Information-Theoretic Document Clustering using Skew Divergence

Toshio UCHIYAMA

Hokkaido Information University

平成28年 3 月

北海道情報大学紀要　第27巻　第 2 号別刷

〈論　文〉

Information-Theoretic Document Clustering using Skew Divergence

TOSHIO UCHIYAMA*

Abstract

Information-theoretic clustering (ITC) and the divergence algorithm for ITC have been proposed for analyzing text data, but only in cases where ITC was applied to feature/word clustering to reduce dimensionality. Therefore, the effectiveness and usefulness of ITC for document clustering, such as finding topics in documents, have not been well evaluated. Moreover, the divergence algorithm for ITC can be affected by the zero-frequency problem, because it uses Kullback-Leibler (KL) divergences. This paper proposes novel algorithms for ITC to solve the problem and evaluate ITC as a document clustering method. The proposed algorithms use skew divergence instead of KL-divergence to avoid the zero-frequency problem. Internal and external evaluation measures, such as purity and Normalized Mutual Information (NMI), are used to compare ITC with spherical clustering. This paper further shows competitive learning algorithms that outperform the k-means algorithm using skew divergence. Experimental results for text data sets are presented to show the effectiveness and usefulness of ITC and proposed algorithms.

要旨

文書データの分析などに用いられる，情報理論的クラスタリング（ITC）およびその学習アルゴリズムであるダイバージェンスアルゴリズムが提案されている．しかし，ITC は次元削減のために特徴あるいは単語に対して適用されている．したがって，文書群に存在するトピックの発見など，文書クラスタリングとしての有用性については，十分に評価されて来なかった．さらに，ダイバージェンスアルゴリズムは，Kullback-Leibler (KL) ダイバージェンスを用いるため，ゼロ頻度問題の悪影響を受けるという問題がある．本論文は，この問題を解決するためのアルゴリズムを提案し，ITC をドキュメントクラスタリング手法として評価する．提案アルゴリズムは，KL ダイバージェンスの代わりに歪ダイバージェンスを用い，ゼロ頻度問題を回避する．内部および外部基準（たとえば純度や正規化相互情報量 (NMI)）に基づき，球面クラスタリングとの比較により ITC を評価する．さらに，同じく歪ダイバージェンスを用いた競合学習アルゴリズムが，前述の k-means タイプのアルゴリズムよりも優れることを示す．文書データセットに対する実験により，ITC と提案アルゴリズムの有効性と有用性を示す．

Keywords

Information-Theoretic Clustering, Skew Divergence, Document Clustering, Competitive Learning

1. Introduction

Clustering is the task of partitioning objects into clusters on the basis of certain criteria so that objects in the same cluster are similar. It is a fundamental procedure to analyze data [Jain et al. (1999); Jain (2010)], such as images and text. Since clustering results depend on criteria and algorithms for optimization, appropriately selecting them is an essential problem. When focusing on partitioning documents into disjoint clusters, cosine similarity (spherical clustering) and the spherical k-means algorithm [Dhillon and Modha (2001)] are known to be successful. On the other hand, information-theoretic clustering (ITC) and the divergence algorithm were also proposed for analyzing text data, but ITC was applied to feature/word clustering [Dhillon et al. (2003)] to reduce dimensionality. Therefore, effectiveness and usefulness of ITC for document clustering, such as finding topics in documents, have not been well evaluated. Moreover, the divergence algorithm for ITC can be affected by the **zero-frequency problem**, because it uses Kullback-Leibler (KL) divergences. This suggests that smoothing is necessary, but an appropriate method for ITC is not easy to find. For example, adding some value to probability distributions to be clustered does not work well. Without solving

* Department of Systems and Informatics, Hokkaido Information University

this problem, it is impossible to reveal any characteristics of ITC as a document clustering method.

This paper proposes novel algorithms for ITC to solve this problem and evaluate ITC as a document clustering method. The proposed algorithms use skew divergence [Lee (1999)] instead of KL-divergence to avoid the zero-frequency problem. Internal and external evaluation measures, such as purity and Normalized Mutual Information (NMI), are used to compare ITC with spherical clustering. This paper further introduces competitive learning algorithms that outperform the k-means algorithm using skew divergence. Experimental results for text data sets are presented to show the effectiveness and usefulness of ITC and proposed algorithms.

2. Related Work

Information-theoretic clustering (ITC) [Dhillon et al. (2003)] is closely related to works about "distributional clustering" [Pereira et al. (1993); Baker and McCallum (1998); Slonim and Tishby (2000)] and uses "within-cluster Jensen-Shannon divergence" as objective functions. In those works, objects are grouped by similarity of distribution (KL-divergence). Since a multinominal Naive Bayes classifier distinguishes objects on the basis of cross entropy decomposed into entropy (constant) and KL-divergence, ITC and Naive Bayes have a close relationship [McCallum and Nigam (1998)]. Although Naive Bayes simply assumes independence of features in a given class, it is a successful classifier.

Skew divergence [Lee (1999)] enables ITC to avoid the zero-frequency problem by smoothing probability distributions and seems to be indispensable for document clustering, although at least one KL-divergence of clusters is guaranteed to be finite [Dhillon et al. (2003)].

This paper uses competitive learning for the clustering, with additional mechanisms to improve its performance, referring to the algorithm based on the least sum of squares criterion Uchiyama and Arbib (1994).

The main contributions of this paper are as follows:

(1) It proposes k-means and competitive learning algorithms for ITC to overcome the zero-frequency problem.
(2) It provides empirical evidence of the effectiveness and usefulness of ITC as a document clustering method and the proposed algorithms.

3. Information-theoretic Clustering

Let $m(m = 1,\ldots,M)$ be a finite number of features (words) of data (documents), P and Q be probability distributions of a discrete random variable, and p_m and q_m be probabilities of when a random variable equals m. The Kullback-Leibler (KL) divergence to Q from P is defined to be

$$D_{\mathrm{KL}}(P||Q) = \sum_{m=1}^{M} p_m \log \frac{p_m}{q_m}, \tag{1}$$

and the generalized Jensen-Shannon (JS) divergence of a finite set of probability distributions $\{P^i : i = 1,\ldots,n\}$ can be expressed as the weighted sum of KL-divergences to the (weighted) mean [Dhillon et al. (2003)]:

$$D_{\mathrm{JS}}(\{P^i : i = 1,\ldots,n\}) = \sum_{i=1}^{N} \pi^i D_{\mathrm{KL}}(P^i||\bar{P}), \tag{2}$$

where π^i is the probability of P^i to be selected and $\bar{P}$ shows the (weighted) mean distribution $\sum_i^n \pi^i P^i$.

Let $P^i(i = 1,\ldots,N)$, $C^k(k = 1,\ldots,K)$, and $Q^k(k = 1,\ldots,K)$ be finite numbers of probability distributions $\mathcal{P}$ corresponding to data (document), clusters, and probability distributions $\mathcal{Q}$ corresponding to the mean distribution of cluster C^k, respectively. The "within-cluster" Jensen-Shannon divergence is defined to be

$$JS_W = \sum_{k=1}^{K} \frac{N_k}{N} D_{\mathrm{JS}}\left(\{P^i | P^i \in C^k\}\right), \tag{3}$$

$$= \frac{1}{N} \sum_{k=1}^{K} \sum_{P^i \in C^k} D_{\mathrm{KL}}(P^i||Q^k), \tag{4}$$

$$= \frac{1}{N} \sum_{k=1}^{K} \sum_{P^i \in C^k} \sum_{m=1}^{M} p_m^i \log \frac{p_m^i}{q_m^k}, \tag{5}$$

where N_k is the number of distributions $P^i(\in C^k)$, and JS_W is the objective function of information-theoretic clustering (ITC) to be minimized [Dhillon et al. (2003)]. Using the relation $\sum_{P^i \in C^k} p_m = N_k q_m^k$, we have

$$JS_W = \frac{1}{N} \sum_{k=1}^{K} N_k \sum_{m=1}^{M} -q_m^k \log q_m^k - \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{m=1}^{M} -p_m^i \log p_m^i \right). \tag{6}$$

Since the second term is constant for any partitioning, the optimization depends on the first term, which is the weighted sum of entropies for the mean probability distribution of cluster Q^k. Therefore, the objective of ITC can be considered to be minimizing those entropies for clusters. It may also be to show the characteristic of ITC on the basis of the information theory.

4. Algorithms

This section first shows a k-means algorithm for ITC. As the equation (4) and the divergence algorithm [Dhillon et al. (2003)] show, the k-means algorithm can be modified for ITC when KL-divergence is used to update the cluster labels. The main difference from the divergence algorithm is to use the skew-divergence [Lee (1999)] instead of KL-divergence to avoid the zero-frequency problem. The skew-divergence is defined as

$$s_\alpha(P, Q) = \mathrm{KL}(P||\alpha Q + (1-\alpha)P), \qquad (7)$$

where $\alpha(0 \leq \alpha \leq 1)$ is the mixture ratio of distributions. The skew divergence is exactly the KL divergence at $\alpha = 1$. The skew-divergence is very simple to use, but it works well for reducing the objective function JS_W. For example, use of $\alpha = 0.99$ may produce a good performance [Lee (1999)].

Algorithm 1 describes the skew-divergence k-means (sdKM) algorithm. In the initialization step, cluster label $k(1 \leq k \leq K)$ is randomly assigned to each distribution P^i. This initialization results in setting the mean distributions $Q^k(k = 1, \ldots, K)$ near the global mean distribution in M dimensional space, and JS-divergence JS_W is large. Note that distributions P^i and Q^k are also treated as vectors, and vice versa. The situation can be considered to be far from all local minima. Since it is impossible to choose good local minima at the beginning, this can be a good strategy for optimization algorithms. Additionally, this initialization guarantees at least one KL-divergence of clusters is finite. This may not affect algorithms using skew divergence.

The mixture ratio (coefficient) for skew divergence $\alpha(0 \leq \alpha \leq 1)$ should be close to 1 when considering reducing distortion from the true objective. However, larger $\alpha(< 1)$ accidentally causes a large penalty for $s_\alpha(P, Q)$ to mean distributions of clusters Q that do not have certain features and makes algorithms stop quickly at bad solutions. By taking these features of α into account, the proposed algorithm 1 includes a mechanism, "gradual change of α", such as $0.99, 0.999, 0.9999, \ldots$, to improve its performance. The initial ratio $\alpha = 0.99$ and the maximum ratio $\alpha_{\max} = 0.999999$ are used in the experiments.

The condition of convergence for a certain α is

$$\frac{\text{previous}JS_W - \text{current}JS_W}{\text{current}JS_W} < \text{epsilon}, \qquad (8)$$

where the small value "epsilon" is 10^{-8}, for example.

Algorithm 2 describes a simple skew-divergence competitive learning (sdCL) algorithm, which has no additional mechanism. In the initialization step, cluster label $k(1 \leq k \leq K)$ is randomly assigned to each distribution P^i in the same way as the k-means algorithm, and the mean probability distributions of clusters Q^k are computed. In each repetition, a winner is decided from $\mathcal{Q}$ for randomly selected distribution P, and only the winner's distribution Q^c is updated. In the update-step, it uses the learning rate γ. In the experiments, the skew-divergence k-means algorithm (sdKM) with $\alpha = 0.999$ is applied for post-processing to remove fluctuation caused by stochastic procedures. Note that sdCL itself can derive better results than sdKM.

For further improvement, this paper proposes a skew-divergence competitive learning algorithm with an additional mechanisms [Uchiyama and Arbib (1994)], which gradually generates units (= probability distributions Q of clusters) on the basis of *wincount*, which shows how many times each distribution Q wins. Details are shown in Algorithm 3. Since the added rule generates or splits distributions Q where the density of distributions P^i is high and it is reasonable to optimize the objective function, we can expect it to outperform usual initialization, which easily drops into bad solutions [Uchiyama and Arbib (1994)].

Algorithm 1 Skew-divergence k-means (sdKM)

Input: $\mathcal{P}$: the set of probability distributions,
K: the number of clusters,
α: the initial ratio for skew divergence.
Output: $\mathcal{C}$: the set of document clusters.
Initialization For each distribution P^i, randomly assign P^i cluster label $k(1 \leq k \leq K)$, to which P^i belongs.
while $\alpha \leq \alpha_{\max}$ **do**
repeat
For each cluster C^k, compute its mean probability distribution Q^k as

$$Q^k \leftarrow \frac{1}{N_k} \sum_{P^i \in C^k} P^i, \qquad (9)$$

where N_k is the number of distributions $P^i (\in C^k)$.
For each distribution P^i, update the cluster label, to which P^i belongs, on the basis of the skew divergences:

$$\arg\min_k s_\alpha(P^i, Q^k). \qquad (10)$$

If there are several candidates, select the smallest k.
until the change of JS-divergence JS_w is small.
Update α by

$$\alpha \leftarrow 1 - (1-\alpha)/10. \qquad (11)$$

end while

5. Experiments

This section provides empirical evidence of the effectiveness and usefulness of ITC as a docu-

Algorithm 2 Skew-divergence competitive learning (sdCL)

Input: $\mathcal{P}$: the set of probability distributions,
K: the number of clusters,
α: the mixture ratio for skew divergence,
γ: the learning rate of competitive learning,
N_r is the number of repetitions.
Output: $\mathcal{C}$: the set of document clusters.
Initialization For each distribution P^i, assign P^i cluster label k randomly. Compute the mean probability distributions of clusters Q^k by $(Q^k \leftarrow \frac{1}{N_k}\sum_{P^i \in C^k} P^i)$.
for $r = 1$ to N_r **do**
Select one distribution P randomly from $\mathcal{P}$ and decide a winner Q^c from $\mathcal{Q}$ by

$$c \leftarrow \arg\min_k s_\alpha(P, Q^k). \tag{12}$$

If there are several candidates, select the smallest k.
Update the winner's distribution Q^c as

$$Q^c \leftarrow (1-\gamma)Q^c + \gamma P. \tag{13}$$

end for
For each distribution P^i, compute the cluster label by the equation (12).

Algorithm 3 Skew-divervence competitive learning with splitting

Input: $\mathcal{P}$: the set of probability distributions,
K: the number of clusters,
α: the mixture ratio for skew divergence,
γ: the learning rate of competitive learning,
N_r is the number of repetitions,
θ: the threshold of times.
Output: $\mathcal{C}$: the set of document clusters.
Initialization Set one probability distribution Q^1 of C^1 by randomly selecting distribution P and attach a variable *wincount* to Q^1 that is initialized to 0 and that shows how many times each distribution wins. Consider a subset of distributions $\mathcal{Q}'$, where $\mathcal{Q}' \subseteq \mathcal{Q}$ and $Q^1 \in \mathcal{Q}'$. Let u be the number of elements in $\mathcal{Q}'$.
for $r = 1$ to N_r **do**
Select one distribution P randomly from $\mathcal{P}$ and decide a winner Q^c from $\mathcal{Q}'$ by

$$c \leftarrow \arg\min_k s_\alpha(P, Q^k).$$

If there are several candidates, select the smallest k.
Update the winner's distribution Q^c as

$$Q^c \leftarrow (1-\gamma)Q^c + \gamma P.$$

Add 1 to *wincount* of the winner.
If the winner's *wincount* equals θ and $u < K$, then add a new distribution Q^{u+1} $(= Q^c)$ to $\mathcal{Q}'$ and clear the *wincount* of both to 0.
end for
For each distribution P^i, compute the cluster label by the equation (12).

ment clustering method and the proposed algorithms. Results of a spherical k-means algorithm [Dhillon and Modha (2001)] (spKM), which uses the same initialization step as skew-divergence k-means (sdKM), are shown for comparison. For each experiment, an operation was iterated 30 times with different initial random seeds for a given set of parameters. For competitive learning algorithms, the learning rate $\gamma = 0.01$, the number of maximum repetitions $N_r = 1,000,000$ and the threshold of times $\theta = 1000$ are used.

5-1 Data Sets

Two data sets are used for experiments. The properties of text data sets are shown in Table 1. Words in the "stop list" [Lewis et al. (2004)] are eliminated.

20Newsgroups data set contains about 20,000 articles evenly divided among 20 UseNet Discussion groups [1].

RCV1 (Reuters Corpus Volume1) data set [Lewis et al. (2004)] contains about 800,000 news articles. I used the second level of RCV1 topic hierarchy as the class label and removed multi-labeled documents.

Table 1 Properties of text data sets

Data	Size(N)	Feature(M)	Class(K)
20Newsgroups	18,774	60,698	20
RCV1	534,135	216,503	53

5-2 Evaluation Measures

In the experiments, I set the number of clusters K to equal the true number of classes in Table 1. I compare the clusters generated by algorithms with the true classes, produced on the basis of human judges, by computing the following two external evaluation measures.

Purity is measured by counting the number of documents from the most frequent class in each cluster. Purity can be computed as

$$\text{purity} = \frac{1}{N}\sum_{k=1}^{K} \max_j T(C^k, A^j), \tag{14}$$

where A^j denotes the j-th class, $T(C^k, A^j)$ is the number of documents that belong to C^k and A^j, and N is the number of documents.

NMI (Normalized Mutual Information) [Manning et al. (2008)] is defined as

$$\text{NMI} = \frac{I(C;A)}{(H(C)+H(A))/2}, \tag{15}$$

where, $I(C;A)$ is mutual information and $H()$ is entropy.

[1] http://qwone.com/ jason/20Newsgroups/

I also used objective functions of clustering algorithms as an internal evaluation measure. These are "within-cluster Jensen-Shannon divergence" JS_W (3) and cosine similarity:

$$\text{cosine} = \frac{1}{N}\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i \in C^k} \boldsymbol{x}_i \cdot \frac{\boldsymbol{\mu}_k}{||\boldsymbol{\mu}_k||},$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{\boldsymbol{x}_i \in C^k} \boldsymbol{x}_i, \tag{16}$$

where $\boldsymbol{x}$ is a vector representation of a document (M dimensional vector) and N_k is the number of documents $\boldsymbol{x}(\in C^k)$.

5-3 Experimental Results

Figures 1-2 display relationships between purity and internal criteria for k-means algorithms. Tables 2-3 show averages of internal and external evaluation measures, where spKM, sdKM, sdCL, and sdCLS correspond to "spherical k-means", "skew-divergence k-means", "skew-divergence competitive learning", and "skew-divergence competitive learning with splitting mechanism", respectively. These figures and tables illustrate the following:

(1) Skew-divergence algorithms yield better solutions in terms of JS-divergence than spherical clustering algorithms (spKM). Hence, the proposed algorithms work for optimizing the objective function of information-theoretic clustering (ITC) as a document clustering method.

(2) Skew-divergence algorithms yield better solutions in terms of purity and NMI than spherical clustering.

(3) Among skew-divergence algorithms, competitive learning algorithms (sdCL and sdCLS) can outperform spKM in most cases.

Figures 1 and 2 clearly show that ITC outperforms spherical clustering in terms of purity. These figures illustrate the usefulness of ITC. Furthermore, we can find a correlation between external evaluation measures (purity and NMI) and JS-divergence in Figure 3. Here, the correlation coefficient between purity and JS-divergence is -0.982 and that between NMI and JS-divergence is -0.985. Therefore, the smaller the JS-divergence a solution has, the larger its external evaluation measures become. Although limited in 20Newsgroups data, the relationships suggest that JS-divergence can be an important factor for the quality of document clustering. For RCV1 data, there is no obvious correlation in results of skew-divergence algorithms (Figure 4), but Figure 2 and Table 3, which include results of spKM, illustrate the effectiveness of JS-divergence at least.

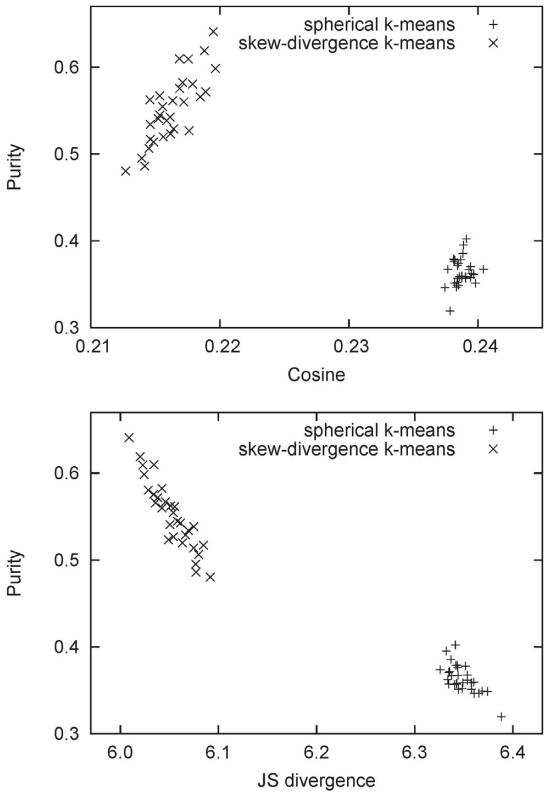

Figure 1 Purity vs. cosine (upper) and JS-divergence (lower) for 20Newsgroups data.

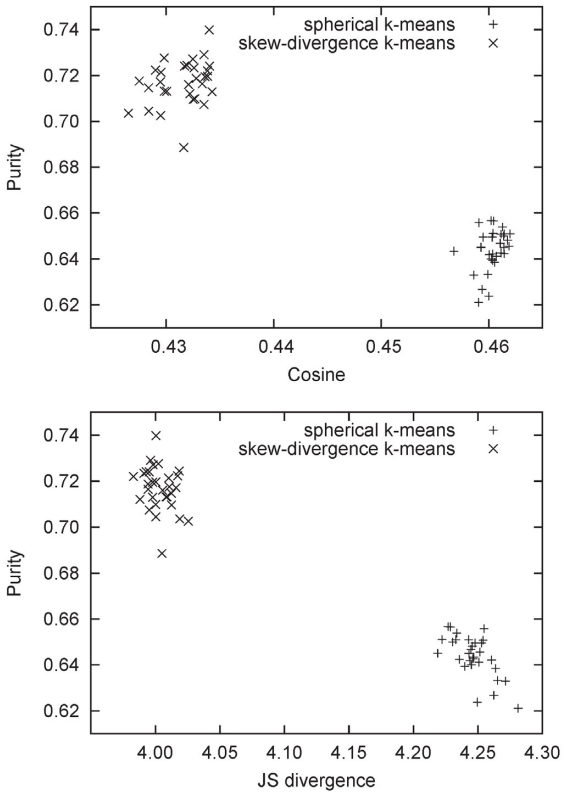

Figure 2 Purity vs. cosine (upper) and JS-divergence (lower) for RCV1 data.

6. Discussion

Figures 1-2 and Tables 2-3 show that skew-

Table 2 Results of evaluation measures for 20Newsgroups data

Algorithm	cosine	JS div.	purity	NMI
spKM	**0.2387**	6.348	0.3650	0.3259
sdKM	0.2153	6.052	0.5520	0.5585
sdCL	0.2200	5.983	0.6839	0.6503
sdCLS	0.2204	**5.976**	**0.6982**	**0.6558**

Table 3 Results of evaluation measures for Rcv1 data

Algorithm	cosine	JS div.	purity	NMI
spKM	**0.4604**	4.246	0.6439	0.4765
sdKM	0.4315	4.003	0.7167	0.5740
sdCL	0.4264	4.005	0.7279	**0.5887**
sdCLS	0.4351	**3.977**	**0.7424**	0.5866

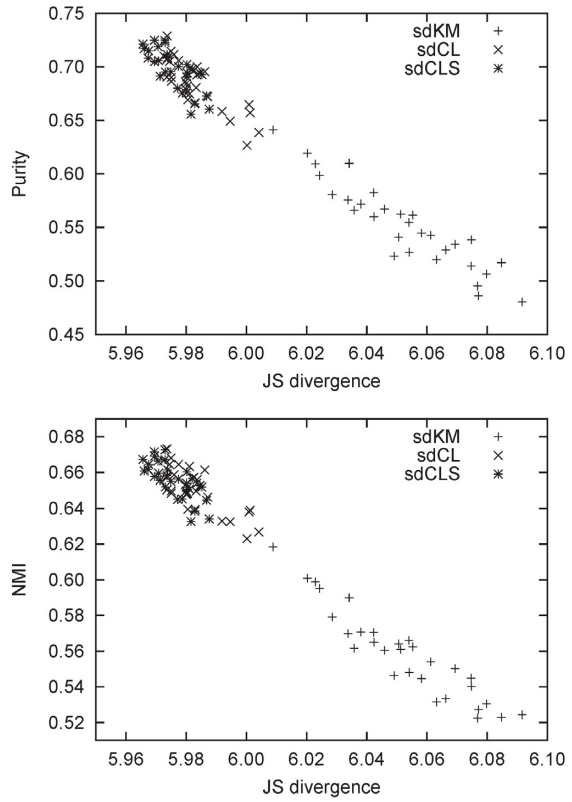

Figure 3 Relationship between external evaluation measures and JS-divergence for 20Newsgroups data.

divergence algorithms outperform a spherical k-means algorithm (spKM). These algorithms drastically improve purity (0.365 to 0.698, 0.643 to 0.742) and NMI. For example, purities yielded by spKM and sdCLS for 20Newsgroups are 0.365 and 0.698, respectively. To reveal such an important characteristic of ITC, the proposed algorithms play an indispensable role. These show the effectiveness and usefulness of ITC and the proposed algorithms. As a document clustering method, ITC may be much more useful than spherical clustering. I beleive the effectiveness of ITC should be derived from Naive Bayes related to ITC.

There are obvious correlations in Figure 3, but no such correlation can be found in Figure 4, where a correlation coefficient between purity and JS-divergence is -0.595 and that between NMI and JS-

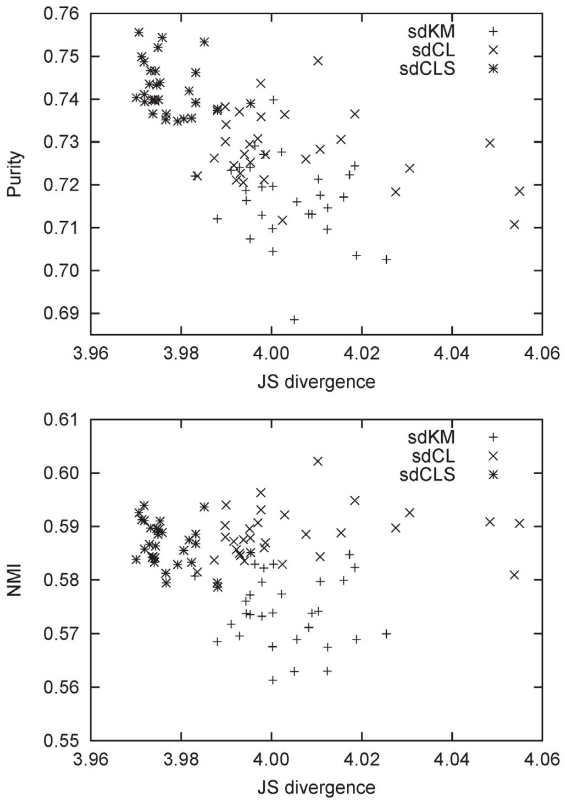

Figure 4 Relationship between external evaluation measures and JS-divergence for Rcv1 data.

divergence is -0.174. This difference may come from the complexity of the clustering problem for RCV1 data. As shown in Table 1, the numbers of documents (18,774 vs. 534,135) and features (60,698 vs. 215,603) are much bigger than those of 20Newsgroups data. CPU time of ICT for RCV1 data is about 2 to 4 hours, where that for 20Newsgroups data is within a few minutes (using a core i7-4770). Clustering problems with large scale data likely have a lot of local minima to which algorithms converge. It is possible that they take various values for internal and external evaluation measures, as shown in Figure 4. These figures also suggest that there are factors that have not been analyzed. I beleive analyzing these factors is an important future pursuit.

7. Conclusion

This paper proposed novel algorithms for information-theoretic clustering (ITC) to utilize ITC as a document clustering method. The proposed algorithms use skew divergence instead of KL-divergence to avoid the zero-frequency problem. Internal and external evaluation measures were used to compare ITC with spherical clustering. This paper further introduced competitive learning algorithms that outperform the k-means algorithm using skew-divergence. Experimental results for text data sets were presented to show the effec-

tiveness and usefulness of ITC and the proposed algorithms.

References

[1] Baker, L Douglas and Andrew Kachites McCallum (1998) "Distributional clustering of words for text classification," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 96–103, ACM.

[2] Dhillon, Inderjit S, Subramanyam Mallela, and Rahul Kumar (2003) "A divisive information theoretic feature clustering algorithm for text classification," *The Journal of Machine Learning Research*, Vol. 3, pp. 1265–1287.

[3] Dhillon, Inderjit S and Dharmendra S Modha (2001) "Concept decompositions for large sparse text data using clustering," *Machine learning*, Vol. 42, No. 1-2, pp. 143–175.

[4] Jain, Anil K (2010) "Data clustering: 50 years beyond K-means," *Pattern recognition letters*, Vol. 31, No. 8, pp. 651–666.

[5] Jain, Anil K, M Narasimha Murty, and Patrick J Flynn (1999) "Data clustering: a review," *ACM computing surveys (CSUR)*, Vol. 31, No. 3, pp. 264–323.

[6] Lee, Lillian (1999) "Measures of distributional similarity," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 25–32, Association for Computational Linguistics.

[7] Lewis, David D, Yiming Yang, Tony G Rose, and Fan Li (2004) "Rcv1: A new benchmark collection for text categorization research," *The Journal of Machine Learning Research*, Vol. 5, pp. 361–397.

[8] Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schütze (2008) *Introduction to information retrieval*, Vol. 1: Cambridge university press Cambridge.

[9] McCallum, Andrew and Kamal Nigam (1998) "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, Vol. 752, pp. 41–48, Citeseer.

[10] Pereira, Fernando, Naftali Tishby, and Lillian Lee (1993) "Distributional clustering of English words," in *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pp. 183–190, Association for Computational Linguistics.

[11] Slonim, Noam and Naftali Tishby (2000) "Document clustering using word clusters via the information bottleneck method," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 208–215, ACM.

[12] Uchiyama, Toshio and Michael Arbib (1994) "Color image segmentation using competitive learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 16, No. 12, pp. 1197–1206.

Acknowledgments This work was supported by JSPS KAKENHI Grant Number 26330259.