Deterministic Annealing Framework in MMMs-Induced Fuzzy Co-Clustering and Its Applicability

Shunnya Oshio†, Katsuhiro Honda†, Seiki Ubukata†, and Akira Notsu†

Osaka Prefecture University, Sakai, Osaka, JAPAN

Summary
Initialization problem is a significant issue in FCM-type clustering models, in which alternative optimization is often started with random initial partitions and can be trapped into local optima caused by bad initialization. The deterministic clustering approach is a practical procedure for utilizing a robust feature of very fuzzy partitions and tries to converge the iterative FCM process to a plausible solution by gradually decreasing the fuzziness degree. In this paper, the initialization sensitivity issue is considered in multinomial mixture models-induced fuzzy co-clustering context and a new approach for implementing the deterministic annealing mechanism to fuzzy co-clustering is proposed. The advantages of the proposed approach against the conventional statistical co-clustering model are demonstrated through some numerical experiments.

Key words:
Fuzzy co-clustering, Multinomial mixture, Deterministic annealing, Initialization problem.

1. Introduction

Fuzzy c-Means (FCM) and its variants [1,2] (in the following, they are called as FCM-type clustering models) are shown to be the basic technique in effectively achieving unsupervised classification in simple iterative processes, in which alternative optimization process generally starts with random initialization. Besides its simple scheme, however, they often suffer from the initialization problem, in which the algorithms may converge to inappropriate local solutions caused by bad initialization.

Deterministic annealing (DA) [3] is a possible way for avoiding local solutions in fuzzy clustering. An advantage of fuzzy partition is its robust feature to noise or outliers and very fuzzy partition may conceal undesirable distortion of prototype assignment. Then, utilizing the robustness of very fuzzy partition, DA starts the FCM process with very fuzzy situation and gradually degrades fuzziness degrees of partitions until it reaches to intended fuzziness degrees. In [3], the entropy-based fuzzification term [4] first plays a role for regularizing the k-Means objective function with a very large fuzzification weight so that it brings a unique solution regardless random initialization, and then, the fuzzification weight is gradually degraded in order to find a plausible solution with the intended fuzzy degree.

This paper considers the implementation of the DA scheme to fuzzy co-clustering, which is a fundamental technique for analyzing cooccurrence relational data such as document-keyword frequencies. Fuzzy Clustering for Categorical Multivariate data (FCCM) [5] is an FCM variant, in which co-clusters of object-item pairs are extracted by estimating two different kinds of fuzzy memberships: object memberships and item memberships. With the goal of extracting familiar object-item pairs, the FCM-type objective function is defined by the aggregation degree of objects and items instead of the within-cluster-error measure of FCM. The fuzzy partition nature of the prototype-less aggregation criterion was achieved by the entropy-based regularization approach [4]. Although the iterative algorithm has similar form to the conventional FCM, the dual fuzzification model has no comparative statistical models, which can be utilized as a guideline, and often needs very careful tuning of two penalty weights in trial and error manner.

Considering the similarity between the FCCM objective function and the pseudo-log-likelihood function of Multinomial Mixture Models (MMMs) [6], Honda et al. proposed Fuzzy Co-Clustering induced by MMMs (FCCMM) [7, 8], where the fuzziness degree can be tuned under comparison with MMMs. The object and item memberships are identified with the probability of each generative class for an object and the probability of appearance of an item in a class, respectively, and they are fuzzified based on different fuzzification principles.

This paper proposes a DA framework for FCCMM, where the fuzziness degree of object memberships are tuned by the K-L information-based regularization approach [9] while the fuzziness degree of item memberships are tuned by a weighting exponent approach [10].

The remaining parts of this paper are organized as follows: Section 2 presents a brief review on the DA scheme in FCM clustering. In Section 3, the DA scheme is implemented to FCCMM. The characteristic features are demonstrated through numerical experiments in Section 4 and a summary conclusion is given in Section 5.

Manuscript received January 5, 2016
Manuscript revised January 20, 2016

2. FCM and Deterministic Annealing

2.1 FCM clustering

Assume that we have n objects with their m-dimensional vector observations xi , i = 1,...,n. In FCM [1,2], the objects are partitioned into C fuzzy clusters with their prototypical centroids bc , c = 1,...,C, in such a way that the within-cluster-errors are minimized:

$$L_{fcm} = \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}^{\theta}\left\|\boldsymbol{x}_i - \boldsymbol{b}_c\right\|^2 , \quad (1)$$

where uci is the fuzzy membership of object i to cluster c and is constrained such that $\sum_{c=1}^{C} u_{ci} = 1$. θ is an exponential weight for fuzzification. The linear objective function of crisp k-Means ($\theta = 1$) was fuzzified by introducing non-linear nature with respect to uci in the FCM objective function.

Miyamoto and Mukaidono [4] introduced the entropy-based regularization concept to fuzzification of k-Means objective function as follows:

$$L_{efcm} = \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\left\|\boldsymbol{x}_i - \boldsymbol{b}_c\right\|^2 + \lambda_u \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci} , \quad (2)$$

where λ_u tunes the degree of fuzziness of object memberships. The larger the λ_u , the fuzzier the partition.

Besides the k-Means clustering concept, Eq. (2) is also identified with a negative log-likelihood function of Gaussian Mixture Models (GMMs), which is composed of C spherical Gaussian components and fixed variances. Then, the fuzziness degree of Eq. (2) can be tuned through comparison with the soft nature of GMMs as a guideline.

2.2 Deterministic Annealing

A significant problem of FCM-type clustering is its sensitivity to initial partitions and can be often trapped into local minima from bad initializations. Deterministic approach is a possible way for avoiding local solutions in fuzzy clustering. Rose et al. introduced a deterministic annealing (DA) [3] concept to a probabilistic data clustering, whose updating formula is equivalent to that of the entropy-based FCM. The fuzzification penalty λ_u is regarded as the temperature parameter and the FCM cost function is deterministically optimized at each temperature sequentially, starting at high temperature. Very fuzzy model with a huge λ_u often brings a unique solution with a smoothened cost function such as uci =1/C for all objects and clusters, and the fuzziness degree is gradually decreased until intended fuzziness.

3. Fuzzy Co-clustering Models and DA-based Fuzzy Co-clustering

3.1 FCCM and Statistical Consideration

Co-clustering is a technique for capturing the intrinsic cluster structures from cooccurrence information among objects and items. Assume that we have a cooccurrence matrix $R = \{r_{ij}\}$ on objects $i = 1,...,n$ and items $j = 1,...,m$, and each element r_{ij} shows the similarity degree among user i and item j , such as frequency of keyword j in document i . Oh et al. [5] proposed FCCM by replacing the within-cluster-deviation of FCM with the within-cluster aggregation degree of each cluster:

$$L_{fccm} = \sum_{c=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{m} u_{ci} w_{cj} r_{ij} - \lambda_u \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci} - \lambda_w \sum_{c=1}^{C}\sum_{j=1}^{m} w_{cj}\log w_{cj} . \quad (3)$$

$U = \{u_{ci}\}$ and $W = \{w_{cj}\}$ are the fuzzy memberships of object i and item j to cluster c , respectively. The entropy terms are the fuzzification penalty in the entropy-based fuzzification approach. u_{ci} and w_{cj} become large in a same co-cluster if object i and item j are highly relevant.

While the sum of u_{ci} is constrained to be 1 in a similar manner to FCM, w_{cj} is constrained as $\sum_{j=1}^{m} w_{cj} = 1$. So, w_{cj} represents the relative responsibility of item j in cluster c . Although this partition concept has close relation with statistical co-clustering such as MMMs, Eq.(3) has no comparative statistical model. Then, fuzzification penalties λ_u and λ_w have no guideline for parameter tuning and must be tuned by trials and errors.

3.2 Fuzzy Co-clustering Induced by MMMs and Deterministic Annealing

Considering the structural similarity between the FCCM objective function and the pseudo-log-likelihood of MMMs, Honda et al. proposed a fuzzy counterpart of MMMs with the K-L information regularization concept [9]. Fuzzy Co-Clustering induced by MMMs (FCCMM) [8] introduced the objective function as:

$$L_{fccmm} = \sum_{c=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\lambda_w}u_{ci}r_{ij}\left(\left(w_{cj}\right)^{\lambda_w}-1\right) + \lambda_u\sum_{c=1}^{C}\sum_{i=1}^{n}u_{ci}\log\frac{\alpha_c}{u_{ci}}, \quad (4)$$

where λ_u and λ_w are the fuzzification penalty weight for object and item memberships, respectively. The additional parameter α_c represents the cluster volume such that $\sum_{c=1}^{C}\alpha_c = 1$.

When $\lambda_u = 1$ and $\lambda_w \to 0$, Eq. (4) is reduced to the pseudo-log-likelihood function of MMMs with:

$$\log w_{cj} \approx \frac{1}{\lambda_w}\left(\left(w_{cj}\right)^{\lambda_w}-1\right). \quad (5)$$

From the object partition viewpoint, the K-L information-based penalty term is a direct extension of the MMMs-based soft object partition. The larger the λ_u, the fuzzier the object partition.

From the item partition viewpoint, because $\lambda_w = 1$ reduces $\frac{1}{\lambda_w}\left(\left(w_{cj}\right)^{\lambda_w}-1\right)$ to a linear function of $\left(w_{cj}-1\right)$, the log function of MMMs likelihood is interpreted as achieving the fuzzification of FCCM aggregation criterion with the non-linear nature of log function. Then, in the same manner with the standard FCM, the fuzziness degree of item partition can be tuned by changing the non-linear degree of $\frac{1}{\lambda_w}\left(\left(w_{cj}\right)^{\lambda_w}-1\right)$ and λ_w plays a role for fuzzification penalty. As λ_w is smaller, $\frac{1}{\lambda_w}\left(\left(w_{cj}\right)^{\lambda_w}-1\right)$ becomes much more non-linear and we have fuzzier item memberships. When $0 < \lambda_w < 1$, the fuzziness degree is smaller than MMMs. Furthermore, when $\lambda_w < 0$, we have much fuzzier item partitions than MMMs.

In the following parts of this paper, two implementation frameworks of the DA scheme in FCCMM are proposed considering the above fuzziness tuning mechanisms. Here, initialization sensitivity on object partition is mainly discussed because item memberships are not essentially responsible for representing cluster partition but just for characterizing each co-cluster.

3.3 DA Implementation by Tuning Object Partition Fuzziness

First, DA implementation is considered by tuning object partition fuzziness λ_u. In [8], it was demonstrated that the frequency of the best object partition becomes larger as the fuzziness degrees of object memberships are larger while too much larger or smaller penalty weights cause computational instabilities. Then, a possible DA process starts from slightly fuzzier situation than the intended fuzziness degrees and is degraded until the model is reduced to the intended one. In general simulated annealing approaches [11], a practical way for decreasing the temperature parameter T_k with iteration index k is:

$$T_{k+1} = \gamma T_k \ (0.8 \le \gamma < 1), \quad (6)$$

where γ is the depletion rate. Based on the same concept, the fuzzification parameters are adjusted. Because the object membership fuzzifier λ_u is directly identifiable with the temperature parameter of the conventional DA clustering model, it can be degraded as:

$$\lambda_{u,k+1} = \max\{\gamma_u \lambda_{u,k}, \lambda_u^{\min}\}, \quad (7)$$

where $0 < \gamma_u < 1$, and $\lambda_u = 1$ corresponds to MMMs.

3.4 DA Implementation by Tuning Item Partition Fuzziness

Although the item fuzziness degree λ_w is designed for tuning the partition fuzziness of item memberships $W = \{w_{cj}\}$, Ref. [8] reported a side effect of λ_w on object partition $U = \{u_{ci}\}$. A smaller λ_w brings a fuzzier item partition but can derive a more crisp object partition. It may be because a fuzzier item memberships can contribute for concealing noise but emphasize the object cluster boundaries, i.e., object partition can be trapped into local maxima. Then, a more crisp item partition is expected to bring a fuzzier object partition and a DA process can be implemented by starting with a larger λ_w and degrading it to a (smaller) intended fuzziness penalty.

In adopting the DA degradation with λ_w, it should be noted that λ_w can take both positive and negative values in contrast with the general annealing parameters, which does not have negative values, i.e., the annealing schedule of Eq. (7) is designed only for positive values. A possible way for tuning λ_w among $\left[\lambda_w^{\min}, \lambda_w^{\max}\right]$ is:

$$\lambda_{w,k+1} = \max\{\gamma_w(\lambda_{w,k} - 2\times\lambda_w^{\min} + \lambda_w^{\max}) + 2\times\lambda_w^{\min} - \lambda_w^{\max}, \lambda_w^{\min}\}, \quad (8)$$

so that $T_{k+1} = \gamma^k \times T_k$ in the interval $\left[0, T^{\max}\right]$ is virtually realized in $\left[\lambda_w^{\min} - \left(\lambda_w^{\max} - \lambda_w^{\min}\right), \lambda_w^{\max}\right]$ with the center $\lambda_w^{\min}$.

3.5 Possible Algorithmic Procedure of FCCMM with DA Implementation

Then, a sample procedure of the proposed algorithm is written as follows:

Algorithm: Fuzzy Co-Clustering induced by Multinomial Mixture models with Deterministic Annealing (FCCMM-DA)

Step 1. Initialize fuzzy memberships u_{ci} and w_{cj} such that they satisfy $\sum_{c=1}^{C} u_{ci} = 1$, $^{\forall} i$ and $\sum_{j=1}^{m} w_{cj} = 1$, $^{\forall} c$. Choose the possible interval of fuzziness penalty weight $\lambda_u \in \left[\lambda_u^{\min}, \lambda_u^{\max}\right]$ and $\lambda_w \in \left[\lambda_w^{\min}, \lambda_w^{\max}\right]$, and termination criterion ε. Let the initial penalties be $\lambda_u = \lambda_u^{\max}$ and $\lambda_w = \lambda_w^{\max}$.

Step 2. Update cluster volumes α_c, $c = 1,...,C$ by

$$\alpha_c = \frac{1}{n}\sum_{i=1}^{n} u_{ci} . \tag{9}$$

Step 3. Update w_{cj}, $c = 1,...,C$, $j = 1,...,m$ by

$$w_{cj} = \left(\sum_{l=1}^{m} \left(\frac{\sum_{i=1}^{n} r_{ij} u_{ci}}{\sum_{i=1}^{n} r_{il} u_{ci}} \right)^{\frac{1}{\lambda_w - 1}} \right)^{-1} . \tag{10}$$

Step 4. Update u_{ci}, $c = 1,...,C$, $i = 1,...,n$ by followings:

For $\lambda_w \neq 0$,

$$u_{ci} = \frac{\alpha_c \exp\left(\frac{1}{\lambda_u \lambda_w} \sum_{j=1}^{m} r_{ij} \left(w_{cj}\right)^{\lambda_w} \right)}{\sum_{l=1}^{C} \alpha_l \exp\left(\frac{1}{\lambda_u \lambda_w} \sum_{j=1}^{m} r_{ij} \left(w_{lj}\right)^{\lambda_w} \right)} . \tag{11}$$

For $\lambda_w = 0$,

$$u_{ci} = \frac{\alpha_c \prod_{j=1}^{m} \left(w_{cj}\right)^{r_{ij}/\lambda_u}}{\sum_{l=1}^{C} \alpha_l \prod_{j=1}^{m} \left(w_{lj}\right)^{r_{ij}/\lambda_u}} . \tag{12}$$

Step 5. Update λ_u and λ_w by Eqs. (7) and (8).

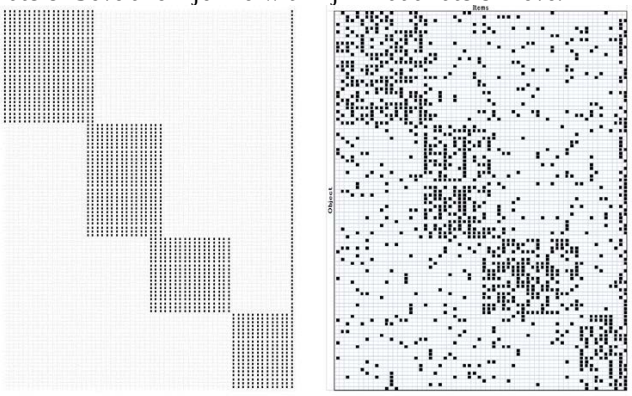
Step 6. If $\max_{c,i} \left| u_{ci}^{NEW} - u_{ci}^{OLD} \right| < \varepsilon$, then stop. Otherwise, return to Step 2.

4. Numerical Experiments

In this section, the characteristic features of the proposed method are demonstrated in numerical experiments.

4.1 Artificial Data Set

A noisy 100×60 artificial cooccurrence matrix R shown in Fig. 1-(b) was used in [7, 8], which was generated from a noise-less R0 with 100 objects (n = 100) and 60 items (m = 60) shown in Fig. 1-(a). R0 = {rij0} and R = {rij} are the base matrix without noise and its noisy variant, whose elements are depicted by black and white cells as rij = 1 and rij = 0, respectively. The noisy matrix R, which includes roughly 4 co-clusters (C = 4) in diagonal blocks while some items are shared by multiple clusters, was generated from R0 by replacing rij0 = 1 with rij = 0 at a rate of 50% and rij0 = 0 with rij = 1 at a rate of 10%.



(a) noiseless R_0 (b) noisy R

Fig.1 Artificial cooccurrence matrices [7,8]

Table 1. Comparison of initialization sensitivity without DA in artificial data: the frequencies of RIu > 0.9 in 210 different trials [8]

		Object penalty λ_u		
		0.5	1.0	2.0
Item Penalty λ_w	0.3	185(88%)	202(96%)	—
	0.0	40(19%)	108(51%)	171(81%)
	-0.3	—	0(0%)	21(10%)

The initialization sensitivity of the FCCMM algorithm is compared with and without annealing mechanisms, where the initial item membership vectors $w_c = \left(w_{ci},..., w_{cm}\right)^T$ of C = 4 clusters were constructed from normalized cooccurrence information vectors $r_i = \left(r_{i1},\ldots, r_{im}\right)^T$ of 4 objects such that $w_c = \left(r_{i1}^*,..., r_{im}^*\right)^T$ and $\sum_j r_{ij}^* = 1$. The FCCMM algorithm with various penalty weight values was applied to the noisy cooccurrence matrix R with initial partitions given by 210 different 4-objects combinations constructed from 10 pre-selected objects, i.e., all trials started from common 210 initialization candidates for fair comparisons. The partition quality is compared with Rand Index (RI) of maximum membership partitions, where RIu

implies the ratio of matching with the ideal object partition of Fig. 1-(a) after maximum membership object partition.

In the previous research [8], without the DA scheme, the frequencies of the clustering results with RIu > 0.9 object partition was reported as Table 1, where "--" means that the algorithm cannot work because of overflow with too fuzzy or too crisp penalty settings. The initialization sensitivity was reduced with larger λ_u and λ_w, i.e., a fuzzier object partition and a crisper item partition can contribute to stable co-clustering.

In the following, the proposed annealing scheme is introduced with the goal of achieving the stable co-clustering features by gradual tuning of fuzziness degrees. First, annealing of object membership fuzziness is considered with fixed item fuzziness, where the fuzziness penalty of object partition is reduced as:

$$\lambda_u = \max\left\{\lambda_u^{\min}, 0.99^k \lambda_u^{\max}\right\}, \quad (13)$$

where k is the iteration index, and the final value was always be guaranteed as $\lambda_u = \lambda_u^{\min}$ in this experiment for comparison purposes. Table 2 shows that the initialization sensitivity of the FCCMM algorithm was efficiently reduced by introducing annealing of object fuzziness degrees and higher quality was achieved with smaller fuzziness degrees compared with Table 1.

Table 2. Comparison of initialization sensitivity with object fuzziness annealing in artificial data: the frequencies of RIu > 0.9 in 210 different trials

		Object penalty λ_u		
	$\lambda_u^{\max}$ ↓ $\lambda_u^{\min}$	1.0 ↓ 0.5	2.0 ↓ 1.0	3.0 ↓ 2.0
Item Penalty λ_w	0.3	200(95%)	207(99%)	—
	0.0	86(41%)	161(77%)	198(94%)
	-0.3	—	21(10%)	29(14%)

Table 3. Comparison of initialization sensitivity with item fuzziness annealing in artificial data: the frequencies of RIu > 0.9 in 210 different trials

		Object penalty λ_u		
	$\lambda_w^{\max} \rightarrow \lambda_w^{\min}$	0.5	1.0	2.0
Item Penalty λ_w	0.5→0.3	196(93%)	208(99%)	—
	0.3→0.0	182(87%)	196(93%)	207(99%)
	0.0→-0.3	—	51(24%)	90(43%)

Second, annealing of item membership fuzziness is considered with fixed object fuzziness, where the fuzziness penalty of item partition is tuned as:

$$\lambda_w = \max\left\{\lambda_w^{\min}, 0.99^t \times 2(\lambda_w^{\max} - \lambda_w^{\min}) + \lambda_w^{\min} - (\lambda_w^{\max} - \lambda_w^{\min})\right\}. \quad (14)$$

Here, λ_w was replaced with $\lambda_w = 0$ in case of $|\lambda_w| < 0.05$ for avoiding computational overflow. t is the iteration index, and the final value was always be guaranteed as $\lambda_w = \lambda_w^{\min}$ in this experiment for comparison purposes. Table 3 shows again that the initialization sensitivity of the FCCMM algorithm was efficiently by introducing annealing of item fuzziness degrees and higher quality was achieved with higher fuzziness degrees compared with Table 1.

4.2 Document Clustering: *citeseer*

4.2.1 Comparison of Initialization Sensitivities

A document clustering experiment was performed with a benchmark document data set. *citeseer* is a famous benchmark document data set, which includes 3312 text documents composed of 3703 terms, and is available from LINQS webpage of Statistical Relational Learning Group UMD (http://linqs.cs.umd.edu/projects//index.shtml). This dataset consists of 6 different document collections: *Agents, AI, DB, IR, ML, and HCI*, and the goal is to classify the documents into the 6 intrinsic classes by unsupervised clustering. The observed values r_{ij} indicate whether each term appears in each document, where $r_{ij} = 1$ implies a term j appears in a document i, on the other hand, $r_{ij} = 0$ implies a term j doesn't appear in a document i.

Table 4. citeseer (3312 documents × 3703 terms)

Class name	Objects
DB	701
IR	668
Agents	596
ML	590
HCI	508
AI	249

First, the initialization sensitivity was compared. Table 5 compared the average RI_u in 50 different random initialization without DA.

Table 5. Comparison of initialization sensitivity without DA in citeseer: the average RIu in 50 different trials

		Object penalty λ_u		
		0.5(0.9)	1.0	2.0
Item Penalty λ_w	0.3	0.875	0.790	0.738
	0.0	(0.765)	0.769	0.800
	-0.3	—	—	—

Table 5 implies that better performances were achieved with a larger λ_u (fuzzier *object* memberships) or smaller λ_w (fuzzier *item* memberships, which also bring more crisp *object* memberships) in this experiment. In case of $(\lambda_u, \lambda_w) = (2.0,0.3)$, we can obtain only degraded solutions because it caused only one cluster (all objects were equally shared by all clusters) and could not find any cluster co-structures. Moreover, in case of $\lambda_u = -0.3$, some memberships were overflowed in Eq. (10) or Eq. (11). Likewise, in case of $(\lambda_u, \lambda_w) = (0.5,0.0)$, we could find any cluster co-structure, so the result of $(\lambda_u, \lambda_w) = (0.9,0.0)$ instead of $(\lambda_u, \lambda_w) = (0.5,0.0)$ is presented.

Next, the initialization sensitivity is also compared by introducing annealing schemes. First, the effect of annealing of object membership fuzziness is considered. Table 6 shows that the average RI_u of the clustering results was improved rather than Table 5. The effect of bad initialization was reduced by introducing annealing of object fuzziness degrees, however, in case of $(\lambda_u, \lambda_w) = (2.0 \rightarrow 1.0,0.0)$, the value became worse than $(\lambda_u, \lambda_w) = (1.0,0.0)$ in Table 5. It may be because the result in $(\lambda_u, \lambda_w) = (2.0,0.0)$ is very fuzzy, so it wasn't improved by introducing annealing scheme.

Table 6. Comparison of initialization sensitivity with object fuzziness annealing in citeseer: the average RIu in 50 different trials

		Object penalty λ_u		
	$\lambda_u^{\max}$ ↓ $\lambda_u^{\min}$	1.0 ↓ 0.5(0.9)	2.0 ↓ 1.0	3.0 ↓ 2.0
Item Penalty λ_w	0.3	0.880	0.738	0.738
	0.0	(0.766)	0.789	0.810
	-0.3	—	—	—

Second, we consider the effect of annealing of item membership fuzziness. Here, λ_w was replaced with $\lambda_w = 0$ in case of $|\lambda_w| < 0.1$ for avoiding computational overflow in this experiment.

Table 7. Comparison of initialization sensitivity with item fuzziness annealing in citeseer: the average RIu in 50 different trials

		Object penalty λ_u		
	$\lambda_w^{\max} \rightarrow \lambda_w^{\min}$	0.5(0.9)	1.0	2.0
Item Penalty λ_w	0.5→0.3	0.881	0.738	0.738
	0.3→0.0	(0.865)	0.868	0.840
	0.0→-0.3	—	—	—

Table 7 shows again that the average RIu of the clustering results was improved rather than Table 5. The effect of bad initialization was also reduced by introducing annealing of item fuzziness degrees.

4.2.2 Performance Comparative Evaluation

Next, performance evaluation was studied with F-measure. F-measure is a measure of test's accuracy, which is used often in document classification. The score can be interpreted as weighted average of the precision and recall, where it reaches its best value at 1 and worst at 0. The meaning of each value is as follows:

TP: A success in data reporting in which a test result properly indicates presence of a condition

FP: An error in data reporting in which a test result improperly indicates presence of a condition

FN: An error in data reporting in which a test result improperly indicates no presence of a condition

TN: A success in data reporting in which a test result properly indicates no presence of a condition

Table 8. A simple example of F-measure

		Observed value (cluster label)	
		1	0
True value (class label)	1	*TP*	*FN*
	0	*FP*	*TN*

Precision, Recall and F-measure are calculated as:

$$\Pr ecision = \frac{TP}{TP + FP}, \qquad \mathrm{Re}\, call = \frac{TP}{TP + FN},$$

$$F - measure = \frac{2 \times \Pr ecision \times \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call}.$$

In this experiment, the binary classification criterion is extended to a multi-class one. The following two experimental results are compared, which corresponds to RI = 0.741 in Table 5 for (λ_u, λ_w) = (1.0 , 0.0) and RI = 0.767 in Table 6 for (λ_u, λ_w) = (2.0→1.0 , 0.0). Although these values are almost same from the RI viewpoint, the confusion matrices are slightly different as follows. Here, a confusion matrix was constructed such that each of objects with a class label are classified into each of clusters. Additionally, they used same initialization conditions.

The purpose of this experiment is to obtain high quality co-cluster structure by gradual tuning of fuzziness degrees. Table 9 and Table 11 show the cross tabulation of object class-cluster matching, where ideal results are the number of diagonal components (TP).

First, Table 9 implies objects with class 2 (IR) , class 3 (Agents) and class 4 (ML) were shared by multiple clusters, and the result is not an ideal one.

Table 9. Correlation between class label and cluster label for (λ_u , λ_w) = (1.0 , 0.0)

	cluster	1	2	3	4	5	6	sum
Class label	1	**210**	174	39	93	54	131	701
	2	126	**149**	67	137	124	65	668
	3	164	93	**138**	75	48	78	596
	4	67	125	16	**179**	82	121	590
	5	91	89	122	60	**121**	25	508
	6	38	55	21	49	30	**56**	249

Table 10 implies a confusion matrix in conventional FCCMM and the result of F-measure.

Table 10. Confusion matrix for (λ_u , λ_w) = (1.0 , 0.0)

		Observed value (cluster label)	
		1	0
True value (class label)	1	853	491
	0	486	1482

$$\Pr ecision = \frac{853}{853+486} = 0.637, \quad \mathrm{Re}\, call = \frac{853}{853+491} = 0.635$$

$$F-measure = \frac{2\times 0.637\times 0.635}{0.637+0.635} = 0.636$$

853 in Table 10 (TP) means the sum of the matching class label with maximum membership object partition in all clusters.

Second, Table 11 implies the result with deterministic annealing.

Table 11. Correlation between class label and cluster label for (λ_u , λ_w) = (2.0→1.0 , 0.0)

	cluster	1	2	3	4	5	6	sum
Class label	1	**235**	191	37	36	37	165	701
	2	92	**232**	52	163	90	39	668
	3	164	53	**241**	49	37	52	596
	4	49	105	10	**220**	88	118	590
	5	89	27	201	34	**142**	15	508
	6	38	44	26	43	27	**71**	249

The number of diagonal components is improved comparing with Table 9. Especially, many objects with class 3 (Agents) are included in cluster 3. Moreover, in Table 9, the number of objects assigned to cluster 3 with class label 3 is not maximum in all clusters. On the other hand, in Table 11, the number is maximum in all clusters, i.e., higher quality was achieved with DA.

Table 12 implies a confusion matrix in FCCMM with DA and the result of F-measure.

Table 12. Confusion matrix for (λ_u , λ_w)=(2.0→1.0, 0.0)

		Observed value (cluster label)	
		1	0
True value (class label)	1	1141	466
	0	432	1273

$$\Pr ecision = \frac{1141}{1141+432} = 0.725, \quad \mathrm{Re}\, call = \frac{1141}{1141+466} = 0.710$$

$$F-measure = \frac{2\times 0.725\times 0.710}{0.725+0.710} = 0.718$$

1141 in Table 12 (TP) means the sum of the matching class label with maximum membership objet partition in all clusters and it is possible to confirm the validity compared with Table 10.

These results indicate the availability of deterministic annealing in document co-cluster analysis. However, optimal fuzzy degrees are dependent on dataset, and a future work includes the criterion for selecting the plausible fuzziness degrees.

5. Conclusion

In this paper, a novel DA framework for MMMs-induced fuzzy co-clustering was proposed, where the fuzziness degree of object partition is gradually degraded so that the robust feature of fuzzier partition is exploited in more crisp situations. The DA-based FCM concept was realized in two approaches: direct tuning of object partition fuzziness and indirect tuning through a side effect of item membership fuzzification, i.e., crisper item memberships can contribute to stable object partition. The experimental results demonstrated that the both DA approaches work well for deriving appropriate solutions more often than the conventional model without DA schemes.

A possible future work includes the development of a better design of annealing schedules for achieving more effective operation of the DA framework. Another direction of future study is to investigate the influences of the DA schemes on the interpretability of co-cluster solutions especially from the item fuzziness tuning view point. Finally, in order to overcome several drawbacks of statistical co-clustering models, MMMs has been extended to Dirichlet mixture [12, 13, 14] in Statistics society. It is also a promising challenge to extend the proposed FCCMM algorithm to a Dirichlet mixture induced co-clustering model.

Acknowledgments

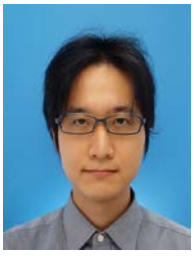
This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology, Japan, under Grant-in-Aid for Scientific Research (26330281).

References

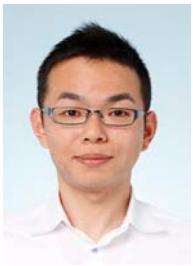
[1] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, 1981.
[2] S. Miyamoto, H. Ichihashi and K. Honda, Algorithms for Fuzzy Clustering, Springer, 2008.
[3] K. Rose, E. Gurewitz and G. Fox, "A deterministic annealing approach to clustering", Pattern Recognition Letter, vol. 11, pp. 589-594, 1990.
[4] S. Miyamoto and M. Mukaidono, "Fuzzy c-means as a regularization and maximum entropy approach", Proc. of the 7th International Fuzzy Systems Association World Congress, vol. 2, pp. 86-92, 1997.
[5] C.-H. Oh, K. Honda and H. Ichihashi, "Fuzzy clustering for categorical multivariate data", Proc. of Joint 9th IFSA World Congress and 20th NAFIPS International Conference, pp. 2154-2159, 2001.
[6] L. Rigouste, O. Cappé and F. Yvon, "Inference and evaluation of the multinomial mixture model for text clustering", Information Processing and Management, vol. 43, no. 5, pp. 1260-1280, 2007.
[7] K. Honda, S. Oshio and A. Notsu, "FCM-type fuzzy co-clustering by K-L information regularization", Proc. of 2014 IEEE International conference on Fuzzy Systems, pp. 2505-2510, 2014.
[8] K. Honda, S. Oshio and A. Notsu, "Fuzzy co-clustering induced by multinomial mixture models", Journal of Advanced Computational Intelligence and Intelligent Informatics, vol. 19, no. 6, pp. 717-726, 2015.
[9] K. Honda and H. Ichihashi, "Regularized linear fuzzy clustering and probabilistic PCA mixture models", IEEE Transactions on Fuzzy Systems, vol. 13, no. 4, pp. 508-516 2005.
[10] K. Honda, S. Oshio and A. Notsu, "Item membership fuzzfication in fuzzy co-clustering based on multinomial mixture concept", Proc. of 2014 IEEE International Conference on Granular Computing, pp. 94-99, 2014.
[11] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, "Optimization by simulated annealing," Science, vol. 4598, no. 220, pp. 671-680, 1983.
[12] I. Holmes, K. Harris and C. Quince, "Dirichlet multinomial mixtures: generative models for microbial metagenomics", PLoS ONE, vol. 7, no. 2, e30126, 2012.
[13] K. Sjölander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I. Saira Mian and D. Haussler, "Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology", Computer Applications in the Biosciences, vol. 12, no. 4, pp. 327-345, 1996.
[14] X. Ye, Y.-K. Yu and S. F. Altschul, "Compositional adjustment of Dirichlet mixture priors", Journal of Computational Biology, vol. 17, no. 12, pp. 1607-1620, 2010.

Shunnya Oshio is currently the graduate student of Faculty of Engineering in the Department of Computer Science and Intelligent Systems. His research interests include fuzzy clustering and data mining.

Katsuhiro Honda received the B.E., M.E., and D.Eng. degrees in industrial engineering from Osaka Prefecture University, Osaka, Japan in 1997, 1999, and 2004, respectively.
From 1999 to 2013, he was a Research Associate, Assistant Professor and Associate Professor at Osaka Prefecture University, where he is currently a Professor in the Department of Computer Science and Intelligent Systems. His research interests include hybrid techniques of fuzzy clustering and multivariate analysis, data mining with fuzzy data analysis, and neural networks.



Seiki Ubukata received the B.E., M.I.S., and D. Information Science degrees from Hokkaido University in 2007, 2009, and 2014, respectively.
From 2014 to 2015, he was an Assistant Professor at Osaka University. From 2015, he is currently an Assistant Professor in the Department of Computer Science and Intelligent Systems at Osaka Prefecture University. His research interests include fuzzy clustering, data mining, rough set theory and agent-based simulation.



Akira Notsu received the B.E., M.I., and D. Informatics degrees from Kyoto University in 2000, 2002, and 2005, respectively.
From 2005 to 2012, he was a Research Associate and Assistant Professor at Osaka Prefecture University, where he is currently an Associate Professor in the Department of Computer Science and Intelligent Systems. His research interests include agent-based social simulation, communication networks, game theory, human-machine interface, and cognitive engineering.