A Noise Rejection Mechanism for pLSA-induced Fuzzy Co-clustering

Katsuhiro Honda*, Keita Hoshii*, Seiki Ubukata* and Akira Notsu†
*Graduate School of Engineering
Osaka Prefecture University, Sakai, Osaka 599-8531, Japan
Email: { honda, subukata } @cs.osakafu-u.ac.jp
†Graduate School of Humanities and Sustainable System Sciences
Osaka Prefecture University, Sakai, Osaka 599-8531, Japan
Email: notsu@cs.osakafu-u.ac.jp

***Abstract*—Noise fuzzy clustering is a useful scheme for analyzing intrinsic data structures through robust estimation of fuzzy c-partition. In this paper, a novel noise rejection scheme for improving fuzzy co-clustering is proposed, which is useful in such cooccurrence information analysis as document classification, where multi-topic co-cluster structure extraction by probabilistic latent semantic analysis (pLSA) is achieved through rejection of the influences of noise objects. Supported by the uniform noise distribution concept in noise fuzzy clustering, a noise cluster having uniform item occurrence probabilities is newly introduced into the pLSA-induced fuzzy co-clustering model. Several numerical experiments demonstrate the advantage of tuning the noise sensitivity of the pLSA-induced objective function.**

I. Introduction

Fuzzy clustering is an effective method for capturing characteristics of data sets by revealing their intrinsic structures, and enables more flexible data analysis than crisp partitioning through estimation of ambiguous classification boundaries utilizing fuzzy memberships [1]. Fuzzy c-Means (FCM) and its variants [2], [3] incorporate the concept of fuzzy partitioning into k-Means [4]. By updating cluster centroids and object memberships using a process similar to the EM algorithm [5], such as k-Means, we can obtain classification results as the local optimal solutions of objective function minimization.

Since FCM adopts the clustering criterion of squared Euclidian distances among objects and cluster centroids, FCM sometimes yields poor results under the influences of noise or outliers in a manner similar to other least square models [6]. Noise fuzzy clustering [7], [8] has been proposed for obtaining robust cluster centroids without noise influences, where the unfavorable influences are reduced by introducing an additional noise cluster for rejecting noise objects in extraction of normal clusters. Assuming that the noise cluster is equidistant from all objects, noise objects distant from all centroids are dumped into the noise cluster.

In the analysis of cooccurrence information of $object \times item$ such as frequency information of $document \times keyword$ in document analysis and purchase history transactions of $customer \times product$ in market analysis, the purpose is not only the extraction of clusters of similar objects but also the feature analysis of clusters by correlating remarkable items in each cluster. Fuzzy c-Means-based fuzzy co-clustering [9], [10] extracts co-clusters, which are composed of the set of strongly related pairs of objects and items, by estimating not only object memberships but also item memberships through maximization of the cluster-wise aggregation degrees instead of the FCM clustering criterion. Fuzzy co-clustering induced by Multinomial Mixture Models (FCCMM) [11] is another direction of co-clustering, which is theoretically supported by the similarity between the aggregation degree criterion of FCM-based fuzzy co-clustering and the pseudo-likelihood function for the maximum likelihood estimation in Multinomial Mixture Models (MMMs) [12]. With the goal of reducing the influences of noise objects in fuzzy co-clustering, FCCMM was extended to a robust model [13], where the noise clustering concept is utilized with equal occurrence probabilities instead of uniform noise distribution.

In addition to uni-topic partition models, such as MMMs and Dirichlet mixtures [14], *topic models* have been shown to be more powerful in revealing intrinsic topic structures in document analysis. Probabilistic Latent Semantic Analysis (pLSA) [15] is a representative topic model, and shows high classification performances by enhancing the occurrence probabilities of objects and items with MMMs from the viewpoint of latent topic estimation. Recently, pLSA was further extended to pLSA-induced fuzzy co-clustering [16], [17], which improves the interpretability and the initialization robustness of pLSA solutions by adjusting the intrinsic fuzziness degree of the dual partition nature. Additionally, it was also reported that the partition quality of co-clusters can be improved compared to pLSA under the deterministic annealing approach [18].

In this paper, a novel noise rejection scheme for pLSA-induced fuzzy co-clustering is proposed for further improving its partition quality by introducing the noise fuzzy clustering concept.

The main contribution of this paper is summarized as follows:

- The conventional pLSA-induced fuzzy co-clustering model is enhanced into a robust clustering model, where a noise cluster is introduced in a similar manner to the noise rejection model for FCCMM.

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

- By introducing a noise cluster having uniform item occurrence probabilities, noise objects are shown to be dumped into the noise cluster.
- The partition quality is demonstrated to be improved through several numerical experiments.

The remaining parts of this paper are organized as follows: Section II presents a brief review of several fuzzy clustering models induced by statistical clustering concepts. Section III introduces noise fuzzy clustering schemes and proposes a novel noise rejection scheme for pLSA-induced fuzzy co-clustering. The characteristic features of the proposed model are demonstrated through several numerical experiments in Section IV and the summary conclusion is given in Section V.

II. Probabilistic Mixture-induced Fuzzy Clustering Models

A. C-means Clustering Models

Assume that the goal is to partition n objects $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})^\top$, $i = 1, \ldots, n$ with m-dimensional observation vectors into C clusters. In C-means clustering, the prototype of cluster c is defined as the mean vector $\boldsymbol{b}_c = (b_{c1}, \ldots, b_{cm})^\top$, and the fuzzy membership of object i to cluster c is represented by u_{ci} ($u_{ci} \in [0, 1]$), which generally follows the probabilistic constraint of $\sum_{c=1}^{C} u_{ci} = 1$. Under different partition assumptions, we have several variants of the C-means clustering family, where their clustering criteria are constructed using squared Euclidean distances among objects and mean vectors.

1) Crisp Partition: In k-Means (Hard c-Means) [4], the objective function to be minimized is defined as a linear function with respect to memberships u_{ci}, and crisp memberships $u_{ci} \in \{0, 1\}$ are then given even if these memberships are generalized into such fuzzy values as $u_{ci} \in [0, 1]$:

$$L_{km} = \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \|\boldsymbol{x}_i - \boldsymbol{b}_c\|^2, \tag{1}$$

where Voronoi regions are obtained by dividing the data space according to perpendicular bisectors of the line segments connecting cluster centroids. In crisp partitions, since noise and outliers are also assigned to either of clusters with the highest memberships ($u_{ci} = 1$), cluster centroids can often be strongly distorted due to noise influences, which also give rise to many local minima, compared to the case where no outliers are present.

2) Probabilistic Fuzzy Partition: Fuzzy partition is an extension of crisp partition, where the responsibility of each object in cluster centroid calculation is fairly shared with multiple clusters considering fuzzy memberships. In Bezdek's Fuzzy c-Means (FCM) [2], the k-means objective function was nonlinearized with respect to membership u_{ci} for realizing fuzzy C-partitions by introducing weighting exponent θ ($\theta > 1$) on memberships u_{ci} as:

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

where $\theta = 2$ is generally recommended. A smaller θ yields a crisper partition, and the model is reduced to the conventional k-means with $\theta = 1$. On the other hand, a larger θ yields a fuzzier partition. The object with $\boldsymbol{x}_i = \boldsymbol{b}_c$ has the maximum membership $u_{ci} = 1$ to the cluster c, whereas outliers distant from all centroids tend to have relatively smaller memberships $u_{ci} \to 1/C$ to all clusters. Then, the unfavorable influences of noise to cluster centroids can be reduced.

In addition to Bezdek-type fuzzification, Miyamoto *et al.* [19] proposed a regularization approach by introducing a nonlinear penalty into the k-Means objective function. When the (negative) entropy term is added with adjustable weight λ, the entropy regularized FCM (eFCM) objective function is given as:

$$L_{efcm} = \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \|\boldsymbol{x}_i - \boldsymbol{b}_c\|^2 + \lambda \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \log u_{ci}, \tag{3}$$

where, as the penalty weight λ ($\lambda > 0$) increases, the partition becomes fuzzier. In contrast to the Bezdek-type model, no object has the maximum membership $u_{ci} = 1$ at the point of $\boldsymbol{x}_i = \boldsymbol{b}_c$ while the membership at a point far from $\boldsymbol{b}_c$ can approach $u_{ci} \to 1$ [20].

The objective function implies a close connection with Gaussian Mixture Models (GMMs) [21]. Hathaway [22] interpreted the pseudo-log-likelihood function in GMMs as the modified hard c-means objective function with an entropy penalty term such that GMMs is a C-means family with soft partition natures. From the GMMs viewpoint, the fuzzification weight λ is identified with the double variances of Gaussian components, and we can then adjust the fuzziness degree in eFCM through comparison with GMMs. For example, an FCM model can be fuzzier than GMMs when the weight λ is larger than twice the cluster-wise variances.

B. Fuzzy Co-clustering Induced by Probabilistic Concepts

Given an $n \times m$ cooccurrence matrix $R = \{r_{ij}\}$ of cooccurrence information among n objects and m items, instead of the coordinate values in the multidimensional data space, the goal of cluster structure analysis can be to extract C pairwise co-clusters of familiar objects and items.

1) Fuzzy Clustering of Categorical Multivariate Data (FCCM): In order to extend the FCM clustering concept to co-clustering tasks, Fuzzy Clustering for Categorical Multivariate data (FCCM) [9] adopts two different types of fuzzy memberships: u_{ci} for the membership of object i to cluster c, and w_{cj} for the membership of item j to cluster c. With the goal of extracting dense co-clusters, the FCM clustering criterion was replaced with the following aggregation degree of objects and items:

$$J = \sum_{c=1}^{C} \sum_{i=1}^{n} \sum_{j=1}^{m} u_{ci} w_{cj} r_{ij}, \tag{4}$$

where the criterion becomes large when the familiar pair of object i and item j having large cooccurrence r_{ij} belongs to the same cluster, i.e., cluster c.

The original FCCM objective function was defined by adopting entropy-based regularization as [9]:

$$L_{fccm} = \sum_{c=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{m} u_{ci} w_{cj} r_{ij} - \lambda_u \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci} - \lambda_w \sum_{c=1}^{C}\sum_{j=1}^{m} w_{cj}\log w_{cj}. \quad (5)$$

Here, λ_u and λ_w are the penalty weights, which tune the fuzziness of the dual C-partition of objects and items. As the weights increase, the partition becomes fuzzier. Here, u_{ci} and w_{cj} should follow different types of constraints in order to avoid a trivial *whole data cluster*. Object memberships u_{ci} can be constrained under $\sum_{c=1}^{C} u_{ci} = 1$ likewise FCM, while item memberships should be constrained under $\sum_{j=1}^{m} w_{cj} = 1$ in order to represent the relative importance among items in each cluster.

Although the FCCM model is a simple extension of the FCM clustering concept to co-clustering, there is no counterpart of probabilistic mixture models, i.e., we have no theoretical supports of corresponding statistical models in fuzziness tuning. Then, it is often difficult to tune fuzziness weights λ_u and λ_w by trial and error in real implementation.

2) Fuzzy Co-Clustering Induced by Multinomial Mixture Models (FCCMM): Honda *et al.* [11] proposed Fuzzy Co-Clustering induced by Multinomial Mixture models (FCCMM), which is a fuzzy co-clustering model induced from a statistical co-clustering model of Multinomial Mixture Models (MMMs) [12]. FCCMM can achieve fuzzy co-clustering by introducing an adjustable fuzziness weight to the pseudo-log-likelihood functions of MMMs as follows:

$$L_{fccmm} = \sum_{c=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{m} u_{ci} r_{ij} \log w_{cj} + \lambda_u \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log \frac{\alpha_c}{u_{ci}}, \quad (6)$$

where λ_u is the penalty weight, which tunes the fuzziness degree of the object partition, and the K-L information term contributes to fuzzification of object memberships u_{ci}, whereas, unlike FCCM, fuzzification of item memberships w_{cj} is realizes by the non-linearity of the $\log$ function. α_c represents the volume of cluster c, i.e., the priori probability for cluster c. The model is reduced to the conventional MMMs when $\lambda_u = 1$. As the weight increases, the object partition becomes fuzzier. Besides the fuzziness degree of object partition, the item partition fuzziness can also be tuned by modifying the non-linearity of the $\log$ function.

In Ref. [11], it was demonstrated that the interpretability of co-cluster partition can be improved, as compared to MMMs, by properly tuning the degree of the partition fuzziness.

3) pLSA-induced Fuzzy Co-clustering Model: Recently, another statistical co-clustering model of *topic model* has become popular, as compared to MMMs and its variants, for document analysis. In the mixture of uni-topic models such as MMMs and Dirichlet mixtures [14], maximum likelihood estimation is performed by assuming that each object belongs to a single class, where item j occurs with a probability of w_{cj}. On the other hand, probabilistic Latent Semantic Analysis (pLSA) [15] assumes that each object i can be connected with certain C topics at ratios of u_{ci}. Then, the probability p_{ij} of occurring item j in conjunction of object i is given as:

$$p_{ij} = \sum_{c=1}^{C} u_{ci} w_{cj}. \quad (7)$$

For example, in document classification, a document (object) can be connected with multiple topics (clusters), where certain words (items) occur with a topic, but other words may be responsible for different topics.

Then, the log-likelihood function to be maximized is defined as:

$$L_{plsa} = \sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} \log \sum_{c=1}^{C} u_{ci} w_{cj}. \quad (8)$$

Since the updating rules cannot be directly obtained by differentiating this likelihood function, the lower bound of L_{plsa} is estimated by maximizing the following pseudo-log-likelihood function under the support of Jensen's inequality [23]:

$$L_{plsa'} = \sum_{c=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{m} \phi_{cij} r_{ij} \log(u_{ci} w_{cj}) - \sum_{c=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{m} \phi_{cij}\log \phi_{cij}, \quad (9)$$

where ϕ_{cij} corresponds to a posteriori probability of class c given item j with object i, and is a latent variable used to update u_{ci} and w_{cj}. Since the second term is interpreted as the entropy penalty term, which adjusts the fuzziness degree of latent variable ϕ_{cij}, the intrinsic fuzziness degree of the pLSA partition can be adjusted by tuning the responsibility of the entropy term.

Then, the pLSA-induced Fuzzy Co-clustering Model (FCCpLSA) [16], [17] introduced the fuzzification weight λ_ϕ to the entropy term, and the objective function to be maximized is defined as:

$$L_{fccplsa} = \sum_{c=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{m} \phi_{cij} r_{ij} \log(u_{ci} w_{cj}) - \lambda_\phi \sum_{c=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{m} \phi_{cij}\log \phi_{cij}, \quad (10)$$

where a larger λ_ϕ yields a fuzzier dual partition of objects and items.

In Ref. [17], it was demonstrated that the partition quality of co-clusters can be improved as compared to pLSA by starting with a higher fuzziness degree and by gradually decreasing under the deterministic annealing approach [18].

III. Noise Rejection in Fuzzy Clustering and Fuzzy Co-clustering

Noise rejection is an important issue in applying fuzzy clustering to real world problems. According to the least squares principle, the conventional frameworks of C-means clustering often suffers from the unfavorable influences of noise and outliers. In this section, the noise fuzzy clustering scheme proposed by Davé [7], [8] is reviewed and is extended to robust fuzzy co-clustering.

A. Noise Fuzzy Clustering

In order to reject the influences of noise in FCM clustering, Davé [7] proposed noise fuzzy clustering, which introduced a noise cluster for dumping noise objects far from all cluster centroids. Assuming that the Cth cluster is a noise cluster, to which all objects are equidistant (γ), the FCM objective function of Eq.(2) is modified as:

$$L_{nfcm} = \sum_{c=1}^{C-1}\sum_{i=1}^{n} u_{ci}^{\theta}\|\boldsymbol{x}_i - \boldsymbol{b}_c\|^2 + \gamma\sum_{i=1}^{n} u_{Ci}^{\theta}, \qquad (11)$$

where cluster C is the noise cluster, whereas other clusters, i.e., clusters $1,\dots,C-1$, are normal clusters. Under the constraint of $\sum_{c=1}^{C} u_{ci} = 1$, the sum of memberships to normal clusters is $\sum_{c=1}^{C-1} u_{ci} < 1$ and u_{Ci} is the membership degree to the noise cluster. Since each object should belong to at most one cluster including the noise cluster, all clusters are mutually exclusive and noise objects have almost zero memberships to the normal clusters when they are more distant than γ from all normal clusters. Then, the centroids of the normal clusters are robustly estimated without influences of noise.

In Ref. [8], noise distance γ is recommended to be calculated considering the variance determined by the following equation:

$$\gamma = \delta\left[\frac{\sum_{c=1}^{C}\sum_{i=1}^{n}\|\boldsymbol{x}_i - \boldsymbol{b}_c\|^2}{n\times C}\right], \qquad (12)$$

where δ is the noise sensitivity weight and is often set as $\delta = 1$.

Here, from the statistical clustering viewpoint, the noise clustering model can be regarded as estimation of the mixture distributions considering Gaussian components and a uniform noise distribution. As shown in Fig.1, each normal cluster is assumed to have a Gaussian peak at its centroid, whereas the noise cluster has equal probabilities over all regions [24]. If the Gaussian probability of the occurrence of an object is larger than the noise probability, then the object is assigned to a normal cluster.

In the remaining part of this paper, noise rejection schemes are designed for fuzzy co-clustering following the above statistical interpretation of noise fuzzy clustering.

B. Noise Rejection in FCCMM

First, let us review the noise rejection scheme in FCCMM, which is the fuzzy co-clustering based on the concept of uni-topic mixture models. Honda *et al.* [13] proposed a noise

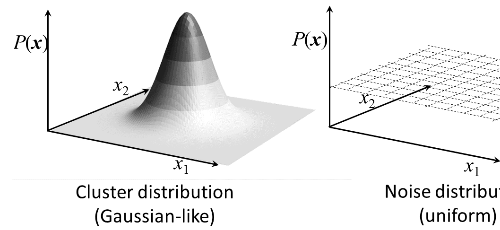


Fig. 1. Image of noise distribution [24]

fuzzy co-clustering model supported by the uniform noise distribution. Although each co-cluster is assumed to have high aggregation of objects and items with a certain bias of item cooccurrence probabilities, noise objects are expected to cooccur with random items under uniform item occurrence probabilities. Then, assuming the Cth cluster to be a noise cluster, item occurrence probabilities are uniform for all items, i.e., $w_{Cj} = \frac{1}{m}$, and the aggregation degree in the noise cluster can be represented by $\left(\sum_{j=1}^{m} r_{ij}\log\frac{1}{m}\right)$.

The FCCMM objective function of Eq.(6) is modified as:

$$\begin{aligned} L_{nfccmm} = &\sum_{c=1}^{C-1}\sum_{i=1}^{n} u_{ci}\left(\sum_{j=1}^{m} r_{ij}\log w_{cj}\right) \\ &+ \gamma\sum_{i=1}^{n} u_{Ci}\left(\sum_{j=1}^{m} r_{ij}\log\frac{1}{m}\right) \\ &- \lambda_u\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci}, \end{aligned} \qquad (13)$$

where γ tunes the noise sensitivity. A smaller γ implies a higher probability of noise distribution and more objects are removed as noise.

In Ref. [13], it was demonstrated that the noise rejection scheme worked to improve the pureness of co-cluster cores by rejecting the unfavorable influences of noise objects in document clustering tasks.

C. Noise Rejection in pLSA-induced Fuzzy Co-clustering

In this paper, the above scheme is further enhanced to the multi-topic mixture models and a novel noise rejection model for pLSA-induced fuzzy co-clustering is proposed. For example, in document classification, some documents (objects) can be noises, which are not relevant to any dominant topics, and should be rejected from topic extraction. Then, in the proposed method, a noise cluster, i.e., a *noise topic*, is introduced such that the Cth cluster is the noise cluster where all items have uniform occurrence probabilities of $w_{Cj} = \frac{1}{m}$.

The aggregation degree of the noise cluster is $\sum_{i=1}^{n}\sum_{j=1}^{m}\phi_{Cij}r_{ij}\log\frac{u_{Ci}}{m}$, and the objective function of Eq.(10) is modified as:

$$L_{nfccplsa} = \sum_{c=1}^{C-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\phi_{cij}r_{ij}\log(u_{ci}w_{cj}) + \gamma\sum_{i=1}^{n}\sum_{j=1}^{m}\phi_{Cij}r_{ij}\log(\frac{u_{Ci}}{m}) - \lambda_\phi\sum_{c=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{m}\phi_{cij}\log\phi_{cij}, \quad (14)$$

where object and item memberships are derived under the same constraint with the conventional pLSA-induced fuzzy co-clustering as $\sum_{c=1}^{C} u_{ci} = 1$ and $\sum_{j=1}^{m} w_{cj} = 1$, respectively. Here, if object i is a noise, then u_{Ci} of noise membership becomes large and the remaining memberships follow $\sum_{c=1}^{C-1} u_{ci} < 1$ in normal clusters. Then, robust topic estimation is expected to be achieved by rejecting the influences of noise objects.

The penalty weight γ is adopted in the same manner as that of the FCCMM type, so that γ can tune the noise sensitivity. A smaller γ implies a higher probability of noise distribution and more objects are removed as noise.

Since the latent variable ϕ_{cij} represents the posteriori probability of component c given u_{ci} and w_{cj}, its constraint is still $\sum_{c=1}^{C}\phi_{cij} = 1$ without modification.

Following a similar process as the EM algorithm for pLSA, the updating formulas in maximizing the objective function of Eq. (14) are given as follows:

1) Updating Formula for Latent Variable ϕ_{cij}: For normal clusters,

$$\phi_{cij} = \frac{(u_{ci}w_{cj})^{(r_{ij}/\lambda_\phi)}}{\sum_{k=1}^{C-1}(u_{ki}w_{kj})^{(r_{ij}/\lambda_\phi)} + (\frac{u_{Ci}}{m})^{(\gamma r_{ij}/\lambda_\phi)}}. \quad (15)$$

For the noise cluster,

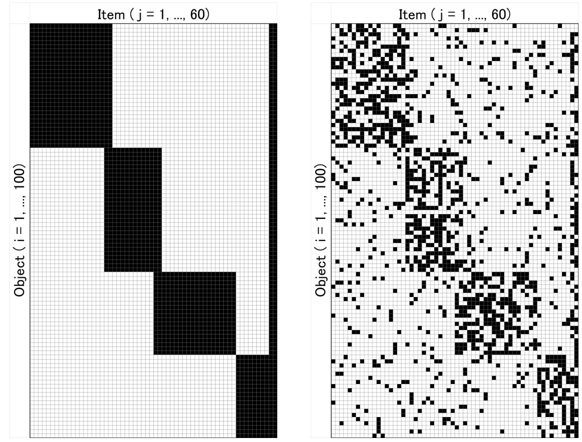
$$\phi_{Cij} = \frac{(\frac{u_{Ci}}{m})^{(\gamma r_{ij}/\lambda_\phi)}}{\sum_{k=1}^{C-1}(u_{ki}w_{kj})^{(r_{ij}/\lambda_\phi)} + (\frac{u_{Ci}}{m})^{(\gamma r_{ij}/\lambda_\phi)}}. \quad (16)$$

2) Updating Formula for Object Memberships u_{ci}: For normal clusters,

$$u_{ci} = \frac{\sum_{j=1}^{m} r_{ij}\phi_{cij}}{\sum_{k=1}^{C-1}\sum_{j=1}^{m} r_{ij}\phi_{kij} + \gamma\sum_{j=1}^{m} r_{ij}\phi_{Cij}}. \quad (17)$$

For the noise cluster,

$$u_{Ci} = \frac{\gamma\sum_{j=1}^{m} r_{ij}\phi_{Cij}}{\sum_{k=1}^{C-1}\sum_{j=1}^{m} r_{ij}\phi_{kij} + \gamma\sum_{j=1}^{m} r_{ij}\phi_{Cij}}. \quad (18)$$



(a) Noiseless R_0 (b) Noisy R_1

Fig. 2. Artificial cooccurrence matrices

3) Updating Formula for Item Memberships w_{cj}: For normal clusters,

$$w_{cj} = \frac{\sum_{i=1}^{n} r_{ij}\phi_{cij}}{\sum_{\ell=1}^{m}\sum_{i=1}^{n} r_{i\ell}\phi_{ci\ell}}. \quad (19)$$

For the noise cluster,

$$w_{Cj} = \frac{1}{m}. \quad (20)$$

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

[pLSA-induced Fuzzy Co-clustering Based on Noise Clustering (NFCCpLSA)]

Step 1. Initialize fuzzy memberships u_{ci}, $c = 1,\dots,C$, $i = 1,\dots,n$ and w_{cj}, $c = 1,\dots,C$, $j = 1,\dots,m$ 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 fuzziness penalty weight λ_ϕ, noise sensitivity γ, number of clusters C and termination criterion ε.

Step 2. Update latent variable ϕ_{cij}, $c = 1,\dots,C$, $i = 1,\dots,n$, $j = 1,\dots,m$ by Eqs.(15) and (16).

Step 3. Update w_{cj}, $c = 1,\dots,C$, $j = 1,\dots,m$ by Eqs.(19) and (20).

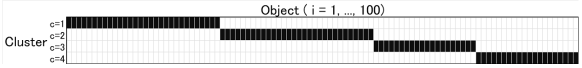
Step 4. Update u_{ci}, $c = 1,\dots,C$, $i = 1,\dots,n$ by Eqs.(17) and (18).

Step 5. If $\max_{c,i} \mid u_{ci}^{NEW} - u_{ci}^{OLD} \mid < \varepsilon$, then stop. Otherwise, return to Step 2.

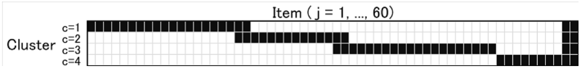
IV. Numerical Experiments

A. Artificial Data Sets

First, the proposed model was applied to an artificially generated data set composed of 100 objects ($n = 100$) and 60 items ($m = 60$). Figure 2 shows visual images of the 100×60 cooccurrence matrices R_0 and R_1, where black and white cells



(a) Ideal object membership u_{ci}



(b) Ideal item membership w_{cj}

Fig. 3. Ideal memberships derived from R_0 without noise

TABLE I
RAND INDEX IN CONJUNCTION WITH REJECT RATE (ARTIFICIAL DATA)

(a) FCCMM and NFCCMM (with γ)

		λ_u			
		0.5	1.0	2.0	3.0
FCCMM		0.950	0.940	0.940	0.950
γ	0.98	0.977 (0.14)	**0.988** (0.16)	0.976 (0.15)	0.977 (0.12)
	1.00	0.958 (0.05)	0.978 (0.10)	0.978 (0.09)	0.957 (0.06)
	1.02	0.959 (0.02)	0.959 (0.03)	0.948 (0.03)	0.949 (0.02)

(b) FCCpLSA and NFCCpLSA (with γ)

		λ_ϕ			
		1.0	1.1	1.3	1.5
pLSA		0.960	0.970	0.960	0.960
γ	0.98	0.966 (0.12)	**0.989** (0.10)	0.968 (0.05)	0.958 (0.04)
	1.00	0.969 (0.04)	0.969 (0.03)	0.970 (0.00)	0.960 (0.01)
	1.02	0.980 (0.00)	0.970 (0.01)	0.970 (0.00)	0.960 (0.00)

depict $r_{ij} = 1$ and $r_{ij} = 0$, respectively. The base matrix R_0 without noise includes roughly four co-clusters ($C = 4$) in diagonal blocks, where some items are shared by multiple clusters. The noisy matrix R_1 to be analyzed was generated from R_0 by replacing $r_{ij} = 1$ with $r_{ij} = 0$ at a rate of 50% and $r_{ij} = 0$ with $r_{ij} = 1$ at a rate of 10%. The goal is to estimate the plausible memberships from the noisy data set, whose ideal values are shown in Fig. 3 by grayscale.

The performances of the proposed NFCCpLSA are compared with the conventional FCCMM, NFCCMM and FCCpLSA. The number of clusters was $C = 4$ for FCCMM and FCCpLSA, and $C = 5$ for NFCCMM and NFCCpLSA including a noise cluster. The algorithms were implemented with $\lambda_u \in \{0.5, 1.0, 2.0, 3.0\}$ and $\gamma \in \{0.98, 1.0, 1.02\}$ for FCCMM and NFCCMM, and with $\lambda_\phi \in \{1.0, 1.1, 1.3, 1.5\}$ and $\gamma \in \{0.98, 1.0, 1.02\}$ for FCCpLSA and NFCCpLSA. Figures 4 and 5 compare the derived object memberships through grayscale, where black and white indicate maximum and zero values, respectively. In the figures, the partition fuzziness was demonstrated to be weak to heavy through upper to lower subfigures while the noise rejection was heavier enforced from left to right subfigures.

In addition, the ratio of matching between the derived partition and the ideal object partition under maximum membership assignment (Rand Index) are compared in Table I, where the best performances in 10 trials with different initializations are presented in conjunction with the noise rejection rate (Reject Rate) shown in brackets.

The partition quality was shown to be improved by properly adjusting the noise sensitivity and fuzziness weights. In the proposed method, the noise sensitivity weight successfully worked for tuning the rejection rate such that as the noise sensitivity γ decreased, the number of objects removed as noise increased. Moreover, a much higher fuzziness weight such as $\lambda_u = 3.0$ and $\lambda_\phi = 1.5$ can cause poor local solutions, where structural information was violated.

TABLE II
RAND INDEX IN CONJUNCTION WITH REJECT RATE (CORA DATASET)

(a) FCCMM and NFCCMM (with γ)

		λ_u		
		1.5	2.0	3.0
FCCMM		0.491	0.495	0.513
γ	0.98	0.532 (0.013)	0.534 (0.009)	0.555 (0.008)
	1.00	0.528 (0.008)	0.545 (0.004)	**0.557** (0.003)
	1.02	0.534 (0.003)	0.525 (0.003)	0.552 (0.002)

(b) FCCpLSA and NFCCpLSA (with γ)

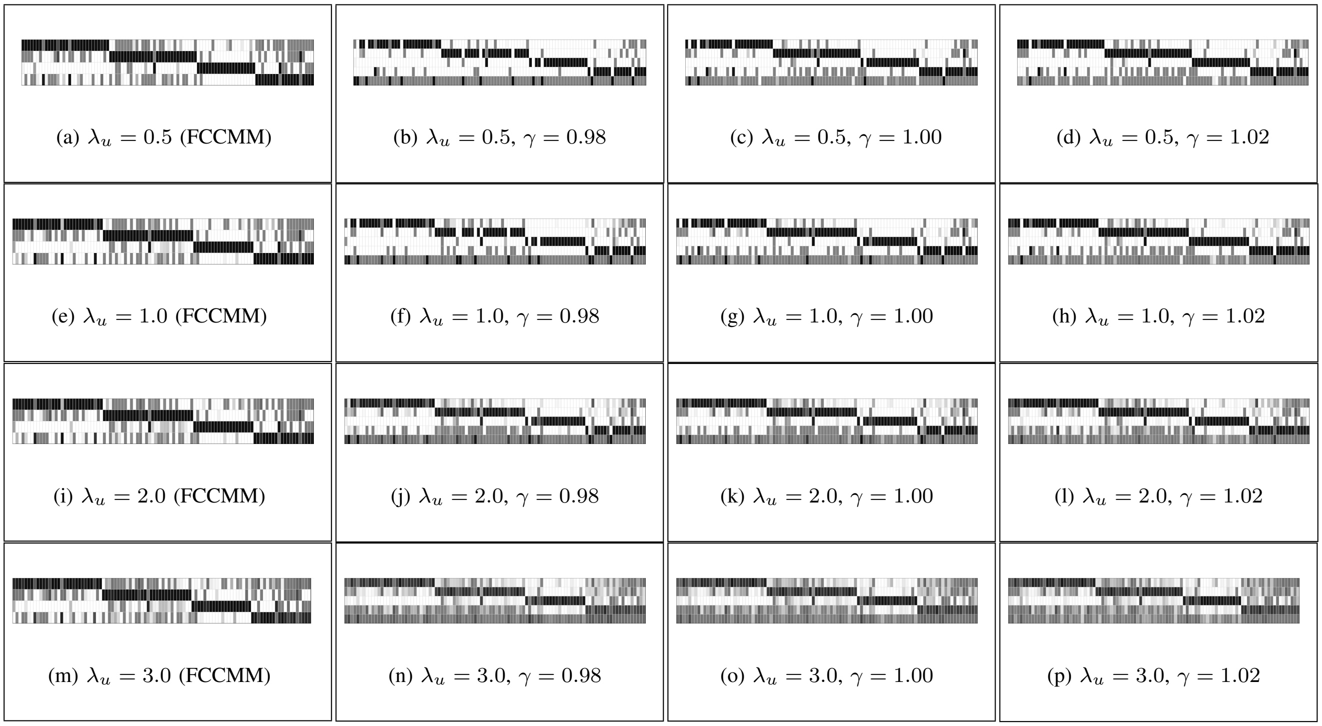
		λ_ϕ		
		1.0	1.1	1.3
pLSA		0.528	0.520	0.556
γ	0.92	0.558 (0.278)	**0.605** (0.260)	0.596 (0.210)
	0.94	0.528 (0.200)	0.552 (0.155)	0.579 (0.129)
	0.96	0.522 (0.008)	0.541 (0.082)	0.571 (0.066)

B. Document Classification Benchmark Data Sets

Next, the proposed model was applied to two document classification benchmark data sets: Cora and CiteSeer, which are available from LINQS webpage of Statistical Relational Learning Group @ UMD (http://linqs.cs.umd.edu/projects//index.shtml). Cora is a 2708×1433 cooccurrence matrix, which consists of 2708 machine learning papers and 1433 feature words, and is classified into one of seven classes. CiteSeer is a 3312×3703 cooccurrence matrix, which consists of 3312 scientific journals and 3703 feature words, and is classified into one of six classes. Cooccurrence information is presented as $r_{ij} = 1$ if the document i contains the feature word j and is presented as $r_{ij} = 0$ if the document i does not contain the feature word j. The number of documents in each class is { 298, 418, 818, 426, 217, 180, 351 } in Cora, and { 701, 668, 596, 590, 508, 249 } in CiteSeer.

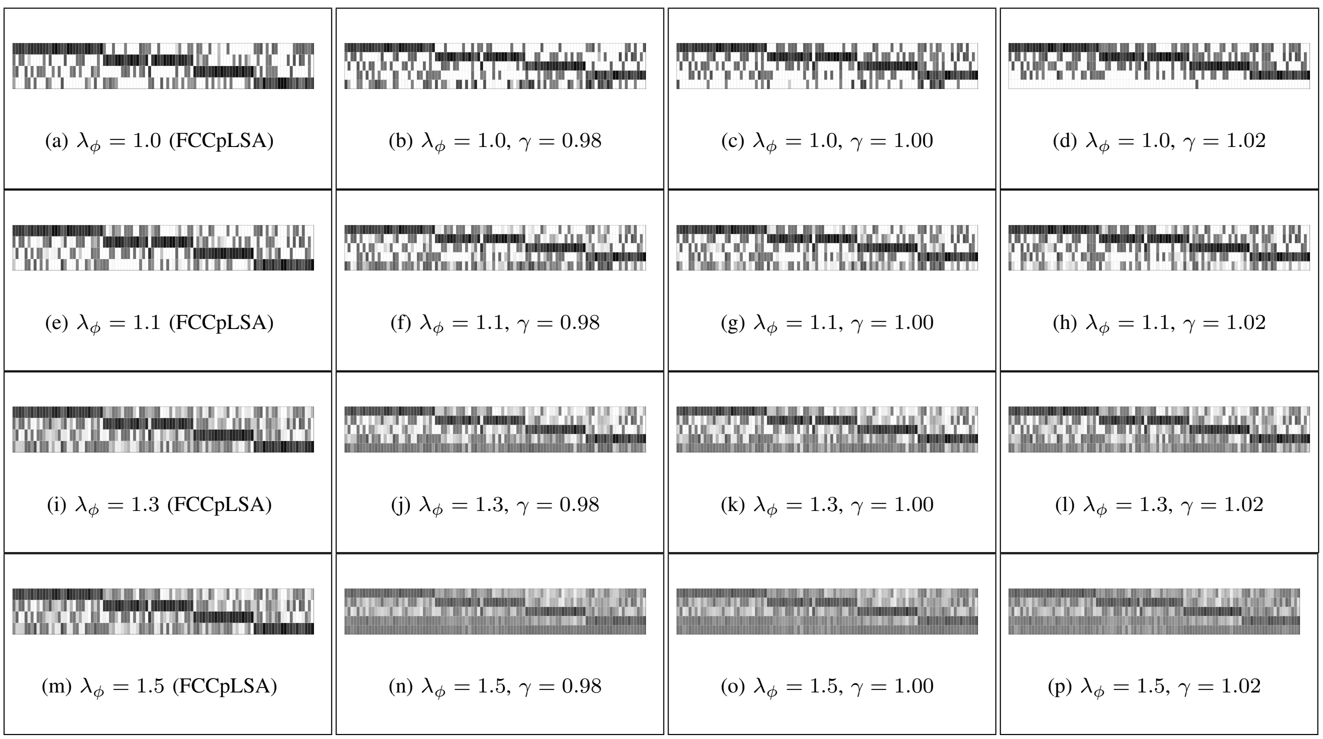
1) Cora dataset: The number of clusters was $C = 7$ for FCCMM and FCCpLSA, and $C = 8$ for NFCCMM and NFCCpLSA including a noise cluster. The algorithms were implemented with $\lambda_u \in \{1.5, 2.0, 3.0\}$ and $\gamma \in \{0.98, 1.0, 1.02\}$ for FCCMM and NFCCMM, and with $\lambda_\phi \in \{1.0, 1.1, 1.3\}$ and $\gamma \in \{0.92, 0.94, 0.96\}$ for FCCpLSA and NFCCpLSA. The ratio of matching between the derived partition and the ideal object partition under maximum membership assignment (Rand Index) are compared in Table II, where the best performances in 10 trials with different initializations are presented in conjunction with the noise rejection rates (Reject Rate) shown in brackets.

Table II shows that the classification performance of pLSA-induced fuzzy co-clustering was better than that of FCCMM. Moreover, introducing the noise clustering schemes, the classification performances of FCCMM and the proposed model



(a) $\lambda_u = 0.5$ (FCCMM) (b) $\lambda_u = 0.5$, $\gamma = 0.98$ (c) $\lambda_u = 0.5$, $\gamma = 1.00$ (d) $\lambda_u = 0.5$, $\gamma = 1.02$
(e) $\lambda_u = 1.0$ (FCCMM) (f) $\lambda_u = 1.0$, $\gamma = 0.98$ (g) $\lambda_u = 1.0$, $\gamma = 1.00$ (h) $\lambda_u = 1.0$, $\gamma = 1.02$
(i) $\lambda_u = 2.0$ (FCCMM) (j) $\lambda_u = 2.0$, $\gamma = 0.98$ (k) $\lambda_u = 2.0$, $\gamma = 1.00$ (l) $\lambda_u = 2.0$, $\gamma = 1.02$
(m) $\lambda_u = 3.0$ (FCCMM) (n) $\lambda_u = 3.0$, $\gamma = 0.98$ (o) $\lambda_u = 3.0$, $\gamma = 1.00$ (p) $\lambda_u = 3.0$, $\gamma = 1.02$

Fig. 4. Object memberships u_{ci} in FCCMM and NFCCMM



(a) $\lambda_\phi = 1.0$ (FCCpLSA) (b) $\lambda_\phi = 1.0$, $\gamma = 0.98$ (c) $\lambda_\phi = 1.0$, $\gamma = 1.00$ (d) $\lambda_\phi = 1.0$, $\gamma = 1.02$
(e) $\lambda_\phi = 1.1$ (FCCpLSA) (f) $\lambda_\phi = 1.1$, $\gamma = 0.98$ (g) $\lambda_\phi = 1.1$, $\gamma = 1.00$ (h) $\lambda_\phi = 1.1$, $\gamma = 1.02$
(i) $\lambda_\phi = 1.3$ (FCCpLSA) (j) $\lambda_\phi = 1.3$, $\gamma = 0.98$ (k) $\lambda_\phi = 1.3$, $\gamma = 1.00$ (l) $\lambda_\phi = 1.3$, $\gamma = 1.02$
(m) $\lambda_\phi = 1.5$ (FCCpLSA) (n) $\lambda_\phi = 1.5$, $\gamma = 0.98$ (o) $\lambda_\phi = 1.5$, $\gamma = 1.00$ (p) $\lambda_\phi = 1.5$, $\gamma = 1.02$

Fig. 5. Object memberships u_{ci} in FCCpLSA and NFCCpLSA.

were improved by properly adjusting the noise sensitivity and fuzziness weights. Then, the advantage of the proposed model was demonstrated in a real-world application, where the numbers of class elements are not constant. Moreover, it was also confirmed that the partition quality of both models can be improved with slightly fuzzier settings than the statistical MMMs and pLSA, but can be degraded with heavily fuzzier settings.

2) CiteSeer: The number of clusters was $C = 6$ for FCCMM and FCCpLSA, and $C = 7$ for NFCCMM and NFCCpLSA including a noise cluster. The algorithms were implemented with $\lambda_u \in \{1.0, 1.5, 2.0, 3.0\}$ and $\gamma \in \{0.90, 0.92, 0.94, 0.96\}$ for FCCMM and NFCCMM, and with $\lambda_\phi \in \{1.0, 1.1, 1.3, 1.5\}$ and $\gamma \in \{0.92, 0.94, 0.96, 0.98\}$ for FCCpLSA and NFCCpLSA. The ratio of matching between the derived partition and the ideal object partition under

TABLE III
RAND INDEX IN CONJUNCTION WITH REJECT RATE (CITESEER DATASET)

(a) FCCMM and NFCCMM (with γ)

		λ_u		
		1.0	2.0	3.0
FCCMM		0.358	0.505	0.546
γ	0.90	0.392 (0.151)	0.533 (0.137)	**0.591** (0.133)
	0.92	0.371 (0.066)	0.517 (0.050)	0.565 (0.046)
	0.94	0.362 (0.021)	0.508 (0.019)	0.557 (0.015)
	0.96	0.359 (0.008)	0.504 (0.005)	0.552 (0.006)

(b) FCCpLSA and NFCCpLSA (with γ)

		λ_ϕ		
		1.0	1.1	1.3
pLSA		0.584	0.623	0.634
γ	0.92	0.653 (0.235)	**0.666** (0.223)	**0.666** (0.196)
	0.94	0.637 (0.106)	0.652 (0.114)	0.657 (0.103)
	0.96	0.628 (0.043)	0.638 (0.049)	0.648 (0.046)
	0.98	0.642 (0.016)	0.628 (0.018)	0.644 (0.018)

maximum membership assignment (Rand Index) are compared in Table III, where the best performances in 10 trials with different initializations are presented in conjunction with the noise rejection rates (Reject Rate) shown in brackets.

Table III shows that the classification performance of pLSA-induced fuzzy co-clustering was better than that of FCCMM. In the same manner as the Cora data set, the advantages of the proposed noise rejection scheme were confirmed again such that the classification performances of FCCMM and the proposed model were improved by properly adjusting the noise sensitivity and fuzziness weights.

V. CONCLUSION

This paper proposed a robust model for pLSA-induced fuzzy co-clustering by introducing a noise rejection mechanism based on noise fuzzy clustering. Numerical experiments showed that the classification performance of the proposed model can be better than that of the conventional model by properly adjusting noise sensitivity γ and fuzziness weight λ_ϕ. Additionally, slightly fuzzier models were demonstrated to be better than the statistical pLSA model, even if the noise rejection scheme is adopted. In an experimental result shown in Figs. 4 and 5, plausible results could be given with certain ranges of γ and λ_ϕ.

However, the partition quality can be degraded when the fuzziness degree is much larger than a plausible setting and the noise sensitivity is over-emphasized. Possible future research includes the development of a systematic process for automatically selecting the plausible settings for such parameters. Another possible study may be the introduction of possibilistic partition [25], [26], instead of the noise fuzzy clustering concept, which is expected to be useful for independently extracting intrinsic topics. In a previous study on NFCCMM [13], possibilistic noise rejection scheme worked for selecting the optimal cluster number through a sequential implementation with different cluster numbers. A similar process is expected to be available in the proposed method.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI Grant Number JP18K11474.

REFERENCES

[1] E. H. Ruspini, "A new approach to clustering," *Information and Control*, vol. 15, no. 1, pp. 22-32, July 1969.
[2] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, 1981.
[3] S. Miyamoto, H. Ichihashi and K. Honda, *Algorithms for Fuzzy Clustering*, Springer, 2008.
[4] J. B. MacQueen, "Some methods of classification and analysis of multivariate observations," *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pp. 281-297, 1967.
[5] A. P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. of the Royal Statistical Society*, Series B, vol. 39, pp. 1–38, 1977.
[6] K. Honda, A. Notsu and H. Ichihashi, "Fuzzy PCA-guided robust k-means clustering," *IEEE Trans. on Fuzzy Systems*, vol. 18, no. 1, pp. 67-79, 2010.
[7] R. N. Davé, "Characterization and detection of noise in clustering," *Pattern Recognition Letters*, vol. 12, no. 11, pp. 657-664, 1991.
[8] R. N. Davé and R. Krishnapuram, "Robust clustering methods: a unified view," *IEEE Trans. on Fuzzy Systems*, vol.5, pp.270-293, 1997.
[9] 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.
[10] K. Kummamuru, A. Dhawale and R. Krishnapuram, "Fuzzy co-clustering of documents and keywords," *Proc. 2003 IEEE Int'l Conf. Fuzzy Systems*, vol. 2, pp. 772-777, 2003.
[11] 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.
[12] 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.
[13] K. Honda, N. Yamamoto, S. Ubukata and A. Notsu, "Noise rejection in MMMs-induced fuzzy co-clustering," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 21, no. 7, pp. 1144-1151, 2017.
[14] I. Holmes, K. Harris and C. Quince, "Dirichlet multinomial mixtures: generative models for microbial metagenomics," *PLoS ONE*, vol. 7, issue 2, e30126, 2012.
[15] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, no. 1-2, pp. 177-196, 2001.
[16] K. Honda, T. Goshima, S. Ubukata and A. Notsu, "A fuzzy co-clustering interpretation of probabilistic latent semantic analysis," *Proc. 2016 IEEE Int'l Conf. Fuzzy Systems*, pp. 718-723, 2016.
[17] T. Goshima, K. Honda, S. Ubukata and A. Notsu, "Deterministic annealing process for pLSA-induced fuzzy co-clustering and cluster splitting characteristics," *International Journal of Approximate Reasoning*, vol. 95, pp. 185-193, 2018.
[18] K. Rose, E. Gurewitz, and G. Fox, "A deterministic annealing approach to clustering," *Pattern Recognition Letters*, vol. 11, pp. 589-594, 1990.
[19] 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.
[20] Z.-Q. Liu, S. Miyamoto (eds.), Soft Computing and Human-Centered Machines, Springer-Verlag, 2000.
[21] K. Honda and H. Ichihashi, "Regularized linear fuzzy clustering and probabilistic PCA mixture models," *IEEE Trans. on Fuzzy Systems*, vol. 13, no. 4, pp. 508-516, 2005.
[22] R. J. Hathaway, "Another interpretation of the EM algorithm for mixture distributions," *Statistics & Probability Letters*, vol. 4, pp. 53-56, 1986.
[23] T. Needham, "A visual explanation of Jensen's inequality," *American Mathematical Monthly*, vol. 100, no. 8, 99.768-771, 1993.
[24] N. Yamamoto, K. Honda, S. Ubukata and A. Notsu, "Noise rejection schemes for FCM-type co-clustering based on uniform noise distribution," *Proc. 2017 IEEE Int'l Conf. Fuzzy Systems*, #F-0264, 2017.
[25] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Trans. on Fuzzy Systems*, vol.1, pp.98-110, 1993.
[26] F. Masulli and S. Rovetta, "Soft transition from probabilistic to possibilistic fuzzy clustering," *IEEE Trans. on Fuzzy Systems*, vol. 14, no. 4, pp. 516-527, Aug. 2006.