Empirical Comparison of External Evaluation Measures for Document Clustering by Using Synthetic Data

Kazuaki Kishida[1,a)]

Abstract: In order to develop effective and efficient techniques or algorithms for document clustering, it is indispensable to explore methodologies for evaluating experiments in which clustering validity has to be measured precisely. This paper reports on an experimental result of empirically comparing external evaluation measures for unsupervised clustering. For computing actual values of purity, inverse purity, F-measure, mutual information, V-measure, Jaccard coefficient, Rand statistic, adjusted Rand index, Fowlkes-Mallows coefficient, BCubed, and van Dongen criterion, a spherical k-means algorithm was repeatedly executed for synthetic data, which were randomly generated under some assumptions on term occurrences in documents. Based on the values obtained by the clustering operations, the experiment revealed 'closeness' between measures in scoring the appropriateness of clustering results. Also, changes of the values with decrement or increment of generated clusters were analyzed to examine the effectiveness of the measures in a situation in which the number of generated clusters was different from that of 'true' topic classes inherent in the document set.

Keywords: Document clustering, Evaluation measure, Generation of synthetic data

1. Introduction

Document clustering (DC), by which a topically heterogeneous set of documents (news articles, books, web pages, and so on) is partitioned into several homogeneous subsets in an unsupervised manner, plays an important role in text mining applications. Hence, many sophisticated algorithms or techniques for DC have been explored, and experiments for examining their effectiveness have been reported in the literature.

However, evaluation measures used for examining the validity of clustering results often differ between experiments (e.g., mutual information, F-measure, purity, Rand statistic, and so on), which makes it difficult to compare the results. In order to understand precisely the advantages or disadvantages of individual DC methods in such situations, it is necessary to have sufficient knowledge of the nature or characteristics of the evaluation measures in scoring the appropriateness of each clustering result.

This paper reports on an experiment that statistically compared the values of some external evaluation measures by using clustering results obtained from a spherical k-means algorithm. The comparison yields a deeper insight on the measures and will be useful for developing more effective DC methods.

The target document sets employed for the experiment were artificially generated from a model including several parameters for determining the random occurrence frequency of each term in a document, thus enabling the 'behavior' of the measures to be observed under various conditions of the target document sets by intentionally adjusting the parameters. For instance, it was possible to control the degree to which better clustering results were easily obtained for the synthetic data by the algorithm. Also, the numbers of documents belonging to topic classes in the artificial document sets can be changed in order to examine how values of the measures vary in situations where documents are distributed equally or unequally across topic classes. Although such synthetic data have often been exploited in research on evaluation measures for clustering, the datasets in the experiment were especially tailored to DC problems by a relatively complicated model.

The rest of this article is organized as follows. First, previous researches on evaluation methodologies for unsupervised clustering are reviewed. Next, some principal evaluation measures based on an external criterion provided by human experts are discussed. After an explanation on the model used for randomly generating artificial document sets, the results of an experiment to compare the evaluation measures are reported and discussed.

2. Related Work

The goodness or validity of clustering results can be evaluated according to internal or external criteria. This paper focuses on only measures for the external evaluation. Although 'indirect' external evaluation in DC situations may be feasible (e.g., if a distributed information retrieval system includes a DC module for dividing the target document set into topically homogeneous parts in its process, then the clustering validity can be indirectly inferred from the search performance of the system), the indirect method is not discussed here. For 'direct' external evaluation, 'ground truth' clusters obtained by a method (e.g., annotation by human experts) are usually employed.

Actually, many external measures (or metrics) for general-purpose unsupervised clustering have been proposed and ex-

[1] School of Library and Information Science, Keio University, Minato-ku, Tokyo 108–8345, Japan
[a)] kz_kishida@z8.keio.jp

plored (see Wu et al., 2009 [20] and Kremer et al., 2011 [11], who enumerated them exhaustively). Brun et al.(2007) [3] calculated Kendall's correlation between several external measures from clustering results obtained iteratively for synthetic data, which were generated based on six simple models. The external measures are Hubert's correlation, Rand statistic, Jaccard coefficient and Fowlkes & Mallows (FM) coefficient. As clustering algorithms, k-means, fuzzy c-means, self-organizing map, hierarchical agglomerative clustering (HAC) and probabilistic clustering based on an EM algorithm were employed. Also, Song & Zhang(2008) [18] compared other external evaluation measures (purity, cluster-based cross entropy, class-based cross entropy, V-measure, homogeneity, completeness and variation of information) by using simple synthetic data.

Wu et al.(2009) [20] intensively examined 13 evaluation measures (entropy, purity, F-measure, variation of information, mutual information, Rand statistic, Jaccard coefficient, FM coefficient, two kinds of Hubert's statistic, Minkowski score, classification error and van Dongen criterion) within the scope of assessing the validity of k-means clustering, which tends to produce clusters with relatively uniform sizes. Both synthetic and real datasets were used in the experiment, which suggested the suitability of the normalized van Dongen criterion.

On the other hand, it is possible to assess the appropriateness of the external measures based on 'formal' conditions. For instance, Meilă(2005) [14] discussed characteristics of external measures according to some mathematical properties such as symmetry, additivity, convex additivity and scale. Also, Rosenberg & Hirschberg(2007) [17] extended the list of desirable properties proposed by Dom(2001) [5], and examined whether several external measures satisfy them or not. More recently, Amigó et al.(2009) [1] focused on four formal constraints (which concern cluster homogeneity, cluster completeness, 'miscellaneous' clusters, and clusters size versus quantity), and reported that only BCubed passed the test of constraints.

There are some external measures not covered by the above empirical and analytical studies. For instance, the fuzzy set theory has been applied to the definition of external measures (e.g., see Campello, 2010 [4]). Also, Kremer et al.(2011) [11] proposed a new measure tailored to evaluation of data stream clustering, and Günnemann et al.(2011) [6] discussed some external measures ('RNIA' and 'E4SC') to assess results of subspace clustering. Further, Hassani et al.(2013) [8] designed a new measure for evaluating effectively subspace clustering of data stream. In Amigó et al.(2011), a novel combination of two measures, which can be used instead of F-measure, was developed (it was named 'unanimous improvement ratio (UIR)').

3. External Evaluation Measures for Unsupervised Clustering

According to Amigó et al.(2009) [1], external evaluation measures for unsupervised clustering can be divided into three categories: 1) set matching, 2) pair counting, and 3) entropy. Suppose that a collection of clusters, which is denoted by $C = \{C_1, C_2, \ldots, C_L\}$, was obtained by applying a clustering algorithm to a document set in which a topic label is assigned to each document by a human expert beforehand. The labels allow us to partition the set into topic classes $\mathcal{A} = \{A_1, A_2, \ldots, A_H\}$, which can be used as a correct answer to the clustering operation (i.e., 'ground truth' classes). In this paper, n_{mk}, n_m and n_k denote the number of documents belonging to both A_m and C_k, the number of documents in A_m and the number of documents in C_k, respectively ($m = 1, \ldots, H; k = 1, \ldots, L$).

3.1 Measures by set matching

By matching two sets C_k and A_m, precision and recall can be computed for each pair, which lead to two measures, called *purity* and *inverse purity*, respectively. More specifically, purity [21] is defined as a weighted average of the maximum values of precision in clusters, namely $pur = \sum_{k=1}^{L}(n_k/N)\max_{\{m=1,\ldots,H\}}(n_{mk}/n_k)$. In contrast, inverse purity is computed by $invp = \sum_{m=1}^{H}(n_m/N)\max_{\{k=1,\ldots,L\}}(n_{mk}/n_m)$ based on the recall ratio of each topic class [1].

For a clustering result, *F-measure* (or 'FScore') [12] can be similarly defined as $F = \sum_{m=1}^{H}(n_m/N)\max_{\{k=1,\ldots,L\}} f_h(A_m, C_k)$ where $f_h(A_m, C_k)$ denotes the harmonic mean of precision and recall for A_m and C_k. Actually, $f_h(A_m, C_k) = (2xy)/(x + y)$ where $x = n_{mk}/n_m$ and $y = n_{mk}/n_k$.

On the other hand, Amigó et al.(2009) [1] recommended employing *BCubed* measures, which are composed of precision and recall versions defined as $Bp = N^{-1}\sum_k\sum_m n_{mk}^2/n_k$ and $Br = N^{-1}\sum_m\sum_k n_{mk}^2/n_m$, respectively. The harmonic mean of them can be computed as $BCF = (2 \times Bp \times Br)/(Bp + Br)$, which is called 'BCubed-F' in this paper.

3.2 Measures by pair counting

Generally, relatedness between two collections of subsets (i.e., C and $\mathcal{A}$) can be measured by interpreting the collection as a set of edges in a graph when individual elements are linked to some other elements. For instance, suppose that C is a graph in which two documents are connected by an edge if and only if they belong to the same cluster. The number of edges (i.e., pairs) in C is computed as $|\Gamma_C| = \sum_{k=1}^{L} {}_{n_k}C_2 = \sum_k n_k(n_k - 1)/2 = (\sum_k n_k^2 - N)/2$ where Γ_C means a set of edges in C. Similarly, it becomes that $|\Gamma_{\mathcal{A}}| = (\sum_m n_m^2 - N)/2$ for $\mathcal{A}$. Because cardinality of the intersection of Γ_C and $\Gamma_{\mathcal{A}}$ amounts to $a = (\sum_k\sum_m n_{km}^2 - N)/2$, the *Jaccard coefficient* can be defined as $Jacrd = a/(|\Gamma_C| + |\Gamma_{\mathcal{A}}| - a)$ [15]. As similar measures based on pair counting, the *Rand statistic* and *Fowlkes-Mallows coefficient* are well known, which are given by the following formulas, $Rand = 1 - [(|\Gamma_C| + |\Gamma_{\mathcal{A}}| - 2a)/(N(N - 1)/2)]$ and $FM = a/\sqrt{|\Gamma_C||\Gamma_{\mathcal{A}}|}$, respectively [15]. Another form of the Rand statistic is

$$ARI = \frac{a - |\Gamma_C||\Gamma_{\mathcal{A}}|/(N(N-1)/2)}{|\Gamma_C|/2 + |\Gamma_{\mathcal{A}}|/2 - |\Gamma_C||\Gamma_{\mathcal{A}}|/(N(N-1)/2)},$$

which is usually called the *adjusted Rand index* (ARI) [15].

3.3 Entropy-based measures

If conditional probability $P(A_m|C_k)$ is operationally defined as $P(A_m|C_k) = n_{mk}/n_k$, then it is possible to calculate *entropy* for a given cluster such that $E_k = -\sum_{m=1}^{H}(n_{mk}/n_k)\log(n_{mk}/n_k)$ according to information theory. Thus 'overall' entropy can be computed as a weighted average, namely $En = \sum_{k=1}^{L}(n_k/N)E_k$ [21].

Another information theoretic measure for gaging closeness between two sets is *mutual information* (MI),

$$M_I(C,\mathcal{A}) = \sum_{m=1}^{H}\sum_{k=1}^{L} P(A_m, C_k)\log\frac{P(A_m, C_k)}{P(A_m)P(C_k)},$$

where $P(A_m)$, $P(C_k)$ and $P(A_m, C_k)$ are empirically estimated by n_m/N, n_k/N and n_{mk}/N, respectively [*1]. If C is completely independent of $\mathcal{A}$, then mutual information $M_I(C,\mathcal{A})$ amounts to zero, which is the minimum. In contrast, the maximum of MI is $\min[E(C), E(\mathcal{A})]$ where $E(C)$ is entropy of the set C, namely $E(C) = -\sum_k P(C_k)\log P(C_k)$, and $E(\mathcal{A})$ is defined similarly. Hence, *normalized mutual information* (nMI), which takes a value in $[0, 1]$, is usually employed for evaluating clustering results. Because $M_I(C,\mathcal{A})$ is bounded by some quantities as $M_I(C,\mathcal{A}) \leq \min[E(C), E(\mathcal{A})] \leq \sqrt{E(C)E(\mathcal{A})} \leq [E(C) + E(\mathcal{A})]/2 \leq \max[E(C), E(\mathcal{A})] \leq E(C,\mathcal{A})$ where $E(C,\mathcal{A}) = -\sum_m\sum_k (n_{mk}/N)\log(n_{mk}/N)$, it is feasible to consider five types of nMI such as $nMI_{min} = M_I(C,\mathcal{A})/\min[E(C), E(\mathcal{A})]$, $nMI_{sqrt} = M_I(C,\mathcal{A})/\sqrt{E(C)E(\mathcal{A})}$, $nMI_{sum} = 2M_I(C,\mathcal{A})/[E(C)+E(\mathcal{A})]$, $nMI_{max} = M_I(C,\mathcal{A})/\max[E(C), E(\mathcal{A})]$ and $nMI_{joint} = M_I(C,\mathcal{A})/E(C,\mathcal{A})$ (see [19] for the source of each measure).

Also, *V-measure* [17] is defined as the harmonic mean of two quantities computed based on information theory. The first quantity is called *homogeneity*, which is defined as

$$homo = \begin{cases} 1, & \text{if } E(\mathcal{A}, C) = 0 \\ 1 - E(\mathcal{A}|C)/E(\mathcal{A}), & \text{otherwise} \end{cases},$$

where $E(\mathcal{A}|C) = -\sum_k\sum_m (n_{mk}/N)\log(n_{mk}/n_k)$, and the second one is *completeness*, which is given by

$$comp = \begin{cases} 1, & \text{if } E(C,\mathcal{A}) = 0 \\ 1 - E(C|\mathcal{A})/E(C), & \text{otherwise} \end{cases},$$

where $E(C|\mathcal{A}) = -\sum_m\sum_k (n_{mk}/N)\log(n_{mk}/n_m)$. Actually, V-measure is computed as $Vm = (2\times homo\times comp)/(homo+comp)$.

3.4 Other measures

A normalized version of the *van Dongen criterion* (Wu et al., 2009 [20]) is defined as

$$vDon = \frac{2N - \sum_k \max_{\{m=1,\dots,H\}} n_{mk} - \sum_m \max_{\{k=1,\dots,L\}} n_{mk}}{2N - \max_{\{k=1,\dots,L\}} n_k - \max_{\{m=1,\dots,H\}} n_m}.$$

Note that a smaller value of the criterion indicates better results of clustering, and vice versa.

4. Framework of Experiment

4.1 Generation of synthetic data

For empirically comparing external evaluation measures of DC, synthetic data generated randomly under several assumptions were used as target document sets in the experiment because the 'ease' or 'difficulty' of producing valid clustering results can be adjusted by shifting parameters in the process of generating synthetic data, which may yield deeper insights on the evaluation. Of course, it would be better to employ real document data (e.g., the RCV1 test collection [13]) to obtain more reliable findings on external evaluation measures. The simulation in this experiment should be considered as a preliminary step for understanding sufficiently the characteristics of the measures.

4.1.1 Random selection of term frequency

The model for generating the synthetic data in this experiment was constructed by modifying and extending a model in Jing et al.(2007) [9], which was originally exploited for examining performance of a subspace clustering technique. Basically, after predefining the number of different index terms (denoted by M), the numbers of 'ground truth' classes (i.e., H) and the numbers of documents belonging to H classes (i.e., n_m; $m = 1,\dots,H$), respectively, occurrence frequencies of M terms in each document were randomly determined under some assumptions on the probabilistic distributions used for the random generation. The number of classes were always fixed to 10 (i.e., $H = 10$) in this experiment.

The random generation was done by selecting a real value from the Normal distribution $\mathcal{N}(\mu,\sigma)$. Before the selection, the set of M terms was divided into M' 'specific' terms and M'' 'general' terms (M' and M'' were also fixed in all generations such that $M' = 140$ and $M'' = 60$, i.e., $M = 200$). In the case of general terms, the value of parameter μ for a particular term was sampled from $\mathcal{N}(\mu_g,\sigma_g)$ where $\mu_g = 3.0$ and $\sigma_g = 3.0$ in this experiment [*2]. The randomly selected value of parameter μ for term t_j is denoted by μ_j here ($j = M'+1,\dots,M$). For each document, frequencies of M'' general terms were sequentially generated from $\mathcal{N}(\mu_j,\sigma_g)$. Actually, after real number x was obtained from the Normal distribution, term frequency was determined as $\lfloor x\rfloor$. Also, if $x < 0$, then term frequency was always set to zero [*3].

The system of generation for specific terms was more complicated because each specific term was assumed to belong inherently to one or more classes in $\mathcal{A} = \{A_1,\dots,A_H\}$. The classes of each term were randomly determined by H Bernoulli trials with parameter $p = 0.1$ (i.e., the probability that the term was assigned to a class was 0.1). Namely, the probability that a specific term belongs to x classes follows a binomial distribution $\mathcal{B}(H,p)$. Note that x was intentionally changed to one when $x = 0$ by allocating the term randomly to a particular class.

After assigning each specific term to one or more classes, parameters μ_j ($j = 1,\dots,M'$) were determined by Normal distributions in a similar way to that for general terms. In the case of specific terms, two values must be selected as μ_j (denoted by $\hat{\mu}_j$ and $\tilde{\mu}_j$, respectively). The value of $\hat{\mu}_j$ for documents in the class to which term t_j belongs was sampled from $\mathcal{N}(4.0, 3.0)$. Instead, $\mathcal{N}(0.5, 1.0)$ was used for sampling a value of $\tilde{\mu}_j$ which is a parameter for the other classes. Finally, the frequency of t_j in each document was again selected from $\mathcal{N}(\hat{\mu}_j, 5.0)$ or $\mathcal{N}(\tilde{\mu}_j, 1.0)$ depending on the class to which the target document belongs ($j = 1,\dots,M'$).

4.1.2 Incorporating error factors

In order to control 'difficulty' of DC, an error factor ϵ was in-

[*1] If $n_{mk} = 0$, then it is assumed that $P(A_m, C_k) = P(A_m)P(C_k)$.

[*2] These values of M', M'', μ_g and σ_g were arbitrarily selected without any special reason. Other parameters described below were set in the same way. Inevitably, the experiment should be considered as a case study.

[*3] Note that the operation was applied to all random generations of term frequency in this experiment.

troduced in the process of generating synthetic data. Namely, after all term frequencies in every document were randomly selected by the procedure explained above, each frequency y was further changed to $\lfloor y + \epsilon \rfloor$ where ϵ was a negative or positive real number sampled from $\mathcal{N}(\mu_\eta, 1.0)$ for each y. If parameter μ_η is large, then it becomes more difficult to correctly partition documents into $\{A_1, \ldots, A_H\}$ because differences of term frequencies between specific and general terms or between specific terms belonging to the class and not to it would become small. In this experiment, values of μ_η were selected from $\{0.0, 0.5, 1.0, 1.5, 2.0\}$ for respective generation of the synthetic data (0.0 is the easiest, and 2.0 is the most difficult).

The overall procedure for generating the synthetic data is summarized in Figure 1.

Set: The number of classes ($H = 10$), the number of specific terms ($M' = 140$), the number of general terms ($M'' = 60$), the number of documents ($N = 10$), the number of documents in classes ($n_1, \ldots, n_H$), the parameter of Bernoulli trials for determining whether a specific term belongs to the class or not ($p = 0.1$), and parameter μ_η for incorporating error factors.

1) Sample parameter μ_j from Normal distribution $\mathcal{N}(3.0, 3.0)$ for general term t_j ($j = M' + 1, \ldots, M$ where $M = M' + M''$).
2) Determine whether t_j belongs to the mth class or not by comparing a randomly generated number in $[0, 1]$ with p ($m = 1, \ldots, H$) for specific term t_j ($j = 1, \ldots, M'$).
3) Sample $\hat{\mu}_j$ and $\tilde{\mu}_j$ from $\mathcal{N}(4.0, 3.0)$ and $\mathcal{N}(0.5, 1.0)$, respectively, for specific term t_j ($j = 1, \ldots, M'$).
4) Allocates arbitrarily each document d_i to a class so that the resulting distribution becomes identical to $n_1, \ldots, n_H$ ($i = 1, \ldots, N$).
5) For t_j in d_i ($j = 1, \ldots, M$; $i = 1, \ldots, N$), sample its frequency from $\mathcal{N}(\mu_j, 3.0)$ (if t_j is a general term), $\mathcal{N}(\hat{\mu}_j, 5.0)$ (if t_j is a specific term belonging to the class of d_i), or $\mathcal{N}(\tilde{\mu}_j, 1.0)$ (if t_j is a specific term not belonging to the class of d_i).
6) For t_j in d_i ($j = 1, \ldots, M$ and $i = 1, \ldots, N$), sample real number ϵ from $\mathcal{N}(\mu_\eta, 1.0)$ and add it to the term frequency ($-\infty < \epsilon < \infty$).

- Note: For real number x (> 0), term frequency is set to $\lfloor x \rfloor$, and if $x < 0$, then term frequency is set to zero.

Fig. 1 Algorithm for randomly generating a document set

4.2 Clustering operation

In this experiment, only a spherical k-means (sk-means) algorithm was used in all clustering operations. For executing the sk-means clustering, the top L documents in the sequence were automatically chosen as seeds, and the well-known Hartigan-Wong algorithm [7] was used to find an optimal allocation of documents to clusters (see Appendix). Note that each element of document vectors was calculated as a simple tf-idf weight, which was actually $x_{ij} \log(N/n_j)$ where x_{ij} denotes occurrence frequency of t_j in d_i and n_j indicates the number of documents with $x_{ij} > 0$ ($i = 1, \ldots, N$; $j = 1, \ldots, M$), and that any feature selection was not executed in the experiment.

The number of documents was always set to 100 (i.e., $N = 100$), and two types of distribution of documents over classes were supposed:

(1) $n_1 = n_2 = \ldots = n_H = 10$, and
(2) $n_1 = 50$, $n_2 = 20$, $n_3 = n_4 = n_5 = 5$, $n_6 = \ldots = n_H = 3$,

where $H = 10$ as mentioned before. Whereas the first type indicates a balanced distribution, the distribution of the second type is skewed. In this experiment, the clustering operation was executed independently for the two types of distribution (which are called ‘balanced’ and ‘skewed’ in this paper).

Table 1 Averages of measures: Balanced ($L = 10$)

	Error factor: μ_η				
Measure	0.0	0.5	1.0	1.5	2.0
pur	.867	.834	.716	.431	.301
invpur	.955	.941	.906	.749	.567
F	.886	.857	.749	.449	.302
BCF	.890	.858	.752	.452	.296
Rand	.965	.956	.919	.765	.698
ARI	.817	.771	.622	.242	.068
Jacrd	.728	.671	.508	.210	.107
FM	.841	.802	.681	.382	.222
NMI-min	.950	.931	.874	.604	.339
NMI-sqrt	.926	.901	.818	.504	.274
NMI-sum	.925	.901	.815	.494	.267
NMI-max	.903	.873	.765	.425	.226
NMI-joint	.866	.825	.697	.337	.157
Vm	.925	.901	.815	.494	.267
vDon	.106	.134	.229	.541	.757

Note: Averages of 1000 iterations when $H = 10$.

In each execution of the sk-means clustering, the number of generated clusters (i.e., L) was selected from { 3, 5, 10, 15, 20, 30, 40 }, respectively. Since $H = 10$, the clustering operation with $L = 10$ would be normal. However, in the case of DC, the ‘true’ number of clusters (i.e., the number of classes) is often unknown. Based on the assumption that H is unknown, the experiment tried to examine the effects of conditions $L < H$ and $L > H$ by varying the number of clusters predefined in the sk-means algorithm.

4.3 Iteration of generating data and clustering

In order to compare empirically the external evaluation measures for various clustering results, it was necessary to iterate a pair of (1) generating N documents and (2) clustering them, and to compute the measures for each pair. Hence, the pair of operations was iterated 1000 times independently for a given set of parameters, and the average of values in each run with 1000 iterations was calculated for every measure in the comparison.

5. Results

5.1 Closeness between measures

Table 1 shows average values of the external evaluation measures obtained by repeating the 1000 iterations with varying parameter μ_η on error factor (i.e., $\mu_\eta = 0.0, 0.5, 1.0, 1.5, 2.0$) for the ‘balanced’ document distribution when $L = 10$ and $H = 10$. As expected, all measures except the van Dongen criterion decrease monotonically as μ_η increases (i.e., as clustering becomes more difficult due to the error factor). Similarly, the van Dongen criterion increases monotonically, which indicates that clustering validity becomes lower as μ_η increases.

Consistent with the theory, nMI versions maintain the descending order of $nMI_{min} \geq nMI_{sqrt} \geq nMI_{sum} \geq nMI_{max} \geq nMI_{joint}$ in Table 1. In the case that $\mu_\eta = 2.0$, the valuers of inverse purity and Rand statistic remain relatively high compared with other measures.

If a set of values of measures in an iteration is considered as an observed record, then the Pearson product-moment correlation coefficient r within 1000 records can be computed from the data used to compile Table 1. By converting the van Dongen criterion to $1 - vDon$, it is possible to group the external evaluation measures based on correlation r. Figure 2 shows a dendrogram produced by applying a hierarchical clustering algorithm (average-linkage) to data of $\mu_\eta = 1.0$ (hclust() function of R [16] was employed after each score of r was converted as $1 - r$). Clearly, three main groups, namely $\{BCF, nMI_{joint}, nMI_{sqrt}, nMI_{sum}, V_m\}$, $\{Jacrd, ARI, FM\}$ and $\{F, nMI_{max}, vDon\}$, can be observed in the dendrogram. The correlation within each group was very high as shown by the left-side scale of $1 - r$ in the figure.

It would be easy to conjecture that the Jaccard coefficient, ARI and FM coefficient are closely gathered because they are computed from pair counting. Also, three versions on nMI (nMI_{joint}, nMI_{sqrt} and nMI_{sum}) and Vm make up a group to which BCubed-F is added. Other entropy-based measures nMI_{min} and nMI_{max} are separate from this group, and in particular, nMI_{max} is closer to the F-measure and van Dongen criterion in the dendrogram.

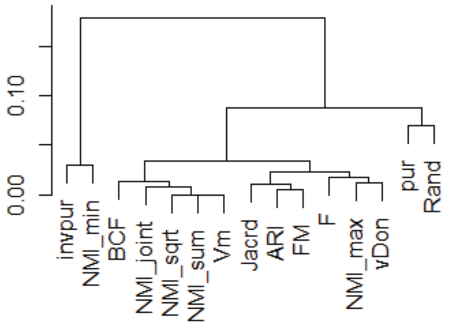

Fig. 2 Dendrogram (1): Balanced ($L = 10, \mu_\eta = 1.0$)

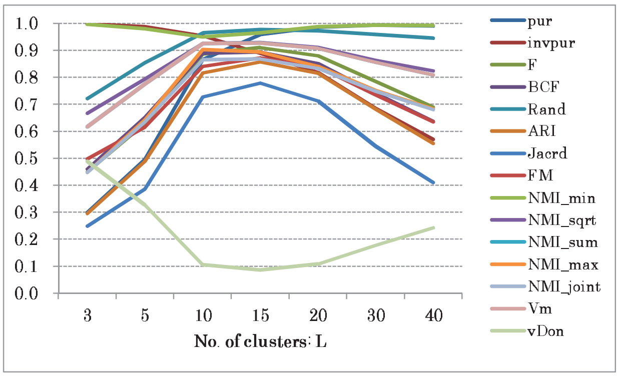

Fig. 3 Values of measures by the number of clusters: Balanced ($\mu_\eta = 0.0$)

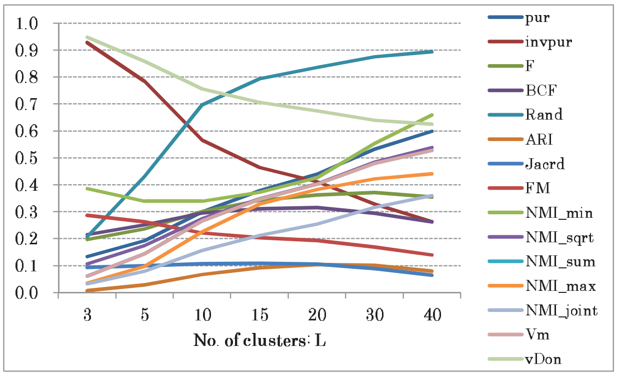

Fig. 4 Values of measures by the number of clusters: Balanced ($\mu_\eta = 2.0$)

5.2 Effect of variation in the number of generated clusters

In most DC experiments, it is usually assumed that the number of generated clusters equals the number of 'ground truth' classes (i.e., $L = H$). However, clustering results must be evaluated sometimes under the condition that $L \neq H$ as pointed out above.

Figure 3 indicates changes of the evaluation measures with variations in the number of clusters ($L = 3, 5, 10, 15, 20, 30, 40$) when $\mu_\eta = 0.0$ (i.e., 'easy' cases). The change in many evaluation measures becomes a bell-shaped curve, in which its peak is at $L = 10$ or $L = 15$. Because $H = 10$, the bell-shaped curve would be interpreted to show a valid trend. On the other hand, purity and inverse purity change monotonically. More precisely, purity increases and inverse purity decreases always as the number of clusters becomes large, which is intuitive from their characteristics.

However, in the case of 'difficult' clustering, the values of some measures do not change like a bell-shaped curve. Figure 4 shows the curves when $\mu_\eta = 2.0$. In particular, all nMI versions and Vm appear to increase monotonically with increment of generated clusters, which implies that nMI versions and Vm may provide a higher score for a set of 'fragmented' clusters in which documents from different classes are mixed. The use of nMI versions or Vm may be risky when $L > H$ and the values are relatively lower.

Although other measures indicate a similar tendency, the curve of Jaccard coefficient reaches the maximum at $L = 15$ as an exception. Since the peaks of ARI and BCubed-F are also at $L = 20$, it is considered that measures by pair counting and BCubed-F behave differently with nMI versions in such situations. The tendency is also observed in a dendrogram for $L = 40$ and $\mu_\eta = 2.0$ (Figure 5), which was obtained in a similar way as Figure 2.

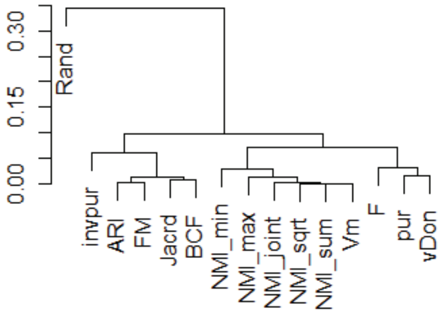

Fig. 5 Dendrogram (2): Balanced ($L = 40, \mu_\eta = 2.0$)

5.3 Effect of biased distribution of documents

Table 2 indicates differences of averages of the external evaluation measures between biased and balanced distributions when $L = 10$. Since the values in the balanced distribution were subtracted from those in the biased distribution, a positive difference in the table means that the value in the biased one is larger. Clearly, in 'difficult' cases, the external evaluation measures in the biased distribution are higher possibly because larger classes prevent the values of the measures from decreasing excessively. This tendency would be remarkable in the measures

Table 2 Difference of values between 'biased' and 'balanced': $L = 10$

	Error factor: μ_η				
Measure	0.0	0.5	1.0	1.5	2.0
pur	0.009	0.015	0.078	0.268	0.315
invpur	-0.163	-0.153	-0.134	-0.074	0.002
F	-0.117	-0.112	-0.057	0.130	0.168
BCF	-0.102	-0.089	-0.037	0.122	0.165
Rand	-0.100	-0.096	-0.088	-0.030	-0.038
ARI	-0.186	-0.153	-0.062	0.098	0.089
Jacrd	-0.159	-0.112	0.009	0.149	0.136
FM	-0.109	-0.081	-0.002	0.143	0.168
NMI-min	-0.092	-0.114	-0.155	-0.089	0.013
NMI-sqrt	-0.122	-0.130	-0.139	-0.029	0.048
NMI-sum	-0.124	-0.131	-0.138	-0.022	0.053
NMI-max	-0.149	-0.144	-0.124	0.016	0.072
NMI-joint	-0.193	-0.195	-0.180	-0.021	0.038
Vm	-0.124	-0.131	-0.138	-0.022	0.053
vDon	0.170	0.170	0.146	0.011	-0.033

Note: value = 'biased' - 'balanced'.

by pair counting (e.g., FM and Jaccard) and BCubed-F (also F-measure).

Figure 6 shows a dendrogram in the case that $L = 10$ and $\mu_\eta = 1.0$ for the biased distribution. The hierarchical structure does not appear to be largely different from that in Figure 2 except for some small changes.

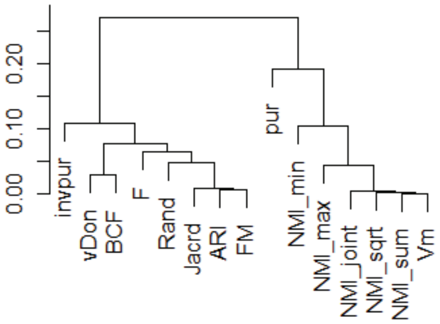

Fig. 6 Dendrogram (3): Biased ($L = 10, \mu_\eta = 1.0$)

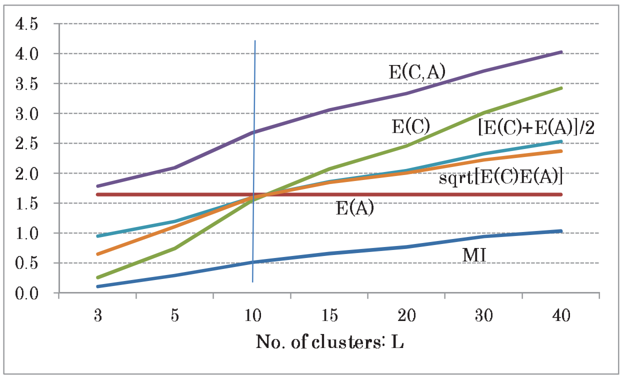

Fig. 7 Values of MI and entropy: Biased ($\mu_\eta = 2.0$)

5.4 Difference of nMI versions

Only the denominator consisting of $E(C)$ and $E(\mathcal{A})$, or $E(C, \mathcal{A})$, is different between nMI versions as discussed above. The three expectations are displayed in Figure 7 when $\mu_\eta = 2.0$ for the biased distribution.

Because the number of classes is fixed (i.e., $H = 10$), $E(\mathcal{A})$ is inevitably constant in the figure. On the other hand, $E(C)$ and $E(C, \mathcal{A})$ increase gradually as L becomes large. The boundary at which $E(C)$ crosses over $E(\mathcal{A})$ is near the point of $L = H = 10$. The curve of $E(C)$ rises more steeply than that of $E(C, \mathcal{A})$, which means that nMI_{max} has the effect of controlling the excessive increases in the normalized value when the number of generated clusters is larger than that of 'ground truth' classes (note that the curve of an average of $E(C)$ and $E(\mathcal{A})$ can not rise more steeply than that of $E(C)$ because $E(A)$ is constant).

6. Discussion

In actual clustering experiments, it is clearly unrealistic to compute and examine all the external evaluation measures. A practical strategy would be to select at least one measure from respective groups of the measures, and to evaluate a clustering result from different perspectives. The dendrograms obtained in this experiment suggest that measures should be selected from at least two groups of pair counting based measures (Jaccard coefficient, ARI, and FM coefficient) and of nMI versions, respectively. Of course, other measures (e.g., BCubed-F or the normalized van Dongen criterion) may provide further evidence on the validity or invalidity of clustering results.

When the number of generated clusters equals that of 'ground truth' classes, there is no functional difference between nMI versions as an evaluation measure to be used for comparing some clustering results. However, in the case of that $L < H$ or $L > H$, it is necessary to use nMI versions carefully. If there are many more generated clusters than classes, then nMI_{max} may be better simply because its value does not increase excessively compared with the other versions. This suggestion would be effective in the case that $L < H$ since $E(\mathcal{A})$ is larger than $E(C)$ in this area (see Figure 7) and nMI_{max} uses $E(\mathcal{A})$ as its denominator.

Even though $L = H$, when comparing two clustering results, one obtained from a document set with a 'balanced' distribution and the other obtained from that with a 'biased' distribution, it is important to pay attention to the external evaluations. For instance, measures by pair counting tend to overestimate the goodness of clustering validity compared with nMI versions when clustering is difficult as exemplified in Table 2.

7. Conclusion

This paper reported an experiment which attempted to compare empirically external evaluation measures for unsupervised clustering. The target datasets were randomly generated based on a model including some assumptions on occurrence frequencies of terms in documents. By using such synthetic data in the experiment, it was possible to observe actual values of individual measures under various states of the document set.

Based on a discussion on the observations, it was suggested that at least one measure should be selected from measures by pair counting and nMI versions, respectively. However, when the number of generated clusters is largely different from that of 'ground truth' classes, it is necessary to pay attention to evaluation by nMI versions (nMI_{max} may be better). Also, it may be possible that measures by pair counting overestimate clustering validity compared with nMI versions when clustering is difficult and the document distribution is biased.

Because this experiment used only synthetic data as target document datasets, further exploration based on real data is needed to extend our knowledge on the external evaluation measures. Even if synthetic data are employed, it may be better to enhance the model used to generate the data. For instance, probabilistic generative models such as latent Dirichlet allocation (LDA) may be a good tool for creating the synthetic data. Also, in this experiment, only spherical k-means clustering was applied; other clustering algorithms or techniques should be examined in future researches.

References

[1] Amigó, E. et al.: A comparison of extrinsic clustering evaluation metrics based on formal constraints, *Inf. Retr.*, Vol.12, pp. 461–486 (2009).

[2] Amigó, E. et al.: Combining evaluation metrics via the unanimous improvement ratio and its application to clustering tasks, *J. Artif. Intell. Res.*, Vol.42, pp. 689–718 (2011).

[3] Brun, M. et al.: Model-based evaluation of clustering validation measures, *Pattern Recognit.*, Vol.40, pp. 807–824 (2007).

[4] Campello, R. J. G. B.: Generalized external indexes for comparing data partitions with overlapping categories, *Pattern Recognit. Lett.*, Vol. 31, pp. 966–975 (2010).

[5] Dom, B. E.: *An Information-theoretic External Cluster-validity Measure*, Technical Report RJ10219, IBM (2001).

[6] Günnemann, S. et al.: External evaluation measures for subspace clustering, *Proc. of 20th ACM Intel. Conf. on Information and Knowledge Management (CIKM '11)*, pp. 1363–1372 (2011).

[7] Hartigan, J. A. and Wong, A.: A k-means clustering algorithm, *Appl. Stat.*, Vol. 28, pp. 100–108 (1979).

[8] Hassani, M. et al.: Effective evaluation measures for subspace clustering of data streams, *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2013 International Workshops*, (Li, J. et al. eds.), Springer, Berlin, pp. 342 – 353 (LNAI 7867) (2013).

[9] Jing, L.: An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data, *IEEE Trans. Knowl. Data Eng.*, Vol. 19, pp. 1026–1041 (2007).

[10] Kishida, K.: Experiment of document clustering by triple-pass leader-follower algorithm without any information on threshold of similarity, *IPSJ SIG Technical Report*, Vol. 2013-IFAT-111, No.23, pp.1–6 (2013).

[11] Kremer, H. et al.: An effective evaluation measure for clustering on evolving data streams, *Proc. of 17th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 868–876 (2011).

[12] Larsen, B. and Aone, C.: Fast and effective text mining using linear-time document clustering, *Proc. of 5th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 16–22 (1999).

[13] Lewis, D. D. et al.: RCV1: A new benchmark collection for text categorization research, *J. Mach. Learn. Res.*, Vol. 5, pp. 361–397 (2004).

[14] Meilă, M.: Comparing clusterings: An axiomatic view, *Proc. of 22nd Intel. Conf. on Machine Learning (ICML '05)*, pp. 577–584 (2005).

[15] Mirkin, B.: *Clustering: A Data Recovery Approach*, CRC Press, Boca Raton, Florida, 2nd edition (2013).

[16] R Development Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna (2010).

[17] Rosenberg, A. and Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure, *Proc. of 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP)*, pp. 410–420 (2007).

[18] Song, M. and Zhang, L.: Comparison of cluster representations from partial second- to full fourth-order cross moments for data stream clustering, *Eighth IEEE Intel. Conf. on Data Mining, 2008 (ICDM '08)*, pp. 560 – 569 (2008).

[19] Vinh, N. X., et al.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance, *J. Mach. Learn. Res.*, Vol. 11, pp. 2837–2854 (2010).

[20] Wu, J. et al.: Adapting the right measures for k-means clustering, *Proc. of 15th ACM SIGKDD Intel. Conf. on Knowledge Discovery and Data Mining*, pp. 877–885 (2009).

[21] Zhao, Y. and Karypis, G.: *Criterion Functions for Document Clustering: Experiments and Analysis*, University of Minnesota (2002).

Appendix

A.1 Spherical K-means Algorithm

Suppose that documents are represented by M-dimensional vectors $\mathbf{d}_i$ $(i = 1, \ldots, L)$. In the Hartigan-Wong algorithm [7], the density of generated clusters is used as an objective criterion for clustering. If the density increases by moving a document to another cluster, then the document is reallocated to the cluster in an iterative procedure. For the case of cosine measure, the 'density' of cluster C_k is reasonably defined such that $J_k = \sum_{i:d_i \in C_k} \mathbf{v}_i^T \mathbf{c}_k / ||\mathbf{c}_k||$ where $\mathbf{v}_i \equiv \mathbf{d}_i / ||\mathbf{d}_i||$ and $\mathbf{c}_k = \sum_{i:d_i \in C_k} \mathbf{v}_i$. From simple manipulation, the increase of J_k by moving document d_* to cluster C_k can be represented by

$$\Delta^+ J_k = \left(\frac{||\mathbf{c}_k||}{||\mathbf{c}_k + \mathbf{v}_*||} - 1 \right) J_k + \frac{2\mathbf{v}_*^T \mathbf{c}_k + 1}{||\mathbf{c}_k + \mathbf{v}_*||},$$

where $\mathbf{v}_* \equiv \mathbf{d}_* / ||\mathbf{d}_*||$. On the other hand, when document d_* is removed from cluster C_k, the decrease of J_k becomes

$$\Delta^- J_k = \left(1 - \frac{||\mathbf{c}_k||}{||\mathbf{c}_k - \mathbf{v}_*||} \right) J_k + \frac{2\mathbf{v}_*^T \mathbf{c}_k - 1}{||\mathbf{c}_k - \mathbf{v}_*||}.$$

By changing the Euclidean distance to the cosine measure and incorporating straightforwardly $\Delta^+ J_k$ and $\Delta^- J_k$ into the original Hartigan-Wong algorithm, it is possible to execute an effective spherical k-means algorithm (first L documents were automatically selected as initial seeds) [10].