



Using cluster validation criterion to identify optimal feature subset and cluster number for document clustering [☆]

Zheng-Yu Niu ^{a,*}, Dong-Hong Ji ^a, Chew Lim Tan ^b

^a *Institute for Infocomm Research, Mail Box B023, 21 Heng Mui Keng Terrace, Singapore 119613, Singapore*

^b *Department of Computer Science, National University of Singapore, 3 Science Drive 2, Singapore 117543, Singapore*

Received 6 April 2006; received in revised form 10 July 2006; accepted 16 July 2006

Abstract

This paper presents a cluster validation based document clustering algorithm, which is capable of identifying an important feature subset and the intrinsic value of model order (cluster number). The important feature subset is selected by optimizing a cluster validity criterion subject to some constraint. For achieving model order identification capability, this feature selection procedure is conducted for each possible value of cluster number. The feature subset and the cluster number which maximize the cluster validity criterion are chosen as our answer. We have evaluated our algorithm using several datasets from the 20Newsgroup corpus. Experimental results show that our algorithm can find the important feature subset, estimate the cluster number and achieve higher micro-averaged precision than previous document clustering algorithms which require the value of cluster number to be provided.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: Document clustering; Cluster validation; Feature selection; Cluster number estimation

1. Introduction

Document clustering is a central problem in information retrieval, which can be defined as grouping documents into clusters according to their topics or main contents in an unsupervised manner. It has been used as a way for improving retrieval performance and navigating large document collection (Cutting, Karger, Pederson, & Tukey, 1992; Zamir & Etzioni, 1998).

A large variety of algorithms have been suggested for document clustering, which can be categorized as hierarchical clustering algorithms (Cutting et al., 1992; El-Hamdouchi & Willett, 1986; Schütze & Silverstein, 1997; Slonim & Tishby, 2000), partitional clustering algorithms (Dhillon, Mallela, & Modha, 2003; El-Yaniv & Souroujon, 2002; Pantel & Lin, 2002; Slonim, Friedman, & Tishby, 2002; Willett, 1980), spectral clustering

[☆] A previous version of this paper appears in the Proceedings of ACM 13th Conference on Information and Knowledge Management.

* Corresponding author. Tel.: +65 68748541/93612080; fax: +65 67755014.

E-mail addresses: zniu@i2r.a-star.edu.sg, niu Zheng@comp.nus.edu.sg (Z.-Y. Niu), dhji@i2r.a-star.edu.sg (D.-H. Ji), tancl@comp.nus.edu.sg (C.L. Tan).

algorithms (Dhillon, 2001; Ding, He, Zha, Gu, & Simon, 2001; Li, Ma, & Ogihara, 2004; Zha, He, Ding, Gu, & Simon, 2001), and matrix factorization algorithms (Xu, Liu, & Gong, 2003; Xu & Gong, 2004), etc.

One common characteristic of most of these methods is that the number of clusters is required to be provided by users. But in practice, this knowledge is usually unknown in advance. Hierarchical clustering techniques do not require a stated number of clusters as input. However, it is desirable to automatically identify the value of cluster number for uncovering the intrinsic structure in document set.

The other observation is that for most of document clustering algorithms, feature words are selected by simple ranking scheme, without considering their effect on clustering procedure. To effectively identify document clusters with different topics, it is better to select the important feature subset with consideration of feedback from clustering procedure.

For achieving feature selection and model order identification capabilities, we propose a cluster validation based model selection algorithm. Firstly salient words are selected from all the words occurred in a document set. Then an unsupervised feature selection algorithm is introduced to refine the salient word set. For achieving model order identification capability, we run this feature selection procedure for each possible value of cluster number. We choose the feature subset and cluster number which maximize the cluster validity criterion as our answer. In this paper, sequential Information Bottleneck (sIB) algorithm (Slonim et al., 2002) is employed to estimate cluster structure since it has been shown to be the state of the art method for document clustering.

This paper is organized as follows. Section 2 provides a brief review of related efforts on document clustering with feature selection and order identification capabilities. In Section 3, we present the document clustering algorithm that determines the feature subset and cluster number by optimizing the cluster validation criterion. In Section 4, we provide the experimental results of our algorithm and discuss some findings from these results. Section 5 concludes our work and suggests some possible improvements.

2. Related work

Vaithyanathan and Dom (1999) proposed a Bayesian approach that used stochastic complexity to determine the important feature subset and the intrinsic value of cluster number. Their feature selection algorithm grouped feature words into clusters and then selected the important feature subset by removing clusters containing noise words. But this procedure made the performance of feature selection algorithm dependant on how well the feature candidates were clustered. If noise words were not put together into a few clusters, feature selection algorithm would fail to remove them.

The other method presented by Liu, Gong, Xu, and Zhu (2002) is to employ Gaussian mixture modeling (GMM) with EM algorithm to conduct an initial clustering. Then document clusters were refined by voting on cluster labels using discriminative features which were identified from the clusters themselves. For inferring the number of clusters, they introduced randomness in the clustering initialization, then determined the intrinsic value of cluster number by which running the document clustering process on the full data set for a fixed number of times yielded the most similar results.

Compared with the Bayesian statistical estimation approach (Vaithyanathan & Dom, 1999), we adopt a different model selection measure, cluster validation. In addition, the feature selection is conducted on features, rather than feature clusters. In contrast with the other approach based on GMM and cluster refinement (Liu et al., 2002), we conduct feature selection before clustering, which will help to reduce the negative effects of the noise words on the clustering procedure. Furthermore, the clustering analysis algorithm (sIB) which we adopt makes no assumption about the structure of data distribution.

3. Proposed algorithm

Given a collection of documents, for revealing the underlying structure of documents, we need to discover the important feature subset and infer the number of clusters. In fact, these two tasks are related, since different feature subsets will uncover different cluster structures. Due to the interplay between feature selection and clustering solution, we should define a unified objective function to evaluate both the feature subset and the cluster number.

Cluster validation is a commonly used method for the problem of model order identification (Lange, Braun, Roth, & Buhmann, 2002; Levine & Domany, 2001; Tibshirani, Walther, Botstein, & Brown, 2001).

For the problem of document clustering, we extend the cluster validation strategy further to address both feature selection and model order identification. The assumption here becomes that if the model order is identical with the intrinsic value and the selected feature subset is important and complete, then the cluster structure estimated from the data is stable against resampling, otherwise, it is more likely to be the artifact of sampled data. This assumption is reasonable since noise features will blur the true cluster structures, and make the estimated cluster structure more likely be the artifact of the data generated by resampling.

From the point of the view of feature selection, this combination can be seen as a wrapper strategy: the important feature set is determined by measuring its impact on the performance of clustering (Dy & Brodley, 2000; Law, Figueiredo, & Jain, 2002; Modha & Spangler, 2003; Vaithyanathan & Dom, 1999).

Table 1 presents our model selection algorithm. The objective function $M_{F_k, k}$ is relevant with both the feature subset and the cluster number. Clustering solution which is stable against resampling will give rise to a local optimum of $M_{F_k, k}$, which indicates both the important feature subset and the intrinsic value of cluster number. We use the sIB algorithm to perform clustering analysis here (described in Section 3.3).

3.1. Preprocessing for feature selection

Before feature selection, we try to remove some noise words using a saliency based criterion. This may help to improve the efficiency of the feature selection. Let $W = \{w_1, w_2, \dots, w_N\}$ represent all the words occurred in a document set D , where $D = \{d_1, d_2, \dots, d_M\}$. The frequency of the word w_i in the document d_j is denoted by $n_{i,j}$, $1 \leq i \leq N$, $1 \leq j \leq M$. The frequency of the word w_i in a large reference corpus \mathcal{R} is represented by $n_{i,R}$. In this paper, we use New York Times News data (July 1994–December 1996) as the reference corpus. The length (the number of occurrences of words) of the document d_j and the corpus \mathcal{R} are denoted by L_j and L_R , respectively. Then we measure the salience of the word w_i in the document d_j by

$$s_{d_j}(w_i) = \frac{n_{i,j}/L_j}{n_{i,R}/L_R}. \quad (1)$$

Higher values of $s(w)$ correspond to more salient words w .

After measuring the salience of all the words in the entire document set, we use the following conditions to identify a salient word set, denoted as W_S :

$$\exists d_j, s_{d_j}(w_i) \geq \tau, \quad \text{and} \quad n_{i,j} \geq 2, \quad (2)$$

$$\sum_j 1\{n_{i,j} > 0\} \geq 2, \quad (3)$$

where τ is a given threshold. In this paper, $\tau = 5.0$, which is set based on our experience in another information retrieval task.

Then each document d_j is represented by a vector v_j , which is defined as

$$v_j = (n_{1,j}, n_{2,j}, \dots, n_{|W_S|,j}), \quad (4)$$

where $n_{i,j}$ is the frequency of the word w_i occurred in the document d_j , and $w_i \in W_S$.

3.2. Feature subset selection

When selecting salient words from the initial vocabulary set W , the ability of the words to discriminate the documents with different topics is not considered. It is necessary to refine this salient word set W_S by

Table 1

Model selection algorithm for document clustering

1	Remove noise words using the method described in Section 3.1
2	Set lower bound K_{\min} and upper bound K_{\max} for the cluster number k
3	Set $k = K_{\min}$
4	Conduct feature selection using the algorithm presented in Section 3.2
5	Record \hat{F}_k, k and the value of objective function $M_{\hat{F}_k, k}$
6	Set $k = k + 1$. If $k \leq K_{\max}$, go to step 4, otherwise go to step 7
7	Choose the value \hat{k} and the feature subset $\hat{F}_{\hat{k}}$ which maximize the value of objective function as estimated cluster number and selected feature subset

unsupervisedly removing noise words which do not help or even deteriorate the discrimination of documents with different topics. This problem can be generalized as selecting an important feature subset from W_S in an unsupervised manner.

Since for each document there should exist some words which can represent its topic, it is reasonable to suppose that the selected features should cover all the documents. Formally, let $\text{coverage}(D, F)$ be the coverage rate of the feature set F with respect to the document set D , i.e., the ratio of the number of documents with occurrence of at least one feature against the total number of documents, then it is assumed that $\text{coverage}(D, F) = 1$.

Our feature subset selection procedure is formulated as

$$\hat{F}_k = \arg \max_{F \subseteq W_S} \{\text{criterion}(F, k, D, q)\}, \quad (5)$$

subject to $\text{coverage}(D, F) = 1$.

\hat{F}_k is the selected feature subset, criterion is the cluster validation based measure described in Table 2, F and k are the feature subset and the value of cluster number to be evaluated. D represents the document set, and q is the sampling frequency for the estimation of cluster validation criterion. q is set as 20 in this work.

Cluster validation process works as follows: (1) randomly sample a subset from the full dataset; (2) group the documents in the sampled subset and the full dataset into k clusters respectively; (3) measure the proportion of document pairs in each cluster computed on the full dataset that are also assigned into the same cluster by clustering solution on the subset. Intuitively, if cluster number k is identical with the intrinsic value, then clustering results on the different subsets generated by sampling should be similar with that on the full dataset. In other words, the clustering solution with the intrinsic cluster number as the parameter is robust against resampling, which gives rise to a local maximum of the function “criterion”.

In this paper we consider the sequential backward floating search (Pudil, Novovicova, & Kittler, 1994) to find an important feature subset on a sorted feature list. The sequential backward floating search starts from a ranked full feature list, and then (1) tries to remove a feature from the bottom of the list to maximize the function “criterion”, and repeat this process for l times till the value of criterion cannot be optimized, and (2) tries to add a feature back to the feature set to maximize the function criterion, and repeat this process for m times till the value of criterion cannot be optimized, and (3) runs the steps (1)–(2) till a local maximum of the function “criterion” is reached. Mutual information between feature candidates and all the documents is used as the criterion for sorting feature candidates. In this paper, $l = 2$, $m = 1$, where l is the number of take-away steps in step (1), and m is the number of plus steps in step (2).

This constrained optimization process results in an important feature subset which maximizes the criterion and meets the given constraint at the same time. For each possible value of cluster number, we can get a

Table 2
Unsupervised algorithm for the evaluation of feature subset and cluster number

	<i>Function:</i> $\text{criterion}(F, k, D, q)$
	<i>Input:</i> feature subset F , cluster number k , document set D , and sampling frequency q
	<i>Output:</i> the score of the merit of F and k
1	Perform clustering analysis using sIB on the full data set D^F with k as input
2	Construct connectivity matrix $C_{F,k}$ based on above clustering solution on D^F
3	Use a random predictor ρ_k to assign uniformly drawn labels to each document in D^F
4	Construct connectivity matrix C_{F,ρ_k} using above clustering solution on D^F
5	For $\mu = 1$ to q do
5.1	Randomly sample a subset $(D^\mu)^F$ with size $\alpha D $ from D^F , $0 \leq \alpha \leq 1$
5.2	Perform clustering analysis using sIB on $(D^\mu)^F$ with k as input
5.3	Construct connectivity matrix $C_{F,k}^\mu$ using above clustering solution on $(D^\mu)^F$
5.4	Use ρ_k to assign uniformly drawn labels to each document in $(D^\mu)^F$
5.5	Construct connectivity matrix C_{F,ρ_k}^μ using above clustering solution on $(D^\mu)^F$
	Endfor
6	Evaluate the merit of F and k using the following objective function
	$M_{F,k} = \frac{1}{q} \sum_{\mu} M(C_{F,k}^\mu, C_{F,k}) - \frac{1}{q} \sum_{\mu} M(C_{F,\rho_k}^\mu, C_{F,\rho_k})$
	where $M(C^\mu, C)$ is given by Eq. (6)
7	Return $M_{F,k}$

corresponding feature subset. In all the pairs of feature subset and cluster number, we choose the pair that maximizes the objective function “criterion” as our answer.

The function $M(C^\mu, C)$ in Table 2 is given by Levine and Domany (2001):

$$M(C^\mu, C) = \frac{\sum_{i,j} 1\{C_{i,j}^\mu = C_{i,j} = 1, d_i \in D^\mu, d_j \in D^\mu\}}{\sum_{i,j} 1\{C_{i,j} = 1, d_i \in D^\mu, d_j \in D^\mu\}}, \quad (6)$$

where D^μ is a subset with size $\alpha|D|$ sampled from the full data set D , C and C^μ are $|D| \times |D|$ connectivity matrices based on the clustering solutions computed on D and D^μ , respectively, and $0 \leq \alpha \leq 1$. The connectivity matrix C is defined as: $C_{i,j} = 1$ if d_i and d_j belong to the same cluster, otherwise $C_{i,j} = 0$. C^μ is calculated in the same way. α is set as 0.90 in this paper.

$M(C^\mu, C)$ measures the proportion of document pairs in each cluster computed on D that are also assigned into the same cluster by clustering solution on D^μ . Clearly, $0 \leq M \leq 1$. Intuitively, if the cluster number k is identical with the intrinsic value, and the selected feature subset contains no noise words, then clustering results on D^μ should be similar with that on D . In other words, the clustering solution with the intrinsic cluster number and the important feature subset as parameters is robust against resampling, which gives rise to a local maximum of $M(C^\mu, C)$.

In our algorithm, we normalize $M(C_{F,k}^\mu, C_{F,k})$ using the equation in step 6 of Table 2, which makes our objective function different from the figure of merit (Eq. (6)) proposed in Levine and Domany (2001). The reason to normalize $M(C_{F,k}^\mu, C_{F,k})$ is that $M(C_{F,k}^\mu, C_{F,k})$ tends to decrease when increasing the value of k . Therefore for avoiding the bias that smaller value of k is to be selected as the cluster number, we use the cluster validity of a random predictor to normalize $M(C_{F,k}^\mu, C_{F,k})$.

In later experiments, we provide the results of cluster number estimation procedure using $M_{F,k}$ or $M_{F,k}^{\text{unnorm}}$ ($M_{F,k}^{\text{unnorm}} = \frac{1}{q} \sum_{\mu} M(C_{F,k}^\mu, C_{F,k})$) as the evaluation function. Our results confirm the necessity to normalize $M(C_{F,k}^\mu, C_{F,k})$.

3.3. Clustering procedure

In this paper we use the sIB algorithm (Slonim et al., 2002) to estimate cluster structure, which measures the similarity of documents according to the similarity of their word conditional distribution. sIB is a simplified “hard” variant of the information bottleneck method (Tishby, Pereira, & Bialek, 1999), which has been shown to be the state of the art method for document clustering task (Slonim et al., 2002).

Let d represent a document, and w represent a feature word, $d \in D$, $w \in F$. Given the joint distribution $p(d, w)$, the document clustering problem is formulated as looking for a compact representation T for D , which preserves as much information as possible about F . T is the document clustering solution. This optimization problem can be solved by the sIB algorithm (Slonim et al., 2002), which found a local maximum of $I(T, F)$ by: given an initial partition T , iteratively drawing a $d \in D$ out of its cluster $t(d)$, $t \in T$, and merging it into t^{new} such that $t^{\text{new}} = \arg\min_{t \in T} \mathbf{d}(d, t)$. $\mathbf{d}(d, t)$ is the change of $I(T, F)$ due to merging d into the cluster t^{new} , which is given by

$$\mathbf{d}(d, t) = (p(d) + p(t))\text{JS}(p(w|d), p(w|t)). \quad (7)$$

$\text{JS}(p, q)$ is the Jensen–Shannon divergence (Lin, 1991), which is defined as

$$\text{JS}(p, q) = \pi_p D_{\text{KL}}(p||\bar{p}) + \pi_q D_{\text{KL}}(q||\bar{p}), \quad (8)$$

$$D_{\text{KL}}(p||\bar{p}) = \sum_y p \log \frac{p}{\bar{p}}, \quad (9)$$

$$D_{\text{KL}}(q||\bar{p}) = \sum_y q \log \frac{q}{\bar{p}}, \quad (10)$$

$$\{p, q\} \equiv \{p(w|d), p(w|t)\}, \quad (11)$$

$$\{\pi_p, \pi_q\} \equiv \left\{ \frac{p(d)}{p(d) + p(t)}, \frac{p(t)}{p(d) + p(t)} \right\}, \quad (12)$$

$$\bar{p} = \pi_p p(w|d) + \pi_q p(w|t). \quad (13)$$

4. Experiments and results

4.1. Test data

Following (Dhillon et al., 2003; El-Yaniv & Souroujon, 2002; Slonim et al., 2002; Slonim & Tishby, 2000), we constructed several subsets from 20Newsgroup corpus (NG20) for the evaluation of our algorithm. The NG20 data contains about 20,000 articles evenly distributed among 20 Usenet discussion groups, which is a widely used benchmark corpus for supervised text categorization task. For our test, we constructed nine datasets by randomly selecting 500 documents evenly distributed among categories in each dataset. The details of these datasets are described in Table 3. Our preprocessing included removing file headers but retaining subject lines, lowering the upper case characters, ignoring all the words that contained digits or non alpha-numeric characters, removing words from a stop-word list containing 599 words, and filtering out low frequency words which appeared only once in the entire corpus. We did not use stemming procedure.

4.2. Evaluation method

When assessing the agreement between the clustering result and the known ground truth, we will encounter the difficulty that there is no label value for each cluster.

To solve this problem, the authors in Dhillon et al. (2003), El-Yaniv and Souroujon (2002), Slonim et al. (2002), Slonim and Tishby (2000) proposed to assign documents in each cluster t , $t \in T$, with the most dominant class label in that cluster, and then conducted evaluation on these labelled documents. In this work, we followed their method to assign labels to document clusters.

Given this uni-labelled data, we define $\alpha(c, T)$ as the number of documents correctly assigned to class c , $\beta(c, T)$ as the number of documents incorrectly assigned to class c and $\gamma(c, T)$ as the number of documents incorrectly not assigned to class c . Following (Dhillon et al., 2003; El-Yaniv & Souroujon, 2002; Slonim et al., 2002; Slonim & Tishby, 2000), we use micro-averaged precision and micro-averaged recall as evaluation measure, which are given by

$$P(T) = \frac{\sum_c \alpha(c, T)}{\sum_c (\alpha(c, T) + \beta(c, T))}, \quad (14)$$

$$R(T) = \frac{\sum_c \alpha(c, T)}{\sum_c (\alpha(c, T) + \gamma(c, T))}. \quad (15)$$

If corpus and algorithm are both uni-labelled then $P(T) = R(T)$. So we report only $P(T)$ for our uni-labelled data.

4.3. Experiments and results

For comparison, we tested both sIB and FSCV on nine datasets from NG20 corpus, created by ourselves following (Slonim et al., 2002). We set $n = 10$, $\epsilon = 0$, $\max L = 50$ for all the runs of sIB algorithm, where n is the number of initializations for each run, ϵ is used for assessing the convergence of iteration procedure, and $\max L$ is the maximum of the number of loops in sIB clustering process. $\epsilon = 0$ means that the clustering process will stop if the local maxima of the objective function in sIB is reached.

Table 3
Description of the test data in our experiments

Dataset	Newsgroup included	# Doc per group	Total # doc
Binary _{1,2,3}	talk.politics.mideast, talk.politics.misc	250	500
Multi _{5,1,2,3}	comp.graphics, rec.motorcycles, rec.sport.baseball, sci.space, talk.politics.mideast	100	500
Multi _{10,1,2,3}	alt.atheism, misc.forsale, comp.sys.mac.hardware, rec.autos, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, talk.politics.guns	50	500

Table 4

Micro-averaged precision of different document clustering algorithms over the datasets from NG20 data

Dataset	DC	IDC	CoC	sIB	FSCV
Binary ₁	70.0	85.0	96.0	72.2	85.8
Binary ₂	68.0	83.0	–	91.6	92.2
Binary ₃	75.0	80.0	–	94.4	93.0
Multi5 ₁	59.0	86.0	89.0	93.2	93.4
Multi5 ₂	58.0	88.0	–	90.6	90.8
Multi5 ₃	53.0	86.0	–	93.4	95.6
Multi10 ₁	35.0	56.0	54.0	63.6	61.8
Multi10 ₂	35.0	49.0	–	68.4	61.6
Multi10 ₃	35.0	55.0	–	65.0	68.8
Average	54.2	74.2	79.7	81.4	82.6

Table 5

Results of model order identification procedure over the datasets from NG20 data

Dataset	# Real categories	$M_{F,k}^{\text{unnorm}}$	$M_{F,k}$
Binary ₁	2	2 ✓	3 ×
Binary ₂	2	2 ✓	2 ✓
Binary ₃	2	2 ✓	4 ×
Multi5 ₁	5	2 ×	5 ✓
Multi5 ₂	5	2 ×	5 ✓
Multi5 ₃	5	2 ×	5 ✓
Multi10 ₁	10	2 ×	12 ×
Multi10 ₂	10	2 ×	10 ✓
Multi10 ₃	10	2 ×	13 ×

✓ and × denote correct and wrong results, respectively.

sIB: For the sIB algorithm, top 2000 words were selected for each dataset according to words' contribution to the mutual information about the documents. Then the sIB algorithm¹ was conducted on word document co-occurrence matrix. The number of clusters was taken to be identical with the number of real categories (ground truth classes).

FSCV: We employed the algorithm presented in Section 3 to determine both the feature subset and the cluster number. The values of K_{\min} and K_{\max} in our algorithm were set as 2 and 15. The final clustering result was computed on full data set D in feature space \hat{F}_k using the sIB algorithm with the cluster number \hat{k} as input.

In Table 4 we present the results of micro-averaged precision for above two procedures (sIB and FSCV) over nine datasets. Moreover, it also summarizes micro-averaged precision results of double clustering algorithm (DC), iterative double clustering algorithm (IDC) and co-clustering algorithm (CoC) taken from (Dhillon et al., 2003; El-Yaniv & Souroujon, 2002; Slonim & Tishby, 2000). The value of cluster number is required to be provided in other clustering algorithms (e.g., DC, IDC, CoC, and sIB).

The size of selected feature subsets with $M_{F,k}$ as the evaluation function for datasets Binary_{1,2,3}, Multi5_{1,2,3}, and Multi10_{1,2,3} are 2252, 2313, 2454, 1591, 1665, 1633, 1513, 1361, and 1488, respectively. Table 5 provides the results of model order identification procedure with different objective functions, $M_{F,k}^{\text{unnorm}}$ and $M_{F,k}$.

Results in Table 4 show that FSCV outperforms DC, IDC, CoC, and sIB if we use average precision to assess their performance. Specifically, FSCV achieved 82.6% average precision, while sIB, CoC, IDC, and DC achieved 81.4%, 79.7%, 74.2%, and 54.2% average precision, respectively. Specifically, when FSCV found

¹ Source code of sIB algorithm is available at <http://www.cs.huji.ac.il/~noamm/>.

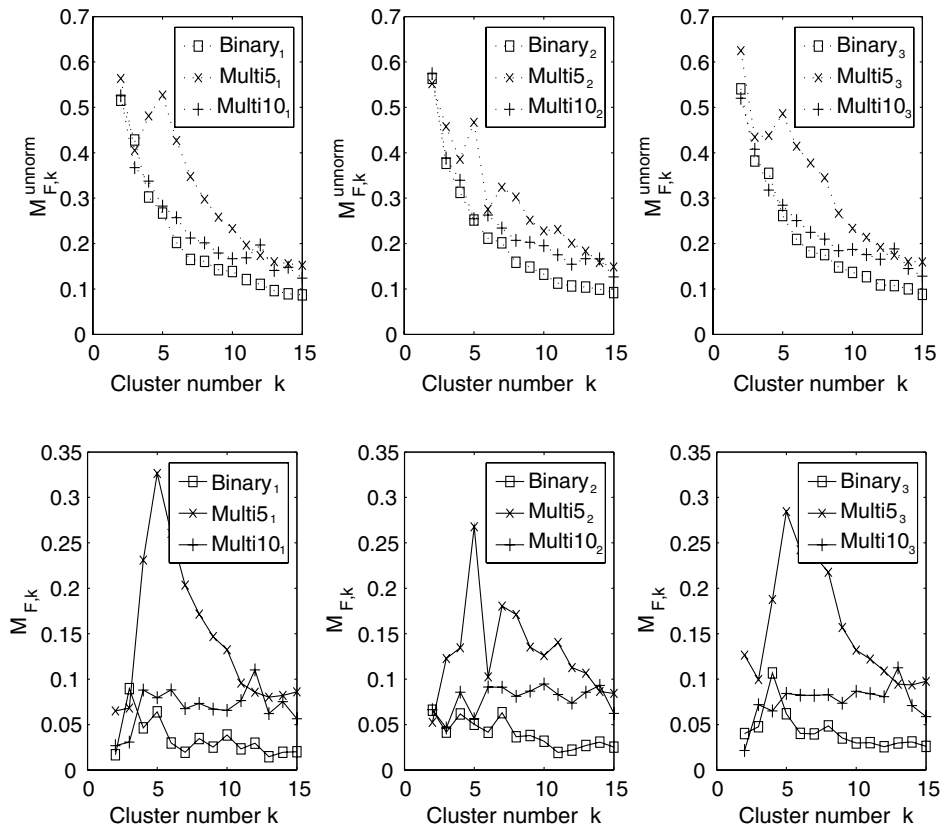


Fig. 1. Results of FSCV algorithm in terms of the score of objective function as a function of cluster number over nine datasets. Top three figures summarize the results using $M_{F,k}^{\text{unnorm}}$ as objective function, and bottom three figures give out the results using $M_{F,k}$ as objective function.

the ground truth cluster number, FSCV performs better than other document clustering algorithms in most of cases. Taking into account no requirement of the predefinition of cluster number in FSCV, we can see that these results are encouraging.

In Table 5, it is shown that the number of real categories can be correctly identified using normalized measure $M_{F,k}$ over five datasets (Binary₂, Multi5₁, Multi5₂, Multi5₃, Multi10₂). On the other four datasets (Binary₁, Binary₃, Multi10₁, Multi10₃), the estimated cluster numbers are very close to the intrinsic values. But the model order identification algorithm with $M_{F,k}^{\text{unnorm}}$ as the evaluation function failed to identify the number of real categories, which chose 2 as the cluster number over all datasets. Comparing the results of $M_{F,k}$ with $M_{F,k}^{\text{unnorm}}$ in Table 5, we can see that $M_{F,k}$ clearly outperforms $M_{F,k}^{\text{unnorm}}$.

Fig. 1 presents the detailed result in terms of the scores of two evaluation criteria as functions of cluster numbers. The scores of $M_{F,k}^{\text{unnorm}}$ decreased while increasing the cluster number k although there was small increase around $k = 5$ on datasets Multi5_{1,2,3}. This indicates that it is necessary to normalize the cluster validity score for avoiding the bias that smaller value of k is to be selected as the cluster number.

5. Conclusions and future work

This paper attacked the problem of feature selection and model order identification for document clustering. The important feature subset and the intrinsic cluster number were determined by optimizing a model selection criterion that evaluated the validity of clustering solutions computed on data subsets generated by resampling. We demonstrated that our algorithm can automatically estimate the cluster number, and the clus-

ter number identified by our algorithm was identical with or very close to the intrinsic value in benchmark data. The overall performance of our algorithm in terms of micro-averaged precision was better than previous document clustering algorithms.

Efficient search strategy is very important for feature selection here, considering the high dimensional search space (about 1300–2600 words) in feature selection procedure. Future work includes the investigation of other search methods, e.g., evolutionary algorithms, to find local optima more efficiently.

Given a large document collection, it is difficult to identify the number of real categories for users when they attempt to categorize the documents. Our algorithm can help to estimate the number of real categories using cluster validation method. After document clustering, it becomes easier to navigate the contents of document collection.

References

- Cutting, D. R., Karger, D. R., Pederson, J. O., & Tukey, J. W. (1992). Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval*.
- Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the 7th ACM SIGKDD conference on knowledge discovery and data mining*.
- Dhillon, I. S., Mallela, S., & Modha, S. (2003). Information-theoretic co-clustering. In *Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining*.
- Ding, C., He, X., Zha, H., Gu, M., & Simon, H. (2001). A min–max cut algorithm for graph partitioning and data clustering. In *Proceedings of the IEEE international conference on data mining*.
- Dy, J. G., & Brodley, C. E. (2000). Feature subset selection and order identification for unsupervised learning. In *Proceedings of the 17th international conference on machine learning*.
- El-Yaniv, R., & Souroujon, O. (2002). Iterative double clustering for unsupervised and semi-supervised learning. *Advances in Neural Information Processing Systems*, 15, 2002.
- El-Hamdouchi, A., & Willett, P. (1986). Hierarchic document classification using Ward's clustering method. In *Proceedings of the 9th annual international ACM SIGIR conference on research and development in information retrieval*.
- Lange, T., Braun, M., Roth, V., & Buhmann, J. M. (2002). Stability-based model selection. *Advances in Neural Information Processing Systems*, 15.
- Law, M. H., Figueiredo, M., & Jain, A. K. (2002). Feature selection in mixture-based clustering. *Advances in Neural Information Processing Systems*, 15.
- Levine, E., & Domany, E. (2001). Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13, 2573–2593.
- Li, T., Ma, S., & Ogihara, M. (2004). Document clustering via adaptive subspace iteration. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval*.
- Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1), 145–150.
- Liu, X., Gong, Y., Xu, W., & Zhu, S. (2002). Document clustering with cluster refinement and model selection capabilities. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval*.
- Modha, D. S., & Spangler, W. S. (2003). Feature weighting in *k*-means clustering. *Machine Learning*, 52(3), 217–237.
- Pantel, P., & Lin, D. (2002). Document clustering with committees. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval*.
- Pudil, P., Novovicova, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, 15, 1119–1125.
- Schütze, H., Silverstein, C. (1997). Projections for efficient document clustering. In *Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval*.
- Slonim, N., Friedman, N., & Tishby, N. (2002). Unsupervised document classification using sequential information maximization. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval*.
- Slonim, N., & Tishby, N. (2000). Document clustering using word clusters via the information bottleneck. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval*.
- Tibshirani, R., Walther, G., Botstein, D., & Brown, P. (2001). Cluster validation by prediction strength. *Technical Report*, Statistics Department, Stanford University.
- Tishby, N., Pereira, F., & Bialek, W. (1999). The information bottleneck method. In *Proceedings of the 37th Allerton conference on communication, control and computing*.
- Vaithyanathan, S., & Dom, B. (1999). Model selection in unsupervised learning with applications to document clustering. In *Proceedings of the 16th international conference on machine learning*.
- Willett, P. (1980). Document clustering using an inverted file approach. *Journal of Information Science*, 2, 223–231.
- Xu, W., & Gong, Y. (2004). Document clustering by concept factorization. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval*.
- Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval*.

- Zha, H., He, X., Ding, C., Gu, M., & Simon, H. D. (2001). Bipartite graph partitioning and data clustering. In *Proceedings of the 10th ACM conference on information and knowledge management*.
- Zamir, O., & Etzioni, O. (1998). Web document clustering: A feasibility demonstration. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*.