Hindawi Publishing Corporation
Advances in Fuzzy Systems
Volume 2015, Article ID 729072, 8 pages
http://dx.doi.org/10.1155/2015/729072

Research Article

A Collaborative Framework for Privacy Preserving Fuzzy Co-Clustering of Vertically Distributed Cooccurrence Matrices

Katsuhiro Honda, Toshiya Oda, Daiji Tanaka, and Akira Notsu

Osaka Prefecture University, 1-1 Gakuen-cho, Nakaku, Sakai, Osaka 599-8531, Japan

Correspondence should be addressed to Katsuhiro Honda; honda@cs.osakafu-u.ac.jp

Received 10 February 2015; Revised 12 March 2015; Accepted 12 March 2015

Academic Editor: Rustom M. Mamlook

In many real world data analysis tasks, it is expected that we can get much more useful knowledge by utilizing multiple databases stored in different organizations, such as cooperation groups, state organs, and allied countries. However, in many such organizations, they often hesitate to publish their databases because of privacy and security issues although they believe the advantages of collaborative analysis. This paper proposes a novel collaborative framework for utilizing vertically partitioned cooccurrence matrices in fuzzy co-cluster structure estimation, in which cooccurrence information among objects and items is separately stored in several sites. In order to utilize such distributed data sets without fear of information leaks, a privacy preserving procedure is introduced to fuzzy clustering for categorical multivariate data (FCCM). Withholding each element of cooccurrence matrices, only object memberships are shared by multiple sites and their (implicit) joint co-cluster structures are revealed through an iterative clustering process. Several experimental results demonstrate that collaborative analysis can contribute to revealing global intrinsic co-cluster structures of separate matrices rather than individual site-wise analysis. The novel framework makes it possible for many private and public organizations to share common data structural knowledge without fear of information leaks.

1. Introduction

Data mining is a powerful tool for many private and public organizations in supporting efficient decision making, and they have been utilizing various databases, which are independently and securely stored in each organization. However, it is often quite expensive or impossible to store enough data by each of themselves and many analysts believe that we can get much more useful knowledge by utilizing multiple databases stored in different organizations. In these collaborative data analysis, a significant problem is the privacy issue. For example, in many corporations, customer segmentation by clustering is a fundamental approach in possible marketing while their customer privacy must be securely protected and each data record such as purchase history and personal profiles must not be published to other corporations or organizations. Similar situations are found in many other organizations such as hospitals with clinical records and governments with military intelligences.

Privacy preserving data mining (PPDM) [1] is a fundamental approach for utilizing multiple databases including personal or sensitive information without fear of information leaks. A possible approach is a priori k-anonymization of databases for secure publication [2, 3], but such anonymization can bring information losses. Another approach for utilizing all distributed information is to analyze the information without revealing each element. In k-means clustering, several secure processes for estimating cluster centers were proposed [4, 5], in which the mean vector of each cluster is calculated with an encryption operation.

In this paper, a novel collaborative framework for utilizing vertically partitioned cooccurrence matrices in fuzzy co-cluster structure estimation is proposed, where cooccurrence information among objects and items is separately stored in several sites. In vertically distributed databases, it is assumed that all sites share common objects but they are characterized with different independent items in each site. The goal is to reveal the global co-cluster structures varied in whole separate databases without publishing each element of independent databases to other sites.

The remaining parts of this paper are organized as follows: Section 2 gives a brief review on related works and Section 3 shows their problems and possible solutions. Section 4 provides explanations on the conventional fuzzy co-clustering model and Section 5 proposes a novel collaborative framework for applying fuzzy co-clustering considering privacy issues. In Section 6, several experimental results demonstrate that collaborative analysis can contribute to revealing global intrinsic co-cluster structures of separate matrices rather than individual site-wise analysis. Finally, a summary conclusion is given in Section 7.

2. Background

Co-clustering is a fundamental technique for summarizing mutual cooccurrence information among objects and items. For example, in document clustering, mutual cooccurrence information of documents and keywords are utilized for revealing intrinsic document clusters with their keywords summaries. In purchase history analysis, mutual connections among customers and their promising products are investigated considering purchase preferences. Co-clustering provides pairwise cluster structures among objects and items and has been widely investigated in both probabilistic [6] and heuristic contexts [7]. In this paper, fuzzy clustering approaches are focused on.

Fuzzy clustering has been proved to have many advantages against hard ones from such view points as noise and initialization sensitivities. Fuzzy variants of co-clustering have also been demonstrated to be useful in such applications as document analysis [8] and collaborative filtering [9, 10]. The goal of fuzzy co-clustering is to simultaneously estimate memberships of both objects and items from a cooccurrence information matrix. For example, in document analysis, each document (object) is characterized by several keywords (items) with their appearance frequencies (degree of cooccurrences), and the goal is to extract document-keyword clusters with their fuzzy memberships for analyzing their contents.

Fuzzy clustering for categorical multivariate data (FCCM) [11] is a Fuzzy c-Means- (FCM-) type [12] co-clustering model, in which a co-cluster aggregation criterion is maximized supported by entropy-based membership fuzzification [13, 14] in FCM-like iterative optimization algorithm. Several fuzzy co-clustering models were proposed based on similar concepts with FCCM, in which other fuzzification mechanisms were adopted [8, 15–18].

In order to analyze distributed databases in k-means-type clustering, several secure processes for estimating cluster centers were proposed [4, 5], in which the mean vector of each cluster is calculated with an encryption operation. However, in fuzzy co-clustering, the clustering criteria of cluster aggregation degrees were defined without cluster centers and the conventional secure framework cannot be adopted. Then, a novel secure mechanism is needed, where the main problems to be solved remained as summarized in the next section.

3. Problems and Solution

In the k-means-type secure clustering model for vertically distributed data [4, 5], multiple sites share common objects, such as customers and patients, while having their own vector observations only, such as customer profiles of their own stores and clinical records in their own hospitals. In order to reveal the intrinsic object clusters without publishing each observation, each coordinate of cluster centers is separately calculated in each site and the derived coordinates are shared by all sites.

On the other hand, fuzzy co-clustering does not use cluster centers as cluster prototypes and utilizes two types of fuzzy memberships only. Then, the conventional secure framework for k-means-type clustering cannot be adopted, and a secure process for calculating the fuzzy memberships must be developed.

In the following, in this paper, a novel framework for calculating fuzzy memberships in fuzzy co-clustering of vertically distributed cooccurrence matrices is proposed following a brief review on the conventional fuzzy co-clustering models. In order to calculate *object* memberships, the sum of products of *item* memberships and cooccurrence observations are needed, and vice versa. In the proposed secure process, the sum calculation is securely achieved through an encryption operation, in which the sum can be calculated by concealing each value.

The novel framework is constructed in the FCCM context only, which is the basic model of fuzzy co-clustering. However, it is easily expected that a similar extension is directly applicable to the other FCCM variants without discussions because all the FCCM variants are based on the FCCM updating process.

4. Methodology of Fuzzy Co-Clustering

Assume that we have a cooccurrence matrix $R = \{r_{ij}\}$ on objects $i = 1, \ldots, n$ and items $j = 1, \ldots, m$, in which r_{ij} represents the degree of cooccurrence of item j with object i. The goal of co-clustering is to simultaneously partition objects and items into C co-clusters by estimating two types of fuzzy memberships. Object partitions are represented by object memberships u_{ci}, which is the memberships degree of object i to cluster c and is forced to be exclusive in the same way with FCM such that $\sum_{c=1}^{C} u_{ci} = 1$. On the other hand, in order to avoid trivial solutions, item partitions are represented by item memberships w_{cj}, which are mostly responsible for representing the mutual typicalities in each cluster such that $\sum_{j=1}^{m} w_{cj} = 1$.

Oh et al. [11] proposed the FCM-type co-clustering model, which is called FCCM, by modifying the FCM algorithm for handling cooccurrence information, where the cluster aggregation degree of each cluster is maximized:

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

The first term to be maximized measures the aggregation degree of objects and items in cluster c, such that it becomes larger when mutually familiar objects and items having a large r_{ij}, simultaneously, have large memberships in a cluster. Here, this aggregation degree is only designed for hard partition because the term is a linear function with respect to both of u_{ci} and w_{cj}, where we have always $u_{ci} \in \{0, 1\}$ and $w_{cj} \in \{0, 1\}$. Then, in order to derive fuzzy memberships $u_{ci} \in [0, 1]$ and $w_{cj} \in [0, 1]$, the aggregation measure must be nonlinearized.

In FCCM, the entropy-based fuzzification method [13, 14] was adopted instead of the standard approach in FCM because the exponential weight in FCM can work only in the minimization framework of positive objective functions. λ_u and λ_w tune the degree of fuzziness of memberships, where a larger λ brings fuzzier partitions while a smaller λ brings crisp partitions.

The clustering algorithm is an iterative process of updating u_{ci} and w_{cj} using the following rules:

$$u_{ci} = \frac{\exp\left(\lambda_u^{-1} \sum_{j=1}^{m} w_{cj} r_{ij}\right)}{\sum_{\ell=1}^{C} \exp\left(\lambda_u^{-1} \sum_{j=1}^{m} w_{\ell j} r_{ij}\right)}, \qquad (2)$$

$$w_{cj} = \frac{\exp\left(\lambda_w^{-1} \sum_{i=1}^{n} u_{ci} r_{ij}\right)}{\sum_{\ell=1}^{m} \exp\left(\lambda_w^{-1} \sum_{i=1}^{n} u_{ci} r_{i\ell}\right)}. \qquad (3)$$

This FCCM process was also reconstructed with other fuzzification mechanisms. For example, Fuzzy CoDoK [8] utilized the quadric term-based regularization [19] for avoiding calculation overflows. Honda et al. [15] adopted K-L information-based regularization [20] for handling unbalanced cluster sizes. As discussed in Section 3, these extended models generally follow the original FCCM procedure and have similar characteristics. So, in this paper, the novel collaborative framework is described in the FCCM context only.

5. Fuzzy Co-Clustering with Privacy Consideration

5.1. Privacy Consideration in k-Means Clustering. When each object is characterized by m-dimensional observation $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})^T$, k-means algorithm tries to minimize the within-cluster errors by iterating cluster center updating and nearest prototype assignment. Let $\mathbf{b}_c = (b_{c1}, \ldots, b_{cm})^T$ be the center of cluster c. In cases of distributed databases, we must care about privacy issues in either of the two phases by adopting such a technique as encryption operation [5].

For vertically distributed databases, where the elements of $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})^T$ are separately stored in several sites, distances between object i and C cluster centers are calculated under collaboration of all sites. Here, the clustering criterion is the sum of squared errors $\sum_{j=1}^{m} |x_{ij} - b_{cj}|^2$ and should be calculated by concealing each value of $|x_{ij} - b_{cj}|^2$ from other sites. Once we find the nearest prototype assignment of each object, we can independently calculate new $\mathbf{b}_c = (b_{c1}, \ldots, b_{cm})^T$ in each site by sharing the object membership information.

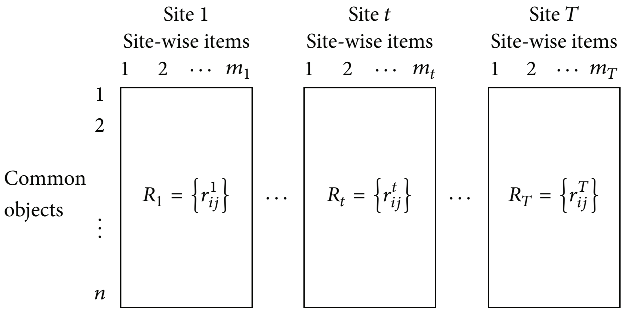


Figure 1: Vertically distributed cooccurrence matrices.

Although the above secure framework is also useful in many other k-means-type clustering algorithms such as FCM, it cannot be directly adopted to co-clustering ones because co-clustering does not use cluster prototypes but considers two types of memberships.

In this paper, similar ideas are adopted to fuzzy co-clustering tasks.

5.2. Fuzzy Co-Clustering with Privacy Consideration. Assume that T sites ($t = 1, \ldots, T$) share common n objects ($i = 1, \ldots, n$) and have different cooccurrence information on different items, which are summarized into $n \times m_t$ matrices $R_t = \{r_{ij}^t\}$, where m_t is the number of items in site t and $\sum_{t=1}^{T} m_t = m$. Figure 1 shows a visual image of vertically distributed cooccurrence matrices. For example, we have a group of T corporations (or hospitals, countries, etc.) and each of them has its independent customer purchase history $R_t = \{r_{ij}^t\}$ (or patients' records, military intelligence, etc.).

If we do not care about the privacy issues, the distributed matrices should be gathered into a full $n \times m$ matrix to be analyzed in a single process without information losses. Taking the privacy preservation into account, however, each matrix should be processed in each site without broadcasting personal information although the reliability of each co-cluster structure may not be enough satisfied because of information losses. Then, the goal of the collaborative fuzzy co-clustering analysis is to estimate object and item memberships as similar to the full-data case as possible by sharing object partition information without broadcasting cooccurrence information $R_t = \{r_{ij}^t\}$.

Object memberships u_{ci} to be shared by sites are common and are defined in the same manner with the conventional FCCM. On the other hand, item memberships w_{cj} are somewhat different because they follow the within-cluster sum constraint. In this paper, it is assumed that item memberships are independently estimated in each site following the site-wise constraint $\sum_{j=1}^{m_t} w_{cj}^t = 1$, where w_{cj}^t is the item membership on item j in site t. Be noted that the item memberships w_{cj}^t should not be opened to other sites from privacy consideration.

In applying FCCM clustering to distributed cooccurrence matrices, (2) implies that each object membership function is dependent on $\sum_{j=1}^{m} w_{cj} r_{ij}$, which is the sum of site-wise

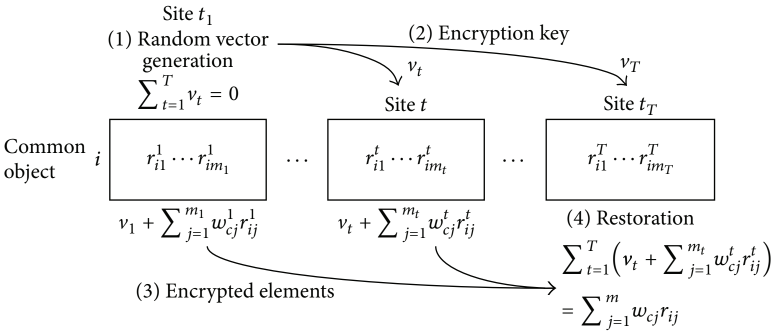

Site t_1
(1) Random vector generation
$\sum_{t=1}^{T} v_t = 0$
(2) Encryption key
v_t
v_T
Site t
Site t_T
Common object
i
$r_{i1}^{1} \cdots r_{im_1}^{1}$
$\cdots$
$r_{i1}^{t} \cdots r_{im_t}^{t}$
$\cdots$
$r_{i1}^{T} \cdots r_{im_T}^{T}$
$v_1 + \sum_{j=1}^{m_1} w_{cj}^{1} r_{ij}^{1}$
$v_t + \sum_{j=1}^{m_t} w_{cj}^{t} r_{ij}^{t}$
(4) Restoration
$\sum_{t=1}^{T}\left(v_t + \sum_{j=1}^{m_t} w_{cj}^{t} r_{ij}^{t}\right)$
$= \sum_{j=1}^{m} w_{cj} r_{ij}$
(3) Encrypted elements

FIGURE 2: Calculation of $\sum_{j=1}^{m} w_{cj} r_{ij}$ with encryption operation.

independent information $\sum_{j=1}^{m_t} w_{cj}^{t} r_{ij}^{t}$. In order to share object partition considering personal privacy, we must calculate $\sum_{j=1}^{m} w_{cj} r_{ij}$ without broadcasting each site-wise information $\sum_{j=1}^{m_t} w_{cj}^{t} r_{ij}^{t}$. A promising approach of secure calculation of $\sum_{j=1}^{m} w_{cj} r_{ij}$ is based on an encryption operation.

Assume that we have at least three sites, that is, $T > 2$, and two sites of t_1 and t_T are selected as representative sites. Figure 2 summarizes the process for secure calculation of $\sum_{j=1}^{m} w_{cj} r_{ij}$ as follows.

(1) Site t_1 generates length C random vectors $\mathbf{v}_t = (v_{t1}, \ldots, v_{tC})^T$, $t = 1, \ldots, T$, such that $\sum_{t=1}^{T} \mathbf{v}_t = \mathbf{0}$.

(2) Site t_1 sends the encryption key vector $\mathbf{v}_t = (v_{t1}, \ldots, v_{tC})^T$ to each of the other sites.

(3) Sites $t_1 \cdots t_{T-1}$ send their encrypted information $v_{tc} + \sum_{j=1}^{m_t} w_{cj}^{t} r_{ij}^{t}$ to site t_T.

(4) Their total amount $\sum_{t=1}^{T}(v_{tc} + \sum_{j=1}^{m_t} w_{cj}^{t} r_{ij}^{t})$ is calculated for estimating u_{ci} in site t_T. Then, site t_T broadcasts u_{ci} to all sites.

$\sum_{t=1}^{T} \mathbf{v}_t = \mathbf{0}$ implies that the total amount $\sum_{t=1}^{T}(v_{tc} + \sum_{j=1}^{m_t} w_{cj}^{t} r_{ij}^{t})$ is equivalent to $\sum_{t=1}^{T} \sum_{j=1}^{m_t} w_{cj}^{t} r_{ij}^{t}$ although the individual value of each site is concealed by v_{tc}. In this scheme, no site can reveal the actual value of $\sum_{j=1}^{m_t} w_{cj}^{t} r_{ij}^{t}$ on other sites.

Once object memberships u_{ci} are broadcasted to all sites, each item membership w_{cj}^{t} is calculated by (3) in each site using in-site information only, where site-wise item memberships w_{cj}^{t} follow site-wise normalization constraints $\sum_{j=1}^{m_t} w_{cj}^{t} = 1$.

It should be noted that, in this algorithm, item memberships are independently estimated in each site under the assumption that each site does not have any information on the items, which other sites deal with, such as the number of items and the degree of fuzziness of item memberships. Additionally, the algorithm cannot exactly reconstruct the equivalent co-clustering result to the whole data case, where all cooccurrence information is shared without care for privacy issues, even if we use the same parameter setting in all sites. It is because the piecewise constraint of $\sum_{j=1}^{m_t} w_{cj}^{t} = 1$ is independently forced to item memberships in each site while we just consider $\sum_{j=1}^{m} w_{cj} = 1$ in the whole data case.

6. Numerical Experiments

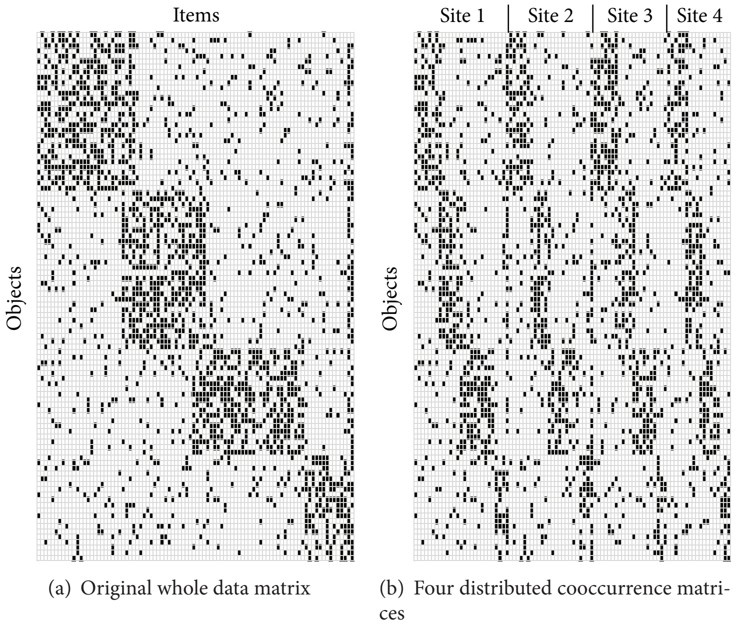
In this section, three experimental results are shown for demonstrating the characteristics of the proposed algorithm. Section 6.1 demonstrates the basic features of the proposed framework with a simple data set and Section 6.2 discusses the applicability to more realistic situations with a data set having unbalanced cluster structure. Then, an applicational experiment is shown in Section 6.3, where a virtual alliance of military sections is simulated using a real world benchmark data set.

6.1. Data Set 1: Homogeneous Cluster Partition.

An artificially generated 100 × 90 cooccurrence matrix $R = \{r_{ij}\}$ was used in this experiment, where 100 objects and 90 items form roughly 4 co-clusters. Figure 3(a) shows the original whole data matrix, where black and white cells depict $r_{ij} = 1$ and $r_{ij} = 0$, respectively.

Vertically distributed cooccurrence submatrices were generated by arranging the 100 × 90 noisy matrix into four sites. Figure 3(b) shows the arranged cooccurrence matrix, where $m = 90$ items were divided into $(m_1, m_2, m_3, m_4) = (27, 24, 21, 18)$. Then, four co-cluster structures are very weakly implied in each site and the global co-cluster structure is only expected to be revealed in collaboration by all sites. This is a virtual situation of a group of four corporations, where they share 100 customers but have independent purchase history data on their own products. Here, the goal of collaborative fuzzy co-clustering is to reveal the intrinsic four customer clusters associated with their familiar products, which can be captured in the whole data strategy without privacy consideration but cannot be found in the site-wise independent analysis.

The co-clustering results of the distributed matrices are compared with that of whole data case, where the conventional FCCM algorithm was applied to the original 100 × 90 cooccurrence matrix $R = \{r_{ij}\}$ without privacy consideration. Figure 4 shows the item membership vectors given in the



(a) Original whole data matrix

(b) Four distributed cooccurrence matrices

Figure 3: Artificially generated data with homogeneous cluster partition.

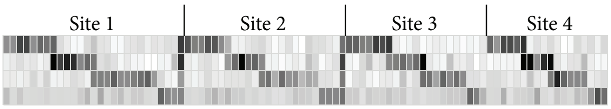


Figure 4: Item memberships of each cluster in full matrices case.

Table 1: Comparison of partition quality measured by correlation coefficients among item memberships (homogeneous partition case).

	Site 1	Site 2	Site 3	Site 4
Proposed model				
Best (Max.)	0.998	0.998	0.997	0.999
Mean	0.945	0.949	0.943	0.947
Site-wise FCCM				
Best (Max.)	0.913	0.889	0.935	0.946
Mean	0.718	0.677	0.851	0.903

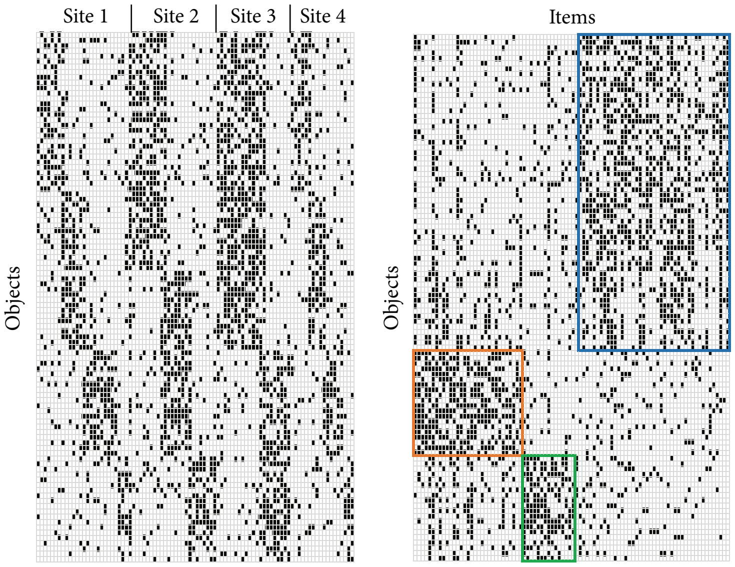
whole data case, where each row depicts 90-dimensional item membership vectors $\mathbf{w}_c = (w_{c1}, \ldots, w_{cm})^T$ of cluster c, $c = 1, \ldots, 4$. Each grayscale cell depicts the fuzzy membership $w_{cj} \in [0, w_{c\cdot}^{\max}]$, where black and white are $w_{cj} = w_{c\cdot}^{\max}$ and $w_{cj} = 0$, respectively. The goal is to estimate site-wise item memberships w_{cj}^t, which are as similar to the original w_{cj} as possible. Then, in this experiment, the similarity between original w_{cj} and site-wise w_{cj}^t is measured by their correlation coefficient.

Table 1 compares the correlation coefficients between the site-wise or proposed item memberships and the original result, where the best and the mean values in 50 trials with different initializations are depicted. In the site-wise FCCM, the conventional FCCM was applied to each submatrix (each small chunk) in each site. The fuzzification weights were set as $\lambda_u = 0.001$ and $\lambda_w = 100.0$, respectively. The table indicates that the proposed framework is useful for estimating reliable item memberships under collaboration of all sites while the derived item membership vectors are not necessarily equivalent to those of the whole data case because of site-wise independent constraints.

6.2. Data Set 2: Heterogeneous Cluster Partition. Next, the applicability of the proposed framework is investigated in a heterogeneous cluster partition case. The second artificial 100 × 90 cooccurrence matrix $R = \{r_{ij}\}$ was vertically distributed into 4 sites as shown in Figure 5(a), where $(m_1, m_2, m_3, m_4) = (27, 24, 21, 18)$. In contrast to the previous experiment, each site has different numbers of virtual co-clusters such that $(C_1, C_2, C_3, C_4) = (4, 3, 2, 4)$. This situation is similar to the case where four corporations in the group have different products characteristics and cannot have the real customer features without their collaboration.

The goal of collaborative co-cluster analysis is to reveal the intrinsic global co-cluster structures, which can be found only with global whole data. Applying the proposed secure framework with various cluster numbers, the FCCM algorithm could derive at most $C = 3$ co-clusters; that is, when $C > 3$, the 4th or later clusters consisted of a few noise objects only.

In order to intuitively validate the $C = 3$ co-clusters derived by the proposed framework, Figure 5(b) provides the arranged whole data matrix, where the all 90 items were first resorted in descending order of item fuzzy memberships of the first cluster in order to extract items of first cluster, and then, the remaining items were second resorted in descending order of the second cluster. Be noted that, in real applications, we cannot construct such whole data



(a) Four distributed cooccurrence matrices

(b) Arranged whole data matrix

Figure 5: Artificially generated data with heterogeneous cluster partition.

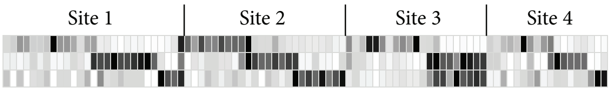


Figure 6: Item memberships of each cluster in heterogeneous partition case.

Table 2: Comparison of partition quality measured by correlation coefficients among item memberships (heterogeneous partition case).

	Site 1	Site 2	Site 3	Site 4
Proposed model				
Best (Max.)	0.998	0.998	0.997	0.998
Mean	0.810	0.825	0.978	0.796
Site-wise FCCM				
Best (Max.)	0.970	0.950	0.972	0.877
Mean	0.768	0.947	0.972	0.640

summary because of privacy issues but the figure was virtually constructed only for validation purposes in this experiment. This figure clearly supports the $C = 3$ co-clusters although it can be revealed only in collaborative analysis among multiple sites.

Figure 6 compares the item memberships derived by the proposed secure framework. Although sites 1 and 3 had different numbers of co-clusters from the global co-cluster structures, that is, $(C_1, C_3) = (4, 2)$, their co-cluster structures were also summarized into $C = 3$. In site 1, the first 2 co-clusters were merged into a solo co-cluster. On the other hand, in site 3, the second co-cluster was shared by two co-clusters because they cannot be distinguished in the global whole co-cluster structure.

Table 3: Comparison of partition quality measured by correlation coefficients among item memberships (terrorist attacks).

	Site 1	Site 2	Site 3	Site 4
Proposed model				
Best (Max.)	0.983	0.817	0.996	0.988
Mean	0.636	0.788	0.863	0.826
Site-wise FCCM				
Best (Max.)	0.969	0.644	0.805	0.544
Mean	0.617	0.601	0.644	0.477

Finally, the derived item memberships are compared with the whole data case, where we do not care about privacy issues. Table 2 compares the correlation coefficients between the site-wise or proposed item memberships and the whole data result. In the similar manner to the previous experiment, the table also supports the high performance of the proposed method in collaborative fuzzy co-cluster analysis.

6.3. Data Set 3: Terrorist Attacks. Third, the proposed secure framework is applied to a social network dataset. Terrorist attacks data set, which is available from LINQS webpage of Statistical Relational Learning Group @ UMD (http://linqs.cs.umd.edu/projects//index.shtml), consists of 1293 terrorist attacks each assigned to one of 6 labels indicating the type of the attack. Each attack is characterized by 106 distinct features with a 0/1-valued vector of attributes whose entries indicate the absence/presence of a feature. The goal of this experiment is to extract the structural knowledge on the terrorist attacks from the 1293 × 106 cooccurrence matrix.

In this experiment, a virtual situation of four allied states is considered, where the 106 distinct features are separately

Table 4: Comparison of cross tabulation tables of object partition (terrorist attacks).

(a)

	Cluster	Whole data FCCM			Proposed model		
		$c = 1$	$c = 2$	$c = 3$	$c = 1$	$c = 2$	$c = 3$
Class	Bombing	274	40	248	278	32	252
	Kidnapping	51	2	126	51	2	126
	Weapon-Attack	407	14	77	400	13	85

(b)

Cluster	Site-wise 1			Site-wise 2			Site-wise 3			Site-wise 4		
	$c = 1$	$c = 2$	$c = 3$	$c = 1$	$c = 2$	$c = 3$	$c = 1$	$c = 2$	$c = 3$	$c = 1$	$c = 2$	$c = 3$
Bombing	149	43	370	44	30	488	103	250	209	179	259	124
Kidnapping	51	1	127	22	50	107	49	32	98	37	73	69
Weapon-Attack	127	13	358	243	71	184	327	82	89	238	103	157

observed in the four states and they want to get a collaborative knowledge on the terrorist attacks without publishing their observed features such as military intelligences. The 106 features were distributed to the four states such as $(m_1, m_2, m_3, m_4) = (26, 26, 27, 27)$; that is, each state has only a part of the whole features ($1293 \times m_*$ matrices) but the states want to get a knowledge, which is given from the whole data case. Because three of six labeled classes have fewer numbers of objects (attacks), the characteristics of major three classes (bombing, kidnapping, and Weapon-Attack) are mainly discussed with $C = 3$.

First, the item memberships derived from the distributed matrices are compared with the whole data result. The whole data result was given by applying the conventional FCCM algorithm with $(\lambda_u, \lambda_w) = (0.001, 180.0)$. The goal is to estimate similar fuzzy memberships to the whole case result from the distributed matrices. The proposed framework and the site-wise FCCM were applied with $(\lambda_u, \lambda_w) = (0.0035, 100.0)$ and $(\lambda_u, \lambda_w) = (0.01, 100.0)$, respectively.

Table 3 compares the correlation coefficients between the site-wise or proposed item memberships and the whole data result. In a similar manner to the previous experiments, the collaborative knowledge is much more efficient than the site-wise one. This result implies the applicability of the proposed framework in strategic collaboration of allied states.

Next, the cross tabulations of the labeled class and clusters are compared for validating the utility of object partitions. In Table 4, the three main classes are compared with the maximum membership cluster assignment. Although the site-wise models derived quite degraded object partitions only, the proposed collaborative model could reconstruct almost equivalent result to the whole data case.

These results show the proposed model efficiently achieves secure co-clustering from both object and item partitions view points and is suitable for co-clustering tasks.

7. Conclusions

In this paper, a novel framework for collaborative fuzzy co-cluster analysis was proposed, in which vertically distributed cooccurrence matrices can be jointly analyzed with personal privacy preservation. In joint calculation of object fuzzy memberships, a secure encryption operation was adopted for calculating cluster-wise typicalities without broadcasting each element of individual cooccurrence matrices. Then, item fuzzy memberships are securely estimated in each site. Several experimental results demonstrated that collaborative analysis can contribute to revealing global intrinsic co-cluster structures of separate matrices rather than individual site-wise analysis.

The proposed framework is expected to enhance the collaborative utilization of many distributed databases, such as strategic marketing in corporation groups, collaborative medical development in hospitals, and strategic military actions in allied countries because they have a potential of sharing common knowledge withholding their independent sensitive information.

A possible future work is to evaluate the responsibility (utility) degree of each site. In the present model, each site is equally responsible for clustering estimation while some sites may have unreliable independent information only. Because the site-wise sum-to-one condition on item memberships can bring an undesirable influence of sites with low confidences, the responsibility of each site should be evaluated considering their confidences and should be fairly reflected in object membership calculation. Noise rejection mechanism [21, 22] would be promising in removing unreliable sites.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

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] C. C. Aggarwal and P. S. Yu, *Privacy-Preserving Data Mining: Models and Algorithms*, Springer, New York, NY, USA, 2008.

[2] P. Samarati, “Protecting respondents’ identities in microdata release,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001.

[3] L. Sweeney, “*k*-anonymity: a model for protecting privacy,” *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[4] J. Vaidya and C. Clifton, “Privacy-preserving *K*-means clustering over vertically partitioned data,” in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03)*, pp. 206–215, Washington, DC , USA, August 2003.

[5] T.-K. Yu, D. T. Lee, S.-M. Chang, and J. Zhan, “Multi-party *k*-means clustering with privacy consideration,” in *Proceedings of the International Symposium on Parallel and Distributed Processing with Applications*, pp. 200–207, 2010.

[6] R. A. McAllister and R. A. Angryk, “Abstracting for dimensionality reduction in text classification,” *International Journal of Intelligent Systems*, vol. 28, no. 2, pp. 115–138, 2013.

[7] T. C. Havens and J. C. Bezdek, “A new formulation of the coVAT algorithm for visual assessment of clustering tendency in rectangular data,” *International Journal of Intelligent Systems*, vol. 27, no. 6, pp. 590–612, 2012.

[8] K. Kummamuru, A. Dhawale, and R. Krishnapuram, “Fuzzy co-clustering of documents and keywords,” in *Proceedings of the IEEE International Conference on Fuzzy Systems*, vol. 2, pp. 772–777, May 2003.

[9] K. Honda, A. Notsu, and H. Ichihashi, “Collaborative filtering by sequential user-item co-cluster extraction from rectangular relational data,” *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 2, no. 4, pp. 312–327, 2010.

[10] K. Honda, M. Muranishi, A. Notsu, and H. Ichihashi, “FCM-type cluster validation in fuzzy co-clustering and collaborative filtering applicability,” *International Journal of Computer Science and Network Security*, vol. 13, no. 1, pp. 24–29, 2013.

[11] C.-H. Oh, K. Honda, and H. Ichihashi, “Fuzzy clustering for categorical multivariate data,” in *Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, pp. 2154–2159, July 2001.

[12] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, 1981.

[13] S. Miyamoto and M. Mukaidono, “Fuzzy *c*-means as a regularization and maximum entropy approach,” in *Proceedings of the 7th International Fuzzy Systems Association World Congress*, vol. 2, pp. 86–92, 1997.

[14] S. Miyamoto and K. Umayahara, “Methods in hard and fuzzy clustering,” in *Soft Computing and Human-Centered Machines*, Z.-Q. Liu and S. Miyamoto, Eds., Computer Science Workbench, pp. 85–129, Springer, Tokyo, Japan, 2000.

[15] K. Honda, S. Oshio, and A. Notsu, “FCM-type fuzzy co-clustering by K-L information regularization,” in *Proceedings of the IEEE International Conference on Fuzzy Systems*, pp. 2505–2510, 2014.

[16] Y. Kanzawa and Y. Endo, “On FNM-based and RFCM-based fuzzy co-clustering algorithms,” in *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE '12)*, pp. 1–8, Brisbane, Australia, June 2012.

[17] Y. Kanzawa, “Fuzzy co-clustering algorithms based on fuzzy relational clustering and TIBA imputation,” *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 18, no. 2, pp. 182–189, 2014.

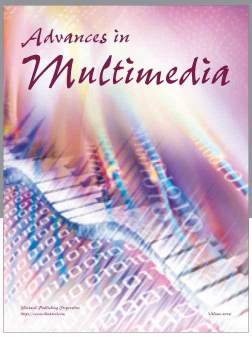
[18] Y. Kanzawa, “On Bezdek-type fuzzy clustering for categorical multivariate data,” in *Proceedings of the Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS)*, pp. 694–699, Kita-Kyushu, Japan, December 2014.

[19] S. Miyamoto and K. Umayahara, “Fuzzy clustering by quadratic regularization,” in *Proceedings of the IEEE International Conference on Fuzzy Systems and IEEE World Congress on Computational Intelligence*, vol. 2, pp. 1394–1399, May 1998.

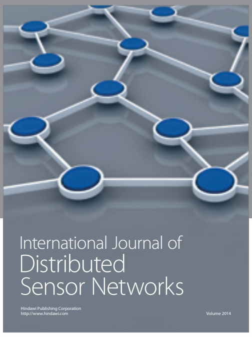
[20] K. Honda and H. Ichihashi, “Regularized linear fuzzy clustering and probabilistic PCA mixture models,” *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, pp. 508–516, 2005.

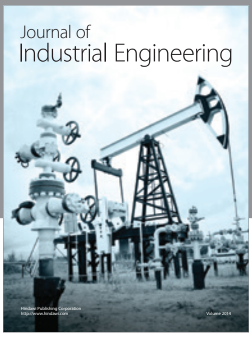
[21] R. N. Davé, “Characterization and detection of noise in clustering,” *Pattern Recognition Letters*, vol. 12, no. 11, pp. 657–664, 1991.

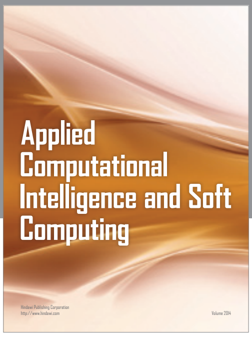
[22] R. N. Davé and R. Krishnapuram, “Robust clustering methods: a unified view,” *IEEE Transactions on Fuzzy Systems*, vol. 5, no. 2, pp. 270–293, 1997.

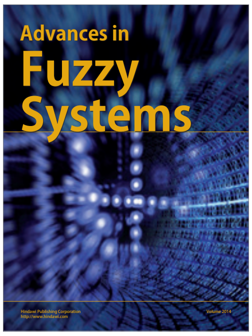

Advances in
Multimedia

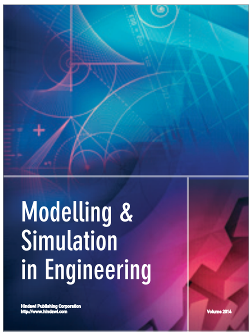
The Scientific
World Journal

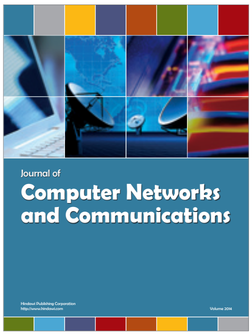

International Journal of
Distributed
Sensor Networks

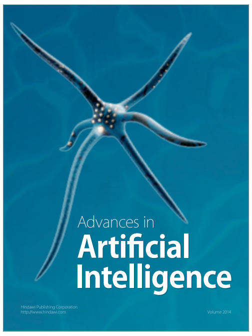

Journal of
Industrial Engineering

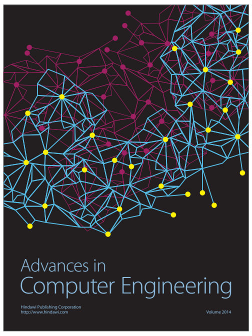

Applied
Computational
Intelligence and Soft
Computing


Advances in
Fuzzy
Systems

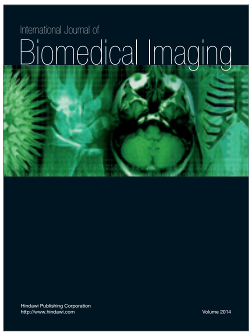

Modelling &
Simulation
in Engineering

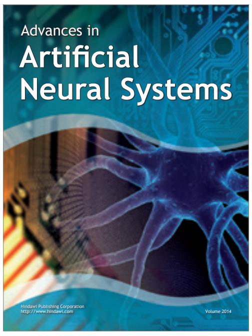

Journal of
Computer Networks
and Communications

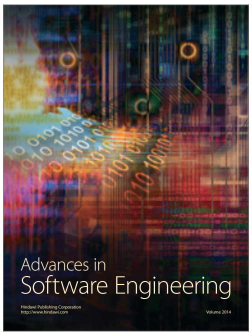

Advances in
Artificial
Intelligence


Advances in
Computer Engineering

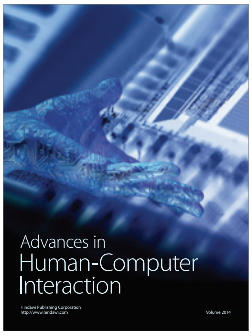
International Journal of
Computer Games
Technology

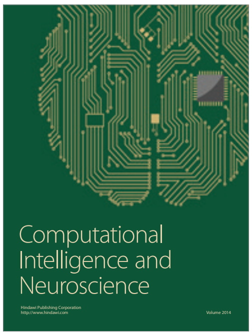

International Journal of
Biomedical Imaging


Advances in
Artificial
Neural Systems


Advances in
Software Engineering

Journal of
Robotics


Advances in
Human-Computer
Interaction


Computational
Intelligence and
Neuroscience

International Journal of
Reconfigurable
Computing

Journal of
Electrical and Computer
Engineering