Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/100428

Reviews on Determining the Number of Clusters

Shuo Xu[1,*], *Xiaodong Qiao*[1], *Lijun Zhu*[1], *Yunliang Zhang*[1], *Chunxiang Xue*[2] *and Lin Li*[3,*]

[1] Information Technology Supporting Center, Institute of Scientific and Technical Information of China, No. 15 Fuxing Rd., Haidian District, Beijing 100038, P.R. China
[2] Department of Information Management, Nanjing University of Science and Technology, No. 200 Xiaolingwei Street, Xuanwu District, Nanjing 210094, P.R. China
[3] College of Information and Electrical Engineering, China Agricultural University (East Campus), No. 17 Qinghua East Rd., Haidian District, Beijing 100083, P.R. China

Received: 10 Apr. 2016, Revised: 27 May 2016, Accepted: 28 May 2016
Published online: 1 Jul. 2016

Abstract: Clustering analysis seeks to partition a given dataset into groups or clusters so that the data objects within a cluster are more similar to each other than the objects in different clusters. A very rich literature on clustering analysis has developed over the past three decades. But a crucial question still remains unanswered: how many clusters are contained in the population on earth when only an observed set of samples is available? The goal of this paper is to provide a comprehensive review of approaches on determining the ”correct” number of clusters. In particular, we divide these approaches into three categories: internal measures, external measures, and clustering stability based methods. Then, we introduce several representative examples, and present specific challenges pertinent to each category. Finally, the promising trends are suggested in this field.

Keywords: Internal Measures, External Measures, Clustering Stability, Axioms, Cluster Validation.

1. Introduction

As the amount of data we nowadays have to deal with becomes larger and larger, clustering analysis is the formal study of algorithms and methods that help us to detect structures in the data and to identify interesting groups or clusters so that the data objects within a cluster are more similar to each other than the objects in different clusters [35]. Adopting a machine learning perspective, clusters correspond to *hidden patterns*, and the search for clusters is *unsupervised learning*. Algorithms and methods for clustering analysis provide core techniques for exploratory data analysis and play an outstanding role in numerous applications, such as information retrieval and text mining [14], web log analysis [82], and many others.

While people are extremely good at pointing out the relevant structure in the data just by looking at the 2-D plots, it is not easy to automatically reorganize underlying clusters from the data. A major challenge is to estimate the ”correct” number of clusters in the population as well as the interpretation of the clusters. The idea of directly asking the question has its origins in population statistics, and some early papers are [22,12,26] and references therein. In fact, part of the difficulty [57] comes from the absence, in general, of an objective way to assess the clustering quality and to compare two clusterings of the data. The goal of this survey is to provide a comprehensive review on how to determine the ”correct” number of clusters in the population when only an observed set of samples is available, also known as *cluster validation* [35,34].

The broad question ”How to determine the number of clusters” is addressed in two ways: (1) to run clustering algorithms with different number of clusters, and use cluster/model validity indexes to select one of them, and (2) to automatically fit a particular number of clusters. The paper mainly focuses on the first way. Our main contributions in the paper include: (1) This paper provides a comprehensive review of approaches on determining the ”correct” number of clusters with a goal of providing useful advice and references to broad community of clustering practitioners. (2) This paper presents a taxonomy of corresponding approaches, introduces several representative examples and challenges & recent advances.

* Corresponding author e-mail: xush@istic.ac.cn , lilincau@gmail.com

The organization of the rest of this paper is as follows. After the related concepts and notations are introduced in Section 2, we divide cluster validation into three categories: the internal measures in Section 3, the external measures in Section 4, and the clustering stability based methods in Section 5. In respective section, we introduce several representative examples, and present specific challenges pertinent to each category. Finally, we conclude this paper and suggest the promising trends in this field.

Note that strictly speaking, the internal measures should not belong to cluster validation, since a general principle for cluster validation should not be restricted to a specific group of clustering algorithms, that is, model free, but internal measures usually assume compact clusters tightly packed around cluster centroids [45]. The internal measures are contained here only for completeness. Apart from the cluster validation approaches described in this study, the Akaike information criterion (AIC) [2], the Bayesian information criterion (BIC) [62], the minimum description length (MDL) [59], and so on, are also often used to score each model, and then the appropriate model is selected according to corresponding scores. But due to space limit, they are excluded in this review.

2. Definitions and Notations

Given a set of n objects $S = \{o_1, o_2, \cdots, o_n\}$. Suppose the objects o_i can be described by m explanatory variables, denoted as $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,m})^t, i = 1, 2, \cdots, n$. Let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)^t$ be data matrix for n objects. Here t denotes the transpose of a vector or matrix. A *clustering* is a set of non-empty disjoint subsets, called as *cluster*, of S such that their union equals S. Of course, clusters need not be disjoint. Soft-cluster memberships [14], fuzzy clustering [5] and overlapping clustering [6] are instances where each object can actually belong to two different clusters, and are often used in cluster analysis. Though in this study we restrict ourselves to hard-cluster case, some methods can be applied directly to soft-cluster case, such as variation of information (see further), etc. Additionally, one can also convert soft-clusters later into disjoint subsets, then utilize the corresponding methods in the work.

For $a, b \in S$ and a clustering $\mathscr{C}$ of S, we write $a \sim_{\mathscr{C}} b$ whenever a and b are in the same cluster of clustering $\mathscr{C}$ and $a \not\sim_{\mathscr{C}} b$, otherwise. The set of all clusterings of S is denoted by $\mathscr{P}(S)$. In addition, for any clustering $\mathscr{C} \in \mathscr{P}(S)$, one can define a discrete random variable $X_{\mathscr{C}}$ as follows:

$$X_{\mathscr{C}} : \begin{pmatrix} 1 & 2 & \cdots & k \\ |C_1|/n & |C_2|/n & \cdots & |C_k|/n \end{pmatrix}. \tag{1}$$

Suppose $\mathscr{C} = \{C_1, C_2, \cdots, C_k\} \in \mathscr{P}(S)$ and $\mathscr{C}' = \{C'_1, C'_2, \cdots, C'_l\} \in \mathscr{P}(S)$ represent two different clusterings of S. Of course, both k and l must be less than or equal to n. Let $n_{i,j}$ denote the number of objects that are common to clusters C_i in $\mathscr{C}$ and C'_j in $\mathscr{C}'$, viz.,

Clustering $\mathscr{C}$

Clustering $\mathscr{C}'$	C'_1	C'_2	$\cdots$	C'_l	$\sum$
C_1	$n_{1,1}$	$n_{1,2}$	$\cdots$	$n_{1,l}$	$\lvert C_1\rvert$
C_2	$n_{2,1}$	$n_{2,2}$	$\cdots$	$n_{2,l}$	$\lvert C_2\rvert$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_k	$n_{k,1}$	$n_{k,2}$	$\cdots$	$n_{k,l}$	$\lvert C_k\rvert$
$\sum$	$\lvert C'_1\rvert$	$\lvert C'_2\rvert$	$\cdots$	$\lvert C'_l\rvert$	n

Figure 1: The Contingency Table of the Pair $\mathscr{C}, \mathscr{C}'$.

$$n_{i,j} = |C_i \cap C'_j|, 1 \le i \le k \wedge 1 \le j \le l. \tag{2}$$

A *trivial clustering* is either the one-clustering, denoted as $\hat{1}$, that consist of just one cluster or the singleton clustering, denoted as $\hat{0}$, in which every element forms its own cluster. In fact, all criteria for comparing clustering can be described using the so-called *confusion matrix*, or *association matrix* or *contingency table* of the pair $\mathscr{C}, \mathscr{C}' \in \mathscr{P}(S)$. The contingency table is a $k \times l$ matrix, whose (i, j)-th entry is $n_{i,j}$, as shown in Fig. 1.

For a given clustering of S into $1 \le k \le n$ clusters, $\mathscr{C} = \{C_1, C_2, \cdots, C_k\}$, each cluster with covariance Γ, $\mathbf{B}_k$ and $\mathbf{W}_k$ are defined to be the $m \times m$ matrices of between and within k-clusters sums of squares and cross-products.

$$\mathbf{W}_k = \sum_{r=1}^{k} \sum_{\mathbf{x}_i \in C_r} (\mathbf{x}_i - \bar{\mathbf{x}}_r)(\mathbf{x}_i - \bar{\mathbf{x}}_r)^t \tag{3}$$

$$\mathbf{B}_k = \sum_{r=1}^{k} |C_r| (\bar{\mathbf{x}}_r - \bar{\mathbf{x}})(\bar{\mathbf{x}}_r - \bar{\mathbf{x}})^t, \tag{4}$$

where $\bar{\mathbf{x}}_r$ and $\bar{\mathbf{x}}$ denote centroid or medoid of cluster r and the whole data set, respectively. Note that $\mathbf{B}_1$ is not defined.

The clustering $\mathscr{C}' \in \mathscr{P}(S)$ is a *refinement* of $\mathscr{C} \in \mathscr{P}(S)$ (or $\mathscr{C}$ is a *coarsening* of $\mathscr{C}'$), if each cluster of $\mathscr{C}'$ is contained in a cluster of $\mathscr{C}$, formally:

$$\forall C'_j \in \mathscr{C}', \exists C_i \in \mathscr{C} \text{ s.t. } C'_j \subseteq C_i. \tag{5}$$

The *product* $\mathscr{C} \times \mathscr{C}'$ of two clusterings $\mathscr{C}, \mathscr{C}' \in \mathscr{P}(S)$ is the coarsest common refinement of the two clusterings:

$$\mathscr{C} \times \mathscr{C}' = \{C_i \cap C'_j | C_i \in \mathscr{C}, C'_j \in \mathscr{C}', C_i \cap C'_j \neq \emptyset\}. \tag{6}$$

The product $\mathscr{C} \times \mathscr{C}'$ is again a clustering in $\mathscr{P}(S)$, and if $\mathscr{C}'$ is a refinement of $\mathscr{C}$, then $\mathscr{C} \times \mathscr{C}' = \mathscr{C}'$.

3. Internal Measures

To estimate the number of clusters K on the data set S, one intuitive approach is to look for k that provides the strongest significant evidence against the null hypothesis H_0 of $k = 1$, that is, "no clusters" in S. Two popular null

hypotheses are unimodality hypothesis [61] and uniformity hypothesis [35,13,30]. Under the former hypothesis, the data are thought to be a random sample from a multivariate normal distribution. Under the latter hypothesis, the data are sampled from a uniform distribution in m-dimensional space. For both types of hypotheses, evidence against H_0 can be summarized formally under probability models for S or more informally by using internal measures as described here. By internal measures, we mean that they are calculated from the same data set that are used to create the clusterings.

Many approaches have been put forward for testing H_0 and estimating the number of clusters in a data set. Jain & Dubes [35] provided a general overview of such methods and Milligan [52] and Milligan & Cooper [53] conducted an extensive Monte Carlo evaluation of 30 internal measures. However, the majorities of existing methods do not attempt to formally test H_0, but rather look for the clustering structure under which a summary statistic of interest is optimal, being large or small depending on the statistic. The following 6 internal measures are commonly used to estimate the number of clusters in a data set.

3.1. CH Index

For each number of clusters $k \geq 2$, Calinski & Harabasz [17] define the index:

$$CH(k) = \frac{tr(\mathbf{B}_k)/(k-1)}{tr(\mathbf{W}_k)/(n-k)}, \tag{7}$$

where $tr(\cdot)$ denotes the trace of a matrix, that is, the sum of the diagonal entries. The value of k, which maximizes $CH(k)$, is regarded as specifying the number of clusters. Note that $CH(1)$ is not defined and hence cannot be used for testing one cluster versus more than one. Even if it were modified by replacing $k-1$ with k, its value at 1 would be zero. Since $CH(k) > 0$ for $k \geq 2$, the maximum would never occur at $k = 1$.

3.2. KL Index

For each number of clusters $k \geq 2$, Krzanowski & Lai [41] define the index:

$$KL(k) = \left| \frac{DIFF(k)}{DIFF(k+1)} \right|, \tag{8}$$

where

$$DIFF(k) = (k-1)^{2/m} tr(\mathbf{W}_{k-1}) - k^{2/m} tr(\mathbf{W}_k). \tag{9}$$

The value of k, which maximizes $KL(k)$, is regarded as specifying the number of clusters. Note that $KL(k)$ is not defined for $k = 1$.

3.3. H Index

For each number of clusters $k \geq 1$, Hartigan [29] defines the index:

$$H(k) = (n-k-1)\left(\frac{tr(\mathbf{W}_k)}{tr(\mathbf{W}_{k+1})} - 1 \right). \tag{10}$$

The idea is to start with $k = 1$ and to add a cluster as long as $H(k)$ is sufficiently large. One can use an approximate F-distribution cut-off, instead Hartigan suggested that a cluster be added if $H(k) > 10$. Hence, the smallest value of $k \geq 1$, such that $H(k) \leq 10$, is regarded as specifying the number of clusters. This estimate is defined for $k = 1$ and can potentially discriminate between one versus more than one cluster.

3.4. Silhouette Statistic

For object i, let $a(i)$ be the average dissimilarity between object i and all other objects in the cluster to which object i belongs. For any other cluster C, let $d(i,C)$ denote the average dissimilarity of object i to all objects of C and let $b(i)$ denote the smallest of these $d(i,C)$. Then the silhouette statistic [38,60] of object i is defined by

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \tag{11}$$

And the overall average silhouette statistic is simply the average of $s(i)$ over all objects, namely,

$$\bar{s} = \frac{1}{n} \sum_{i=1}^{n} s(i). \tag{12}$$

Intuitively, objects with large silhouette statistic are well clustered, whereas those with small silhouette statistic tend to lie between clusters. Kaufman & Rousseeuw [38] proposed to choose the value of k, which maximizes the $\bar{s}$, as specifying the number of clusters. Note that $s(i)$ is not defined for $k = 1$.

3.5. Gap and GapPC Statistic

The Gap statistic [73] investigates the relationship between the $\log(tr(\mathbf{W}_k))$ for different values of k and the expectation of $\log(tr(\mathbf{W}_k))$ for a suitable null reference distribution, which is defined:

$$Gap(k) = \mathbb{E}[\log(tr(\mathbf{W}_k))] - \log(tr(\mathbf{W}_k)). \tag{13}$$

Here $\mathbb{E}$ denotes the expectation under the null distribution. To estimate the expectation of $\log(tr(\mathbf{W}_k))$, generate B reference data sets under the null distribution and apply the clustering algorithm to each, calculating the within-cluster sums of squares $tr(\mathbf{W}_k^1), tr(\mathbf{W}_k^2), \cdots, tr(\mathbf{W}_k^B)$. Thus, compute the estimated Gap statistic

$$\widehat{Gap}(k) = \frac{1}{B} \sum_{b=1}^{B} \log(tr(\mathbf{W}_k^b)) - \log(tr(\mathbf{W}_k)). \tag{14}$$

Let $sd(k)$ denote the standard deviation of $\log(tr(\mathbf{W}_k^1)), \log(tr(\mathbf{W}_k^2)), \cdots, \log(tr(\mathbf{W}_k^B))$ and define

$$s(k) = sd(k)\sqrt{1+1/B}. \quad (15)$$

The smallest value of k, such that $\widehat{Gap}(k) \geq \widehat{Gap}(k+1) - s(k+1)$, is regarded as specifying the number of clusters.

Tibshirani et al. [73] chose the uniform distribution as null distribution and considered two approaches for constructing the region of support of the distribution. In the first approach, the support for j-th explanatory variable, $1 \leq j \leq m$, is the range of the observed valued for that variable. In the second approach, the variables are sampled from a uniform distribution in a box aligned with the principal components of the centered designed matrix. Specifically, suppose that the columns of $\mathbf{X}$ have mean 0 and compute the singular value decomposition $\mathbf{X} = \mathbf{UDV}^t$. Tibshirani et al. [73] transformed via $\mathbf{X}' = \mathbf{XV}$ and then drew uniform features $\mathbf{Z}'$ over the ranges of the column of $\mathbf{X}'$, as in the first approach. Finally to back-transform via $\mathbf{Z} = \mathbf{Z}'\mathbf{V}^t$ to give reference data set $\mathbf{Z}$. Whereas the first approach has the advantage of simplicity, and the second one takes into account the shape of the data distribution, and makes the procedure rotationally invariant, as long as the clustering method itself is invariant.

Note that in both approaches, the variables are sampled independently. The version of the gap method that uses the original explanatory variables to construct the region of support is referred to as Gap statistic and the second version as GapPC statistic, where "PC" stands for principle components [20].

3.6. Distortion Index

Motivated by ideas from *rate distortion theory* [18], Sugar & James [71] define a quantity that measures the average distance, per dimension, between each object and its closest cluster centroid or medoid, named as distortion. Formally, for each number of clusters $k \geq 1$, the distortion is defined as follows.

$$\hat{d}_k = \frac{1}{m}\sum_{r=1}^{k}\frac{1}{|C_r|}\sum_{\mathbf{x}_i \in C_r}(\mathbf{x}_i - \bar{\mathbf{x}}_r)^t \Gamma^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_r), \quad (16)$$

which is simply the average Mahalanobis distance, per dimension, between each object and its closest cluster centroid or medoid. Note that in the case where Γ is the identity matrix, distortion is simply mean squared error.

Sugar & James [71] show, both theoretically and empirically, that for a large class of distributions the distortion curve, when transformed to an appropriate negative power p (a typical value is $p = m/2$), will exhibit a sharp jump at the "true" number of clusters. Thus, the value of k, which maximizes $J_k = \hat{d}_k^{-p} - \hat{d}_{k-1}^{-p}$, is regarded as specifying the number of clusters. Note that $d_0^{-p} = 0$. Therefore, the distortion index can detect the absence of clustering, i.e., $\hat{1}$.

3.7. General Remarks

The preceding measures all depends on a very strong null hypothesis. However, the unimodality hypothesis typically gives a high probability of rejection of H_0 if the data are sampled from a distribution with a lower kurtosis than the normal distribution, such as the uniform distribution [61]. Measures based on the uniformity hypothesis tend to be conservative, that is, lead to few rejections of H_0, when the data are sampled from a strongly unimodal distribution such as the normal distribution. In two or more dimensions, and depending on the test statistic, the results can be very sensitive to the region of support of the reference distribution [61].

These measures can only make comparisons between clusterings generated using the same model/metric. Furthermore, they often make assumptions about cluster structure. For example, if the particular data set that is being studied consists of several clouds of data point, with each cloud spherically distributed about its center, measures that assume such structure will work well. Otherwise, the same measure will possibly mislead. On the other hand, since these measures are calculated from the same observations that are used to create the clustering. Consequently, the distributions of these measures are intractable. In particular, as clustering methods attempt to maximize the separation between clusters, the ordinary significance tests such as analysis of variance F-tests are not valid for testing differences between the clusters. Although many internal measures have been proposed, none of them is completely satisfactory [20].

3.8. Axiomatic View

The authors of the respective literatures had different motivations for looking for a "good" measure. What's more, now new internal measures still emerge continuously. But what does a "good" internal measure look like? Stated differently, what requirements or axioms should a "good" internal measure meet? One usually considers that a good internal measure should reflect our intuitions, e.g., scale invariance (see further). However it is not easy to formalize the intuitions and design a new measure that meets these intuitions. Fisher & Van Ness [21] was one of the earliest attempts to axiomatize what is a "good" clustering, though it does not explicitly axiomatize a measure of clustering validity. Ackerman & Ben-David [1] proposed 4 axioms, and the set of these axioms is a consistent set of axioms.

3.8.1. Axioms

As described above, internal measures are typically functions of the within-clusters, and possibly between-clusters, sums of squares. Therefore, a distance

function d over S is implicitly defined. In fact, an internal measure *index* is a function that is given a clustering $\mathscr{C} \in \mathscr{P}(S)$ over (S,d) and returns a non-negative real number, namely $index : \mathscr{P}(S) \times S \times \mathscr{D}(S,S) \to \mathbb{R}_0^+$, where $\mathscr{D}(S,S)$ is the set of all interested distance functions over S, and $\mathbb{R}_0^+$ is the set of all real number greater or equal to 0.

Definition 1(Scale Invariance [1]). *Given a distance function $d \in \mathscr{D}(S,S)$ over S and a positive number λ, an internal measure index satisfies* scale invariance *if for every clustering $\mathscr{C} \in \mathscr{P}(S)$ of S, $index(\mathscr{C},S,d) = index(\mathscr{C},S,\lambda d)$.*

This is simply the requirements that the internal measure should not be sensitive to changes in the units of distance measurement; that is to say, it should not have a build-in "length scale" [1,39].

Let $\mathscr{C} \in \mathscr{P}(S)$ be a clustering over S, and d and $d' \in \mathscr{D}(S,S)$ be two distance functions over S, we say d' is a *$\mathscr{C}$-consistent variant* of d, if $d'(a,b) \le d(a,b)$ for all $a \sim_{\mathscr{C}} b$, and $d'(a,b) \ge d(a,b)$ for all $a \not\sim_{\mathscr{C}} b$.

Definition 2(Consistency [1]). *Given two distance functions $d,d' \in \mathscr{D}(S,S)$ over S, an internal measure index satisfies* consistency *if for every clustering $\mathscr{C} \in \mathscr{P}(S)$ of S, whenever d' is a $\mathscr{C}$-consistent variant of d, then $index(\mathscr{C},S,d') \ge index(\mathscr{C},S,d)$.*

Intuitively, consistent changes to d should not hurt the quality of a given clustering. In other words, the clusterings arise from the distance functions d and d' should be same [39]. Though this intuition is captured, it allows the possibility that some clusterings will improve more than others as a result of such change [1].

Definition 3(Richness [1]). *An internal measure index satisfies* richness *if for each non-trivial clustering $\mathscr{C} \in \mathscr{P}(S)$ over S, there exists a distance function $d \in \mathscr{D}(S,S)$ over S such that $\mathscr{C} = \arg\max_{\mathscr{C} \in \mathscr{P}(S)} \{index(\mathscr{C},S,d)\}$.*

Another way to say this is that the output of the clustering function should be "rich"—every clustering in $\mathscr{P}(S)$ is a possible output. In other words, suppose we are only given the objects in S but not the distance between them. Richness requires that for any desired clustering $\mathscr{C}$, it should be possible to construct a distance function d, such that the value of *index* for $\mathscr{C}$ is maximum over $\mathscr{P}(S)$.

Let $\mathscr{C},\mathscr{C}' \in \mathscr{P}(S)$ be two clusterings over S, and $d \in \mathscr{D}(S,S)$ be a distance function over S, we say $\mathscr{C}$ and $\mathscr{C}'$ are *isomorphic*, denoted $\mathscr{C} \cong_d \mathscr{C}'$, if there exists a distance-preserving isomorphism $\varphi : S \to S$, such that for all $a,b \in S, a \sim_{\mathscr{C}} b$ if and only if $\varphi(a) \sim_{\mathscr{C}} \varphi(b)$.

Definition 4(Isomorphism Invariance [1]). *Given a distance function $d \in \mathscr{D}(S,S)$ over S, an internal measure index satisfies* isomorphism invariant *if for all clusterings $\mathscr{C},\mathscr{C}' \in \mathscr{P}(S)$ over S where $\mathscr{C} \cong_d \mathscr{C}'$, $index(\mathscr{C},S,d) = index(\mathscr{C}',S,d)$.*

This is the requirement that clustering should be indifferent to individual identity of clustered objects, that is, *permutation invariance* [56].

3.8.2. Two Novel Internal Measures

Two novel internal measures, *weakest link* and *additive margin*, are proposed by Ben-David & Ackerman [1], which reflect the underlying intuition of center-based and linkage-based clustering, respectively. What's more, both of them satisfy the four axioms as described above, and given a data clustering, can be calculated in polynomial time. Analyzing which internal measures above meet these axioms is the subject of our next work.

(a)Weakest Link Measure

In linkage-based clustering, whenever a pair of objects shares the same cluster they are connected via a tight chain of points in that cluster. The weakest link measure focuses on the longest link in such a chain. Particularly, Let $\mathscr{C} = \{C_1,C_2,\cdots,C_k\} \in \mathscr{P}(S)$ be a clustering over S, and $d \in \mathscr{D}(S,S)$ be a distance function over S, Ben-David & Ackerman [1] define the measure:

$$WL(\mathscr{C}) = \frac{\max_{a \sim_{\mathscr{C}} b} \mathscr{C}\text{-}WL(a,b)}{\min_{a \sim_{\mathscr{C}} b} d(a,b)}, \tag{17}$$

where

$$\mathscr{C}\text{-}WL(a,b) = \min_{i=1}^{k} \max_{x,y \in C_i} \{d(a,x),d(x,y),d(y,b)\}. \tag{18}$$

Note that the range of values of weakest link measure is $(0,\infty)$.

(b)Additive Margin Measure

Let $\mathscr{C} = \{C_1,C_2,\cdots,C_k\} \in \mathscr{P}(S)$ be a clustering over S, $d \in \mathscr{D}(S,S)$ be a distance function over S and $\mathscr{K} \subseteq S$ be a representative set of $\mathscr{C}$. Of course, $|\mathscr{K}| = k$ and for all i, $\mathscr{K} \cap C_i \neq \emptyset$. Ben-David & Ackerman [1] define the measure:

$$AM(\mathscr{C}) = \min \frac{\frac{1}{n}\sum_{x \in S}(d(x,c_x) - d(x,c'_x))}{\frac{1}{\sum_{i=1}^{k}|C_i|(|C_i|-1)}\sum_{a \sim_{\mathscr{C}} b} d(a,b)} \tag{19}$$

where $c_x, c'_x \in \mathscr{K}$ are the closest and second centers to x, respectively.

4. External Measures

The term cluster validation usually refers to the ability of a given clustering approach to recover the true clustering structure in a data set. Bock [13] and Hartigan [31] attempted to assess validity on theoretical ground. However, these methods turn out to be of little applicability in real-life tasks, especially in the context of high-dimensional complex data sets [20]. In many

validation studies, the performance of a clustering approach is evaluated on some data sets with *a priori* known clustering structure. In order to assess the ability of a clustering approach to recover true clustering structure of S, it is necessary to define a measure of agreement between $\mathscr{C} = \{C_1, C_2, \cdots, C_k\} \in \mathscr{P}(S)$ and $\mathscr{C}' = \{C'_1, C'_2, \cdots, C'_l\} \in \mathscr{P}(S)$ of S, where the former is the *a priori* known clustering structure of S, and the latter comes from some clustering approach. In the clustering literature, such measures are referred to as external measures.

Though a common ground of these measures is that they can be calculated from the contingency matrix, they base on different ideas: pair-counting based measures, set-matching based measures and information theoretic based measures. The division also reflects the chronological development of the measures [77]: pair-counting based measures date from the 1970s and 1980s, set-matching based measures from the 1990s and information theoretic based measures have been developed in the 2002/2003.

4.1. Pair-Counting based Measures

A very intuitional approach to comparing clusterings is counting the number of unordered pairs of objects that are (or are not) placed into the same cluster according to $\mathscr{C}$ and $\mathscr{C}$'. Consequently, a 2×2 agreement/disagreement table [16] is formed, as shown in Fig. 2, where $m_{1,1}$ is the number of unordered pairs that are placed in the same cluster according to both $\mathscr{C}$ and $\mathscr{C}'$, $m_{1,0}$ ($m_{0,1}$) is the number of unordered pairs that are placed in the same cluster according to $\mathscr{C}$ ($\mathscr{C}'$) but not according to $\mathscr{C}'$ ($\mathscr{C}$), and finally $m_{0,0}$ is the number of unordered pairs that are not in the same cluster according to either of $\mathscr{C}$ and $\mathscr{C}'$. Types ($m_{1,1}$) and ($m_{0,0}$) are typically interpreted as agreements in the classification of the objects from a pair; types ($m_{1,0}$) and ($m_{0,1}$) represent disagreements [32]. Note that for simplification, $m_{a,b}, a, b \in \{0,1\}$ is also referred to the type that the corresponding unordered pairs belong to.

Since each unordered pair of objects must fall into one of these four types, we have

$$m = m_{1,1} + m_{1,0} + m_{0,1} + m_{0,0} = \binom{n}{2}, \tag{20}$$

where

$$m_{1,1} = \sum_{i=1}^{k}\sum_{j=1}^{l}\binom{n_{i,j}}{2}, \tag{21}$$

$$m_{1,0} = \sum_{i=1}^{k}\binom{|C_i|}{2} - \sum_{i=1}^{k}\sum_{j=1}^{l}\binom{n_{i,j}}{2}, \tag{22}$$

$$m_{0,1} = \sum_{j=1}^{l}\binom{|C'_j|}{2} - \sum_{i=1}^{k}\sum_{j=1}^{l}\binom{n_{i,j}}{2}, \tag{23}$$

$$m_{0,0} = m + \sum_{i=1}^{k}\sum_{j=1}^{l}\binom{n_{i,j}}{2} - \sum_{i=1}^{k}\binom{|C_i|}{2} - \sum_{j=1}^{l}\binom{|C'_j|}{2}. \tag{24}$$

Note that $\binom{a}{2}$ is defined as 0 when $a = 0$ or 1.

4.1.1. Rand Statistic, Mirkin Metric and Hurbert Statistic I

Intuitively, two clusterings that are similar produce relatively large values of $m_{1,1} + m_{0,0}$ and small values for $m_{1,0} + m_{0,1}$. Thus, depending on how $m_{1,1} + m_{0,0}$ and $m_{1,0} + m_{0,1}$ are normalized, different measures are possible, e.g., Rand statistic (denoted as R) [67,58], Mirkin metric (denoted as M) [54,4], and Hurbert statistic I (denoted as $H1$) [33] as follows:

$$R(\mathscr{C}, \mathscr{C}') = \frac{m_{1,1} + m_{0,0}}{m}, \tag{25}$$

$$M(\mathscr{C}, \mathscr{C}') = \frac{m_{1,0} + m_{0,1}}{m}, \tag{26}$$

$$H1(\mathscr{C}, \mathscr{C}') = \frac{m_{1,1} + m_{0,0} - m_{1,0} - m_{0,1}}{m}. \tag{27}$$

All three of these measures have straightforward probabilistic interpretations with respect to picking a pair of objects at random. For example, $R(\mathscr{C}, \mathscr{C}')$ is the probability of an agreement, $M(\mathscr{C}, \mathscr{C}')$ is the probability of a disagreement, and $H1(\mathscr{C}, \mathscr{C}')$ is the difference between the probability of an agreement and a disagreement.

In addition, Mirkin metric [54,4] is also known as *Equivalence Mismatch Distance*, which corresponds to the normalized Hamming distance for binary vectors if the set of all pairs of elements is enumerated and a clustering is represented by a binary vector defined on this enumeration [77,10]. An advantage is the fact that this distance is a metric in $\mathscr{P}(S)$ [77]. As a matter of fact, Mirkin metric is a variation of Rand statistic, since it can be rewritten as $M(\mathscr{C}, \mathscr{C}') = 1 - R(\mathscr{C}, \mathscr{C}')$.

4.1.2. Wallace Index I and II, Fowlkes-Mallows Index

For comparing hierarchical clusterings, Fowlkes & Mallows [23] proposed their index, denoted as FM. However, it can also be used for flat clusterings since it consists in calculating an index for each level of the hierarchies in consideration, which can be easily generalized to a measure for clusterings with different numbers of clusters.

Wallace [78] in commenting on Fowlkes & Mallows' paper [23] suggested two other measures of clustering validation, denoted as $W1$ and $W2$, respectively. In essence, the symmetric measure, Fowlkes-Mallows index, is the simple geometric mean of the two non-symmetric Wallace indices. The definitions of these three measures are as follows:

$$W1(\mathscr{C}, \mathscr{C}') = \frac{m_{1,1}}{m_{1,1} + m_{1,0}}, \tag{28}$$

$$W2(\mathscr{C}, \mathscr{C}') = \frac{m_{1,1}}{m_{1,1} + m_{0,1}}, \tag{29}$$

$$FM(\mathscr{C}, \mathscr{C}') = \frac{m_{1,1}}{\sqrt{(m_{1,1} + m_{1,0})(m_{1,1} + m_{0,1})}}. \tag{30}$$

		Clustering $\mathscr{C}'$		
Clustering $\mathscr{C}$	#unordered pairs	same cluster	different cluster	$\sum$
	same cluster	$m_{1,1}$	$m_{1,0}$	$m_{1,1}+m_{1,0}$
	different cluster	$m_{0,1}$	$m_{0,0}$	$m_{0,1}+m_{0,0}$
	$\sum$	$m_{1,1}+m_{0,1}$	$m_{1,0}+m_{0,0}$	m

Figure 2: 2×2 Agreement/Disagreement Table.

In the context of information retrieval, all three of measures have definite interpretations [77]. For instance, $W1(\mathscr{C},\mathscr{C}')$ can be interpreted as ratio of the number of retrieved relevant documents to the total number of relevant documents, i.e., recall; $W2(\mathscr{C},\mathscr{C}')$ can be interpreted as ratio of the number of retrieved relevant documents to the total number of retrieved documents, i.e., precision; $FM(\mathscr{C},\mathscr{C}')$ can be interpreted as the geometric mean of precision and recall.

4.1.3. Jaccard Index

The Jaccard index [35], also known as *Jaccard similarity coefficient*, is very similar to the Rand index, however it disregards the unordered pairs of objects that are in different clusters for both $\mathscr{C}$ and $\mathscr{C}'$. It is defined as follows:

$$J(\mathscr{C},\mathscr{C}')=\frac{m_{1,1}}{m_{1,1}+m_{1,0}+m_{0,1}}. \tag{31}$$

4.1.4. Hubert Statistic II

In order to define Hubert statistic II [32], one need to define two $n\times n$ partitioned binary matrices, **P** and **Q**, based on the cluster of the n objects in S according to $\mathscr{C}$ and $\mathscr{C}'$, respectively. See Fig. 3 for the partitioned binary matrix **P**, which is assumed to have zeros along its main diagonal with the indicated ones and zeros defining all the entries in the corresponding sub-matrices. The rows and columns of **P** are partitioned according to the row sums of the original contingency table (Fig. 1), whose (i,j)-th entry is denoted $P(i,j)$. **Q** can be defined similarly.

Without loss of generality, we can assume that the objects in S that are indexed by $|C_{i-1}|+1,|C_{i-1}|+2,\cdots,|C_i|$ belong to $C_i(1\le i\le k)$. Note that $|C_0|=0$. Obviously, the n row/column objects of **Q** are the same as the row/column objects of **P** but reordered to be consistent with the partition represented by **Q**. Therefore, let $\pi_0(\cdot)$ denote the permutation on the first n positive integers, such that if $\pi_0(r)=t$, then the r-th row (and column) in **P**, which corresponds to object o_r, is actually the t-th row (and column) in **Q**.

Hurbert statistic II evaluates the similarity between the two clustering $\mathscr{C}$ and $\mathscr{C}'$, based on predicting one matrix from the other. For example, suppose we predict $Q(\pi_0(r),\pi_0(s))$ from $P(r,s)$ using least-squares $(r,s=1,2,\cdots,n)$. The regression coefficient obtained, b_1, can be written as

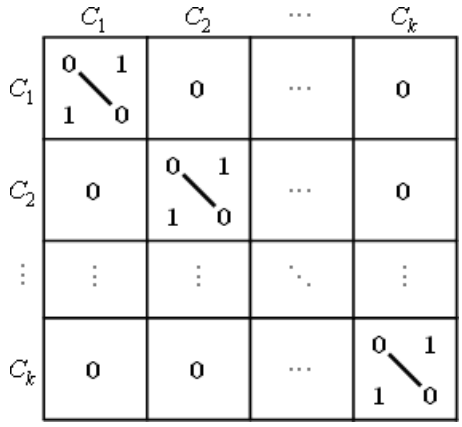

Figure 3: Partitioned Binary Matrix **P**.

$$b_1=\frac{\binom{n}{2}\sum_{i=1}^{k}\sum_{j=1}^{l}\binom{n_{i,j}}{2}-\sum_{i=1}^{k}\binom{|C_i|}{2}\sum_{j=1}^{l}\binom{|C'_j|}{2}}{\sum_{i=1}^{k}\binom{|C_i|}{2}\left[\binom{n}{2}-\sum_{i=1}^{k}\binom{|C_i|}{2}\right]}. \tag{32}$$

Likewise, we also can predict $P(r,s)$ from $Q(\pi_0(r),\pi_0(s))$ using least-squares $(r,s=1,2,\cdots,n)$. The regression coefficient obtained, b_2, has almost the same form with b_1:

$$b_2=\frac{\binom{n}{2}\sum_{i=1}^{k}\sum_{j=1}^{l}\binom{n_{i,j}}{2}-\sum_{i=1}^{k}\binom{|C_i|}{2}\sum_{j=1}^{l}\binom{|C'_j|}{2}}{\sum_{j=1}^{l}\binom{|C'_j|}{2}\left[\binom{n}{2}-\sum_{j=1}^{l}\binom{|C'_j|}{2}\right]}. \tag{33}$$

To obtain a symmetric measure, Hubert [33] took the geometric mean of the two regression coefficients, namely

$$H2(\mathscr{C},\mathscr{C}')=\sqrt{b_1b_2}. \tag{34}$$

4.1.5. Minkowski Score

The Minkowski score [36] calculates the agreement between a reference clustering $\mathscr{C}$ and a clustering result $\mathscr{C}'$, based on their cophenetic matrices, $\mathbf{M}^{\mathscr{C}}$ and $\mathbf{M}^{\mathscr{C}'}$. A cophenetic matrix of $\mathscr{C}$ is a binary matrix with

$M^{\mathscr{C}}_{i,j} = 1 (i, j = 1, 2, \cdots, n)$ if and only if object i and object j are in the same cluster in $\mathscr{C}$ ($i \neq j$). Similarly, one can construct the corresponding cophenetic matrix $\mathbf{M}^{\mathscr{C}'}$ for $\mathscr{C}'$. The Minkowski score is defined as

$$MS(\mathscr{C}, \mathscr{C}') = \frac{||\mathbf{M}^{\mathscr{C}} - \mathbf{M}^{\mathscr{C}}||}{||\mathbf{M}^{\mathscr{C}}||} = \sqrt{\frac{m_{1,0} + m_{0,1}}{m_{1,1} + m_{1,0}}}. \quad (35)$$

Note that it is limited to the interval $[0, +\infty)$

4.1.6. General Remarks

For different reasons, these measures do not seem to be very appealing. Many of them are sensitive to cluster sizes and number of clusters, which are undesirable for a similarity measure. For example, the Rand statistic has been shown to be highly dependent upon the number of clusters [55]. Fowlkes & Mallows [23] showed that in the (unrealistic) case of independent clusterings, the Rand statistic converges to one as the number of clusters increases. As another example, the Mirkin metric is also very sensitive to cluster sizes, such that two clusterings, for which each cluster in one clustering contains the same amount of elements of each of the clusters of the other clustering, are closer to each other than two clusterings for which one is refinement of the other [74].

Other measures, like the Fowlkes-Mallows index, make use of a very strong null hypothesis, that is, independence of the clusterings, fixed number of clusters, and fixed cluster sizes. When comparing clustering results provided by clustering methods, these assumptions (apart from the number of clusters that is fixed for some methods) do not usually hold. None of the algorithms works with fixed cluster sizes. Furthermore, independence of the clusterings is against our intuition when comparing clusterings, since the aim of our comparison is that we suppose a certain relationship between them and we want to know how strong it is [78].

4.2. Set-Matching based Measures

This kind of measures tries to match clusters that have a maximum absolute or relative overlap, which is also a quite intuitional approach. The following 5 measures are popular in the literatures.

4.2.1. F-measure

The F-measure has its origin in the field of document clustering [46, 24, 68]. Each cluster of $\mathscr{C}$ is a (predefined) class of documents and each cluster of $\mathscr{C}'$ is treated as the result of a query. The F-measure for C'_j with respect to C_i, $F_{i,j}$, indicates how "good" C'_j describes C_i, which is calculated with the harmonic mean of precision $p_{i,j} = n_{i,j}/|C'_j|$ and recall, $r_{i,j} = n_{i,j}/|C_i|$, for C'_j and C_i, namely

$$F_{i,j} = \frac{2 r_{i,j} p_{i,j}}{r_{i,j} + p_{i,j}} = \frac{2 n_{i,j}}{|C_i| + |C'_j|}. \quad (36)$$

The overall F-measure is then defined as the weighted sum of the maximum F-measures for the clusters in $\mathscr{C}'$:

$$F(\mathscr{C}, \mathscr{C}') = \frac{1}{n} \sum_{i=1}^{k} |C_i| \max_{j=1}^{l} \{F_{i,j}\}. \quad (37)$$

As we know, the range of the F-measure is $(0, 1]$. Wu et al. [80] proposed a procedure to find a tight lower bound for the F-measure, denoted F_{-}. The readers are invited to consult [80] for details.

4.2.2. Meilă-Heckerman Criterion

Meilă-Heckerman criterion [51], also known as *maximum match measure*, can be calculated as follows: look for the largest entry $n_{a,b}$ of the contingency table and match the corresponding clusters C_a in $\mathscr{C}$ and C'_b in $\mathscr{C}'$, which is the cluster pair with the largest (absolute) overlap. Denote by $match(a)$ the index of the cluster C'_b that matches cluster C_a. Afterward delete the a-th row and the b-th column and repeat this step until the matrix has size 0. Finally, sum up the matches and divide it by the total number of objects:

$$MH(\mathscr{C}, \mathscr{C}') = \frac{1}{n} \sum_{i=1}^{\min\{k,l\}} n_{i,match(i)}. \quad (38)$$

This measure is symmetric and takes value 1 for $\mathscr{C} = \mathscr{C}'$. Note that in the case of $k \neq l$, this measure completely disregards the $|k - l|$ "remaining" clusters in the clustering with the higher cardinality.

4.2.3. Goodman-Kruskal Coefficient

The Goodman-Kruskal coefficient [27, 37] takes a classification view on clustering. Specially, the following classification rule is adopted: (a) In the absence of knowledge about $X_{\mathscr{C}}$, the object $a \in S$ will be classified in the cluster $\arg\max_j P(X_{\mathscr{C}'} = C'_j) = \arg\max_j |C'_j|/n$ in $\mathscr{C}'$; (b) Otherwise, if one has known in advance that the object $a \in S$ belongs to the cluster C_i in $\mathscr{C}$, a will be classified in the cluster $\arg\max_j P(X_{\mathscr{C}'} = C_j | X_{\mathscr{C}} = C_i) = \arg\max_j n_{i,j}/|C_i|$ in $\mathscr{C}'$. The probability of misclassification committed by applying this rule is $1 - \max_j P(X_{\mathscr{C}'} = C_j | X_{\mathscr{C}} = C_i)$. The Goodman-Kruskal coefficient is the expected value of this error:

$$\begin{aligned} & GK(\mathscr{C}, \mathscr{C}') \\ &= \sum_{i=1}^{k} P(X_{\mathscr{C}} = C_i)(1 - \max_{j=1}^{l}(X_{\mathscr{C}'} = C'_j | X_{\mathscr{C}} = C_i)) \\ &= 1 - \frac{1}{n} \sum_{i=1}^{k} \max_{j=1}^{l} n_{i,j}. \end{aligned} \quad (39)$$

$GK(\mathscr{C},\mathscr{C}')$ has an upper bound of $1-(1/n)\times\max_j|C'_j|$. And $GK(\mathscr{C},\mathscr{C}')=0$ if and only if $\mathscr{C}$ is a refinement of $\mathscr{C}'$. Furthermore, $GK(\cdot,\cdot)$ is monotonic increasing in its first argument and monotonic decreasing in its second [37]. In other words, if $\mathscr{C},\mathscr{C}',\mathscr{C}''\in\mathscr{P}(S)$ such that $\mathscr{C}$ is a refinement of $\mathscr{C}'$, then $GK(\mathscr{C},\mathscr{C}'')\leq GK(\mathscr{C}',\mathscr{C}'')$, and if $\mathscr{C},\mathscr{C}',\mathscr{C}''\in\mathscr{P}(S)$ such that $\mathscr{C}'$ is a refinement of $\mathscr{C}''$, then $GK(\mathscr{C},\mathscr{C}')\geq GK(\mathscr{C},\mathscr{C}'')$. In addition, the purity and micro-averaged precision (MAP) in [80] are equivalent to this measure.

4.2.4. van Dongen Criterion

van Dongen criterion [74] is a symmetric measure, which is also based on maximum intersections of clusters. It is defined as follows:

$$VD(\mathscr{C},\mathscr{C}')=2n-\sum_{i=1}^{k}\max_{j=1}^{l}n_{i,j}-\sum_{j=1}^{l}\max_{i=1}^{k}n_{i,j}. \tag{40}$$

It can seen as a symmetric version of $GK(\cdot,\cdot)$, since $VD(\mathscr{C},\mathscr{C}')=n\times(GK(\mathscr{C},\mathscr{C}')+GK(\mathscr{C}',\mathscr{C}))$ [80]. Therefore, $VD(\mathscr{C},\mathscr{C}')\leq 2n-\max_i|C_i|-\max_j|C'_j|$ [80]. Additionally, this measure has a nice property that it is a metric in $\mathscr{P}(S)$ [37]. However, it ignores the parts of the clusters outside the intersections.

4.2.5. Classification Error Metric

Similar to Goodman-Kruskal coefficient, the classification error metric [11,50] takes a classification view on clustering, too. Nevertheless, it tries to map each cluster in one clustering with the lower cardinality to a different cluster in the other clustering in order to minimize the total misclassification rate. In specific, let σ be an injective mapping of $\{1,2,\cdots,\min\{k,l\}\}$ into $\{1,2,\cdots,\max\{k,l\}\}$. Thus, each σ can be seen as a (partial) correspondence between the cluster labels in $\mathscr{C}$ and $\mathscr{C}'$, so one can calculate the "classification error" of one clustering with the lower cardinality with respect to the other clustering. The classification error metric, denoted as ε, is defined as the minimum possible "classification error" under all correspondences:

$$\varepsilon(\mathscr{C},\mathscr{C}')=1-\frac{1}{n}\begin{cases}\max_\sigma\sum_{i=1}^{k}n_{i,\sigma(i)}, & k\leq l\\ \max_\sigma\sum_{j=1}^{l}n_{\sigma(j),j}, & \text{otherwise}\end{cases} \tag{41}$$

$\varepsilon(\mathscr{C},\mathscr{C}')$ has an upper bound of $1-1/\max\{k,l\}$ [80]. Though the number of all correspondences are order $\min\{k,l\}!\times\binom{\max\{k,l\}}{\min\{k,l\}}$, the maximum can be calculated in polynomial time as the solution of a linear program identical to the maximum bipartite matching algorithm in graph theory [25].

4.2.6. General Remarks

It is very easy to see that set-matching based measures have the common property of just taking the overlaps into account and completely disregarding the unmatched parts of the clusters or even complete clusters. Meilă [49] presented a nice example that pointed out the negative effect of this "behavior" of a measure: suppose $\mathscr{C}\in\mathscr{P}(S)$ is a clustering with k equal size clusters. $\mathscr{C}'$ is obtained from $\mathscr{C}$ by shifting a fraction α of the objects in each cluster C_i to the "next" cluster $C_{(i+1)\bmod k}$. The clustering $\mathscr{C}''$ is obtained from $\mathscr{C}$ by reassigning a fraction α of the elements in each cluster C_i evenly between the other clusters. If $\alpha<0.5$, then $F(\mathscr{C},\mathscr{C}')=F(\mathscr{C},\mathscr{C}'')$, $MH(\mathscr{C},\mathscr{C}')=MH(C,C'')$, $VD(\mathscr{C},\mathscr{C}')=VD(\mathscr{C},\mathscr{C}'')$, $GK(\mathscr{C},\mathscr{C}')=GK(\mathscr{C},\mathscr{C}'')$, and $\varepsilon(\mathscr{C},\mathscr{C}')=\varepsilon(\mathscr{C},\mathscr{C}'')$. This contradicts our intuition that $\mathscr{C}'$ is a less disrupted version of $\mathscr{C}$ than $\mathscr{C}''$, which is therefore not desirable.

Another drawback is the asymmetry of some of the measures, such as F-measure, Goodman-Kruskal coefficient. These may be appropriate indices for comparing a clustering with an optimal clustering solution. However, in general the optimal solution is not known, which makes an asymmetric measure hard to interpret.

4.3. Information Theoretic based Measures

Here we first review some of the very fundamental concepts of information theory. For more details we refer the readers to [18]. As stated in the section Introduction, for any clustering $\mathscr{C}\in\mathscr{P}(S)$, one can define a discrete random variable $X_{\mathscr{C}}$, the *entropy* of which is defined as

$$\begin{aligned}H(\mathscr{C})&=-\sum_{i=1}^{k}P(X_{\mathscr{C}}=C_i)\log P(X_{\mathscr{C}}=C_i)\\&=-\sum_{i=1}^{k}\frac{|C_i|}{n}\log\frac{|C_i|}{n},\end{aligned} \tag{42}$$

where log bases 2. We can understand it as follows [49]: assuming that each object of S has the same probability of being picked and choosing an object of S at random, the probability that this object is in cluster $C_i\in\mathscr{C}$ is $P(X_{\mathscr{C}}=C_i)=|C_i|/n$. The uncertainty in this context is equal to the entropy of random variable $X_{\mathscr{C}}$. Usually, $H(\mathscr{C})$ is called the *entropy associated with clustering* $\mathscr{C}$. $H(\mathscr{C})$ is always non-negative, which takes value 0 only when $\mathscr{C}$ is a trivial clustering.

4.3.1. Entropy

To calculate this measure [68,81], for each cluster $C'_j\in\mathscr{C}'$, the conditional probability $p_{i|j}=P(X_{\mathscr{C}}=C_i|X_{\mathscr{C}'}=C'_j)=n_{i,j}/|C_j|$ is first computed, and then the entropy of cluster C'_j using the standard entropy, $E_j=-\sum_i p_{i|j}\log(p_{i|j})$. The

total entropy, denoted as E, is computed as the weighted sum of the entropies of each cluster in $\mathscr{C}'$, namely

$$E(\mathscr{C},\mathscr{C}') = \sum_{j=1}^{l} P(X_{\mathscr{C}'} = C'_i)E_j = -\sum_{j=1}^{l} \frac{|C'_j|}{n}\left(\sum_{i=1}^{k} \frac{n_{i,j}}{|C'_j|}\log\frac{n_{i,j}}{|C'_j|}\right). \quad (43)$$

In fact, this measure is nothing but the *conditional entropy* [18] of $\mathscr{C}$ on $\mathscr{C}'$, $H(\mathscr{C}|\mathscr{C}')$, which implies that if the objects in each large cluster of $\mathscr{C}'$ are mostly from the same cluster in $\mathscr{C}$, this measure tends to be small [80]. Furthermore, this measure is always non-negative, less than or equal to $\log k$ [80], but asymmetric.

4.3.2. Mutual Information

The *mutual information* between two clusterings, denoted as MI, is the information that one clustering has about the other, which is equal to the mutual information between the associated random variables [70,69], namely

$$MI(\mathscr{C},\mathscr{C}') = MI(X_{\mathscr{C}},X_{\mathscr{C}'}) = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{n_{i,j}}{n}\log\frac{n_{i,j}}{|C_i||C'_j|} = H(\mathscr{C})+H(\mathscr{C}')-H(\mathscr{C},\mathscr{C}') \quad (44)$$

where $H(\mathscr{C},\mathscr{C}')$ is the *joint entropy* [18] of the two clusterings. Intuitively, $MI(\mathscr{C},\mathscr{C}')$ can be interpreted as follows [49]: Given an object in S, the uncertainly about its cluster in $\mathscr{C}$ is measured by $H(\mathscr{C})$. Assume that it is known that which cluster the object belongs to in $\mathscr{C}'$. This knowledge often reduces the uncertainty about its cluster in $\mathscr{C}$. This reduction in uncertainty, averaged over all objects in S, is equal to $MI(\mathscr{C},\mathscr{C}')$.

$MI(\mathscr{C},\mathscr{C}')$ is always non-negative and symmetric, and never exceed the total uncertainty in a clustering, so $MI(\mathscr{C},\mathscr{C}') \leq \min\{H(\mathscr{C}),H(\mathscr{C}')\}$. Equality in this formula occurs when one clustering is a refinement of the other. Another way to say this is that if $\mathscr{C}'$ is a refinement of $\mathscr{C}$, then $MI(\mathscr{C},\mathscr{C}') = H(\mathscr{C}) < H(\mathscr{C}')$. And $MI(\mathscr{C},\mathscr{C}') = H(\mathscr{C}) = H(\mathscr{C}')$ if and only if $\mathscr{C} = \mathscr{C}'$.

By simple transformation, $MI(\mathscr{C},\mathscr{C}') = H(\mathscr{C}) - (H(\mathscr{C}) - MI(\mathscr{C},\mathscr{C}')) = H(\mathscr{C}) - H(\mathscr{C}|\mathscr{C}')$, one can easily find that $MI(\mathscr{C},\mathscr{C}')$ is equivalent to $E(\mathscr{C},\mathscr{C}')$ for any given data set S if $\mathscr{C}$ is the *a priori* known clustering structure of S, since $H(\mathscr{C})$ is a constant in this case [80].

4.3.3. Variation of Information

By analogy with the total variation of a function, variation of information [49,50,75] between two clusterings $\mathscr{C},\mathscr{C}' \in \mathscr{P}(S)$ is defined as

$$VI(\mathscr{C},\mathscr{C}') = H(\mathscr{C})+H(\mathscr{C}')-2MI(\mathscr{C},\mathscr{C}') = [H(\mathscr{C})-MI(\mathscr{C},\mathscr{C}')]+[H(\mathscr{C}')-MI(\mathscr{C},\mathscr{C}')] \quad (45)$$

Informally, when going from clustering $\mathscr{C}$ to clustering $\mathscr{C}'$, the first term in the above formula measures the amount of information about $\mathscr{C}$ that we loose, which corresponds to the conditional entropy $H(\mathscr{C}|\mathscr{C}')$, while the second term measures the amount of information about $\mathscr{C}'$ that we have to gain, which corresponds to the conditional entropy $H(\mathscr{C}'|\mathscr{C})$ [49]. This implies that the variation of information is a symmetry version of the entropy measure [80]. Additionally, by Equation (44), $VI(\mathscr{C},\mathscr{C}')$ can be re-expressed as

$$VI(\mathscr{C},\mathscr{C}') = 2H(\mathscr{C},\mathscr{C}')-H(\mathscr{C})-H(\mathscr{C}'). \quad (46)$$

Meilă [49] analyzed in detail the variation of information between two clusterings, and summarized many properties. Here, we briefly review several main ones:

(a)$VI(\mathscr{C},\mathscr{C}')$ is a metric in $\mathscr{P}(S)$.
(b)$VI(\mathscr{C},\mathscr{C}') \leq \min\{\log n, 2\log\max\{k,l\}\}$. This means that for large enough n, clusterings of different data sets, with different numbers of elements, but with bounded numbers of clusters are on the same scale in the metric VI. This allows us to compare, add or subtract VI metric across different clustering space independently of the underlying data set.
(c)The product of two clusterings $\mathscr{C},\mathscr{C}' \in \mathscr{P}(S)$ is *collinear* with these two clusterings, namely

$$VI(\mathscr{C},\mathscr{C}') = VI(\mathscr{C},\mathscr{C}\times\mathscr{C}')+VI(\mathscr{C}\times\mathscr{C}',\mathscr{C}'). \quad (47)$$

This also implies $VI(\mathscr{C},\mathscr{C}') \geq VI(\mathscr{C},\mathscr{C}\times\mathscr{C}')$ with equality only if $\mathscr{C}' = \mathscr{C}\times\mathscr{C}'$. Thus, the nearest neighbor of $\mathscr{C}$ is either a refinement of $\mathscr{C}$ or a clustering whose refinement is $\mathscr{C}$. In essence, the nearest neighbor of a clustering is obtained by splitting one element off the smallest cluster (or by the corresponding merging process). This means that small changes in a clustering result in small VI metrics.
(d)$VI(\mathscr{C},\mathscr{C}') \geq 2/n$. Thus, with increasing n, the space of clusterings gets a finer granularity.
(e)$VI(\mathscr{C},\mathscr{C}')$ can be calculated in $\mathscr{O}(n+k\times l)$: $\mathscr{O}(n)$ for computing the confusion matrix and $\mathscr{O}(k\times l)$ for computing $VI(\mathscr{C},\mathscr{C}')$ from the matrix.

4.3.4. General Remarks

At present, though there is no consensus on which is the best measure, information theoretic based measures have received increasing attention for their solid theoretical background. Another reason that these measures seem to be quite promising is that they do not suffer from the drawbacks that we can find for measures that are based on

counting pairs or on matching set. However, they possibly suffer from other disadvantages that we do not know yet [77].

4.4. Correction for Chance

On close examination, one can find that preceding external measures either do not have a fixed bound, or do not have a constant baseline value, i.e., average value between random clusterings of a data set. Since a measure is meant to provide a comparison mechanism, it is generally preferable that it lies within a predetermined range and has a constant baseline value, so as to facilitate comparison and enhance intuitiveness [75]. Otherwise, if they have a considerable inherent bias attributable solely to chance, it may potentially reduce their usefulness in a number of common situations. Therefore, it is necessary to correct these measures for chance, also known as normalization in the literature.

Generally speaking, normalizing techniques tend to fall into two kinds: one (Type-I in short) is based on a statistical view, which formulates a baseline distribution to correct the measure for randomness; the other (Type-II in short) uses the minimum and maximum values to normalize the measure into the $[0,1]$ range. Fig. 4 illustrates the normalization scheme for various external measures, where $\max(index)$, $\min(index)$ is the maximum, minimum value of the measure *index*, and $\mathbb{E}(index)$ is the expected value of *index* based on the baseline distribution. In this study, we consider 17 external measures in total, as shown in Fig. 5. By positive/negative measures, denoted as $+/-$, respectively, we mean that a higher value indicates a better/worse clustering performance.

As for the baseline distribution, Hurbert & Arabie [32] proposed to use the exact generalized hypergeometric distribution [75,43] as the baseline distribution in which the row and column sums are fixed, but the clusterings are randomly selected. Morey & Agresti [55] suggested an asymptotic form based on the multi-nominal distribution. These lead to the following expected values as follows, respectively.

$$\mathbb{E}\left(\sum_{i=1}^{k}\sum_{j=1}^{l} n_{i,j}^2\right) = \frac{1}{n(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{l}|C_i|^2|C'_j|^2 + \frac{n^2}{n-1} - \frac{1}{n-1}\left(\sum_{i=1}^{k}|C_i|^2 + \sum_{j=1}^{l}|C'_j|^2\right) \tag{48}$$

$$\mathbb{E}\left(\sum_{i=1}^{k}\sum_{j=1}^{l} n_{i,j}^2\right) \approx \frac{1}{n^2}\sum_{i=1}^{k}\sum_{j=1}^{l}|C_i|^2|C'_j|^2. \tag{49}$$

The difference between expectations (48) and (49) pointed out by Hubert & Arabie [32], can be apparent only when the data size n is small; otherwise they are slight. In this study, the expectation (48) is adopted.

Based on (48), one can easily calculate the expectation for $R, M, H1, W1, W2, FM, J', H2$ and MS', since they are linear functions of $\sum_{i,j} n_{i,j}^2$ under the hypergeometric distribution assumption. Just as many authors [80,3,79] observed, after correction for chance, many of these measures become equivalent, e.g., $R_{\text{norm}} = M_{\text{norm}} = H1_{\text{norm}} = J'_{\text{norm}} = MS'_{\text{norm}}$, and $H2_{\text{norm}} = H2 = \sqrt{W1_{\text{norm}} \times W2_{\text{norm}}}$ [32]. It is worth mentioning that it is not trivial to calculate the expectations for MI and VI. Nevertheless, Vinh et al. [75, 76] derived an analytical formula for the expected value of MI and VI under the hypergeometric distribution as follows.

$$\mathbb{E}(MI) = \sum_{i=1}^{k}\sum_{j=1}^{l}\sum_{n_{i,j}=Low}^{High}\frac{n_{i,j}}{n}\log\left(\frac{nn_{i,j}}{|C_i||C'_j|}\right) \frac{|C_i|!|C'_j|!(n-|C_i|)!(n-|C'_j|)!}{n!n_{i,j}!(|C_i|-n_{i,j})!(|C'_j|-n_{i,j})!(n-|C_i|-|C'_j|+n_{i,j})!} \tag{50}$$

where $Low = \max\{0, |C_i| + |C'_j| - n\}$ and $High = \min\{|C_i|, |C'_j|\}$.

$$\mathbb{E}(VI) = H(\mathscr{C}) + H(\mathscr{C}') - 2\mathbb{E}(MI) \tag{51}$$

Note that there exist some criticisms [50] for artificiality of the randomness model in Type-I normalization technique. Since the "amount" of similarity of two clusterings corresponds to the deviation from the expected value under the null hypothesis of independent clusterings with fixed cluster sizes. Again, the strong assumptions on the distribution make the result hard to interpret.

4.5. Axiomatic View

As like internal measures, one usually consider what axioms a "good" external measure should satisfy in order to better understand their properties, their limitations, and the implied assumptions underlying them. Meilă [50] proposed the following 6 axioms, and derived an impossibility result for external measures: no measure in the space of clusterings $\mathscr{P}(S)$ can simultaneously satisfy three desirable properties (see further), each of which makes the measure intuitive in some sense.

Definition 5(Symmetry [80,50]). *An external measure index satisfies* symmetry *if for any two clusterings* $\mathscr{C}, \mathscr{C}' \in \mathscr{P}(S)$, $index(\mathscr{C}, \mathscr{C}') = index(\mathscr{C}', \mathscr{C})$.

In order words, transposing two clustering in the confusion matrix should not bring any difference to the measure value [80]. Obviously, this axiom is not true for the F-measure, the Goodman-Kruskal criterion, the classification error metric and the entropy.

	Positive	Negative
Type-I	I: $\frac{index-\mathbb{E}(index)}{\max(index)-\mathbb{E}(index)}$	II: $\frac{\mathbb{E}(index)-index}{\mathbb{E}(index)-\min(index)}$
Type-II	III: $\frac{index-\min(index)}{\max(index)-\min(index)}$	IV: $\frac{\max(index)-index}{\max(index)-\min(index)}$

Figure 4: The Normalization Scheme for Various External Measures.

Measure	Range	Pos./Neg.	Normalization
R	$(0,1]$	$+$	I [32]
M	$[0,1)$	$-$	II [32]
$H1$	$(-1,1]$	$+$	I [80]
$W1$	$[0,1]$	$+$	I [3]
$W2$	$[0,1]$	$+$	I [3]
FM	$[0,1]$	$+$	I [3]
J'	$[0,1]$	$-$	II [80]
$H2$	$[0,1]$	$+$	I [32,80]
MS'	$[0,\infty)$	$-$	II [80]
F	$[F_-,1]$	$+$	III [80]
MH	$(0,1]$	$+$	III
GK	$[0,1-(1/n)\times\max_j\lvert C'_j\rvert]$	$-$	IV
VD	$[0,2n-\max_i\lvert C_i\rvert-\max_j\lvert C'_j\rvert]$	$-$	IV [80]
ε	$[0,1-1/\max\{k,l\}]$	$-$	IV [80]
E	$(0,\log k]$	$-$	IV
MI	$0,\min\{H(\mathscr{C}),H(\mathscr{C}')\}$	$+$	I [75], III
VI	$[2/n,\min\{\log n,2\log\max\{k,l\}\}]$	$-$	II [75], IV [80]

Figure 5: Summary on External Measures.

Note: Let $J'=(1-J)/(1+J)$ and $MS'=MS^2$ since J and MS are not linear functions of $\sum_{i,j}n_{i,j}^2$, which implies that it is very complex to calculate the expectation for them. But it is easy to see that J' and MS' are equivalent to J and MS, respectively [80].

Definition 6(n-Invariance [80,50]). *Let $\mathscr{C},\mathscr{C}',\mathscr{C}'',\mathscr{C}'''\in\mathscr{P}(S)$ and denote by $\mathbf{M}_1$ and $\mathbf{M}_2$ the confusion matrices of $\mathscr{C}$ and $\mathscr{C}'$, $\mathscr{C}''$ and $\mathscr{C}'''$, respectively. An external measure index satisfies* n-invariance *if for any $\lambda\in\mathbb{R}^+$, $index(\mathscr{C},\mathscr{C}')=index(\mathscr{C}'',\mathscr{C}''')$ whenever $\mathbf{M}_1=\lambda\mathbf{M}_2$.*

Intuitively, an external measure *index* should not directly depend on n, but depends only on the relative values $n_{i,j}/n$. However, some measures cannot fulfill this axiom, such as R_{norm}, FM_{norm} and $H2_{\text{norm}}$.

Definition 7(Additivity w.r.t. Refinement [50]). *An external measure index satisfies* additivity w.r.t. refinement *if for any clustering $\mathscr{C}\in\mathscr{P}(S)$, $index(\hat{0},\mathscr{C})+index(\mathscr{C},\hat{1})=index\ (\hat{0},\hat{1})$.*

Definition 8(Additivity w.r.t. Production [50]). *An external measure index satisfies* additivity w.r.t. production *if for any two clusterings $\mathscr{C},\mathscr{C}'\in\mathscr{P}(S)$, $index(\mathscr{C},\mathscr{C}')=index(\mathscr{C},\mathscr{C}\times\mathscr{C}')+index(\mathscr{C}',\mathscr{C}\times\mathscr{C}')$.*

Intuitively, the preceding two axioms describe the geometric properties of an external measure, i.e., that it is aligned with the lattice of clustering [50].

Definition 9(Convex Additivity [50]). *Let $\mathscr{C},\mathscr{C}'\in\mathscr{P}(S)$ such that $\mathscr{C}'$ be a refinement of $\mathscr{C}$. Denote by $\mathscr{C}'_i$ the partitioning induced by $\mathscr{C}'$ on $C_i(i=1,2,\cdots,k)$. An external measure index satisfies* convex additivity *if*

$$index(\mathscr{C},\mathscr{C}')=\sum_{i=1}^{k}\frac{|C_i|}{n}index(\hat{1}_{|C_i|},\mathscr{C}'_i),\qquad(52)$$

where $\hat{1}_{|C_i|}$ is the one-clustering of the data set C_i.

Definition 10(Non-decreasing [50]). *Denote by $\mathscr{C}_k^U$ the "uniform" clustering, i.e., the clustering with k equal clusters. An external measure index satisfies* non-decreasing *if $f(k)=index(\hat{1},\mathscr{C}_k^U)$ is a non-decreasing function of k whenever $\mathscr{C}_k^U$ exists.*

The preceding two axioms set the scale of an external measure index [50]. Particularly,*convex additivity* requires that *index* should show additivity along the lattice of clustering. Some un-normalized measures meet this axiom, such as the F-measure, van Dongen criterion, the variation of information and the classification error metric. However, none of the normalized measures above satisfies this axiom [80].

Furthermore, Meilă [50] shown any external measure satisfying Axiom 5 and Axiom 7-Axiom 10 is identical to the variation of information up to a multiplicative constant, which is closely matched to the lattice of clusterings. From this point, Meilă [50] obtained the following impossibility result:

There is no index symmetric, n-invariant, with index$(\hat{1}, \mathscr{C}_k^U)$ *non-decreasing, that satisfies simultaneously the following three properties: (a) index is aligned to the lattice of clusterings; (b) index is convexly additive; (c) index is bounded.*

5. Clustering Stability based Methods

Clustering stability based methods are a family of widely used *model selection* techniques. Their unifying theme is that an appropriate model should result in a clustering which is robust with respect to various kinds of perturbations. In other words, the clustering algorithm should be stable with respect to input randomization. In past few years, these methods are often utilized to choose a suitable number of clusters along with stability measures. The rational [9] behind is that when the number of clusters is too large, the algorithm has to "randomly" split some true cluster, and the choice of the cluster it splits might change with the randomness of the sample, in which case instability occurs. On the other hand, when the number of clusters is too small, we have to "randomly" merge several true clusters, the choice of which might similarly change with each particular random sample, resulting in instability again.

Generally speaking, clustering stability based methods can be divided into two categories. One is based on resampling, the basic idea of which clusters non-disjoint sub-sample of S in order to measure the similarity of the clustering solutions obtained for the intersection of both samples. Levine $ Domany's resampling approach [48] and the model explorer algorithm [10] fall into this category. The other is based on prediction, which is pioneered in an early work by Breckenridge [15]. The basic idea is to measure the agreement of clustering solutions generated by a clustering algorithm and by a classifier trained using a second (clustered) sub-sample of S. Though a specific implementable procedure for choosing the number of clusters did not proposed by Breckenridge, his study suggests the usefulness of such kind of approaches. The prediction strength method [72], clest [20], and Lange et al.'s method [45,44] build on the Breckenridge's ideas but generalize his work.

But Lange et al. [45] pointed out that the overlapping sub-samples in the first kind of methods may lead to an undesirable, artificially induced stability. Rakhlin & Caponnetto [57] gave a precise characterization of clustering stability with respect to both complete and partial changes of the data. Specially, for clustering algorithms that minimizes an objective function, such as the squared error in K-means clustering, in the case of a unique global minimizer, the clustering solution is stable with respect of complete changes of the data, while for the case of multiple minimizers, the change of $\Omega(\sqrt{n})$ samples defines the transition between stability and instability.

5.1. Levine & Domany's Resampling Approach

At first, Levine & Domany's resampling approach [48] randomly constructs r sub-samples $S_1, S_2, \cdots, S_r$ of size $\lceil fn \rceil$ ($f \in [0,1]$) from S. And then for S and all sub-samples $S_1, S_2, \cdots, S_r$ of S, clustering solutions are calculated. Finally, a stability measure LD is defined to assess the average similarity of the solutions obtained on the $S_1, S_2, \cdots, S_r$ with the one obtained on S.

In order to define the measure LD, the clustering solutions need to be represented as the *cluster connectivity matrix*, $\mathbf{M}$, whose (i,j)-th entry is denoted $M(i,j)$. Specifically, the matrix $\mathbf{M}$ for S is a binary square matrix of size $n \times n$, where $M(i,j) = 1$ if $i \neq j$ and the i-th and j-th objects are in the same cluster, and zero otherwise. Similarly, $\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_r$ of size $\lceil fn \rceil \times \lceil fn \rceil$ can also be defined for sub-samples $S_1, S_2, \cdots, S_r$. For each number of clusters $k \geq 1$, Levine & Domany [48] define LD as:

$$LD(k) = \frac{1}{k} \sum_{i=1}^{r} \frac{\sum_{a \in S_i} \sum_{b \in \mathscr{N}_{i,a}} \delta(M(a,b), M_i(a,b))}{\sum_{a' \in S_i} |\mathscr{N}_{i,a'}|}, \quad (53)$$

where $\mathscr{N}_{i,a}$ ($a \in S_i$) defines a neighborhood between objects in sub-sample S_i, and $\delta(x,y) = 1$ if $x = y$, and zero otherwise. The neighborhood definition, such as κ-mutual nearest neighbor neighborhood definition in [47], is left as a free parameter, which should be supplied externally by user.

It is easy to see that $LD(k)$ measures the extent to which the clustering calculated on the sub-samples is in agreement with the clustering on the full data set. Therefore, $LD(k) = 1$ for perfect agreement. Levine & Domany [48] suggest that the value of k, which maximizes $LD(k)$, should be regarded as specifying the number of clusters. But when several maxima can occur, it is not clear how to choose a single number of clusters [45].

5.2. Model Explorer Algorithm

Firstly, the model explorer algorithm [10] randomly constructs two sub-samples of size $\lceil fn \rceil$ ($f \in (0.5,1)$) from S. Then, the similarity between the solutions for these sub-samples is calculated at the intersection of the sub-samples. The similarity measure can be set as some external measure, such as the Fowlkes-Mallows index above, which is a free parameter. This procedure is repeated r times.

To estimate the number of clusters, the experimental section of [10] suggests choosing the value where there is a transition from a similarity distribution that is concentrated near 1 to wider distribution. This can be quantified by a "jump" in

$$P(s_k \geq \eta) \approx \frac{1}{r}\sum_{i=1}^{r}\delta(s(i,k) > \eta), \tag{54}$$

where s_k is a random variable that denotes the similarity between clusterings into k clusters, $\delta(\cdot)$ is a Dirac delta function, $s(i,k)$ denotes the empirically measured similarity between clusterings for the two sub-samples into k clusters in the i-th loop, and η is a pre-set constant.

Since looking for a "jump" in the cumulative distribution is qualitative in nature, it is not a well-defined criterion for determining the number of clusters. It is possibly very difficult to choose an appropriate value in some situation [45].

5.3. Prediction Strength Method

The main idea of prediction strength [72] method is to: (a) split randomly the whole data set S into two non-empty disjoint subsets: a training set S_{tr} and a testing set S_{te}. In real-world applications, *r-fold cross-validation* is utilized. That is, S is randomly divided into r subsets of nearly equal size. The first $r-1$ subsets represent S_{tr}, and the last one is S_{te}. (b) the clustering solutions are calculated for S_{tr} and S_{te}, respectively; (c) A nearest class centroid classifier is built using S_{tr} and the clustering solution on S_{tr}; (d) The resulting classifier is employed to predict the clusters of objects in S_{te}. Specially, for each pair of objects in S_{te} that are assigned to the same test cluster, to determine whether they are also assigned to the same cluster based on the classifier. This procedure is repeated multiple times.

In order to assess the similarity of clustering solutions by the classifier and the clustering algorithm on S_{te}, often known as the *predicted labels* and the *clustered labels* respectively, Tibshirani et al. [72] defined a similarity index, named as prediction strength. This index essentially measures the intersection of the two clusters in both solutions that match worst. The largest value of k is regarded as specifying the number of clusters, such that the average similarity score is above a pre-set threshold.

However, this procedure has two severe disadvantages pointed out by Lange et al. [45]: (1) it is reasonably applicable to squared-error clustering algorithm only due to the use of the nearest class centroid classifier; (2) The prediction strength measure can trivially drop to zero for the larger number of clusters. In particular, the latter point severely limits its applicability in practice.

5.4. Clest

The first step in the Clest procedure [20] is similar to (a)-(d) in the prediction strength method. The differences are that the sizes of S_{tr} and S_{te}, the classifier and the similarity measure between two clusterings are all free parameters. Given a fixed k, suppose that the first step is repeated B times, thus B similarity scores $s_{k,1}, s_{k,2}, \cdots, s_{k,B}$ can be obtained. Let $t_k = \text{median}(s_{k,1}, s_{k,2}, \cdots, s_{k,B})$ denote the observed similarity statistic for the clustering of S into k clusters. And then B_0 data sets are drawn from a suitable null reference distribution. Similar to S, one can obtain a similarity score for each data set, denoted $s^0_{k,b}, b = 1,2,\cdots,B_0$, respectively. Let t^0_k be the average of these B_0 similarity scores, namely $t^0_k = (1/B_0)\sum_b t_{k,b}$.

In order to find an appropriate number of clusters in S, let d_k denote the difference between the observed similarity statistic and its estimated expected value under the null hypothesis of one-clustering, namely $d_k = t_k - t^0_k$. And let p_k denote the p-value for t_k, that is, $p_k = (1/B_0)\{b|t_{k,b} \geq t_k\}$. Define the set K^- as

$$K^- = \{2 \leq k \leq M | p_k \leq p_{\max}, d_k \geq d_{\min}\}, \tag{55}$$

where M is some pre-defined upper bound for the number of clusters, $p_{\max}$ and $d_{\min}$ are pre-set thresholds. If K^- is non-empty, the value of k in K^- , which maximizes d_k, is regarded as specifying the number of clusters in S. Otherwise, the number of clusters in S is one. In fact, the set K^- is determined fully by $p_{\max}$ and $d_{\min}$, which can be chosen badly so that K^- is always empty, for example [45].

One can easily see that there are a large number of free parameters in the Clest procedure, which have to be set by the user. But little guidance in [20] is given on how to reasonably select the values for these parameters in real-world applications. Lange et al. [45] pointed out that this lack of parameter specification poses a severe practical problem since the obtained statistics are of little value for poor parameter selection. For example, very unbalanced splitting schemes can lead to unreliable results, since the group structure might no longer be visible for a clustering algorithm if there are too few objects in one of the two subsets [45]. Another example is that an inappropriate classifier may result in decreasing largely the similarity scores. Therefore, Lange et al. [45] consider that the Clest is only a conceptual framework, not a fully specified algorithm.

5.5. Lange et al.'s Method

The first step in Lange et al.'s method [45,44] is still similar to (a)-(d) in the prediction strength method. But there are three differences between these two methods as follows:

(a)S is split into two disjoint subsets, S_{tr} and S_{te}, of approximately equal size.

(b)Intuitively, a good classifier should mimic the clustering algorithm. Based on this point, further guidance is given on how to reasonably choose the classifier in practice. Specially, for clustering algorithms that minimizes an objective function, the classifier, which uses the least-cost increase criterion, can mimic their grouping strategies. The same strategy is applicable for agglomerative algorithms. For K-means clustering, the nearest centroid classifier becomes the classifier of choice up to (negligible) $\mathscr{O}(1/n)$ corrections. For single linkage, this strategy leads to the nearest-neighbor classifier. Of course, there exist some algorithms that cannot be easily understood as mimimizers of a cost function, e.g., CLICK [66]. For these cases, the K-nearest-neighbor classifier can be safety chosen, since it is asymptotically Bayes' optimal [19], at least for metric data.

(c)In order to quantitatively compare the two solutions, the predicted labels and the clustered labels, Lange et al. [44,45] proposed a novel dissimilarity index on the basis of their normalized Hamming distance. Without loss of generality, we can assume that the objects in S_{te} are indexed by $1,2,\cdots,|S_{te}|$. Since either solution can be formally represented by a vector of labels, let the predicted labels and the clustered labels be, respectively

$$\mathbf{Y}^p=(y_1^p,y_2^p,\cdots,y_{|S_{te}|}^p)^t,\mathbf{Y}^c=(y_1^c,y_2^c,\cdots,y_{|S_{te}|}^c)^t, \tag{56}$$

where $y_i^p,y_i^c\in\{1,2,\cdots,k\}$ and $y_i^p/y_i^c=v$ if o_i is predicted/clustered to cluster v. For each number of clusters $k\geq 2$, Lange et al. [44,45] define the dissimilarity index:

$$LRBB_k(\mathbf{Y}^p,\mathbf{Y}^c)=\min_{\pi\in\Pi(k)}\frac{1}{|S_{te}|}\sum_{i=1}^{|S_{te}|}\delta(\pi(y_i^p)\neq y_i^c), \tag{57}$$

where $\Pi(k)$ is the set of all permutations of the elements in $\{1,2,\cdots,k\}$ and $\delta(\cdot)$ is a Dirac delta function. Though the number of all permutations is $k!$, the minimization can be performed in time $\mathscr{O}(|S_{te}|+k^3)$ by using the Hungarian method [42] for minimum weighted bipartite matching, which is guaranteed to find the globally optimal $\pi\in\Pi(k)$, where $\mathscr{O}(|S_{te}|)$ is required for setting up a weight matrix and $\mathscr{O}(k^3)$ for the matching itself.

In order to measure the stability of a clustering algorithm $\mathscr{A}_k$, Lange et al. [44,45] define a stability index as the average similarity between solutions, namely

$$S(\mathscr{A}_k)=\mathbb{E}_{S_{tr},S_{te}}(LRBB_k(\mathbf{Y}^p,\mathbf{Y}^c)), \tag{58}$$

where the expectation is taken with regard to pairs of disjoint subsets, S_{tr} and S_{te}, of approximately equal size. To estimate the expectation, generate r pairs of S_{tr} and S_{te}, and apply the above procedure to each. Thus, an estimate $\hat{S}(\mathscr{A}_k)$ can obtain by averaging r values of $LRBB_k(\cdot,\cdot)$.

However, the range of possible stability values $S(\mathscr{A}_k)\in[0,1-1/k]$ depends on the number of clusters k, which implies that stability indices are not directly comparable for different values of k. To enable comparability, Lange et al. [44,45] normalize the empirical misclassification rate of the clustering algorithm $S(\mathscr{A}_k)$ with the asymptotic misclassification rate of random labeling $S(\mathscr{R}_k)$, where the random labeling algorithm $\mathscr{R}_k$ assigns an object to cluster v with probability $1/k$, namely

$$\bar{S}(\mathscr{A}_k)=S(\mathscr{A}_k)/S(\mathscr{R}_k). \tag{59}$$

Note that the stability measure is not defined for $k=1$. Similar to $\hat{S}(\mathscr{A}_k)$, an estimate $\hat{S}(\mathscr{R}_k)$ for $S(\mathscr{R}_k)$ can be obtained by sampling s random k-labelings and calculating the empirical average of the dissimilarities. Thus one can get an estimate $\hat{\bar{S}}((A)_k)$ for $\bar{S}(\mathscr{A}_k)$ by normalizing $\hat{A}(\mathscr{A}_k)$ with $\hat{S}(\mathscr{R}_k)$. Lange et al. [45] proposed to choose the value of k, which minimizes $\hat{\bar{S}}(\mathscr{A}_k)$, as specifying the number of clusters.

5.6. Theoretical Understanding

Clustering stability based methods have been shown to be rather effective in practice, and gain more and more influence in applications. However, their theoretical foundations are not yet well understood so far. While it is reasonable to require that a clustering algorithm should demonstrate stability in general, it is not obvious whether the one, which is the most stable, also must have the best performance. Over the past few years, related theoretical study has been initiated in a framework, where the data are assumed to be drawn independently from some underlying distribution.

However, a fundamental hurdle is the following observations, made and rigorously analyzed in [9,7] and also pointed out in [40]. Under mild conditions, stability is asymptotically fully determined by the behavior of the objective function which the clustering algorithm attempts to optimize. In particular, the existence of a unique global/local optimum for some model choice implies stability as sample size tends to infinity and instability otherwise. Furthermore, this kind of instability is usually not related to the correct number of clusters, but it might depend on completely unrelated criteria, such as symmetries in the data. Therefore, for large enough samples one might get a stable solution regardless of the chosen model. As a result, it is quite possible that there exists some hard-to-compute sample size, beyond which clustering stability estimators 'break down' and become unreliable in detecting the most stable model.

A possible solution to this difficulty is proposed in [63,64], where the scaling constant in the definition of stability is chosen as $1/\sqrt{n}$ rather than $1/n$. The authors show that the important factor in the way clustering stability based methods work may not be the asymptotic stability of the model, but rather how fast exactly does it converge to this stability. With this more refined analysis, stability criteria might actually be able to discern the stability of different models, no matter how large is the sample, despite the universal convergence to absolute stability.

However, the work in [63,64] only concentrates on specific toy distribution or specific idealized clustering frameworks, which still do not give us general sufficient conditions for the reliability of clustering stability estimators in the large sample regime. Such a set of conditions is presented in [65] with making no such assumptions. The main condition is the existence of a *central limit theorem* for the clustering framework, in an appropriately defined sense. Additionally, non-trivial asymptotic behavior of these estimators is explicitly characterized for any framework satisfying these conditions. A similar characterization was given in [64] for the K-means framework.

Ben-David & von Luxburg [8] relate the stability of clustering algorithms (on finite sample sizes) to properties of the optimal data clustering itself. Specifically, the quantitative value of stability can be upper bounded by the mass in a small tube around the optimal clustering boundary, which has already been implicitly utilized in [63] only in a very simple one-dimensional setting. Unfortunately, the reverse statement is not true in general. That is, there can usually be clusterings whose decision boundary lies in a high density area, but we have high stability.

In fact, as stated in [8], even if one find satisfactory reasons which explain why a certain clustering tends to be more stable than another one, such statements are not very useful for drawing conclusions about stability measures of any given *finite* sample size. The reason [8] is that as opposed to the standard statistical learning theory (SLT) settings, it is impossible to give global convergence guarantees for stability. Thus, while one can use stability criteria in practice, it is impossible to give distribution-free performance guarantees on any of its results. No matter how large the sample size is, we can always find distributions where the stability evaluated on that particular sample size is misleading, in the sense that it is far from the "true stability" [8].

6. Current and Future Research Directions

In a nutshell, the internal measures compare clusterings based on the goodness of fit between each clustering and the data set, and often make assumptions about the distribution of cluster. Hence, they can only make comparisons between clusterings generated using the same model/metric. The external measures assess agreement between a clustering solution generated by a clustering algorithm and a pre-defined reference clustering. But since a pre-defined reference clustering is typically unavailable in real-world unsupervised tasks, they do not directly applicable in practice. Though new internal/external measures still emerge continuously, we think that a "good" measure should meet a set of axioms in [1]/[50] as possible as one can. That is, a set of axioms in [1]/[50] should be helpful in detecting and defining "good" measures.

While the popularity of clustering stability based methods has grown in the past few years, they have an inherent drawback: high computational cost of generating and assessing multiple clusterings of the data set, which prohibit them from being applied to large, high-dimensional data sets, such as text corpora. In our opinion, one of future research directions is to tackle their computational issues. As a first step, Greene & Cunningham [28] present an efficient prediction based cluster validation for kernel clustering algorithm by means of a prototype reduction strategy.

In addition, although the central limit approach in [65] proved to be a convenient framework, it remains an open question how far it is from being *necessary* for stability estimators not to 'break down' in the large sample regime. As we known, the reasons, why a certain clustering tends to be more stable than another one, are not very useful for drawing conclusions about stability measures of any given finite sample size [8]. Nevertheless, better understanding the meaning of the asymptotic value of clustering instability in [65] may help to understand the behavior of clustering stability on finite samples.

Acknowledgement

We thank the financial support from the Foundmental Research Funds for the Central Universities (ID: JGTD 2015-04), National Natural Science Foundation of China (ID: 71403255, 70903032), Social Science Foundation of Jiangsu Province (ID: 09TQC011), and MOE Project of Hummanities and Social Sciences (ID: 09YJC870014). Our gratitude also goes to the anonymous reviewers for their valuable comments.

References

[1] Margareta Ackerman and Shai Ben-David. Measures of clustering quality: A working set of axioms for clustering. In Daphne Koller, Yoshua Bengio, Dale Schuurmans, Léon Bottou, and Aron Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 121–128. MIT Press, Cambridge, MA, 2009.

[2] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[3] Ahmed N. Albatineh, Magdalena Niewiadomska-Bugaj, and Daniel Mihalko. On similarity indices and correction for chance agreement. *Journal of Classification*, 23(2):301–313, 2006.

[4] Phipps Arabie and Scott A. Boorman. Multidimensional scaling of measures of distance between partitions. *Journal of Mathematical Psychology*, 10(2):148–203, 1973.

[5] Sayed F. Bahght, Sultan Aljahdali, E. A. Zanaty, Ahmed S. Ghiduk, and Ashraf Afifi. A new validity index for fuzzy c-means for automatic medical image clustering. *International Journal of Computer Applications*, 38(12):1–8, 2012.

[6] Arindam Banerjee, Chase Krumpelman, Joydeep Ghosh, Sugato Basu, and Raymond J. Mooney. Model-based overlapping clustering. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 532–537, New York, NY, USA, 2005. ACM.

[7] Shai Ben-David, Dávid Pál, and Hans Ulrich Simon. Stability of *k*-means clustering. In *Proceedings of the 20th Annual Conference on Computational Learning Theory*, pages 20–34, San Diego, CA, USA, 2007.

[8] Shai Ben-David and Ulrike von Luxburg. Relating clustering stability to properties of cluster boundaries. In *Proceedings of the 21st Annual Conference on Computational Learning Theory*, pages 379–390, Helsinki, Finland, 2008.

[9] Shai Ben-David, Ulrike von Luxburg, and Dávid Pál. A sober look at clustering stability. In *Proceedings of the 19th Annual Conference on Computational Learning Theory*, pages 5–29, Pittsburgh, PA, USA, 2006.

[10] Asa Ben-Hur, Andre Elisseeff, and Isabelle Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, volume 7, pages 6–17, 2002.

[11] Asa Ben-Hur and Isabelle Guyon. *Functional Geomics: Methods and Protocols*, chapter Detecting Stable Clusters using Principal Component Analysis, pages 159–182. Humana Press, 2003.

[12] D. A. Binder. Bayesian cluster analysis. *Biometrika*, 65(1):31–38, 1977.

[13] H. H. Bock. On some significance tests in cluster analysis. *Journal of Classification*, 2(1):77–108, 1985.

[14] Gloria Bordogna and Gabriella Pasi. Soft clustering for information retrieval applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(2):138–146, 2011.

[15] James N. Breckenridge. Replicating cluster analysis: Method, consistency, and validity. *Multivariate Behavioral Research*, 24(2):147–161, 1989.

[16] Robert L. Brenna and Richard J. Light. Measuring agreement when two observers classify people into categories not defined in advance. *British Journal of Mathematical and Statistical Psychology*, 37:154–163, 1974.

[17] R. B. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.

[18] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[19] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2 edition, 2001.

[20] Sandrine Dudoit and Jane Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7), 2002.

[21] Lloyd Fisher and John W. van Ness. Admissible clustering procedures. *Biometrika*, 58(1):91–104, 1971.

[22] R. A. Fisher, A. Steven Corbet, and C. B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, 12(1):42–58, 1943.

[23] E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983.

[24] Benjamin Chin Ming Fung. Hierarchical document clustering using frequent itemsets. Master's thesis, Simon Fraser University, 2002.

[25] M. Golumbic. *Algorithmic Graph Theory and Perfect Graphs*. Academic Press, New York, 1980.

[26] I. J. Good and G. H. Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

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

[28] Derek Greene and Pádraig Cunningham. Efficient prediction-based validation for document clustering. In *Proceedings of the 17th European Conference on Machine Learning*, pages 663–670, 2006.

[29] J. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.

[30] J. A. Hartigan. Asymptotic distributions for clustering criteria. *The Annals of Statistics*, 6(1):117–131, 1978.

[31] J. A. Hartigan. Statistical theory in clusteirng. *Journal of Classification*, 2(1):63–76, 1985.

[32] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

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

[34] A. K. Jain, M. Murty, and R. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):265–323, 1999.

[35] Anikl K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, New Jersey, 1988.

[36] Nicholas Jardine and Robin Sibson. *Mathematical Taxonomy*. John Wiley & Sons, New York, 1971.

[37] Szymon Jaroszewicz, Dan A. Simovici, Winston P. Kuo, and Lucila Ohno-Machado. The goodman-kruskal coefficient and its applications in genetic diagnosis of cancer. *IEEE Transactions on Biomedical Engineering*, 51(7):1095–1102, 2004.

[38] L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990.

[39] Jon Kleinberg. An impossibility theorem for clustering. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 446–453. MIT Press, Cambridge, MA, 2003.

[40] Abba M. Krieger and Paul E. Green. A cautionary note on using internal cross validation to select the number of clusters. *Psychometrika*, 64(3):341–353, 1999.

[41] W. J. Krzanowski and Y. T. Lai. A criterion for determining the number of groups in a data set using sum of squares clustering. *Biometrics*, 44(1):23–34, 1988.
[42] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.
[43] Henry Oliver Lancaster. *The Chi-Squared Distribution*. John Wiley, New York, 1969.
[44] Tilman Lange, Mikio L. Braun, Volker Roth, and Joachim M. Buhmann. Stability-based model selection. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 617–624. MIT Press, Cambridge, MA, 2003.
[45] Tilman Lange, Volker Roth, Mikio L. Braun, and Joachim M. Buhmann. Stability-based validation of clustering solutions. *Neural Computation*, 16(6):1299–1323, 2004.
[46] Bjornar Larsen and Chinatsu Aone. Fast and effective text mining using linear-time document clustering. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 16–22, San Diego, California, United States, 1999.
[47] Erel Levine. Un-supervised estimation of cluster validity - methods and applications. Master's thesis, Weizmann Institute of Science, 1999.
[48] Erel Levine and Eytan Domany. Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13(11):2573–2593, 2001.
[49] Marina Meilă. Comparing clusterings by the variation of information. In *Proceedings of the 16th Annual Conference on Computational Learning Theory*, pages 173–187, Washionton, DC, USA, 2003.
[50] Marina Meilă. Comparing clusterings – an axiomatic view. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 577–584, Bonn, Germany, 2005.
[51] Marina Meilă and David Heckerman. An experimental comparison of model-based clustering methods. *Machine Learning*, 42(1-2):9–29, 2001.
[52] Glenn W. Milligan. A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46(2):187–199, 1981.
[53] Glenn W. Milligan and Martha C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.
[54] B. G. Mirkin and L. B. Chernyi. Measurement of the distance between distinct partitions of a finite set of objects. *Automation and Remote Control*, 31:786–792, 1970.
[55] Leslie C. Morey and Alan Agresti. The measurement of classification agreements: An adjustment to the rand statistic for chance agreement. *Educational and Psychological Measurement*, 44(1):33–37, 1984.
[56] Jan Puzicha, Thomas Hofmann, and Joachim M. Buhmann. A theory of proximity based clustering: Structure detection by optimization. *Pattern Recognition*, 33(4):617–634, 2000.
[57] Alexander Rakhlin and Andrea Caponnetto. Stability of *k*-means clustering. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1121–1128. MIT Press, Cambridge, MA, 2007.
[58] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.
[59] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.
[60] Peter Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–66, 1987.
[61] W. Sarle. Cubic clustering criterion. Technical report a-108, SAS Institue, Inc., 1983.
[62] Gideon E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.
[63] Ohad Shamir and Naftali Tishby. Cluster stability for finite samples. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1297–1304. MIT Press, Cambridge, MA, 2008.
[64] Ohad Shamir and Naftali Tishby. Model selection and stability in *k*-means clustering. In *Proceedings of the 21st Annual Conference on Computational Learning Theory*, pages 367–378, Helsinki, Finland, 2008.
[65] Ohad Shamir and Naftali Tishby. On the reliability of clustering stability in the large sample regime. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1465–147. MIT Press, Cambridge, MA, 2009.
[66] Roded Sharan and Ron Shamir. Click: A clustering algorithm with applications to gene expression analysis. In *Proceedings of the 8th International Conference on Intelligent System for Molecular Biology*, pages 307–316, Menlo Park, CA, USA, 2000.
[67] R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438, 1958.
[68] Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, Boston, MA, 2000.
[69] Alexander Strehl and Joydeep Ghosh. Cluster ensembles-a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3(Dec):583–617, 2002.
[70] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney. Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search*, pages 58–64. AAAI, 2000.
[71] Catherine A. Sugar and Gareth M. James. Finding the number of clusters in a data set: An information theoretic approach. *Journal of the American Statistical Association*, 98(463):750–763, 2003.
[72] Robert Tibshirani, Guenther Walther, David Botstein, and Patrick Brown. Cluster validation by prediction strength. Technical report, Department of Statistics, Stanford University, 2001.
[73] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society. Series B*, 63(2):411–423, 2001.
[74] Stijn van Dongen. Performance criteria for graph clustering and markov cluster experiments. Technical report ins-r0012, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, 2000.
[75] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the*

26th International Conference on Machine Learning, pages 1073–1080, Montreal, Quebec, Canada, 2009.

[76] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clustering comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

[77] Silke Wagner and Dorothea Wagner. Comparing clusterings - an overview. Technical report 2006-4, Faculty of Informatics, Universität Karlsruhe (TH), 2006.

[78] David L. Wallace. A method for comparing two hierarchical clusterings: Comment. *Journal of the American Statistical Association*, 78(383):569–576, 1983.

[79] Matthijs J. Warrens. On similarity coefficients for 2×2 tables and correction for chance. *Psychometrika*, 73(3):487–502, 2008.

[80] Junjie Wu, Hui Xiong, and Jian Chen. Adapting the right measures for k-means clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 877–886, Paris, France, 2009.

[81] Hui Xiong, Junjie Wu, and Jian Chen. K-means clustering versus validation measures: A data distribution perspective. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 779–784, Philadelphia, Pennsylvania, USA, 2006.

[82] Shuo Xu, Xiaodong Qiao, Lijun Zhu, and Huixia Zheng. Deep analysis on mining frequent & maximal reference sequences with generalized suffix tree. *Journal of Computational Information Systems*, 6(7):2187–2197, 2010.

Shuo Xu works as associate professor and the manager of Text Mining Lab., Information Technology Support Center, at Institute of Scientific and Technical Information of China (ISTIC), China. He obtained his M.S. and Ph.D. from China Agriculture University (CAU). His current research interest includes text mining (TM), machine learning (ML), natural language processing (NLP), science and technology monitoring, and knowledge organization system (KOS), etc.

Xiaodong Qiao works as professor and chief engineer of Institute of Scientific Technical Information of China (ISTIC), China. He obtained his M.S. from university of Sheffield, United Kingdom. His current research interest includes knowledge service technology, digital library, and information resource management, etc.

Lijun Zhu works as associate professor and deputy director of Information Technology Support Center, at Institute of Scientific and Technical Information of China (ISTIC), China. He obtained his M.S. and Ph.D. from China University of Petroleum and China Agriculture University (CAU), respectively. His current research interest includes semantic web, web service and knowledge organization system (KOS), science and technology information service based knowledge technology, etc.

Yunliang Zhang works as associate professor and the manager of Knowledge Engineering Lab., Information Technology Support Center, at Institute of Scientific and Technical Information of China (ISTIC), China. He obtained his B.S. and Ph.D. from Tsinghua University and Chinese Academy of Sciences, respectively. His current research interest includes knowledge organization, text classification, natural language processing (NLP), Hierarchical Network of Concepts (HNC) theory, and knowledge service, etc.

Chunxiang Xue works as associate professor of School of Economic & Management, at Nanjing University of Science and Technology(NUST), China. She obtained her Ph.D. from Nanjing Agricultural University(NAU), China. Her current research interest include information organization, knowledge organization systems, automatic processing of text, etc.

Lin Li works as associate professor of Computer Science and Technology Department, College of Information and Electrical Engineering at China Agriculture University (CAU). She obtained her Ph.D. from China Agriculture University (CAU). Her current research interest includes software engineering (SE), software automation (SA), bioinformatics, etc.