum distance value can be obtained. Finally, we have designed various experiments by exploiting the K-means clustering algorithm to show that 1) the normalized distance measures outperform the original distance measure and 2) the normalized Shannon distance has the best performance among four observed distance measures.

Overview. The remainder of this paper is organized as follows: Section 2 describes the basic denotations and concepts of external clustering validation measures and quasi-distance. In Section 3, we briefly describe our previous works on partition entropy and conditional entropy, which are the bases of the quasi-distances. Section 4 details the uniform framework of quasi-distances between two partitions and presents some examples to show how the proposed framework induces several well-known distances for clustering validation. In Section 5, we describe the importance of the distance normalization when comparing clusterings of different data sets. In Section 6, we theoretically analyze the computation of the maximum value of a distance measure for a data set, which is the key to distance normalization. Section 7 demonstrates the experimental setup and results. Finally, in Section 8, we draw conclusions.

2 Basic Concepts

In this paper, we adopt the notations used in [23] and [18]. The set of reals and the set of natural numbers are denoted by $\mathbb{R}$ and $\mathbb{N}$, respectively. All other sets considered in the following discussion are nonempty and finite. $\pi = \{A_1, \ldots, A_m\}$ is a partition of a set A, iff $\cup_{i=1}^{m} A_i = A$ and $A_i \cap A_j = \emptyset$ $(i \neq j)$. A *block* of a partition refers to any element in a partition of a set A. Let $PART(A)$ be the set of partitions of set A. The class of all partitions of finite sets is denoted by $PART$. If $\pi, \pi' \in PART(A)$, then $\pi \subseteq \pi'$ if every block of π is included in a block of π'. If A, B are two disjoint sets, $\pi \in PART(A)$, $\sigma \in PART(B)$, where $\pi = \{A_1, \ldots, A_m\}$, $\sigma = \{B_1, \ldots, B_n\}$, then the partition $(\pi + \sigma) \in PART(A \cup B)$ is given by

$$\pi + \sigma = \{A_1, \ldots, A_m, B_1, \ldots, B_n\}.$$

Let $\pi \in PART(A)$ and let $C \subseteq A$. The "trace" of π on C is given by $\pi_C = \{A_i \cap C | A_i \in \pi \; such \; that \; A_i \cap C \neq \emptyset\}$. Obviously, $\pi_C \in PART(C)$. When $D \subseteq A$, it is clear that $(\pi_C)_D = \pi_{(C \wedge D)}$.

Let $\pi, \sigma \in PART(A)$ (two partitions defined on the same set A), where $\pi = \{A_1, \ldots, A_m\}$, $\sigma = \{B_1, \ldots, B_n\}$. The partition $\pi \wedge \sigma$ whose blocks consist of the nonempty intersections of the blocks of π and σ can be written as

$$\pi \wedge \sigma = \pi_{B_1} + \cdots + \pi_{B_n} = \sigma_{A_1} + \cdots + \sigma_{A_m}.$$

External measures. When the external information (the class labels of all the objects) is provided, the external measure for clustering validation is actually the distance between two partitions of the data set: one is the partition resulted from a clustering algorithm, the other is the "true" partition generated by the class labels. Thus, given a set A, the external measure for clustering validation is a mapping

$$d : PART(A)^2 \rightarrow \mathbb{R}. \quad (1)$$

$d(\pi, \sigma)$ is used to measure the distance from π to σ. The first argument refers to the output partition $\pi = \{A_1, \ldots, A_m\}$ of A. The second argument refers to the "true" partition $\sigma = \{B_1, \ldots, B_n\}$ of A, where B_i contains all the objects with class label i (for $i = 1, \ldots, n$). The smaller $d(\pi, \sigma)$ is, the better the clustering result π is.

Quasi-distance. Various information-theoretic distance measures, such as Shannon Distance [17], Goodman-Kruskal coefficient [24], [10], the Van Dongen criterion [19], and the Mirkin metric [19], can be used as external measures for clustering validation. Meila [19] provided some basic requirements of external measures for clustering validation, such as *refinement additivity*, *join additivity*, and *convex additivity*. However, in this paper, we show that all these information-theoretic distance measures are actually quasi-distance between two partitions when they are used as external measures for clustering validation. A quasi-distance is defined as follows:

Definition 1. *Let π, σ be two partitions on A. The measure $d(\pi, \sigma)$ is a quasi-distance between these two partitions; that is, for any partitions π, σ, and τ on A, it satisfies*

1. *$d(\pi, \sigma)$ reaches its minimum over both π and σ iff $\pi = \sigma$ (minimum reachable),*
2. *$d(\pi, \sigma) = d(\sigma, \pi)$ (symmetry), and*
3. *$d(\pi, \sigma) + d(\sigma, \tau) \geq d(\pi, \tau)$ (the triangle law).*

Note that $d(\pi, \sigma)$ is minimum reachable. In other words, even if two partitions are the same, $d(\pi, \sigma)$ can reach a minimum value, but this minimum distance value may not be zero. For example, as you will see in Section 4, this situation happens for the distance d_{pal}^1 in Table 1. This is the reason why we define it as quasi-distance.

3 The Concepts of Partition Entropy and Conditional Entropy

In this section, we briefly describe our previous work on partition entropy and conditional entropy [16], which is the basis of the results in Section 4.

3.1 Partition Entropy and Conditional Entropy

Partition entropy is a mapping

$$\mathcal{H} : PART \rightarrow \mathbb{R}, \quad (2)$$

satisfying some additional conditions as described in Section 3.4. A formal definition of partition entropy is also given in Section 3.4.

TABLE 1
Examples of Quasi-Distance with Conditional Entropy $\mathcal{C}^1$

Entropy Name	Entropy	Partition Distance	Distance Minimum
Shannon Entropy [22]	$\mathcal{H}_{sha}(\pi)=\sum_{i=1}^{m} p_i log_2 \frac{1}{p_i}$	$d^1_{sha}(\pi,\sigma)=\sum_{i=1}^{m}\frac{\lvert A_i\rvert}{\lvert A\rvert}\cdot\mathcal{H}_{sha}(\sigma_{A_i})+\sum_{j=1}^{n}\frac{\lvert B_j\rvert}{\lvert A\rvert}\cdot\mathcal{H}_{sha}(\pi_{B_j})$	0
Pal Entropy [20]	$\mathcal{H}_{pal}(\pi)=\sum_{i=1}^{m} p_i e^{1-p_i}$	$d^1_{pal}(\pi,\sigma)=\sum_{i=1}^{m}\frac{\lvert A_i\rvert}{\lvert A\rvert}\cdot\mathcal{H}_{pal}(\sigma_{A_i})+\sum_{j=1}^{n}\frac{\lvert B_j\rvert}{\lvert A\rvert}\cdot\mathcal{H}_{pal}(\pi_{B_j})$	2
Gini Index [4]	$\mathcal{H}_{gin}(\pi)=\sum_{i=1}^{m} p_i(1-p_i)$	$d^1_{gin}(\pi,\sigma)=\sum_{i=1}^{m}\frac{\lvert A_i\rvert}{\lvert A\rvert}\cdot\mathcal{H}_{gin}(\sigma_{A_i})+\sum_{j=1}^{n}\frac{\lvert B_j\rvert}{\lvert A\rvert}\cdot\mathcal{H}_{gin}(\pi_{B_j})$	0
Goodman-Kruskal coefficient [24], [10]	$\mathcal{H}_{goo}(\pi)=1-max_{i=1}^{m} p_i$	$d^1_{goo}(\pi,\sigma)=\sum_{i=1}^{m}\frac{\lvert A_i\rvert}{\lvert A\rvert}\cdot\mathcal{H}_{goo}(\sigma_{A_i})+\sum_{j=1}^{n}\frac{\lvert B_j\rvert}{\lvert A\rvert}\cdot\mathcal{H}_{goo}(\pi_{B_j})$	0

Given a set A, conditional entropy is a mapping

$$\mathcal{C}: PART(A)^2 \to \mathbb{R}. \tag{3}$$

The first argument refers to a condition partition, while the second one refers to a decision partition. If π, σ are two partitions of A, $\mathcal{C}(\pi,\sigma)$ measures the degree of difficulty in predicting σ by π. Based on an existing partition entropy, we give two definitions of conditional entropy as follows:

Definition 2. *Let π, $\sigma \in PART(A)$, $\pi=\{A_1,\ldots,A_m\}$, $\sigma=\{B_1,\ldots,B_n\}$. A conditional entropy $\mathcal{C}^1$ is a function $\mathcal{C}$ in (3) such that*

$$\mathcal{C}^1(\pi,\sigma)=\sum_{i=1}^{m}\frac{|A_i|}{|A|}\cdot\mathcal{H}(\sigma_{A_i}), \tag{4}$$

where σ_{A_i} is the "trace" of σ on A_i.

Definition 2 states that the conditional entropy $\mathcal{C}^1$ is the expected value of the entropies calculated according to conditional distributions, i.e., $\mathcal{C}^1(\pi,\sigma)=E_{A_i}(\mathcal{H}(\sigma_{A_i}))$, $A_i\in\pi$.

Definition 3. *Let π, $\sigma \in PART(A)$, $\pi=\{A_1,\ldots,A_m\}$, $\sigma=\{B_1,\ldots,B_n\}$. A conditional entropy $\mathcal{C}^2$ is a function $\mathcal{C}$ in (3) such that*

$$\mathcal{C}^2(\pi,\sigma)=\mathcal{H}(\pi\wedge\sigma)-\mathcal{H}(\pi). \tag{5}$$

Definition 3 states that the conditional entropy $\mathcal{C}^2$ is the difference between two entropies. The equality $\mathcal{C}^1(\pi,\sigma)=\mathcal{C}^2(\pi,\sigma)$ yields the Shannon entropy [1]. Thus, this axiomatization of the Shannon entropy shows the rationality of these two definitions.

3.2 Equality Properties of Partition Entropy

If $\pi=\{A_1,\ldots,A_n\}$ is a partition of a set A, then the probability distribution vector attached to π is $P(\pi)=(p_1,\ldots,p_n)$, where $p_i=\frac{|A_i|}{|A|}$ for $1\le i\le n$. Thus, it is straightforward to consider the notion of partition entropy via the entropy of the corresponding probability distribution. We define the measure function of $\mathcal{H}$ as a mapping $\mathcal{M}:\Delta\to\mathbb{R}$ such that $\mathcal{H}(\pi)=\mathcal{M}(P(\pi))$ for every $\pi\in PART$, where $\Delta=\{P(\pi)|\pi\in PART\}$. The blocks in a partition π are unordered while the elements in $P(\pi)$ are ordered. Thus, the inherent postulate of $\mathcal{M}$ is that it is symmetric in the sense that

$$\mathcal{M}(P(\pi))=\mathcal{M}(P'(\pi)), \tag{6}$$

where $P'(\pi)$ is any permutation of $P(\pi)$.

The other equality postulate of $\mathcal{M}$ is expansibility in the sense that for every $\vec{p}\in\Delta_m$

$$\mathcal{M}(\vec{p})=\mathcal{M}(\vec{p}'), \tag{7}$$

where $\vec{p}=(p_1,\ldots,p_m)$, $\vec{p}'=(p_1,\ldots,p_m,0)$ and $\Delta_m=\{(p_1,\ldots,p_m):0\le p_i\le 1 \text{ for } i=1,\ldots,m, p_1+\cdots+p_m=1\}$.

3.3 Inequality Postulates of Partition Entropy

We give the inequalities that partition entropy and its corresponding conditional counterpart must satisfy as follows:

Postulate 1. *Let π, $\pi'\in PART(A)$ and $\pi\preceq\pi'$, then*

$$\mathcal{H}(\pi')\le\mathcal{H}(\pi),$$

where $\preceq$ is the majorization relationship (entropically comparable relationship) between two partitions, detailed in [18] and [16].

Postulate 2. *Let π, π', $\sigma\in PART(A)$ and $\pi\subseteq\pi'$, then*

$$\mathcal{C}(\pi,\sigma)\le\mathcal{C}(\pi',\sigma).$$

Postulate 3. *Let π, σ, $\sigma'\in PART(A)$ and $\sigma\subseteq\sigma'$, then*

$$\mathcal{C}(\pi,\sigma')\le\mathcal{C}(\pi,\sigma).$$

A function $\mathcal{H}$, which satisfies Postulate 1, is actually a *Schur-concave* function [18]. Postulates 2 and 3 state that conditional entropy $\mathcal{C}$ should be monotonic in the first argument and dually monotonic in the second argument. Specifically, Postulate 2 shows that finer condition partition leaves less uncertainty about decision partition and thus owns more ability in predicting decision partition. On the other hand, Postulate 3 shows that coarser decision partition relaxes the requirement of precision for predicting and thus contains less uncertainty also. They are the two postulates conditional entropy holds inherently.

3.4 Formal Definition of Partition Entropy and Its Checking Conditions

Definition 4. *When a function defined by (2) satisfies Postulates 1 through 3, and its corresponding measure function $\mathcal{M}$ is symmetric and expansible, this function is partition entropy.*

Considering the two definitions of conditional entropy separately, Luo et al. [16] reduce the redundancies in Postulates 1 through 3, and give the easy-checking conditions for any partition entropy. The main results are summarized as the following Theorems 1 and 2.

Theorem 1. *When conditional entropy is defined as $\mathcal{C}^1$, the measure function $\mathcal{M}$ of $\mathcal{H}$ is symmetric and expansible, if and only $\mathcal{H}$ is concave, it is a partition entropy.*

Theorem 2. *Given a function $f:[0,1]\to\mathbb{R}$, $f(0)=0$, f is continuous on [0, 1], f''' exists in (0,1), $f''(x)\le 0$ and $f'''(x)\le 0$ for any $x\in(0,1)$. Let $\pi=\{A_1,A_2,\ldots,A_m\}$. Then, $\mathcal{H}(\pi)=\sum_{i=1}^{m} f(\frac{|A_i|}{A})$ is a partition entropy when its conditional counterpart is defined as $\mathcal{C}^2$.*

4 From Conditional Entropy to Quasi-Distance between Two Partitions

In this section, we introduce some properties, which can be used to induce the quasi-distance based on the generic form $\mathcal{C}$ of conditional entropy.

Let π, σ be two partitions on a data set A, and $\mathcal{C}$ be a conditional entropy, we consider the following distance between π and σ:

$$d(\pi,\sigma)=\mathcal{C}(\pi,\sigma)+\mathcal{C}(\sigma,\pi), \tag{8}$$

where σ is considered as the "true" partition, $\mathcal{C}(\pi,\sigma)$ is the measure of the purity in π, and $\mathcal{C}(\sigma,\pi)$ is a penalty to the situation that a data cluster in the "true" partition σ is separated into several clusters in π.

The following properties give the conditions, which guarantee that $d(\pi,\sigma)=\mathcal{C}(\pi,\sigma)+\mathcal{C}(\sigma,\pi)$ is a quasi-distance. To show this, we consider the two situations where the conditional entropy $\mathcal{C}$ is defined as $\mathcal{C}^1$ and $\mathcal{C}^2$, respectively.

Lemma 1. *Let π, σ be any two partitions on A. Then, $d(\pi,\sigma)=\mathcal{C}(\pi,\sigma)+\mathcal{C}(\sigma,\pi)$ reaches its minimum if and only if $\pi=\sigma$, where $\mathcal{C}$ is the conditional counterpart of a partition entropy $\mathcal{H}$, defined as $\mathcal{C}^1$ or $\mathcal{C}^2$.*

Proof. We prove this lemma under the situations that the conditional entropy is defined as $\mathcal{C}^1$ and $\mathcal{C}^2$, respectively.

When $\mathcal{C}$ is defined as $\mathcal{C}^1$, $\mathcal{C}^1(\pi,\sigma)$ and $\mathcal{C}^1(\sigma,\pi)$ reach their minimal values when $\pi=\sigma$. Thus, $d^1(\pi,\sigma)$ reaches its minimal value $2\mathcal{M}(0,1)$ when $\pi=\sigma$, where $\mathcal{M}$ is the measure function of the corresponding entropy.

When $\mathcal{C}$ is defined as $\mathcal{C}^2$, $d^2(\pi,\sigma)=2\mathcal{H}(\pi\wedge\sigma)-\mathcal{H}(\pi)-\mathcal{H}(\sigma)$. It is clear that $\mathcal{H}(\pi\wedge\sigma)\ge\mathcal{H}(\pi)$ and $\mathcal{H}(\pi\wedge\sigma)\ge\mathcal{H}(\sigma)$. Thus, d^2 reaches its minimal value 0 when $\pi=\sigma$. □

Lemma 2. *Let π, σ, τ be three partitions on A. If $\mathcal{C}(\pi\wedge\tau,\sigma)+\mathcal{C}(\tau,\pi)\ge\mathcal{C}(\tau,\pi\wedge\sigma)$, then $d(\pi,\sigma)=\mathcal{C}(\pi,\sigma)+\mathcal{C}(\sigma,\pi)$ is a quasi-distance, where $\mathcal{C}$ (defined as $\mathcal{C}^1$ or $\mathcal{C}^2$) is the corresponding conditional entropy of a partition entropy $\mathcal{H}$.*

Proof. By Lemma 1, $d(\pi,\sigma)$ satisfies the condition 1 of a quasi-distance. The symmetry of $d(\pi,\sigma)$ is immediate to see. Next, we prove the triangular property of $d(\pi,\sigma)$:

$$\mathcal{C}(\sigma,\pi)+\mathcal{C}(\tau,\sigma)\ge\mathcal{C}(\sigma\wedge\tau,\pi)+\mathcal{C}(\tau,\sigma) \tag{9}$$

$$\ge\mathcal{C}(\tau,\pi\wedge\sigma) \tag{10}$$

$$\ge\mathcal{C}(\tau,\pi), \tag{11}$$

where (9) follows from Postulate 2, (10) follows from the condition in this lemma, and (11) follows from Postulate 3.

In a similar manner, we prove that

$$\mathcal{C}(\pi,\sigma)+\mathcal{C}(\sigma,\tau)\ge\mathcal{C}(\sigma\wedge\pi,\tau)+\mathcal{C}(\pi,\sigma) \tag{12}$$

$$\ge\mathcal{C}(\pi,\tau\wedge\sigma)\ge\mathcal{C}(\pi,\tau). \tag{13}$$

Then, adding inequalities (11) and (13) together, we have $d(\pi,\sigma)+d(\sigma,\tau)\ge d(\pi,\tau)$. So, $d(\pi,\sigma)$ is a quasi-distance. □

Note that Lemmas 1 and 2 remain true no matter $\mathcal{C}$ is defined as $\mathcal{C}^1$ or $\mathcal{C}^2$.

4.1 When Conditional Entropy Is Defined as $\mathcal{C}^1$

Theorem 3. *Let π, σ be two partitions on a data set A, and the conditional entropy is defined as $\mathcal{C}^1$ based on a partition entropy $\mathcal{H}$. If $\mathcal{H}(\pi\wedge\sigma)\le\mathcal{C}^1(\pi,\sigma)+\mathcal{H}(\pi)$, then $d^1(\pi,\sigma)=\mathcal{C}^1(\pi,\sigma)+\mathcal{C}^1(\sigma,\pi)$ is a quasi-distance.*

Proof. Let $\pi=\{B_1,\ldots,B_l\}$, $\sigma=\{C_1,\ldots,C_m\}$, $\tau=\{D_1,\ldots,D_n\}$. First, we prove that $\mathcal{C}^1(\pi\wedge\tau,\sigma)=\sum_{i=1}^{n}\frac{|D_i|}{|A|}\mathcal{C}^1(\pi_{D_i},\sigma_{D_i})$:

$$\sum_{i=1}^{n}\frac{|D_i|}{|A|}\mathcal{C}^1(\pi_{D_i},\sigma_{D_i})=\sum_{i=1}^{n}\frac{|D_i|}{|A|}\left(\sum_{j=1}^{l}\frac{|B_j\wedge D_i|}{|D_i|}\mathcal{H}\Big((\sigma_{D_i})_{B_j}\Big)\right) \tag{14}$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{l}\frac{|B_j\wedge D_i|}{|A|}\mathcal{H}\Big(\sigma_{(D_i\wedge B_j)}\Big) \tag{15}$$

$$=\mathcal{C}^1(\pi\wedge\tau,\sigma), \tag{16}$$

where (14) follows from the definition of $\mathcal{C}^1$ (Definition 2), (15) follows from $(\sigma_{D_i})_{B_j}=\sigma_{(D_i\wedge B_j)}$ (refer to Section 2 for the definition of the "trace" of a partition), and (16) also follows from Definition 2. Then,

$$\mathcal{C}^1(\pi\wedge\tau,\sigma)+\mathcal{C}^1(\tau,\pi)=\sum_{i=1}^{n}\frac{|D_i|}{|A|}\mathcal{C}^1(\pi_{D_i},\sigma_{D_i})+\sum_{i=1}^{n}\frac{|D_i|}{|A|}\mathcal{H}(\pi_{D_i}) \tag{17}$$

$$=\sum_{i=1}^{n}\frac{|D_i|}{|A|}\left[\mathcal{C}^1(\pi_{D_i},\sigma_{D_i})+\mathcal{H}(\pi_{D_i})\right] \tag{18}$$

$$\ge\sum_{i=1}^{n}\frac{|D_i|}{|A|}\mathcal{H}(\pi_{D_i}\wedge\sigma_{D_i}) \tag{19}$$

$$=\sum_{i=1}^{n}\frac{|D_i|}{|A|}\mathcal{H}\Big((\pi\wedge\sigma)_{D_i}\Big)=\mathcal{C}^1(\tau,\pi\wedge\sigma), \tag{20}$$

where (17) follows from (16), (19) follows from the condition in this theorem, and (20) follows from $\pi_{D_i}\wedge\sigma_{D_i}=(\pi\wedge\sigma)_{D_i}$.

Finally, by Lemma 2, this theorem follows. □

Corollary 1. *Let π, σ be any two partitions on A, $g:[0,1]\to\mathbb{R}$, $\mathcal{M}(p_i,\ldots,p_m)=\sum_{i=1}^{m}p_i g(p_i)$ be the measure function of a partition entropy $\mathcal{H}$. If $g(x)+g(y)\ge g(xy)$ for $0\le x\le 1$ and $0\le y\le 1$, then $d^1(\pi,\sigma)=\mathcal{C}^1(\pi,\sigma)+\mathcal{C}^1(\sigma,\pi)$ is a quasi-distance.*

TABLE 2
Examples of Quasi-Distance with Conditional Entropy $\mathcal{C}^2$

Entropy Name	Entropy	Partition Distance	Distance Minimum
Shannon Entropy [22]	$\mathcal{H}_{sha}(\pi)=\sum_{i=1}^{m} p_i log_2 \frac{1}{p_i}$	$d^2_{sha}(\pi,\sigma)=2\mathcal{H}_{sha}(\pi\wedge\sigma)-\mathcal{H}_{sha}(\pi)-\mathcal{H}_{sha}(\sigma)$	0
Gini Index [4]	$\mathcal{H}_{gin}(\pi)=\sum_{i=1}^{m} p_i(1-p_i)$	$d^2_{gin}(\pi,\sigma)=2\mathcal{H}_{gin}(\pi\wedge\sigma)-\mathcal{H}_{gin}(\pi)-\mathcal{H}_{gin}(\sigma)$	0

Proof. Let $\pi=\{B_1,\ldots,B_l\}$, $\sigma=\{C_1,\ldots,C_m\}$. Then,

$$\mathcal{C}^1(\pi,\sigma)+\mathcal{H}(\pi)=\sum_{i=1}^{l}\sum_{j=1}^{m}\frac{|B_i\cap C_j|}{|A|}\left(g\left(\frac{|B_i\cap C_j|}{|B_i|}\right)+g\left(\frac{|B_i|}{|A|}\right)\right),$$

$$\mathcal{H}(\pi\wedge\sigma)=\sum_{i=1}^{l}\sum_{j=1}^{m}\frac{|B_i\cap C_j|}{|A|}g\left(\frac{|B_i\cap C_j|}{|A|}\right).$$

If $g(x)+g(y)\geq g(xy)$ for $0\leq x\leq 1$ and $0\leq y\leq 1$, then

$$g\left(\frac{|B_i\cap C_j|}{|B_i|}\right)+g\left(\frac{|B_i|}{|A|}\right)\geq g\left(\frac{|B_i\cap C_j|}{|A|}\right)$$

for $i=1,\ldots,l$ and $j=1,\ldots,m$. Then,

$$\mathcal{C}^1(\pi,\sigma)+\mathcal{H}(\pi)\geq\mathcal{H}(\pi\wedge\sigma).$$

From the above and by Theorem 3, this corollary is true.

4.2 When Conditional Entropy Is Defined as $\mathcal{C}^2$

Theorem 4. *Let π, σ be two partitions on A, and conditional entropy is defined as $\mathcal{C}^2$ based on a partition entropy $\mathcal{H}$. Then, $d^2(\pi,\sigma)=\mathcal{C}^2(\pi,\sigma)+\mathcal{C}^2(\sigma,\pi)$ is a quasi-distance.*

Proof.

$$\begin{aligned}&\mathcal{C}^2(\pi\wedge\tau,\sigma)+\mathcal{C}^2(\tau,\pi)\\&\quad=\mathcal{H}(\pi\wedge\tau\wedge\sigma)-\mathcal{H}(\pi\wedge\tau)+\mathcal{H}(\pi\wedge\tau)-\mathcal{H}(\tau)\\&\quad=\mathcal{H}(\pi\wedge\tau\wedge\sigma)-\mathcal{H}(\tau)=\mathcal{C}^2(\tau,\pi\wedge\sigma).\end{aligned}$$

From the above and by Lemma 2, this theorem holds. □

4.3 Examples of Quasi-Distance

Let π, σ be two partitions on a data set A, based on the above discussion, we have the following two methods to induce quasi-distances:

1. Let $\mathcal{H}$ be a partition entropy, and its conditional entropy is defined as $\mathcal{C}^1$. If $\mathcal{H}$ satisfies the conditions in Theorem 3 or Corollary 1, $d^1(\pi,\sigma)=\mathcal{C}^1(\pi,\sigma)+\mathcal{C}^1(\sigma,\pi)$ is a quasi-distance.
2. Let $\mathcal{H}$ be a partition entropy, and its conditional entropy is defined as $\mathcal{C}^2$. Then, $d^2(\pi,\sigma)=\mathcal{C}^2(\pi,\sigma)+\mathcal{C}^2(\sigma,\pi)=2\mathcal{H}(\pi\wedge\sigma)-\mathcal{H}(\pi)-\mathcal{H}(\sigma)$ is a quasi-distance.

Here, we first give some examples of partition entropy, and then induce the corresponding quasi-distances. All these examples, to be proved by the proposed theorems and corollaries, are under the following assumption: let $\pi=\{A_1,\ldots,A_m\}$ and $\sigma=\{B_1,\ldots,B_n\}$ be two partitions of a set A, the probability distribution vector attached to π be $P(\pi)=(p_1,\ldots,p_m)$, where $p_i=\frac{|A_i|}{|A|}$ for $1\leq i\leq m$.

Examples when the conditional entropy is defined as $\mathcal{C}^1$. The examples in Table 1 are partition entropies when their conditional counterparts are defined as $\mathcal{C}^1$ (proved by Theorem 1). It can be proved by Corollary 1 that d^1_{sha}, d^1_{pal}, and d^1_{gin} are quasi-distances. d^1_{goo} is also a quasi-distance, which can be proved by Theorem 3. The details of these proofs are omitted. d^1_{sha} is first proposed in [17], and referred to as *variation of information* in [19]. Additionally, Meila [19] gives an axiomatic method of d^1_{sha}, which is aligned with the lattice of partitions and convexly additive. d^1_{goo} is actually the n-invariant version of the Van Dongen criterion [19].

Examples of the conditional entropy is defined as $\mathcal{C}^2$. The examples in Table 2 are all partition entropies when their conditional counterparts are defined as $\mathcal{C}^2$ (proved by Theorem 2). It can be easily proved by Theorem 4 that d^2_{sha} (d^1_{sha} and d^2_{sha} are the same distance, expressed in two ways) and d^2_{gin} are both quasi-distances. d^2_{gin} is actually the n-invariant version of the Mirkin metric [19].

It should be noted that all the quasi-distances in Tables 1 and 2 except d^1_{pal} are *true* metrics since the minimal values of these distances are all 0. However, the minimum of d^1_{pal} is 2. Thus, it is not a *real* distance.

5 Normalization Issues

In this section, we discuss normalization issues of distance measures. Normalization is critical when distance measures are used to compare clusterings of different data sets.

Different data sets have different data characteristics, and thus have different degree of difficulty for clustering. In general, the bigger the degree of difficulty on the clustering data is, the more possible that a clustering algorithm generates a result with a bigger distance. Thus, the clustering result on a specific data set is affected by both the performance of the clustering algorithm and the degree of clustering difficulty on the data set itself. When comparing the performances of a clustering algorithm on different data sets, since the degrees of difficulty on these data sets are different, the original quasi-distance might be biased. For instance, we assume that $\pi=\{A_1,\ldots,A_m\}$ and $\beta=\{B_1,\ldots,B_n\}$ are "true" partitions for two data sets A and B, respectively. Also, let π' and β' be the clustering results of data set A and B by a specific clustering algorithm, respectively. By a distance measure d, their distances $d(\pi,\pi')$, $d(\beta,\beta')$ and the distance ranges are shown in Fig. 1. As can be seen, the maximum distance d^{max}_{β}[2] is much bigger than d^{max}_{π}, which shows that the degree of difficulty in clustering B is much greater than that for data set A. It also shows that $d(\beta,\beta')>d(\pi,\pi')$, indicating that the clustering performance on A is better than that on B. However, it is clear that π' is a bad result because $d(\pi,\pi')$ is close to its maximum distance d^{max}_{π}. Also, β' is a good clustering because $d(\beta,\beta')$ is close to its minimum distance d^{min}_{β}.

2. The clustering result with the maximal (minimal) distance d^{max} (d^{min}) is the worst (best) result on the corresponding data set. The formal definitions of d^{max} and d^{min} are given in (21).

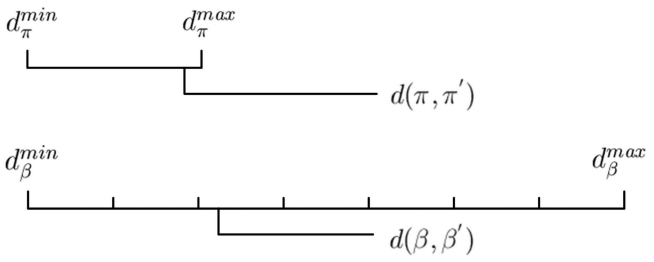


Fig. 1. Comparing clusterings of different data sets.

Therefore, when comparing the performances of a clustering algorithm on different data sets, the distance measure for clustering validation should be irrelevant to the degree of clustering difficulty on a data set. A possible way to solve this problem is the use of the normalized distance, which represents the relative position of the original distance between the minimal and maximal distance. The formal definition of distance normalization is given as follows:

When σ is the fixed "true" partition of a data set A and π is any partition of A, the quasi-distance $d(\pi,\sigma)$ is a function of π, denoted by $d_\sigma(\pi)$. Let d_σ^{max} and d_σ^{min} be the maximal and minimal values of $d_\sigma(\pi)$ ($\pi \in PART(A)$), respectively, the normalized form of this distance is denoted by

$$normd_\sigma(\pi) = \frac{d_\sigma(\pi) - d_\sigma^{min}}{d_\sigma^{max} - d_\sigma^{min}}. \tag{21}$$

After normalization, $normd_\sigma(\pi)$ is in [0,1]. In fact, $normd_\sigma(\pi)$ is the relative position of the original distance in the distance range $[d_\sigma^{min}, d_\sigma^{max}]$.

6 Computation of d_σ^{min} and d_σ^{max}

In this section, we focus on the computation of d_σ^{min} and d_σ^{max} when the conditional entropy used in the quasi-distance is $\mathcal{C}^1$. Let π, $\sigma \in PART(A)$, $\sigma = \{B_1,\ldots,B_n\}$ is the "true" partition, $\pi = \{A_1,\ldots,A_m\}$. In the following, we assume that $d(\pi,\sigma) = \mathcal{C}^1(\pi,\sigma) + \mathcal{C}^1(\sigma,\pi)$ is a quasi-distance, where $\mathcal{C}^1$ is a conditional entropy, $\mathcal{H}$ and $\mathcal{M}$ are its corresponding partition entropy and the measure function, respectively. According to Theorem 1, when $\mathcal{M}$ is symmetric and expansible $\mathcal{M}$ must be a concave function.

It is easy to compute d_σ^{min}. However, the computation of the exact value of d_σ^{max} is rather complicated. Section 6.1 gives the computing methods of d_σ^{min}. In Section 6.2, we give some mathematical facts about the partition entropy and conditional entropy. Based on these facts, we formulate a $\pi_0 \in PART(A)$, for which one might think that $d_\sigma^{max} = d(\pi_0,\sigma)$. However, we give the example to show it is not true. In Section 6.3 we give the explicit expression of $d(\pi_0,\sigma)$. In Section 6.4, we approximate the value of d_σ^{max} in general cases. Finally, Section 6.5 gives the exact value of d_σ^{max} in some special cases.

6.1 The Exact Computation of d_σ^{min}

Theorem 5. *Let σ be the "true" partition of a data set A, π be a partition of A, $\mathcal{H}$ be a partition entropy, and its conditional entropy be defined as $\mathcal{C}^1$, $d^1(\pi,\sigma) = \mathcal{C}^1(\pi,\sigma) + \mathcal{C}^1(\sigma,\pi)$ be a quasi-distance. Then, $d_\sigma^{min} = 2\mathcal{M}(0,1)$, where $\mathcal{M}$ is the measure function of $\mathcal{H}$.*

Proof. $d_\sigma(\pi)$ reaches this minimum when $\pi = \sigma$. This minimal value is actually $2\mathcal{C}^1(\sigma,\sigma) = 2\mathcal{M}(0,1)$. □

6.2 Analysis on d_σ^{max}

Unlike d_σ^{min}, the exact value of d_σ^{max} is usually difficult to obtain. Before we describe our analysis on d_σ^{max}, we first present some mathematical facts:

Fact I. Assuming that the measure function $\mathcal{M}$ is concave and symmetric, then $\mathcal{M}(p_1,\ldots,p_m) \le \mathcal{M}(\frac{1}{m},\ldots,\frac{1}{m})$, $m \in \mathbb{N}$, for any p_i satisfying $\sum_{i=1}^m p_i = 1$ and $0 \le p_i \le 1$, $i = 1,\ldots,m$.

Fact II. Let $\mathcal{M}(p_1,\ldots,p_m) = \sum_{i=1}^m f(p_i)$, f is a continuous function on [0, 1] with a nonpositive second derivative (note that $\mathcal{M}$ is concave) in (0, 1), and $f(0) = 0$ (due to the expansibility of $\mathcal{M}$). Then, we can derive that $\mathcal{M}(\frac{1}{m},\ldots,\frac{1}{m}) \le \mathcal{M}(\frac{1}{n},\ldots,\frac{1}{n})$ if $m \le n$. To prove this result, we define a function $g(x) = \mathcal{M}(\frac{1}{x},\ldots,\frac{1}{x}) = xf(\frac{1}{x})$, $x \ge 1$. We have

$$g'(x) = f(\frac{1}{x}) - \frac{f'(\frac{1}{x})}{x} = \frac{\frac{f(\frac{1}{x})-f(0)}{\frac{1}{x}} - f'(\frac{1}{x})}{x}.$$

By Mean-Value Theorem (see [21, p. 86]), there exists a $\xi \in (0,\frac{1}{x})$ such that $g'(x) = \frac{f'(\xi)-f'(\frac{1}{x})}{x}$. Using the fact $f'' \le 0$, we conclude that $g'(x) \ge 0$, and the above claim is proved.

Fact III. $\mathcal{C}^1(\pi,\sigma) \le \mathcal{H}(\sigma)$, which means that if σ represents a "true" partition, then $\mathcal{C}^1(\pi,\sigma)$ takes on its maximal value $\mathcal{H}(\sigma)$ when π is the trivial partition $\{A\}$ of the data set A. This claim follows from the concavity of $\mathcal{H}$. The proof is as follows:

$$\begin{aligned}
\mathcal{C}^1(\pi,\sigma) &= \sum_{i=1}^m \frac{|A_i|}{|A|}\cdot\mathcal{H}(\sigma_{A_i}) \\
&= \sum_{i=1}^m \frac{|A_i|}{|A|}\cdot\mathcal{M}\left(\frac{|B_1\bigcap A_i|}{|A_i|},\ldots,\frac{|B_n\bigcap A_i|}{|A_i|}\right) \\
&\le \mathcal{M}\left(\sum_{i=1}^m\left(\frac{|A_i|}{|A|}\frac{|B_1\bigcap A_i|}{|A_i|}\right),\ldots,\sum_{i=1}^m\left(\frac{|A_i|}{|A|}\frac{|B_n\bigcap A_i|}{|A_i|}\right)\right) \\
&= \mathcal{M}\left(\frac{|B_1|}{|A|},\ldots,\frac{|B_n|}{|A|}\right) = \mathcal{H}(\sigma).
\end{aligned}$$

Based on the above facts, one might think that the following formulation would probably generate a $\pi_0 \in PART(A)$ such that $d_\sigma^{max} = d^1(\pi_0,\sigma) = \mathcal{C}^1(\pi_0,\sigma) + \mathcal{C}^1(\sigma,\pi_0)$.

Suppose $\sigma = \{B_1,\ldots,B_n\}$ is the "true" partition. Without loss of generality, we sort the elements in σ such that $|B_1| \le |B_2| \le \cdots \le |B_n$. Additionally, $B_j = \{a_1^j, a_2^j, \ldots, a_{|B_j|}^j\}$, $1 \le j \le n$, where $a_i^j \in B_j \subseteq A$. Then, $\pi_0 = \{A_1,\ldots,A_m\}$, where $A_i = \{a_i^j \in B_j \mid |B_j| \ge i, 1 \le j \le n\}$ $(1 \le i \le m)$, and $m = max\{|B_j| \mid 1 \le j \le n\}$. For the easy understanding of the computation of the quasi-distance, we show the partition pair (π,σ) by an intersection matrix in which the element in the ith row and jth column equals $|A_i \cap B_j|$. π_0 is the partition such that the entry in the intersection matrix is either 0 or 1, and in each column of this matrix the entries with the values of 1 always appear above those with the values of 0. The following is an example intersection matrix of π_0 and σ. Since we sort the element in σ, in this matrix the entry values of the rightmost column are all 1, while in the leftmost column only the entry values of the first two rows are 1:

$$m\begin{cases}1&1&1&1&\cdots&1\\1&1&1&1&\cdots&1\\0&1&1&1&\cdots&1\\0&0&1&1&\cdots&1\\0&0&0&1&\cdots&1\\0&0&0&1&\cdots&1\\\vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\0&0&0&0&\cdots&1\\0&0&0&0&\cdots&1.\end{cases}$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{n}$$

Then, it seems that $\pi_0 = \{A_1, \ldots, A_m\} \in PART(A)$ might be a reasonable candidate satisfying $d_\sigma^{max} = d^1(\pi_0, \sigma)$. At first glance, using the above Facts I and II, we observe that $\mathcal{C}^1(\sigma, \pi_0)$ takes on its maximal value among all possible $\pi \in PART(A)$. Also, Fact I appears to suggest that the value $\mathcal{C}^1(\pi_0, \sigma)$ is at least not too small. Furthermore, we can prove the following theorem.

Theorem 6. *Let $d(\pi, \sigma) = \mathcal{C}^1(\pi, \sigma) + \mathcal{C}^1(\sigma, \pi)$ is a quasi-distance, where $\mathcal{C}^1$ is a conditional entropy, $\mathcal{H}$ and $\mathcal{M}$ are its corresponding partition entropy and the measure function, respectively. Also, let $g(x) = x \cdot \mathcal{M}(\frac{1}{x}, \ldots, \frac{1}{x})$. If $g''(x) \geq 0$, $d^1(\pi_0, \sigma) = max\{d^1(\pi', \sigma) \mid \pi' = \{A'_1, \ldots, A'_m\} \in PART(A),\ |A'_i \bigcap B_j| \leq 1, 1 \leq i \leq m, 1 \leq j \leq n, m \in \mathbb{N}\}$.*

Proof. The proof can be reduced to the following claim: for such a $\pi' \in PART(A)$, if $|A'_{i_1}| \leq |A'_{i_2}|$ and $|A'_{i_1} \bigcap B_j| = 1$, $|A'_{i_2} \bigcap B_j| = 0$, for some i_1, i_2, j, then after moving the unique element of $A'_{i_1} \bigcap B_j$ into A'_{i_2}, the quasi-distance $d^1(\pi', \sigma)$ may increase. This fact is shown in the following two matrixes. The distance of the matrix on the right is not smaller than that on the left:

$$\begin{matrix}\vdots&\vdots&\vdots&\vdots&\vdots\\1&1&1&0&1\\1&0&0&\underline{1}&1\\\vdots&\vdots&\vdots&\vdots&\vdots\end{matrix}\quad\Rightarrow\quad\begin{matrix}\vdots&\vdots&\vdots&\vdots&\vdots\\1&1&1&\underline{1}&1\\1&0&0&0&1\\\vdots&\vdots&\vdots&\vdots&\vdots\end{matrix}$$

The proof of the above claim, with the aid of Mean-value theorem (see [21, p. 86]), is straightforward. Note that the difference between the two quasi-distances (before and after the moving) by Δ. Then,

$$\Delta = \frac{\left[g\left(|A'_{i_2}| + 1\right) - g\left(|A'_{i_2}|\right)\right] - \left[g\left(|A'_{i_1}|\right) - g\left(|A'_{i_1}| - 1\right)\right]}{|A|}.$$

Using the Mean-value theorem, there exist $\xi_1 \in (|A'_{i_1}| - 1, |A'_{i_1}|)$, $\xi_2 \in (|A'_{i_2}|, |A'_{i_2}| + 1)$, satisfying

$$g\left(|A'_{i_1}|\right) - g\left(|A'_{i_1}| - 1\right) = g'(\xi_1),\ g\left(|A'_{i_2}| + 1\right) - g\left(|A'_{i_2}|\right) = g'(\xi_2).$$

So, $\Delta = \frac{g'(\xi_2) - g'(\xi_1)}{|A|}$. Since $g''(x) \geq 0$, $g'(x)$ is a nondecreasing function. Therefore, $\Delta \geq 0$, and the quasi-distance may increase after the above adjustment. □

We can further check that the partition entropies in Table 1 satisfy the conditions in Theorem 6, as shown in Table 3 which lists these entropies with the corresponding $g''(x)$. Thus, this theorem holds for all the quasi-distances in Table 1.

TABLE 3
The Partition Entropies with the Corresponding $g''(x)$ Defined in Theorem 6 $(x > 0)$

Entropy Name	$\mathcal{M}(\frac{1}{x}, \cdots, \frac{1}{x})$	$g(x)$	$g''(x)$
Shannon Entropy [22]	$\log_2 x$	$x \cdot \log_2 x$	$\log_2 x + \ln 2$
Pal Entropy	$e^{1-\frac{1}{x}}$	$x \cdot e^{1-\frac{1}{x}}$	$\frac{e^{1-\frac{1}{x}}}{x^3}$
Gini Index	$1 - \frac{1}{x}$	$x - 1$	0
Goodman-Kruskal coefficient	$1 - \frac{1}{x}$	$x - 1$	0

Nevertheless, there is an appreciable difference between d_σ^{max} and $d^1(\pi_0, \sigma)$. Here, we provide an example to show this. In this example, the Shannon Entropy is used in the quasi-distance measure and we use the notations in the formulation of π_0. Specifically, we assume that $n = 17$, $|B_j| = N$ if $1 \leq j \leq 16$, and $|B_{17}| = 2 \cdot N$, where N is an arbitrary positive integer. The left matrix in the diagram below corresponds to π_0. Next, we define a new partition of A, $\pi' = \{A'_1, \ldots, A'_N\}$, where $A'_i = \{a_i^1, a_i^2, \ldots, a_i^{15}, a_i^{16}, a_{2i-1}^{17}, a_{2i}^{17}\}$, $1 \leq i \leq N$. The corresponding intersection matrix is illustrated by the right matrix below:

$$N\begin{cases}1&1&\cdots&1&1\\1&1&\cdots&1&1\\\vdots&\vdots&\vdots&\vdots&\vdots\\1&1&\cdots&1&1\end{cases}\qquad N\begin{cases}1&1&\cdots&1&2\\1&1&\cdots&1&2\\\vdots&\vdots&\vdots&\vdots&\vdots\\1&1&\cdots&1&2.\end{cases}$$
$$\underbrace{\qquad\qquad\qquad}_{17}\qquad\qquad\underbrace{\qquad\qquad\qquad}_{17}$$
$$N\begin{cases}0&0&\cdots&0&1\\0&0&\cdots&0&1\\\vdots&\vdots&\vdots&\vdots&\vdots\\0&0&\cdots&0&1\end{cases}$$
$$\underbrace{\qquad\qquad\qquad}_{17}$$

Then, we can easily verify that $d^1(\pi', \sigma) - d^1(\pi_0, \sigma) = \log_2 17 - \log_2 16$, which is independent of the size of N. This example shows that in the general cases it is really hard to obtain the exact value of d_σ^{max}.

6.3 Computation of $d(\pi_0, \sigma)$

In this section, we give the explicit expression of $d(\pi_0, \sigma)$, which is useful in the approximation of d_σ^{max}. Let $|B_j| = b_j$ $(j = 1, \ldots, n)$ and $\sum_{j=1}^n b_j = b$. The quasi-distance between π_0 and σ can be expressed as

$$d(\pi_0, \sigma) = \mathcal{C}^1(\pi_0, \sigma) + \mathcal{C}^1(\sigma, \pi_0).$$

It is clear that $\mathcal{C}^1(\sigma, \pi_0) = \frac{\sum_{j=1}^n b_j \mathcal{G}(b_j)}{b}$, where $\mathcal{G}(b_j) = \mathcal{H}(\frac{1}{b_j}, \ldots, \frac{1}{b_j})$ (for example, when $\mathcal{H}$ is the Shannon entropy, $\mathcal{G}(b_j) = \log_2 b_j$). However, it will take much efforts to express $\mathcal{C}^1(\pi_0, \sigma)$ analytically.

To this end, we specify all the *change points* $b_{j_1}, \ldots, b_{j_k}$ in the sequence $b_0 \leq b_1 \leq \cdots \leq b_n$ (b_0 is set to 0 for convenience) such that $b_{(j_l - 1)} < b_{j_l}$ $(l = 1, \ldots, k)$. Then, the other conditional entropy is

$$\mathcal{C}^1(\pi_0, \sigma) = \frac{\sum_{l=1}^k \left(b_{j_l} - b_{(j_l - 1)}\right)(n + j_1 - j_l)\mathcal{G}(n + j_1 - j_l)}{b}.$$

TABLE 4
Some Characteristics of Experimental Data Sets

Data set	Source	# of objects	# of features	# of classes	Min class size	Max class size	CV_0
Document Data Sets							
fbis	TREC	2463	2000	17	38	506	0.961
hitech	TREC	2301	126373	6	116	603	0.495
tr23	TREC	204	5832	6	6	91	0.935
tr45	TREC	690	8261	10	14	160	0.669
la2	TREC	3075	31472	6	248	905	0.516
ohscal	OHSUMED-233445	11162	11465	10	709	1621	0.266
re0	Reuters-21578	1504	2886	13	11	608	1.502
re1	Reuters-21578	1657	3758	25	10	371	1.385
k1a	WebACE	2340	21839	20	9	494	1.004
wap	WebACE	1560	8460	20	5	341	1.040
Biomedical Data Sets							
LungCancer	KRBDSR	203	12600	5	6	139	1.363
Leukemia	KRBDSR	325	12558	7	15	79	0.584
UCI Data Sets							
ecoli	UCI	336	7	8	2	143	1.160
pendigits	UCI	10992	16	10	1055	1144	0.042

TABLE 5
Experimental Results for Real-World Data Sets

Data set	Coefficient of Variation		Distances				Normalized Distances			
	CV_0	DCV	d^1_{sha}	d^1_{pal}	d^1_{gin}	d^1_{goo}	$normd^1_{sha}$	$normd^1_{pal}$	$normd^1_{gin}$	$normd^1_{goo}$
fbis	0.96	0.41	3.1379	3.5271	1.0555	0.7808	0.2885	0.5014	0.5755	0.4367
hitech	0.5	0.13	3.3003	3.6902	1.1742	0.8462	0.2999	0.5861	0.6651	0.4876
tr23	0.93	0.51	2.5445	3.3766	0.9773	0.75	0.3462	0.5347	0.6109	0.4920
tr45	0.67	0.23	2.3524	3.1590	0.8055	0.6029	0.2572	0.3916	0.4480	0.3438
la2	0.52	0.14	2.1880	3.0950	0.7821	0.5489	0.1925	0.3838	0.4475	0.3222
ohscal	0.27	-0.17	3.6165	3.6746	1.1387	0.8223	0.2718	0.5365	0.6079	0.4435
re0	1.5	1.11	3.8815	3.8079	1.1976	0.9674	0.3840	0.6641	0.7187	0.6096
re1	1.39	1.06	3.6441	3.6515	1.0926	0.8690	0.3571	0.5502	0.6022	0.4935
k1a	1	0.51	3.1826	3.5024	1.0228	0.7530	0.2951	0.4908	0.5573	0.4229
wap	1.04	0.55	3.2019	3.5144	1.0234	0.7712	0.3144	0.4986	0.5607	0.4360
lungcancer	1.36	0.73	2.1770	3.1860	0.8469	0.6207	0.2968	0.5287	0.6145	0.4809
leukemia	0.58	0.21	2.9684	3.5946	1.1039	0.8554	0.3649	0.5565	0.6257	0.4929
ecoli	1.16	0.66	2.5780	3.3095	0.9209	0.7262	0.3199	0.4992	0.5672	0.4683
pendigits	0.04	-0.53	2.4315	3.2218	0.8578	0.6261	0.1814	0.3854	0.4522	0.3304
Parameters used in CLUTO: -clmethod=rb -sim=cos -crfun=i2 -niter=30										

Altogether, $d(\pi_0, \sigma)$ can be expressed as

$$d(\pi_0,\sigma) = \frac{\sum_{l=1}^{k}\left(b_{j_l} - b_{(j_l-1)}\right)(n + j_1 - j_l)\mathcal{G}(n + j_1 - j_l)}{b} + \frac{\sum_{j=1}^{n} b_j \mathcal{G}(b_j)}{b}.$$

To further clarify the computation of $\mathcal{C}^1(\pi_0, \sigma)$, we give the following example. The 5×7 intersection matrix below is induced by two partitions π_0 and σ:

$$5\left\{\underbrace{\begin{matrix}1&1&1&1&1&1&1\\1&1&1&1&1&1&1\\1&1&1&1&1&1&1\\0&0&0&1&1&1&1\\0&0&0&0&0&1&1\end{matrix}}_{7}\right..$$

In this example, the sequence of $(b_0, b_1, \ldots, b_7)$ is (0, 3, 3, 3, 4, 4, 5, 5), $b = 27$, $n = 7$, the sequence of the change points is (b_1, b_4, b_6), and the index sequence (j_1, j_2, j_3) of the change points is (1, 4, 6). Then, in this example

$$\begin{aligned}\mathcal{C}^1(\pi_0,\sigma) &= \frac{(b_1 - b_0)(n + j_1 - j_1)\mathcal{G}(n + j_1 - j_1)}{b} \\ &\quad + \frac{(b_4 - b_3)(n + j_1 - j_2)\mathcal{G}(n + j_1 - j_2)}{b} \\ &\quad + \frac{(b_6 - b_5)(n + j_1 - j_3)\mathcal{G}(n + j_1 - j_3)}{b} \\ &= \frac{3 \cdot 7 \cdot \mathcal{G}(7) + 1 \cdot 4 \cdot \mathcal{G}(4) + 1 \cdot 2 \cdot \mathcal{G}(2)}{27}.\end{aligned}$$

6.4 Approximation of d_σ^{max} in the General Cases

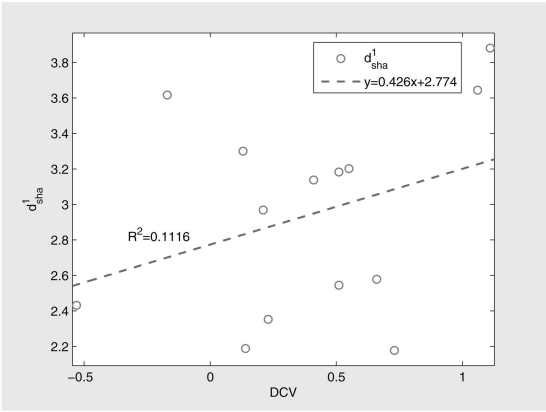
In general, it is still unknown when $d_\sigma^{max} = d^1(\pi, \sigma)$ happens. However, the mathematic facts listed above inspire us that d_σ^{max} has its approximation range, as shown in the following inequality:

$$\underline{d_\sigma^{max}} \le d_\sigma^{max} \le \overline{d_\sigma^{max}}, \tag{22}$$

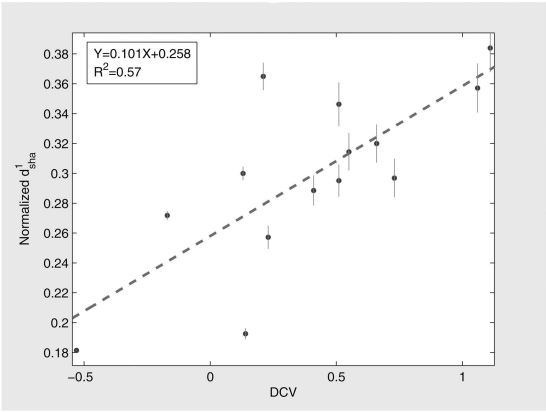
where $\underline{d_\sigma^{max}} = \mathcal{C}^1(\pi_0, \sigma) + \mathcal{C}^1(\sigma, \pi_0)$ and $\overline{d_\sigma^{max}} = \mathcal{H}(\sigma) + \mathcal{C}^1(\sigma, \pi_0)$. They are the tight lower and upper bound of d_σ^{max}, respectively.

This approximation range is reasonable because the following two inequalities always holds:

$$0 \le \overline{d_\sigma^{max}} - d_\sigma^{max} \le \mathcal{H}(\sigma), \tag{23}$$



(a)



(b)

Fig. 2. DCV versus d^1_{sha} and $normd^1_{sha}$. (a) DCV versus d^1_{sha}. (b) DCV versus $normd^1_{sha}$.

$$0 \leq d_\sigma^{max} - \underline{d_\sigma^{max}} \leq \mathcal{H}(\sigma). \tag{24}$$

Inequalities (23) and (24) show that the difference between d_σ^{max} and its upper (lower) bound is smaller than $\mathcal{H}(\sigma)$. When $|A|$ is very large (this is a popular assumption for practical clustering problems), $\mathcal{C}^1(\sigma, \pi_0)$ is much larger than $\mathcal{H}(\sigma)$. Thus, the above upper and lower bounds for d_σ^{max} are effectively estimated. In practice, d_σ^{max} might be approximated by the medium value of the upper and lower bounds

$$\begin{aligned}\widehat{d_\sigma^{max}} &= \frac{\left(\underline{d_\sigma^{max}} + \overline{d_\sigma^{max}}\right)}{2} \\ &= \frac{\left(\mathcal{C}^1(\pi_0,\sigma) + \mathcal{C}^1(\sigma,\pi_0)\right) + \left(\mathcal{H}(\sigma) + \mathcal{C}^1(\sigma,\pi_0)\right)}{2}.\end{aligned} \tag{25}$$

Note that if $\underline{d_\sigma^{max}}$ ($\overline{d_\sigma^{max}}$) is substituted for the d_σ^{max} in (21), the result corresponds to the upper (lower) bounds of the normalized distance.

6.5 Exact Computation of d_σ^{max} in the Special Cases

In this section, we show the special cases where the exact value of d_σ^{max} can be obtained. First, we consider the case when the cluster sizes of the "true" clusters are all equal.

Theorem 7. *When the cluster sizes of the "true" clustering σ are all equal, $d_\sigma^{max} = \underline{d_\sigma^{max}} = \overline{d_\sigma^{max}}$.*

Proof. In this case, one can easily verify that $\mathcal{C}^1(\pi_0, \sigma) = \mathcal{H}(\sigma)$. By Inequality (22), $d_\sigma^{max} = \underline{d_\sigma^{max}} = \overline{d_\sigma^{max}}$ holds directly. □

Next, we analyze the exact computation of d_σ^{max} when d^1_{goo} in Table 1 is adopted.

Theorem 8. *For the distance d^1_{goo}, $d_\sigma^{max} = \underline{d_\sigma^{max}} = \overline{d_\sigma^{max}}$.*

Proof. In this case that $\mathcal{H}_{goo}$ is used in the distance computation, one can easily verify that $\mathcal{C}^1_{goo}(\pi_0, \sigma) = \mathcal{H}_{goo}(\sigma)$. By Inequality (22), $d_\sigma^{max} = \underline{d_\sigma^{max}} = \overline{d_\sigma^{max}}$ holds directly. □

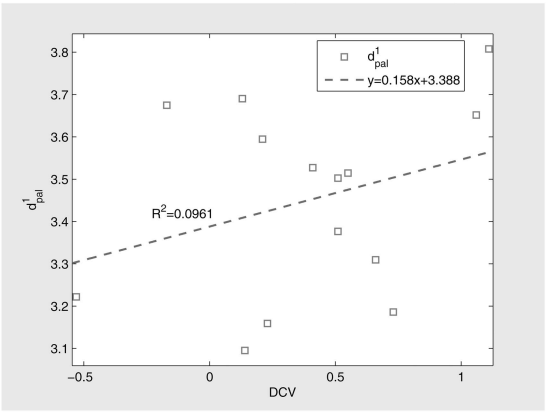
7 Experimental Results

In this section, we present experimental results to illustrate the effectiveness of distance normalization when we use the distance measures in Table 1 for comparing clusterings of different data sets.

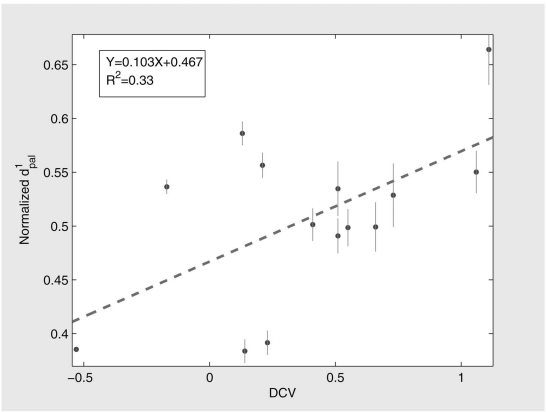
7.1 The Experimental Setup

Experimental tool. Since we aim to compare different clustering validation measures (not the performance of different clustering algorithms), the most popular clustering algorithm K-means is adopted. In our experiments, we used the CLUTO [13] implementation of K-means.

Experimental data sets. For our experiments, we used a number of real-world data sets that were obtained from different application domains. Some characteristics of these data sets are shown in Table 4. In the table, "# of classes" indicates the number of "true" clustering. Please refer to [26] for more details of these data sets.



(a)



(b)

Fig. 3. DCV versus d^1_{pal} and $normd^1_{pal}$. (a) DCV versus d^1_{pal}. (b) DCV versus $normd^1_{pal}$.

7.2 Evaluation Metric

We first introduce the Coefficient of Variation (CV) [6], which is a measure of dispersion of a data distribution. CV is defined as the ratio of the standard deviation to the mean. The larger the CV value is, the greater the variability is in the data. Given a set of data objects $X = \{x_1, x_2, \ldots, x_n\}$, we have $CV = \frac{s}{\bar{x}}$, where

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

and $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$.

Next, we define CV_0 as the CV value on the cluster sizes of the "true" clusters and CV_1 as the CV value on the cluster sizes of the clustering results. Also, $DCV = CV_0 - CV_1$ is the change of CV values before and after clustering. Xiong et al. [26] has shown that DCV can be used to describe how different between the "true" cluster distribution and the distribution of cluster results. DCV owns the property that can be used to indicate bad clusterings when the DCV values are large. In fact, a large DCV value indicates that a clustering result is away from the true cluster distribution. Thus, a good quasi-distance must have large values if the DCV values are large. Therefore, we evaluate the proposed quasi-distances by checking whether any clustering results with larger DCV values will lead to larger quasi-distances. However, we agree that DCV is a necessary but not sufficient condition for the clustering quality. In other words, a large DCV value indicates a bad clustering result, but a small DCV value may not indicate a good clustering result. This is also the reason that we introduce the quasi-distance.

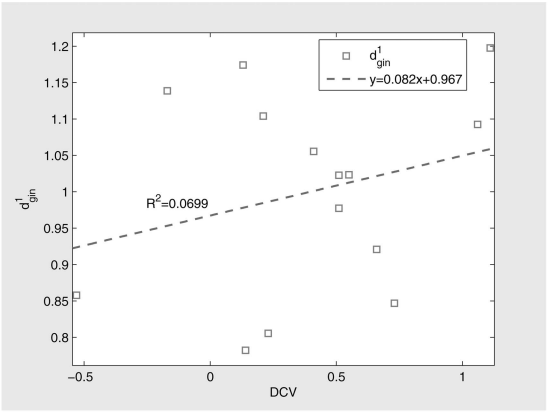
7.3 The Effect of Distance Normalization

In our experiments, we first applied K-means for clustering the input data sets and the number of clusters k was set as the "true" cluster number for the purpose of comparison. Then, we computed the following values:

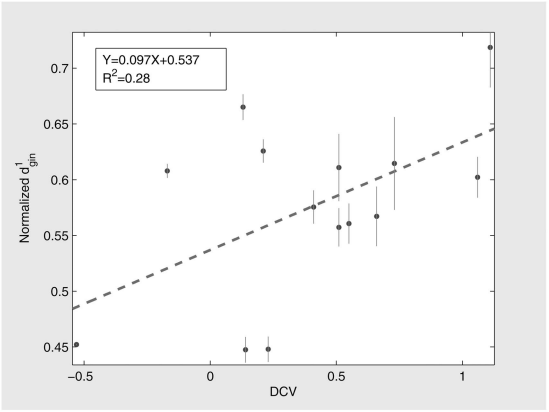
- The values of the four distances d^1_{sha}, d^1_{pal}, d^1_{gin}, d^1_{goo} (as shown Table 1) between the "true" clustering and the clustering results, respectively.
- The values of the four normalized distances $normd^1_{sha}$, $normd^1_{pal}$, $normd^1_{gin}$, $normd^1_{goo}$, respectively.

Note that the normalized distances $normd^1_{sha}$, $normd^1_{pal}$, $normd^1_{gin}$ are approximated using the approximate computation of d^{max}_{σ} in (25). The exact value of $normd^1_{goo}$ is obtained by Theorem 8. Table 5 presents a summary of the experimental results on various real-world data sets.

Also, Figs. 2, 3, 4, and 5 show the DCV values and the corresponding (normalized) distance values of the four distance measures on all the experimental data sets. For the normalized distance on each data set, the range bar shows the upper and lower bounds of the normalized distance, which are computed using $\underline{d^{max}_{\sigma}}$ and $\overline{d^{max}_{\sigma}}$ in (24) and (23), respectively. As can be seen in each subfigure, there is a linear regression fitting line for all the points. The value of R Square (R^2) is also shown in the figure. The R^2 value provides a guide to the "goodness-of-fit" of linear regression.



(a)



(b)

Fig. 4. DCV versus d^1_{gin} and $normd^1_{gin}$. (a) DCV versus d^1_{gin}. (b) DCV versus $normd^1_{gin}$.

In these figures, we can also observe that the R^2 values of the normalized distances are always larger than those of the original distance, and the R^2 value of $normd_{sha}$ is the largest among all observed distance measures. In other words, our experimental results indicate that the normalized distance performs better than the original distance when comparing clustering of different data sets. Also, the normalized distance $normd_{sha}$ performs the best among four distance measures in Table 1.

7.4 The Range of Performance of the Normalized Distance Measures

As can be seen in Figs. 2, 3, 4, and 5, we also use the range bar to indicate the range of performance of the normalized distance measures. The shorter the range bar is, the more precise the approximate normalized distance is to its true value. In the following, we show that the length of the range bar is closely related to the CV_0 value of the data set.

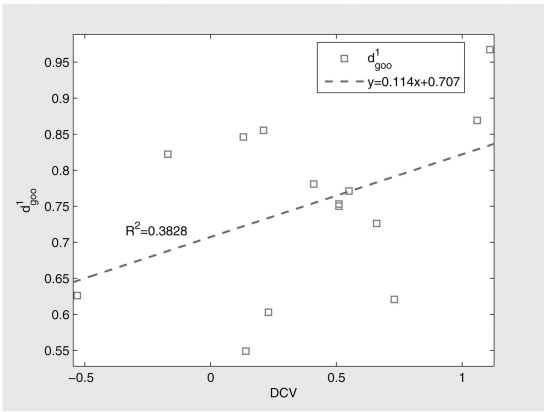
Figs. 6, 7, and 8 show the CV_0 values and the lengths of ranges of the three normalized distances: $normd^1_{sha}$, $normd^1_{pal}$, $normd^1_{gin}$ [3] on all the data sets. In these figures, we can observe that the length of the range bar increases as the increase of CV_0. This indicates that the approximate computation of these three normalized distances performs better when CV_0 is smaller. This result agrees with Theorem 7, which states that the exact normalized distance can be obtained when $CV_0 = 0$.[4]

8 Conclusions

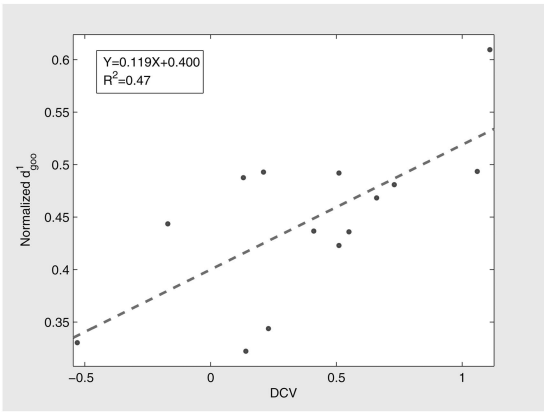
In this paper, we first proposed a uniform representation of quasi-distance, which possesses three properties: symmetry, the triangle law, and the minimum reachable. Several well-known information-theoretic distance measures such as Shannon Distance, Pal Distance, the Van Dongen criterion, and the Mirkin metric can be described by this generalized representation. Also, three properties of the quasi-distance naturally lend itself as the external measure for clustering validation. Furthermore, we highlighted the importance of normalization when applying distance measures to compare the clustering results of different data sets. Along this line, we provided a theoretic analysis of the computation form of the maximum value of a distance measure. This is important for the normalization process. Finally, in order to compare the clustering performances of an algorithm on different data sets, we applied the K-means clustering algorithm to empirically show that 1) the normalized distance measures outperform the original distance measure and 2) the normalized Shannon distance has the best performance among four observed distance measures.

3. Since the exact value of $normd^1_{goo}$ is obtained by Theorem 8, the length of its range bar is always 0. Thus, the corresponding range bar for $normd^1_{goo}$ is omitted.

4. When $CV_0 = 0$, the cluster sizes of the "true" clustering are all equal.



(a)



(b)

Fig. 5. DCV versus d^1_{goo} and $normd^1_{goo}$. (a) DCV versus d^1_{goo}. (b) DCV versus $normd^1_{goo}$.

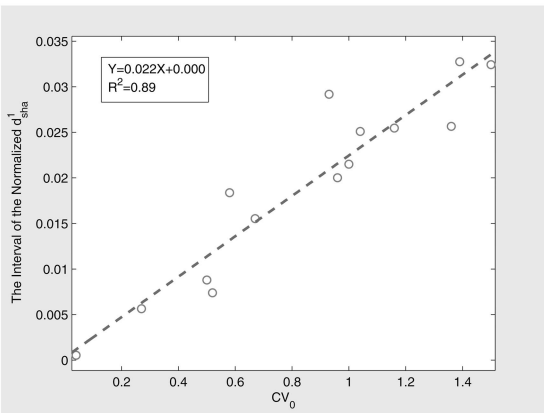


Fig. 6. CV_0 versus the range bar length of $normd^1_{sha}$.

ACKNOWLEDGMENTS

This work is supported by the National Basic Research Priorities Program (2007CB311004 and 2003CB317004), the National Science Foundation of China (60435010, 90604017, 60675010, and 60775035) and the 863 Project (2006AA01Z128 and 2007AA01Z132). Also, this research was supported in part by the Rutgers Seed Funding for Collaborative Computing Research and a Faculty Research Grant from Rutgers Business School—Newark and New Brunswick. Finally, the authors are grateful to the anonymous referees for their constructive comments on this paper.

REFERENCES

[1] J. Aczl and Z. Darczy, *On Measures of Information and Their Characterizations.* Academic Press, 1975.

[2] D. Barbará and P. Chen, "Using Self-Similarity to Cluster Large Data Sets," *Data Mining and Knowledge Discovery,* vol. 7, no. 2, pp. 123-152, 2003.

[3] D. Barbará, Y. Li, and J. Couto, "Coolcat: An Entropy-Based Algorithm for Categorical Clustering," *Proc. 11th ACM Int'l Conf. Information and Knowledge Management (CIKM '02),* pp. 582-589, 2002.

[4] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen, *Classification and Regression Trees.* Wadsworth Int'l Group, 1984.

[5] Y. Chen, Y. Zhang, and X. Ji, "Size Regularized Cut for Data Clustering," *Proc. 18th Conf. Neural Information Processing Systems (NIPS),* 2005.

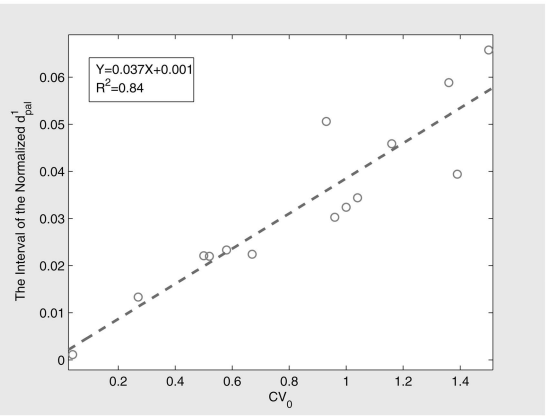


Fig. 7. CV_0 versus the range bar length of $normd^1_{pal}$.

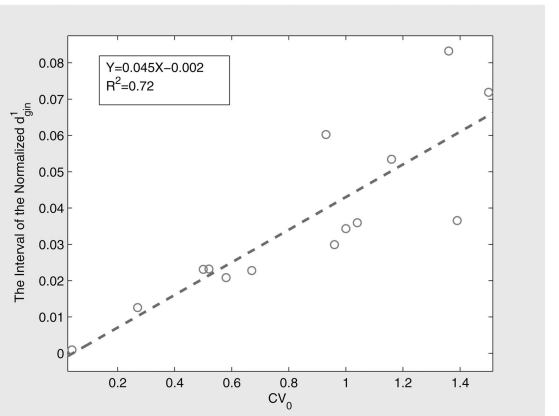


Fig. 8. CV_0 versus the range bar length of $normd^1_{gin}$.

[6] M.H. DeGroot and M.J. Schervish, *Probability and Statistics,* third ed. Addison-Wesley, 2001.

[7] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On Clustering Validation Techniques," *J. Intelligent Information Systems,* vol. 17, nos. 2/3, pp. 107-145, 2001.

[8] M. Halkidi, D. Gunopulos, N. Kumar, M. Vazirgiannis, and C. Domeniconi, "A Framework for Semi-Supervised Learning Based on Subjective and Objective Clustering Criteria," *Proc. Fifth IEEE Int'l Conf. Data Mining (ICDM '05),* pp. 637-640, 2005.

[9] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data.* Prentice Hall, 1998.

[10] S. Jaroszewicz, D.A. Simovici, W. Kuo, and L. Ohno-Machado, "The Goodman-Kruskal Coefficient and Its Applications in the Genetic Diagnosis of Cancer," *IEEE Trans. Biomedical Eng.,* vol. 51, no. 7, pp. 1095-1102, July 2004.

[11] I. Jonyer, D.J. Cook, and L.B. Holder, "Graph-Based Hierarchical Conceptual Clustering," *J. Machine Learning Research,* vol. 2, pp. 19-43, 2001.

[12] I. Jonyer, L.B. Holder, and D.J. Cook, "Graph-Based Hierarchical Conceptual Clustering in Structural Databases," *Proc. 17th Nat'l Conf. Artificial Intelligence and 12th Conf. Innovative Applications of Artificial Intelligence (AAAI '00),* p. 1078, 2000.

[13] G. Karypis, http://glaros.dtc.umn.edu/gkhome/views/cluto, 2008.

[14] J. Li, D. Tao, W. Hu, and X. Li, "Kernel Principle Component Analysis in Pixels Clustering," *Proc. IEEE/WIC/ACM. Int'l Conf. Web Intelligence (WI '05),* pp. 786-789, 2005.

[15] W. Li, W.K. Ng, Y. Liu, and K.-L. Ong, "Enhancing the Effectiveness of Clustering with Spectra Analysis," *IEEE Trans. Knowledge and Data Eng.,* vol. 19, no. 7, pp. 887-902, July 2007.

[16] P. Luo, G. Zhan, Q. He, Z. Shi, and K. Lü, "On Defining Partition Entropy by Inequalities," *IEEE Trans. Information Theory,* vol. 53, no. 9, pp. 3233-3239, 2007.

[17] R. Lopez De Mantaras, "A Distance-Based Attribute Selection Measure for Decision Tree Induction," *Machine Learning,* vol. 6, no. 1, pp. 81-92, 1991.

[18] A.W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications.* Academic Press, 1979.

[19] M. Meila, "Comparing Clusterings: An Axiomatic View," *Proc. 22nd Int'l Conf. Machine Learning (ICML '05),* pp. 577-584, 2005.

[20] N.R. Pal and S.K. Pal, "Entropy: A New Definition and Its Applications," *IEEE Trans. Systems Man and Cybernetics,* vol. 21, no. 5, pp. 1260-1270, 1991.

[21] M.H. Protter and C.B. Morrey Jr., *A First Course in Real Analysis,* second ed. Springer, 1991.

[22] C.E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical J.,* vol. 27, pp. 379-423, 623-656, 1948.

[23] D.A. Simovici and S. Jaroszewicz, "An Axiomatization of Partition Entropy," *IEEE Trans. Information Theory,* vol. 48, no. 7, pp. 2138-2142, 2002.

[24] D.A. Simovici and S. Jaroszewicz, "A Metric Approach to Building Decision Trees Based on Goodman-Kruskal Association Index," *Proc. Eighth Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD '04),* pp. 181-190, 2004.

[25] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining.* Addison-Wesley, 2005.

[26] H. Xiong, J. Wu, and J. Chen, "K-Means Clustering Versus Validation Measures: A Data Distribution Perspective," *Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '06),* pp. 779-784, 2006.

Ping Luo received the PhD degree in computer science from the Chinese Academy of Sciences. He is currently a research scientist in the Hewlett-Packard Labs China, Beijing. His research interests include knowledge discovery and machine learning. He has published several papers in some prestigious refereed journals and conference proceedings, such as the *IEEE Transactions on Information Theory*, the *Journal of Parallel and Distributed Computing*, ACM SIGKDD, and ACM CIKM. He is a recipient of the President's Exceptional Student Award, Institute of Computing Technology, CAS. He is a member of the ACM.

Hui Xiong received the BE degree in automation from the University of Science and Technology of China, China, the MS degree in computer science from the National University of Singapore, Singapore, and the PhD degree in computer science from the University of Minnesota. He is currently an associate professor in the Management Science and Information Systems Department, Rutgers University, Newark, New Jersey. His general area of research is data and knowledge engineering, with a focus on developing effective and efficient data analysis techniques for emerging data intensive applications. He has published more than 60 technical papers in peer-reviewed journals and conference proceedings. He is a coeditor of *Clustering and Information Retrieval* (Kluwer Academic Publishers, 2003), a coeditor-in-chief of *Encyclopedia of GIS* (Springer, 2008), an associate editor of the *Knowledge and Information Systems* journal, and has served regularly on the organization committees and the program committees of a number of international conferences and workshops. He was the recipient of the 2008 IBM ESA Innovation Award, the 2009 Rutgers University Board of Trustee Research Fellowship for Scholarly Excellence, the 2007 Junior Faculty Teaching Excellence Award and the 2008 Junior Faculty Research Award at the Rutgers Business School. He is a senior member of the IEEE and a member of the ACM.

Guoxing Zhan received the master's degree from the Chinese Academy of Sciences. He is currently a PhD student in the Department of Computer Science, Wayne State University, Detroit, Michigan. His research interests include data analysis, wireless sensor networks, and algebraic groups.

Junjie Wu received the BE degree in civil engineering and the PhD degree in management science and engineering from Tsinghua University, China. He is currently an assistant professor in the Department of Information Systems, School of Economics and Management, Beihang University, Beijing, China. His research interests include data mining and statistical modeling, with a special interest on solving the problems raised from the real-world business applications. He has published four papers in KDD and one paper in ICDM. He is a cochair of Data Mining in Business, a special track in AMIGE '08. He has also been a reviewer for the leading academic journals and many international conferences in his area. He is the recipient of the Outstanding Young Research Award of the School of Economics and Management, Tsinghua University. He is a member of the ACM and the AIS.

Zhongzhi Shi is a professor in the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, leading the Research Group of Intelligent Science. His research interests include intelligence science, multiagent systems, semantic Web, machine learning, and neural computing. He has won a Second-Grade National Award at Science and Technology Progress of China in 2002 and two Second-Grade Awards at Science and Technology Progress of the Chinese Academy of Sciences in 1998 and 2001, respectively. He is the chair for the WG 12.2 of IFIP. He serves as a vice president for the Chinese Association of Artificial Intelligence and the executive president of the Chinese Neural Network Council. He is a senior member of the IEEE and a member of the AAAI and the ACM.