A Greedy Algorithm for \( k \)-Member Co-clustering and Its Applicability to Collaborative Filtering

Arina Kawano\textsuperscript{a}, Katsuhiro Honda\textsuperscript{a}, Hirohide Kasugai\textsuperscript{a}, Akira Notsu\textsuperscript{a}

\textsuperscript{a}Osaka Prefecture University, 1-1 Gakuen-cho, Nakaku, Sakai, Osaka 599-8531, Japan

Abstract

Privacy preserving data mining is an important issue in network societies and co-clustering is a basic technique for analyzing intrinsic data structures in cooccurrence information among objects and items. In this paper, a greedy algorithm for \( k \)-member clustering, which achieves \( k \)-anonymity by coding at least \( k \) records into a solo observation, is enhanced to a co-clustering model. In the greedy algorithm, \( k \)-member clusters are sequentially extracted one-by-one, where each cluster is composed of homogeneous objects. In numerical experiments, the applicability of the proposed algorithm to collaborative filtering tasks is discussed.

© 2013 The Authors. Published by Elsevier B.V. Open access under \textit{CC BY-NC-ND} license.

Keywords: Co-clustering; \( k \)-anonymity; Collaborative Filtering

1. Introduction

Privacy preserving data mining [1] is an important issue in exploiting various databases in network societies. \( k \)-anonymity [2, 3] is a qualitative measure for evaluating the degree of secureness, in which an object is indistinguishable from at least \( k \) other objects. \( k \)-anonymity is practically achieved by coding at least \( k \) objects into a solo observation, and the objects should be as similar as possible so that information loss (coding cost) is minimized. A greedy procedure is \( k \)-member clustering [4], which sequentially extracts clusters with predefined sizes.

Co-clustering is a basic technique for summarizing intrinsic co-cluster structures from cooccurrence information among objects and items such as document-keyword cooccurrence and user-item purchase histories. Memberships of objects and items are simultaneously estimated so that each co-cluster is composed of objects and items, which are mutually related. Co-clustering has been achieved based on various frameworks such as spectral clustering [5], neural networks [6], fuzzy clustering [7] and evolutional computation [8]. However, most of these methods can control the number of clusters but cannot tune the size of each cluster, i.e., the number of objects in each cluster.

In order to achieve co-clustering in a privacy preserving manner, a \( k \)-member co-clustering model is considered in this paper, in which all clusters are guaranteed to have their sizes of \( k \) or larger. A greedy algorithm is
constructed so that each co-cluster has at least \( k \) objects and their familiar items. From the privacy preserving view point, the number of objects in each cluster is constrained to be equal to or larger than \( k \) while such size constraint is not force to the number of items. Characteristic features and the applicability to collaborative filtering tasks are demonstrated in numerical experiments. Collaborative filtering \([9, 10]\) is a kind of information filtering techniques, which can be used for reducing information overload, and achieves personalized recommendation by searching for user neighborhood comparing user preference histories such as purchase history data.

The remaining part of this paper is organized as follows: Section 2 briefly reviews \( k \)-member clustering and co-clustering. A novel greedy algorithm for \( k \)-member co-clustering is proposed in Section 3 and several experimental results are shown in Section 4. A summary conclusion is presented in Section 5.

2. Brief Review on Conventional Methods

2.1. \( k \)-Anonymization by \( k \)-Member Clustering

Assume that we have \( m \)-dimensional observations on \( n \) objects \( x_i = (x_{i1}, \ldots, x_{im})^\top, i = 1, \ldots, n \). \( k \)-anonymization \([2, 3]\) is a framework of preserving personal privacy from information leaks, in which an original database is transformed into a secure one that can be published without fears. \( k \)-anonymity is conceptually achieved by forcing at least \( k \) objects to have same observations through generalization and suppression of observations. However, the problem of finding optimal \( k \)-anonymous tables is an NP-hard problem and is computationally expensive \([11]\).

With the goal being to minimize anonymization costs (losses of information), the \( k \) objects coded into a same record should be as homogeneous as possible. Then, the problem can be identified with extraction of clusters having a certain size (\( k \) or larger members). Besides the conventional clustering methods, in which the number of clusters can be controlled, \( k \)-member clustering deals with a predefined size clusters so that the \( k \)-anonymity of each cluster is guaranteed. Byun et al. \([4]\) proposed an efficient greedy algorithm for \( k \)-member clustering, which can be applied to both (or the mixture of) numerical and categorical observations. This merging process is essentially similar to the nearest-neighbor (or single-linkage) principle. First, a starting point is chosen such that it is the furthest objects from a randomly selected object. Then, \( k - 1 \) nearest objects are merged into a \( k \)-member cluster. After extracting a \( k \)-member cluster, next cluster core is chosen such that it is the furthest object from the last merged one. This iterative process is continued until the number of remaining objects is smaller than \( k \), and then, the remaining ones are assigned to their nearest clusters. The prototypical observations of each cluster can be given by interval values for numerical one or mixed categories for categorical one.

In \([4]\), the best neighbor was selected so that the sum of the interval of numerical observation and/or the number of the merged category of categorical observation is minimized after merging. The information loss in cluster \( G_c \), which is a combination of sum of interval for nominal observations and height of categorical taxonomy for nominal observations, is measured as follows:

\[
IL_c = |G_c|\left(\sum_{i} \frac{\text{max}_{ci} - \text{min}_{ci}}{\text{Size}_i} + \sum_{j} \frac{H(\Lambda(\cup C_j))}{H(T_c)}\right),
\]

where \( \text{Size}_i \) is the size of numeric domain of numeric attribute \( i \), and \( \text{max}_{ci} \) and \( \text{min}_{ci} \) are the maximum and minimum values in \( G_c \). \( T_c \) is the taxonomy tree defined for the domain of categorical attribute \( C_j \) and \( H(T) \) is the height of taxonomy tree \( T \). \( \Lambda(\cup C_j) \) measures the deviation in \( G_c \).

A fuzzy variant of \( k \)-member clustering was also proposed \([12]\) and was demonstrated to be able to achieve \( k \)-anonymity with lower information loss. In fuzzy \( k \)-member clustering, each object can belong to multiple clusters with its fuzzy membership degrees. This soft partitioning model seems to be useful for handling ambiguous cluster boundaries in noisy data sets.

2.2. \( k \)-Means-type Clustering and Co-clustering

In the \( k \)-Means clustering \([13]\), the goal is to partition \( n \) objects into \( k \) disjoint clusters, which are represented by prototypical centroids \( b_c \). The clustering criterion is the distance between objects and cluster centroids and the objective function is the within-cluster errors to be minimized. Starting with a random centroids distribution, an
iterative process of updating object assignment and cluster centroids distribution is continued until convergent. In k-Means clustering, only the homogeneity of objects in clusters is considered.

The partition quality should be evaluated comparing the within-cluster compactness and the inter-cluster separateness. A well-established hard cluster validity criterion is the separation index $D_1$ proposed by Dunn [14], which is designed for finding compact and separate (CS) clusters:

$$D_1 = \min_{i \leq i \leq K} \left\{ \min_{i + 1 \leq j \leq K - 1} \left\{ \frac{\text{dis}(G_i, G_j)}{\max_{1 \leq k \leq K} (\text{dia}(G_k))} \right\} \right\},$$

where

$$\text{dia}(G_i) = \max_{i \neq G_k} d(x_i, x_j),$$

$$\text{dis}(G_k, G_c) = \min_{i \in G_k, j \in G_c} d(x_i, x_j).$$

$d$ is a metric between two objects. A similar measure was also considered in the fuzzy clustering context such as Xie-Beni index [15].

In co-clustering tasks, a problem space is given with a cooccurrence matrix among objects and items. Assume that we have an $n \times m$ rectangular relational data matrix $R = \{r_{ij}\}$ composed of cooccurrence of $n$ objects and $m$ items, and the goal is to simultaneously partition objects and items into a certain number of co-clusters, in which objects and items in a same cluster are mutually familiar. In $k$-Means type co-clustering, the clustering criterion is given by the degree of aggregation to be maximized in each cluster.

Fuzzy clustering for categorical multivariate data (FCCM) [7] is a $k$-Means-type fuzzy co-clustering model, which achieves dual partition of objects and items by finding co-clusters having high aggregation degrees. FCCM modified the objective function of Fuzzy c-Means (FCM) [16, 17] using aggregation degrees of clusters to be maximized as:

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

$u_{ci}$ is the fuzzy membership of object $i$ to cluster $c$ while $w_{cj}$ is the fuzzy membership of item $j$ to cluster $c$. Here, based on the probabilistic concept, $u_{ci}$ is constrained to be exclusive such that $\sum_{c=1}^{C} u_{ci} = 1$. On the other hand, another type of constraint is forced to $w_{cj}$, which is $\sum_{j=1}^{m} w_{cj} = 1$ in each cluster. The entropy terms were adopted for membership fuzzification based on the entropy-based regularization approach [17]. Weight parameters $\lambda_u$ and $\lambda_w$ tune the degree of fuzziness of $u_{ci}$ and $w_{cj}$, respectively. The clustering algorithm is an FCM-type iterative optimization procedure for $u_{ci}$ and $w_{cj}$.

In order to evaluate the partition quality in the co-clustering context, a compactness/separateness-based measure was proposed for FCCM [18], which is a co-clustering version of Xie-Beni index:

$$V_{XBCO} = \frac{\text{compactness}}{\text{separateness}} = \frac{(C - 1) \sum_{c=1}^{C} \sum_{i=1}^{n} \sum_{j=1}^{m} u_{ci}w_{cj}(2r_{ij} - 1)}{\sum_{c=1}^{C} \sum_{i=1}^{n} \sum_{j=1}^{m} u_{ci}w_{cj}r_{ij}},$$

where $r_{ij}$ is assumed to be normalized so that $r_{ij} \in [0, 1]$. The larger the value of $V_{XBCO}$, the more compact and separate the fuzzy co-cluster partition is. So, we can find the optimal co-cluster partition by searching for the largest $V_{XBCO}$.

A potential application of co-clustering is collaborative filtering [9, 10], whose goal is to recommend a promising item to users. The problem space of collaborative filtering is given as a user-item relational data matrix and...
personalized recommendation is achieved by predicting the applicability of each item to each user by considering mutual similarities among users and items. Then, the prediction process can be identified with co-clustering, in which familiar users and items are associated into a group. In the previous researches, it has been demonstrated that co-cluster structures with high aggregation degrees are useful for recommending promising items to users [19, 18].

3. A Greedy Algorithm for $k$-Member Co-clustering

In this paper, a novel greedy algorithm for $k$-member co-clustering is proposed. In order to apply to co-clustering tasks, the conventional greedy $k$-member clustering algorithm is modified so that it can be used in conjunction with the aggregation degree to be maximized.

Assume that the cooccurrence relation $r_{ij}$ is drawn from the interval as $r_{ij} \in [0, 1]$, and objects and items are partitioned based on different mechanisms. Objects are partitioned into $k$-member clusters in a crisp manner, in which each object can belong to a solo cluster. On the other hand, items are associated to clusters with their responsibility weights for characterizing each cluster. Then, the within-cluster aggregation degree $p_t$ to be maximized in cluster $G_t$ is defined as:

$$p_t = \sum_{i \in G_t} \sum_{j=1}^{m} w_{ij} r_{ij} - \lambda w \sum_{j=1}^{m} w_{ij} \log w_{ij},$$

where $w_{ij}$ is the membership degree of item $j$ in cluster $G_t$ and is constrained to be $\sum_{j=1}^{m} w_{ij} = 1$. The entropy term is added for fuzzification of memberships so that $w_{ij}$ represent the relative (soft) responsibility degrees. Considering the necessary condition for the optimality, $w_{ij}$ is calculated as:

$$w_{ij} = \frac{\exp\left(\lambda^{-1} \sum_{i \in G_t} r_{ij}\right)}{\sum_{\ell=1}^{m} \exp\left(\lambda^{-1} \sum_{i \in G_t} r_{i\ell}\right)}.$$  

The greedy $k$-member clustering algorithm can be modified as follows:

[Greedy $k$-member co-clustering algorithm]
1. Let $S$ be a set of objects. Choose the anonymity level $k$ and randomly select an object $r$.
2. Let a cluster index $t$ be 0. Repeat the following process while $|S| > k$.
   (a) Replace $r$ with its furthest object and remove $r$ from $S$.
   (b) $t = t + 1$. Generate cluster $G_t$ with a single element $r$.
   (c) Repeat the following process while $|G_t| < k$.
      i. Find the best neighbor object $r$ of cluster $G_t$, which has the largest within-cluster aggregation of Eq.(7) after merging.
      ii. Add $r$ to cluster $G_t$ and remove $r$ from $S$.
3. Repeat the following process while $|S| > 0$.
   (a) Randomly select an object $r$ from $S$.
   (b) Find the best neighbor cluster $G_t$ of $r$, where the object has the largest within-cluster aggregation of Eq.(7) after merging.
   (c) Add $r$ to cluster $G_t$ and remove $r$ from $S$.

4. Numerical Experiments

Numerical experiments were performed with a real-world purchase history data in order to demonstrate the characteristic features of the proposed co-clustering model. The purchase history data set used in [19] was col-
lected by Nikkei Inc. in 2000 and includes the purchase history of 996 users (objects, \( n = 996 \)) on the 18 items (\( m = 18 \)) shown in Table 1, where the number of users who have the item is also presented.

Table 1. 18 items and the number of users who have each item.

<table>
<thead>
<tr>
<th>Item</th>
<th># of owners</th>
<th>Item</th>
<th># of owners</th>
<th>Item</th>
<th># of owners</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car</td>
<td>825</td>
<td>Piano</td>
<td>340</td>
<td>VTR</td>
<td>933</td>
</tr>
<tr>
<td>Air Conditioner</td>
<td>911</td>
<td>PC</td>
<td>588</td>
<td>Word Processor</td>
<td>506</td>
</tr>
<tr>
<td>CD</td>
<td>844</td>
<td>VD</td>
<td>325</td>
<td>Motorcycle</td>
<td>294</td>
</tr>
<tr>
<td>Bicycle</td>
<td>893</td>
<td>Refrigerator</td>
<td>858</td>
<td>Small Refrigerator</td>
<td>206</td>
</tr>
<tr>
<td>Microwave Oven</td>
<td>962</td>
<td>Oven</td>
<td>347</td>
<td>Coffee Maker</td>
<td>617</td>
</tr>
<tr>
<td>Washing Machine</td>
<td>986</td>
<td>Drying Tumbler</td>
<td>226</td>
<td>Dishwasher</td>
<td>242</td>
</tr>
</tbody>
</table>

The element \( r_{ij} \) of 996 × 18 relational data matrix \( R = [r_{ij}] \) is 1 if user \( i \) has item \( j \) while otherwise 0. Here, we should note that ‘0’ elements does not necessarily mean a negative feeling but may mean ‘the user has not bought yet but may buy near future’. So, the goal of collaborative filtering is to predict the possibility of purchase of ‘has-not-bought-yet’ items for recommending the items to be bought in near future.

Randomly selected 1,000 elements of the data matrix were given as a test data set and were withheld from the training data set, i.e., the data set was partitioned into two subsets of the training and test sets. The training set composes the 996 × 18 relational data matrix \( R = [r_{ij}] \), in which all the 1,000 test element \( r_{ij} \) were withheld such that all of them were replaced with 0, i.e., unknown.

In order to evaluate the usefulness of the proposed anonymization model, the applicability in collaborative filtering is compared with the conventional \( k \)-member clustering.

4.1. Comparison of Aggregation Quality

First, the quality of data aggregation is compared. The 996 × 18 relational training data set was anonymized by the conventional \( k \)-member clustering and the proposed \( k \)-member co-clustering models. The conventional model was applied in such a way that the cooccurrence information related to object \( i \) is regarded as an \( m \)-dimensional vector observation \( r_i = (r_{i1}, \ldots, r_{im})^T \) and the \( k \)-member clustering process was performed considering the mutual similarities among objects only. On the other hand, the proposed \( k \)-member co-clustering model was implemented such that each cluster consists of similar objects in conjunction with association of related items, where membership fuzzification parameter was set as \( \lambda_w = 5.0 \). Both two \( k \)-member clustering algorithms were implemented with the same starting object but gave different cluster partitions because of using different similarity measures. \( k \)-member clusters were extracted with various anonymity levels of \( k = \{5, 10, 20, 30, 40, 50, 60, 70, 80\} \) and the average aggregation degrees \( p_i/n \) are compared in Fig. 1.

![Fig. 1. Comparison of average aggregation degrees.](image-url)
The figure indicates that the aggregation degree of object-item cooccurrence in the proposed model is higher than that in the conventional model. The $k$-member co-clusters extracted by the proposed model is useful for revealing the intrinsic co-cluster structures in a privacy preserving manner.

4.2. Comparison of Applicability of $k$-member Clusters in Collaborative Filtering

Second, the applicability of $k$-member clusters in collaborative filtering is compared, where $k$-member clusters are directly used for finding promising items. The recommendation ability for the test data set withheld from the training set was tested, where a test record in the test set is composed of object index, item index and its purchase history, i.e., the third attribute is 1 when the user has the item, while otherwise 0. The applicability of the item was predicted using its membership value in the cluster the object belongs to, i.e., items having large memberships are recommended.

The recommendation ability was assessed by ROC sensitivity [20], which is a true positive rate vs. false positive rate plots drawn by changing the threshold of the applicability level (fuzzy membership) in recommendation. The lower area of the curve becomes large as the recommendation ability is higher, i.e., a higher value means better recommendation ability.

![Fig. 2. Comparison of recommendation ability in direct use of $k$-member clusters.](image)

Figure 2 compares the ROC sensitivity calculated with various anonymity levels $k$. The figure indicates that the recommendation ability of the $k$-member co-clusters given by the proposed model is inferior to that by the conventional one in the case of low anonymity levels ($k \leq 30$) but is superior in the case of high anonymity levels ($k \geq 30$). When the anonymity level is low, the proposed model might extract many small co-clusters having high aggregation degrees, in which only a small number of items having high cooccurrence are selected, and might reject most of items in recommendation process. Because of high rejection rates, $k$-member co-clusters with low anonymity levels are not suitable for direct use in collaborative recommendation. On the other hand, when the anonymity level is high and each $k$-member cluster has many objects, rejection rates become smaller and the plausibility of items in each cluster can be fairly evaluated. So, the recommendation ability of the proposed model is better than that of the conventional one.

However, the above recommendation quality is not necessarily better than that by FCCM, whose best ROC sensitivity is more than 0.83 [18]. It is because the proposed $k$-member co-clustering model is mainly designed for realizing $k$-anonymization of the original data set and should be used for preprocessing in data publication, where the published data are further analyzed for later applications. Then, in the next experiment, the availability of published data sets is considered.

4.3. Comparison of Applicability of Published Data in Hybrid Use with GroupLens Recommendation

Third, the applicability of $k$-member clusters is compared in hybrid use with GroupLens Recommendation [9]. $k$-anonymization is often performed in preprocessing for privacy preserving data mining so that the data are pub-
lished in a secure manner. In this experiment, the situation, in which data sets anonymized by \( k \)-member clustering are utilized in collaborative filtering tasks with GroupLens recommendation algorithm after publication, is considered.

In order to publish the cooccurrence information considering the estimated co-cluster structure, each record of the original training data matrix (a row \( r_i \) of \( 996 \times 18 \)) was replaced with the item membership vector \( w_c = (w_{c1}, \ldots, w_{cm})^\top \) of the cluster, which the object belongs to.

GroupLens is a most well-established algorithm for personalized recommendation. The memory-based algorithm is composed of two phases. In the first phase, neighborhood users of the active user are searched for based on nearest neighbor principle, where the Pearson correlation coefficient among user profiles is often used. In the second phase, the applicability of each item for the active user is predicted by the weighted average in the user neighborhood. This recommendation process can be identified with a pseudo-model of *word-of-mouth* in human societies. A missing element \( y_{ij} \approx r_{ij} \) is predicted by calculating the similarity-weighted average as follows:

\[
y_{ij} = \bar{r}_i + \frac{\sum_{a=1}^n (r_{aj} - \bar{r}_a) \times cor_{ia}}{\sum_{a=1}^n cor_{ia}},
\]

where \( cor_{ia} \) is the Pearson correlation coefficient between users \( i \) and \( a \). \( \bar{r}_i \) is the mean rating value of user \( i \). In GroupLens, it was recommended to use the deviations from mean value of users \( (r_{aj} - \bar{r}_a) \) instead of original values \( (r_{aj}) \) in order to remove the influence of users’ tendencies.

Figure 3 compares the ROC sensitivity in a hybrid use with GroupLens after anonymization with various anonymity levels \( k \). The recommendation ability becomes lower as the anonymity level becomes higher. It is because a higher anonymity level causes a greater loss of original information. The proposed model could achieve higher recommendation ability with lower information loss than the conventional one even when the anonymity level is higher.

These results imply that the proposed \( k \)-member co-clustering model has higher applicability for privacy preserving data mining of cooccurrence information.

5. Conclusions

This paper proposed a novel approach for anonymizing cooccurrence information through \( k \)-member co-clustering. The conventional \( k \)-member clustering for vector observations is modified so that it can handle cooccurrence information, where the goal is to extract co-clusters composed of familiar objects and items. The availability of the proposed model was demonstrated through three types of experiments with a real-world purchase history data set. First, the quality of co-clusters was evaluated considering the aggregation degrees of the extracted clusters. Second, the applicability of the extracted co-cluster structures was evaluated through a direct application to
collaborative recommendation. Third, considering the availability of published data sets after anonymization, the recommendation ability in a hybrid use with GroupLens recommendation algorithm was evaluated. The experimental results fairly demonstrated the higher availability of the proposed co-clustering model than the conventional one.

Possible future work includes application to other tasks such as document-keyword analysis. Extension to fuzzy variants [7, 12] and introduction of exclusive condition [21] are also included in future work.

Acknowledgements

This work was supported in part by the Japan Science and Technology Agency (JST) through Support for Accelerating Utilization of University IP, and by the Ministry of Education, Culture, Sports, Science and Technology, Japan, under Grant-in-Aid for Scientific Research (23500283).

References