Multiple Co-Clusterings

Xing Wang[1], Guoxian Yu[1], Carlotta Domeniconi[2], Jun Wang[1,*], Zhiwen Yu[3], Zili Zhang[1,4]
[1]College of Computer and Information Sciences, Southwest University, China
[2]Department of Computer Science, George Mason University, USA
[3]School of Computer Science and Engineering, South China University of Technology, China
[4]School of Information Technology, Deakin University, Geelong, VIC 3220, Australia
Email: [1]{xingwang,gxyu,kingjun,zhangzl}@swu.edu.cn, [2]carlotta@cs.gmu.edu, [3]zhwyu@scut.edu.cn

***Abstract*—The goal of multiple clusterings is to discover multiple independent ways of organizing a dataset into clusters. Current approaches to this problem just focus on *one-way* clustering. In many real-world applications, though, it's meaningful and desirable to explore alternative *two-way* clustering (or co-clusterings), where both samples and features are clustered. To tackle this challenge and unexplored problem, in this paper we introduce an approach, called Multiple Co-Clusterings (MultiCC), to discover non-redundant alternative co-clusterings. MultiCC makes use of matrix tri-factorization to optimize the sample-wise and feature-wise co-clustering indicator matrices, and introduces two non-redundancy terms to enforce diversity among co-clusterings. We then combine the objective of matrix tri-factorization and two non-redundancy terms into a unified objective function and introduce an iterative solution to optimize the function. Experimental results show that MultiCC outperforms existing multiple clustering methods, and it can find interesting co-clusters which cannot be discovered by current solutions.**

***Index Terms*—multiple clusterings, co-clustering, matrix tri-factorization, redundancy.**

I. Introduction

Clustering is a fundamental problem in unsupervised machine learning. Traditional clustering methods just find a single data partition. However, when clustering complex data, many solutions may exist, and each one may provide a reasonable grouping of the data [1]. Given that, multiple clustering solutions have been developed to explore alternative clusterings. Naive solutions explore alternative clusterings by (i) running a clustering algorithm multiple times, using different parameter values each time; or (ii) running different clustering algorithms; or (iii) running a combination of the above two strategies [2]. These approaches may generate multiple clusterings with high redundancy, since they do not take into account the already explored clusterings. To overcome this drawback, two general strategies are introduced. The first one simultaneously generates multiple clusterings, which are required to be different from each other [3]. The second one generates multiple clusterings in a greedy manner, and forces the new clusterings to be different from the previously generated ones [4]–[6].

However, typically multiple clustering algorithms only consider one-way clustering, i.e., they cluster samples based on their feature similarity. But in many real-world applications, it is meaningful and desirable to explore alternative two-way clusterings (or co-clusterings) [7], where both samples and features are simultaneously clustered. For example, in collaborative filtering [8], [9], the aim is to produce multiple groups of "users" and "items", while in gene expression data analysis, the goal is to output groups of "samples" and "genes" [10].

To uncover multiple co-clusterings from data, we propose a solution called Multiple Co-Clusterings (MultiCC). Motivated by the use of semi-nonnegative matrix factorization in co-clustering [8], [11], MultiCC repeatedly factorizes the data matrix $\mathbf{X}$ into $\mathbf{R}^h\mathbf{S}^h(\mathbf{C}^h)^T$ to obtain alternative co-clusterings, where $\mathbf{R}^h \geq 0$ and $\mathbf{C}^h \geq 0$ correspond to the row-cluster and column-cluster indicator matrices of the h-th co-clustering. $\mathbf{S}^h$ plays the role of absorbing the different scaling factors of $\mathbf{R}^h$ and $\mathbf{C}^h$ to minimize the squared error. MultiCC also defines redundancy measurement terms based on $\mathbf{R}^h$ and $\mathbf{C}^h$, and minimizes them to enforce diversity among alternative co-clusterings. To ensure high quality and diverse co-clusterings, MultiCC integrates the matrix factorization and the terms measuring redundancy into a unified objective function, and uses an iterative solution to optimize $\mathbf{R}^h$ and $\mathbf{C}^h$.

The paper makes the following contributions:

- We study the multiple co-clustering problem to uncover co-clusters of data from different perspectives. To the best of our knowledge, this is a largely *unexplored* topic, and has meaningful applications in real life scenarios.
- We introduce a matrix factorization based approach called MultiCC to find alternative co-clusterings of high quality and diversity.
- Experimental results show that MultiCC significantly outperforms related approaches [3], [5], [6], [12], [13] in the discovery of multiple clusterings, and it can uncover co-clusterings which are more diverse from one another.

The remainder of this paper is organized as follows. We briefly review related work in Section II, and then elaborate on the proposed algorithm and its optimization in Section III. Section IV provides the experimental results and Section V gives the conclusion and future work.

II. Related Work

The problem of multiple clusterings has been studied for one decade [2]. Existing solutions for multiple clustering can be roughly categorized into two groups: unsupervised and semi-supervised. The unsupervised algorithms attempt to seek multiple diverse clusterings without reference to existing

*Corresponding author: kingjun@swu.edu.cn (Jun Wang)

clusterings, while semi-supervised ones sequentially generate multiple clusterings with reference to existing clusterings.

Within the semi-supervised category, COALA (Constrained Orthogonal Average Link Algorithm) [12] uses the existing clusterings to generate a "cannot-link" constraint for each pair of samples in the same cluster, and attempts to achieve a tradeoff between satisfying these cannot-links and ensuring high quality within a hierarchical clustering, but the result heavily depends on the quality of the generated "cannot-link" constraints. MSC (Multiple Stable Clusterings) [5] uses a simplex constraint to generate different sparse weights assigned to features, and then uses spectral clustering [14] to produce multiple stable clusterings. MNMF (Multiple clustering by Nonnegative Matrix Factorization) [6] integrates a diversity regularization term, constructed using the existing clusterings, with the objective function of NMF, to explore multiple clusterings in a sequential manner. Some researchers have explored alternative clusterings from the perspective of feature spaces [4], [13]. ADFT (Alternative Distance Function Transformation) [13] uses must-link and cannot-link constraints between instances to learn a distance function [15]. It then uses the distance function to compute a transformation matrix, and hence a subset of features to produce alternative clusterings.

Meta Clustering [16] is a well-known approach within the category of unsupervised multiple clusterings. It first assigns different weights to features based on the Zipf distribution [17], and then obtains multiple clusterings by applying k-means on the weighted features. This approach often generates redundant clusterings. Other methods attempt to seek alternative clusterings simultaneously by minimizing the correlation between two distinct clusterings. For example, Dec-kmeans (Decorrelated k-means) [3] generates alternative clusterings by minimizing the error terms of two individual k-means, and it uses the dot products of the mean vectors of the two respective clusterings to quantify their dissimilarity.

The aforementioned multiple clustering algorithms focus on finding multiple clusterings from a sample-wise or a feature-wise perspective, and some of them can only explore two alternative clusterings [3], [13]. In many practical applications, e.g. cancer genomic data analysis [10] and collaborative filtering [8], [9]), it is desirable to present multiple alternative co-clusterings to uncover the hidden patterns of the data matrix. We want to remark that exploring multiple co-clusterings is significantly more challenging than the widely studied multiple clustering scenario, since co-clustering must be simultaneously performed along both samples and features, instead of on samples or features alone. To address this challenge, we propose a novel Multiple Co-Clustering (MultiCC) solution. Our empirical study shows that MultiCC can uncover multiple diverse co-clusterings of good quality.

III. Methodology

A. Multiple Co-Clusterings

MultiCC aims at exploring multiple co-clusterings by simultaneously grouping the elements of a matrix along both samples and features, while enforcing the quality and diversity among the co-clusterings as much as possible. As such, there are two questions to be addressed in MultiCC: (1) how to find multiple co-clusterings with good *quality*, and (2) how to reduce the *redundancy* among the co-clusterings.

For the first question, to obtain multiple co-clusterings, MultiCC repeats semi-nonnegative matrix tri-factorization [11] on the data matrix. Co-clusterings are represented by the respective sample-wise and feature-wise co-clustering indicator matrices ($\mathbf{R}^h$ and $\mathbf{C}^h$, $h = 1, 2, \cdots, m$, where m is the target number of alternative co-clusterings) [18]–[20]. MultiCC pursues the relatively good quality by minimizing the residues of matrix factorization. For the second question, MultiCC quantifies the redundancy using the respective sample and feature co-clusters, so that it will keep the diversity among different co-clusterings.

Let $\mathbf{X} = \{\mathbf{x}_{\cdot 1}, \cdots, \mathbf{x}_{\cdot n}\} \in \mathbb{R}^{d \times n}$ be the data matrix for n samples represented by d numeric features. MultiCC repeatedly factorizes $\mathbf{X}$ into the product of three matrices as follows:

$$\Psi_1(\{\mathbf{R}^h\}_{h=1}^m, \{\mathbf{C}^h\}_{h=1}^m) = \frac{1}{m}\sum_{h=1}^{m} \parallel \mathbf{X} - \mathbf{R}^h\mathbf{S}^h(\mathbf{C}^h)^T \parallel^2 \quad s.t. \quad \mathbf{R}^h \geq 0; \mathbf{C}^h \geq 0 \tag{1}$$

where $\mathbf{R}^h \in \mathbb{R}^{d \times k_h}$ is the row-cluster indicator matrix, stating that the h-th co-clustering groups the d features into k_h row-clusters. If feature $\mathbf{x}_{i\cdot}$ belongs to the $k_h^{'}$-th row-cluster of the h-th co-clustering, $\mathbf{R}^h_{ik_h^{'}} = 1$; otherwise, $\mathbf{R}^h_{ik_h^{'}} = 0$. Similarly, $\mathbf{C}^h \in \mathbb{R}^{n \times l_h}$ is the column-cluster indicator matrix, stating that the h-th co-clustering groups the n samples into l_h column-clusters. If $\mathbf{x}_{\cdot j}$ belongs to the $l_h^{'}$-th cluster, $\mathbf{C}^h_{jl_h^{'}} = 1$; otherwise, $\mathbf{C}^h_{jl_h^{'}} = 0$. $\mathbf{S}^h \in \mathbb{R}^{k_h \times l_h}$ is introduced to account for the different number of row-clusters and column-clusters and to minimize the squared error induced by matrix factorization. To enable datasets which include negative values, MultiCC does not require $\mathbf{S}^h$ to be nonnegative.

To enforce dissimilarity among the various co-clusterings, we first define a co-association matrix $\mathbf{W}_r^h = \mathbf{R}^h * (\mathbf{R}^h)^T$ based on $\mathbf{R}^h$, where T denotes a matrix transpose. $(\mathbf{W}_r^h)_{ij}$ is the inner product between the i-th and the j-th rows of $\mathbf{R}^h$. Clearly, $\mathbf{W}_r^h(i,j) = 1$ if $\mathbf{x}_{i\cdot}$ and $\mathbf{x}_{j\cdot}$ are in the same row-cluster, and $\mathbf{W}_r^h(i,j) = 0$ otherwise. We can measure the total feature-wise redundancy, for each pair of row-clusterings, as follows:

$$\Psi_2(\{\mathbf{R}^h\}_{h=1}^m) = \sum_{\substack{h_1,h_2=1 \\ h_1 \neq h_2}}^{m} \sum_{i,j=1}^{d} \Theta_r \mathbf{W}_r^{h_1}(i,j)\mathbf{W}_r^{h_2}(i,j) \tag{2}$$

where $\Theta_r = \frac{1}{d^2} \times \frac{1}{k_{h_1} k_{h_2}}$ is the normalization factor aimed at reducing the influence of different numbers of features (d) and clusters (k_{h_1} and k_{h_2}) of two row clusterings. Eq. (2) measures the redundancy for all pairs of row-clusterings: the smaller the resulting value is, the smaller the frequency at which two features are grouped in the same row-cluster, and therefore the larger the diversity among row-clusterings is. From the

properties of the trace operation, we can reformulate Eq. (2) as follows:

$$\begin{aligned}\Psi_2(\{\mathbf{R}^h\}_{h=1}^m) &= \sum_{\substack{h_1,h_2=1\\h_1\neq h_2}}^{m} \Theta_r tr((\mathbf{R}^{h_1})^T\mathbf{R}^{h_1}\mathbf{W}_r^{h_2}) \\ &= \sum_{\substack{h_1,h_2=1\\h_1\neq h_2}}^{m} \Theta_r \parallel (\mathbf{R}^{h_1})^T\mathbf{R}^{h_2} \parallel^2\end{aligned} \tag{3}$$

Similarly, we can measure the total sample-wise redundancy for each pair of column-clusterings as follows:

$$\Psi_3(\{\mathbf{C}^h\}_{h=1}^m) = \sum_{\substack{h_1,h_2=1\\h_1\neq h_2}}^{m} \Theta_c \parallel (\mathbf{C}^{h_1})^T\mathbf{C}^{h_2} \parallel^2 \tag{4}$$

where $\mathbf{W}_c^h \in \mathbb{R}^{n\times n}$ is the co-association matrix based on $\mathbf{C}^h$, $\Theta_c = \frac{1}{n^2} \times \frac{1}{l_{h_1} l_{h_2}}$ is the normalization factor.

Based on the above, MultiCC integrates Eq. (1) with Eq. (3) and Eq. (4) to pursue m different co-clusterings via the following unified objective function:

$$\begin{aligned}J(\{\mathbf{R}^h\}_{h=1}^m, \{\mathbf{C}^h\}_{h=1}^m) &= \frac{1}{m}\sum_{h=1}^{m} \parallel \mathbf{X} - \mathbf{R}^h\mathbf{S}^h(\mathbf{C}^h)^T \parallel^2 \\ &+\frac{\lambda}{C_m^2} \sum_{\substack{h_1,h_2=1\\h_1\neq h_2}}^{m} (\Theta_r \parallel (\mathbf{R}^{h_1})^T\mathbf{R}^{h_2} \parallel^2 + \Theta_c \parallel (\mathbf{C}^{h_1})^T\mathbf{C}^{h_2} \parallel^2) \\ &s.t. \quad \mathbf{R}^{(h)} \geq 0; \mathbf{C}^{(h)} \geq 0\end{aligned} \tag{5}$$

where the regularization parameter $\lambda \geq 0$ controls the tradeoff between the quality of the m co-clusterings, which is pursued by the matrix tri-factorization, and the dissimilarity among these co-clusterings, which is pursued by the last two terms. To reduce the scale issue, the factors $1/C_m^2$ and $1/m$ are introduced, where $C_m^2 = m(m-1)/2$ is the total number of pairwise co-clusterings.

B. Optimization Algorithm

$\mathbf{R}^h$ and $\mathbf{C}^h$ are binary matrices, so directly minimizing Eq. (5) to find m alternative co-clusterings is very difficult. To address this issue, we relax the elements of $\mathbf{R}^h$ and $\mathbf{C}^h$ to have continuous nonnegative values. Under this relaxed condition, the above equations still hold. Eq. (5) is non-convex with respect to $\mathbf{R}^h$, $\mathbf{S}^h$, and $\mathbf{C}^h$ altogether. As such, it's unrealistic to expect to concurrently find the global optimal values for all variables. Leveraging the multiplicative updating technique [11], [21], which was used in nonnegative matrix factorization and has been proved to converge, we introduce an iterative solution that alternatively optimizes one variable, while fixing the other variables, until convergence. The iterative process is detailed below.

Optimizing J with respect to $\mathbf{S}^h$ is equivalent to optimizing the following function:

$$\begin{aligned}J_1(\mathbf{S}^h) &= \frac{1}{m} \parallel \mathbf{X} - \mathbf{R}^h\mathbf{S}^h(\mathbf{C}^h)^T \parallel^2 \\ &s.t.\ \mathbf{R}^h \geq 0; \mathbf{C}^h \geq 0\end{aligned} \tag{6}$$

Setting the partial derivative $\frac{\partial J_1}{\partial \mathbf{S}^h} = 0$, leads to the following updating formula for $\mathbf{S}^h$:

$$\mathbf{S}^h = [(\mathbf{R}^h)^T\mathbf{R}^h]^{-1}(\mathbf{R}^h)^T\mathbf{X}\mathbf{C}^h[(\mathbf{C}^h)^T\mathbf{C}^h]^{-1} \tag{7}$$

Optimizing J with respect to $\mathbf{R}^h$ is equivalent to optimizing the following function:

$$\begin{aligned}J_2(\mathbf{R}^h) &= \frac{1}{m} \parallel \mathbf{X} - \mathbf{R}^h\mathbf{S}^h(\mathbf{C}^h)^T \parallel^2 \\ &+ \frac{\lambda}{C_m^2} \sum_{\substack{h_2=1\\h_2\neq h}}^{m} \Theta_r \parallel (\mathbf{R}^h)^T\mathbf{R}^{h_2} \parallel^2 \\ &s.t. \quad \mathbf{R}^h \geq 0;\end{aligned} \tag{8}$$

For the constraint $\mathbf{R}^h \geq 0$,we introduce the Lagrangian multiplier $\boldsymbol{\alpha} \in \mathbb{R}^{d\times k_h}$, thus the Lagrangian function is as follows:

$$\begin{aligned}L(\mathbf{R}^h) &= \frac{1}{m} \parallel \mathbf{X} - \mathbf{R}^h\mathbf{S}^h(\mathbf{C}^h)^T \parallel^2 \\ &+ \frac{\lambda}{C_m^2} \sum_{\substack{h_2=1\\h_2\neq h}}^{m} \Theta_r \parallel (\mathbf{R}^h)^T\mathbf{R}^{h_2} \parallel^2 - tr(\boldsymbol{\alpha}(\mathbf{R}^h)^T)\end{aligned} \tag{9}$$

Setting the partial derivative $\frac{\partial L(\mathbf{R}^h)}{\partial \mathbf{R}^h} = 0$, we can get

$$\boldsymbol{\alpha} = -2\mathbf{A} + 2\mathbf{R}^h\mathbf{B} + 2\lambda\boldsymbol{\Gamma}_r \tag{10}$$

where $\mathbf{A} = \mathbf{X}\mathbf{C}^h(\mathbf{S}^h)^T/m$, $\mathbf{B} = \mathbf{S}^h(\mathbf{C}^h)^T\mathbf{C}^h(\mathbf{S}^h)^T/m$ and $\boldsymbol{\Gamma}_r = (\sum_{h_2=1;h_2\neq h}^{m} \Theta_r(\mathbf{R}^{h_2}(\mathbf{R}^{h_2})^T\mathbf{R}^h))/C_m^2$. Using the Karush-Kuhn-Tucker (KKT) [22] complementarity condition $\boldsymbol{\alpha}_{ij}\mathbf{R}_{ij}^h = 0$, we obtain:

$$[\lambda\boldsymbol{\Gamma}_r - \mathbf{A} + \mathbf{R}^h\mathbf{B}]_{ij}\mathbf{R}_{ij}^h = 0 \tag{11}$$

Introducing $\boldsymbol{\Gamma}_r = \boldsymbol{\Gamma}_r^+ - \boldsymbol{\Gamma}_r^-$, $\mathbf{A} = \mathbf{A}^+ - \mathbf{A}^-$, $\mathbf{B} = \mathbf{B}^+ - \mathbf{B}^-$, where $\mathbf{A}_{ij}^+ = (|\mathbf{A}_{ij}| + \mathbf{A}_{ij})/2$ and $\mathbf{A}_{ij}^- = (|\mathbf{A}_{ij}| - \mathbf{A}_{ij})/2$, we can rewrite Eq. (11) as follows:

$$[\lambda(\boldsymbol{\Gamma}_r^+ - \boldsymbol{\Gamma}_r^-) - \mathbf{A}^+ + \mathbf{A}^- + \mathbf{R}^h\mathbf{B}^+ - \mathbf{R}^h\mathbf{B}^-]_{ij}\mathbf{R}_{ij}^h = 0 \tag{12}$$

Eq. (12) leads to the following updating formula for $\mathbf{R}^h$

$$\mathbf{R}_{ij}^h \leftarrow \mathbf{R}_{ij}^h \sqrt{\frac{[\lambda\boldsymbol{\Gamma}_r^- + \mathbf{A}^+ + \mathbf{R}^h\mathbf{B}^-]_{ij}}{[\lambda\boldsymbol{\Gamma}_r^+ + \mathbf{A}^- + \mathbf{R}^h\mathbf{B}^+]_{ij}}} \tag{13}$$

Similarly, we can update $\mathbf{C}^h$ as follows:

$$\mathbf{C}_{ij}^h \leftarrow \mathbf{C}_{ij}^h \sqrt{\frac{[\lambda\boldsymbol{\Gamma}_c^- + \mathbf{P}^+ + \mathbf{C}^h\mathbf{Q}^-]_{ij}}{[\lambda\boldsymbol{\Gamma}_c^+ + \mathbf{P}^- + \mathbf{C}^h\mathbf{Q}^+]_{ij}}} \tag{14}$$

where $\mathbf{P} = \mathbf{X}^T\mathbf{R}^h\mathbf{S}^h/m$, $\mathbf{Q} = (\mathbf{S}^h)^T(\mathbf{R}^h)^T\mathbf{R}^h\mathbf{S}^h/m$ and $\boldsymbol{\Gamma}_c = (\sum_{h_2=1;h_2\neq h}^{m} \Theta_c(\mathbf{C}^{h_2}(\mathbf{C}^{h_2})^T\mathbf{C}^h))/C_m^2$.

IV. EXPERIMENTS

A. Experimental Protocol

To study the performance of MultiCC, we measure the quality and diversity of the discovered clusterings. To measure quality, we adopt the widely used Silhouette Index (SI) and Dunn Index (DI) as the internal index [1]. A large values of SI and DI indicate a high quality clustering. To measure the diversity between alternative clusterings, we adopt Normalized Mutual Information (NMI) and Jaccard Index (JI) as the external index. The *smaller* the values of NMI and JI, the *more diverse* the clusterings are. As such, smaller values are to be preferred.

Since no prior work exists on multiple co-clusterings, we study the performance of MultiCC from two view-points: (1) Finding multiple sample-wise clusterings, and comparing MultiCC against representative and related multi-clustering algorithms; and (2) Finding multiple co-clusterings, and visualizing these co-clusterings. Five datasets collected from the UCI machine learning repository are used for the experiments. These datasets were widely used for multiple clusterings [4], [6] and their statistics is summarized in Table I.

TABLE I
CHARACTERISTICS OF THE DATASETS.

Datasets	Samples	Features	Classes
Glass	214	9	7
Ionosphere	351	34	2
CMUface	640	15360	20\4
Vehicle	846	17	4
Vowel	528	10	10

Besides specifying the target number of alternative co-clusterings (m), MultiCC also needs to choose the number of row-clusters (k) and the number of column-clusters (l) for each co-clustering. For simplicity, we adopt the same values for k and l in each alternative co-clustering. Here we adopt a widely used technique to determine the number of clusters [23], which runs k-means multiple times and then computes the cophenetic correlation coefficient. The larger the coefficient is, the more stable the clustering results are. After applying this technique with $m = 2$ and $\lambda = 100$, MultiCC chooses $l = 3$ for Glass, $l = 7$ for Ionosphere, $l = 6$ for CMUface, $l = 3$ for Vehicle, and $l = 5$ for Vowel, and it directly specifies k as the number of true classes in each dataset listed in Table I. For the CMUface dataset, since the number of persons is 20 and the number of poses is 4, we set $k_1 = 20$ and $k_2 = 4$ for the first and the second co-clusterings, respectively.

B. Parameter Analysis

The regularization parameter λ controls the tradeoff between the quality and the diversity of m alternative clusterings. We investigate the effect by varying λ between $[10^{-4}, 10^4]$ for the six UCI datasets under different input values of λ. As λ increases from 1 to 10^2, quality decreases and diversity increases. Due to the known tradeoff between quality and diversity [6], [24], this trend is expected. Increasing λ enforces more stringent non-redundancy between alternative clusterings and consequently sacrifices the quality of alternative clusterings. The increase in diversity shows that integrating the two non-redundancy terms with matrix tri-factorization indeed contributes to diverse alternative clusterings. With $\lambda \in [10, 10^3]$, we achieve a good and stable tradeoff between quality and diversity; as such, we set $\lambda = 10^2$ in the experiments.

C. Finding Multiple One-way Clusterings

Since no prior work on multiple co-clusterings exists, we compare MultiCC with multiple clustering algorithms from the perspective of clustering data sample-wise. Particularly, we compare MultiCC with COALA [12], ADFT [13], MNMF [6], MSC [5], MetaClustering [16] and Dec-kmeans [3] (all discussed in the related work Section). The first four methods are semi-supervised, and the last two are unsupervised. For COALA, ADFT and MNMF, we use k-means to generate the first clustering ($\mathcal{C}_1$), and then apply their respective solutions to generate the second alternative clustering ($\mathcal{C}_2$). Parameters were specified or optimized as suggested by the authors. Following the experimental protocol adopted by these methods [6], we measure clustering quality on $\mathcal{C}_2$. Table II reports the average results and standard deviation of ten independent runs.

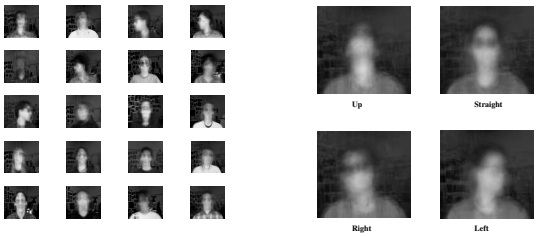
From Table II, we can see that MultiCC achieves at least one best result for both quality and diversity on the first three datasets. MultiCC also obtains the best performance (except for SI on the MetaClustering) on CMUface. MultiCC looses against some of the comparing methods on Vehicle, possibly because MultiCC enforces a high degree of diversity between $\mathcal{C}_1$ and $\mathcal{C}_2$. Similarly, due to the overemphasis of quality on Vowel, MNMF obtains more dissimilar alternative clusterings than MultiCC on Vowel, but its second clustering has lower quality than that of MultiCC. The reason is that MNMF finds multiple clusterings in a greedy manner, and its performance on $\mathcal{C}_2$ heavily depends on $\mathcal{C}_1$. Similarly to MultiCC, Dec-$k$means can also simultaneously find two dissimilar alternative clusterings, but it frequently has lower quality values and higher diversity values than MultiCC. In summary, overall MultiCC holds a good balance between dissimilarity and quality of alternative clusterings ($\mathcal{C}_1$ and $\mathcal{C}_2$).

To verify that MultiCC can group data according to different perspectives, we visualize the clustering results for the CMUface dataset. CMUface contains 640 grey face images of 20 individuals with varying poses (up, straight, right, left), and so it can be clustered either by person or by pose. Fig. 1 shows the respective mean image of 20 individual clusters and 4 different poses. We observe that the first clustering indeed groups the images by person, and the second clustering groups the images according to the pose. These two clusterings provide two meaningful interpretations of the same data. This visual example can also explain why MultiCC obtains a better quality and diversity than the other competing algorithms on the CMUface dataset in Table II. This example suggests that MultiCC can explore meaningful and different clusterings.

TABLE II

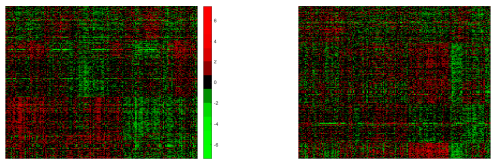
RESULTS OF QUALITY AND DIVERSITY OF THE VARIOUS COMPETING METHODS. ↓ (↑) INDICATES THE DIRECTION OF PREFERRED VALUES FOR THE CORRESPONDING MEASURE. ●/○ INDICATES WHETHER MULTICC IS STATISTICALLY (ACCORDING TO PAIRWISE t-TEST AT 95% SIGNIFICANCE LEVEL) SUPERIOR/INFERIOR TO THE OTHER METHOD.

		MetaClustering	COALA	Dec-$k$means	ADFT	MNMF	MSC	MultiCC
Glass	SI	0.150±0.018●	0.667±0.000○	0.539±0.101○	0.560±0.004○	-0.156±0.058●	0.671±0.010○	0.164±0.009
	DI	0.025±0.009●	0.209±0.000●	0.051±0.014●	0.026±0.004●	0.013±0.003●	0.128±0.013●	0.264±0.012
	NMI	0.533±0.018●	0.176±0.000●	0.023±0.035○	0.887±0.012●	0.082±0.012●	0.314±0.032●	0.041±0.010
	JI	0.473±0.012●	0.446±0.000●	0.405±0.041●	0.885±0.010●	0.181±0.010●	0.748±0.013●	0.082±0.006
Ionosphere	SI	0.372±0.051●	0.393±0.000●	0.259±0.020●	0.414±0.008	0.099±0.016●	0.401±0.007	0.406±0.012
	DI	0.071±0.016	0.041±0.000●	0.090±0.014○	0.071±0.008	0.015±0.007●	0.038±0.014●	0.065±0.010
	NMI	0.276±0.014●	0.363±0.000●	0.554±0.014●	0.803±0.013●	0.315±0.004●	0.594±0.038●	0.110±0.009
	JI	0.564±0.013●	0.501±0.000●	0.584±0.028●	0.782±0.011●	0.342±0.006○	0.748±0.029●	0.457±0.014
CMUface	SI	0.204±0.018○	0.048±0.000●	0.018±0.012●	0.063±0.005●	-0.022±0.013●	0.018±0.010●	0.076±0.013
	DI	0.102±0.015●	0.117±0.000●	0.096±0.026●	0.012±0.003●	0.031±0.011●	0.029±0.005●	0.130±0.008
	NMI	0.509±0.020●	0.088±0.000●	0.038±0.016●	0.640±0.013●	0.051±0.019●	0.554±0.013●	0.020±0.006
	JI	0.181±0.024●	0.167±0.000●	0.158±0.003●	0.531±0.011●	0.158±0.005●	0.523±0.011●	0.046±0.010
Vehicle	SI	0.401±0.022○	0.663±0.000○	0.136±0.055	0.721±0.002○	-0.181±0.013●	0.790±0.011○	0.115±0.012
	DI	0.024±0.011●	0.064±0.000○	0.008±0.001●	0.038±0.001●	0.008±0.002●	0.025±0.006●	0.054±0.006
	NMI	0.471±0.017●	0.701±0.000●	0.178±0.010●	0.980±0.021●	0.135±0.003○	0.910±0.027●	0.142±0.010
	JI	0.442±0.013●	0.724±0.000●	0.293±0.024●	0.991±0.010●	0.247±0.005●	0.952±0.033●	0.144±0.008
Vowel	SI	0.190±0.012●	0.135±0.000●	0.072±0.032●	0.223±0.018	0.005±0.023●	0.243±0.040○	0.222±0.013
	DI	0.030±0.012●	0.082±0.000	0.026±0.003●	0.051±0.014●	0.020±0.004●	0.028±0.003●	0.080±0.004
	NMI	0.364±0.023○	0.167±0.000○	0.031±0.007○	0.605±0.051●	0.005±0.002○	0.436±0.012●	0.416±0.012
	JI	0.165±0.019●	0.215±0.000●	0.130±0.006○	0.332±0.048●	0.117±0.006○	0.406±0.017●	0.147±0.012

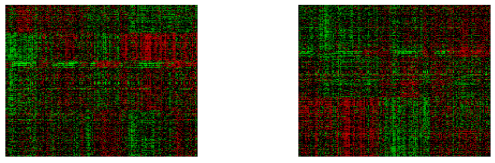


(a) Clustering by persons (b) Clustering by poses

Fig. 1. The mean image of 20 clusters in the first clustering explored by MultiCC from the perspective of 20 persons (a), and 4 clusters in the second clustering from the perspective of 4 different poses (b).



(a) First co-clustering (b) Second co-clustering



(c) Third co-clustering (d) Fourth co-clustering

Fig. 2. Heatmaps of co-clusters for four different co-clusterings.

D. Finding Multiple Co-Clusterings

We apply MultiCC on the Diffuse Large B Cell Lymphoma (DLBCL) gene expression data [25] to investigate the discovery of alternative co-clusterings. We preprocess the DLBCL data by removing the genes that are not expressed or have a small variance, and finally obtain a data matrix with 360 genes and 180 samples (cancer patients). For this investigation, we set the number of alternative co-clusterings to $m = 4$, the regularization parameter to $\lambda = 100$, the number of gene-clusters to $k = 5$, and the number of sample-clusters to $l = 3$. We visualize the results by plotting the heatmap of each co-clustering in Fig. 2. We also measure the diversity between co-clusterings, both gene-wise and sample-wise, using the average co-cluster relevance score [26], and report it in Table III. $\text{CS}_{h_1h_2}$ is the average co-cluster relevance score (CS) between the h_1-th and h_2-th co-clusterings. The smaller $\text{CS}_{h_1h_2}$ is, the larger the diversity between the two co-clusterings is.

In the heatmap, red points indicate that the gene expression is up-regulated (high expression value), while green points indicate down-regulation (low expression values). From the heatmaps of the four alternative co-clusterings, we can clearly see that MultiCC groups genes and samples in multiple red and green blocks, which implies that MultiCC can find co-expression patterns of genes across specific samples. In other words, MultiCC can find multiple high quality co-clusters. In addition, the co-clusters of these co-clusterings contain a different number of samples. For example, the first co-clustering partitions 180 samples into three clusters of sizes 40, 70, and 70, and the second co-clustering groups the samples into three clusters of sizes 100, 40, and 40. From the gene perspective,

TABLE III

AVERAGE CO-CLUSTER RELEVANCE SCORE (CS) OF FOUR CO-CLUSTERINGS FOUND BY MULTICC.

	CS_{12}	CS_{13}	CS_{14}	CS_{23}	CS_{24}	CS_{34}
Gene-wise	0.43	0.38	0.49	0.29	0.34	0.39
Sample-wise	0.28	0.29	0.34	0.29	0.33	0.34

the co-clusters manifest diverse expression profiles in different co-clusterings. This phenomenon can be visually observed in the heatmaps, through the variation of size and distribution of red and green blocks across co-clusterings. In addition, from the CS scores given in Table III, we can observe a low redundancy between co-clusters across co-clusterings, since both gene-wise and sample-wise scores are less than 0.5. Both the visualization and quantitative analyses demonstrate that MultiCC is able to discover multiple diverse co-clusterings of good quality. These alternative co-clusterings provide options for analyzing the same cancer data from different perspectives.

To conclude, we emphasize the two features that contribute to the competitive performance of MultiCC: (i) Simultaneously grouping features and samples via matrix factorization enables the finding of co-clusterings (clusterings) of high quality; (ii) The integration of two terms that measure row-cluster and column-cluster redundancy enforces the exploration of diverse co-clusterings.

V. Conclusions

In this paper, we study how to find multiple co-clusterings in data. The problem is relevant from an application standpoint, is challenging, and seldom studied. We introduce an approach called MultiCC to generate multiple diverse co-clusterings of quality. MultiCC repeatedly factorizes the data matrix to obtain multiple co-clusterings, and it enforces diversity by minimizing the redundancy between row-clusters and column-clusters. Our experimental results demonstrate that MultiCC outperforms state-of-the-art multiple clustering methods, and has the capability of finding multiple diverse co-clusterings. In the future, we will further investigate its potential for cancer subtype categorization. The codes of MultiCC are available at http://mlda.swu.edu.cn/codes.php?name=MultiCC.

Acknowledgments

The authors thank the anonymous reviewers for their helpful comments on improving this paper. This research is supported by NSFC (61873214, 61872300 and 61741217). Natural Science Foundation of CQ CSTC (cstc2018jcyjAX0228 and cstc2016jcyjA0351), the Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing (KLIGIP-2017A05), and Chongqing Graduate Student Research Innovation Project (No. CYS18089).

References

[1] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[2] J. Bailey, "Alternative clustering analysis: A review." *Data Clustering*, p. 535, 2013.

[3] P. Jain, R. Meka, and I. S. Dhillon, "Simultaneous unsupervised learning of disparate clusterings," *Statistical Analysis and Data Mining*, vol. 1, no. 3, pp. 195–210, 2008.

[4] Y. Cui, X. Z. Fern, and J. G. Dy, "Non-redundant multi-view clustering via orthogonalization," in *Proceedings of the 7th International Conference on Data Mining*, 2007, pp. 133–142.

[5] J. Hu, Q. Qian, J. Pei, R. Jin, and S. Zhu, "Finding multiple stable clusterings," in *Proceedings of the 15th International Conference on Data Mining*, 2015, pp. 171–180.

[6] S. Yang and L. Zhang, "Non-redundant multiple clustering by nonnegative matrix factorization," *Machine Learning*, vol. 106, no. 5, pp. 695–712, 2017.

[7] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001, pp. 269–274.

[8] Q. Gu and J. Zhou, "Co-clustering on manifolds," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 359–368.

[9] X. Luo, M. Zhou, Y. Xia, and Q. Zhu, "An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1273–1284, 2014.

[10] Y. Cheng and G. M. Church, "Biclustering of expression data," in *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, 2000, pp. 93–103.

[11] C. H. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 45–55, 2010.

[12] E. Bae and J. Bailey, "Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity," in *Proceedings of the 6th Internaional Conference on Data Mining*, 2006, pp. 53–62.

[13] I. Davidson and Z. Qi, "Finding alternative clusterings using constraints," in *Proceedings of the 8th International Conference on Data Mining*, 2008, pp. 773–778.

[14] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, 2002, pp. 849–856.

[15] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Advances in Neural Information Processing Systems*, 2003, pp. 521–528.

[16] R. Caruana, M. Elhawary, N. Nguyen, and C. Smith, "Meta clustering," in *Proceedings of the 6th Internaional Conference on Data Mining*, 2006, pp. 107–118.

[17] G. K. Zipf, "Human behavior and the principle of least effort." *The Southwestern Social Science Quarterly*, vol. 30, no. 2, pp. 147–149, 1949.

[18] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix tri-factorizations for clustering," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 126–135.

[19] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th ACM SIGIR Conference on Research and Development in Informaion Retrieval*. ACM, 2003, pp. 267–273.

[20] H. Wang, F. Nie, H. Huang, and F. Makedon, "Fast nonnegative matrix tri-factorization for large-scale data co-clustering," in *Proceedings of 21st International Joint Conference on Artificial Intelligence*, vol. 22, no. 1, 2011, p. 1553.

[21] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[22] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[23] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, vol. 52, no. 1, pp. 91–118, 2003.

[24] D. Niu, J. G. Dy, and M. I. Jordan, "Iterative discovery of multiple alternativeclustering views," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1340–1353, 2014.

[25] A. Rosenwald, G. Wright *et al.*, "The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma," *New England Journal of Medicine*, vol. 346, no. 25, pp. 1937–1947, 2002.

[26] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, "A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics*, vol. 22, no. 9, pp. 1122–1129, 2006.