Overlapping Clustering with Sparseness Constraints

Haibing Lu
OMIS, Santa Clara University
hlu@scu.edu

Yuan Hong
MSIS, Rutgers University
yhong@cimic.rutgers.edu

W. Nick Street
MS, The University of Iowa
nick-street@uiowa.edu

Fei Wang
IBM T.J. Watson Research Center
fwang@us.ibm.com

Hanghang Tong
IBM T.J. Watson Research Center
htong@us.ibm.com

I. Abstract

Overlapping clustering allows a data point to be a member of multiple clusters, which is more appropriate for modeling many real data semantics. However, much of the existing work on overlapping clustering simply assume that a data point can be assigned to any number of clusters without any constraint. This assumption is not supported by many real contexts. In an attempt to reveal true data cluster structure, we propose sparsity constrained overlapping clustering by incorporating sparseness constraints into an overlapping clustering process. To solve the derived sparsity constrained overlapping clustering problems, efficient and effective algorithms are proposed. Experiments demonstrate the advantages of our overlapping clustering model.

II. Introduction

Overlapping clustering is a type of clustering technique that allows a data point to be a member of multiple clusters. Compared to partitional clustering techniques, which partition data into non-overlapping regions, overlapping clustering is more appropriate in modeling data relationships for many real applications. In biology, clustering techniques are common approaches to identifying functional groups in gene expression data by clustering genes with similar expression profiles into the same group. It has been known that many genes are multi-functional and they should belong to more than one functional group [1]. Therefore partitional clustering techniques have the limitation in their ability to discover the true cluster structure in gene expression data. Apart from biology, many other domains, including role-based access control and movie recommender systems, also motivate overlapping clustering. Due to its importance, overlapping clustering has received much attention recently.

However, much of the existing work go to the opposite extreme of partitional clustering. They simply allow a data point to belong to as many clusters as needed without considering any contextual information, which may result in too many cluster assignments. To illustrate, consider the biology application. It is true that a gene can participate in multiple processes. However, according to current biological understanding, it is unlikely a gene would participate in many processes. So when an overlapping clustering result assigns many gene to over 20 processes, its correctness would be highly doubted.

In an attempt to discover true overlapping cluster structures, we propose overlapping clustering with sparseness constraints. The basic idea is to incorporate available background knowledge of the dataset to be studied, such as the maximum clusters a data point can belong to, into an overlapping clustering process. Therefore clustering results would not only provide good descriptions on the input data, but also match the prior knowledge on the data.

In this paper, we specifically look at the overlapping clustering technique proposed by Cleuziou [2], which we call the k-extended technique because its solution is derived from the well known k-means algorithm. The k-extended technique can be described as the following: Given a set of data points, group them into overlapping clusters, while minimizing the sum of the distances between each point and the mean of the representatives of clusters to which the point belongs.

Apart from the k-extended technique, there are many other overlapping clustering models, including the plaid model [3], the fuzzy c-means clustering technique [4], and the probabilistic model [1]. We chose k-extended for two reasons. The first reason is that the k-extended technique is a hard overlapping clustering technique, in which a data point is either a member of a cluster or not, while many overlapping clustering techniques are soft (probabilistic) such as fuzzy c-means clustering [4] in which a data point belongs to a cluster with some probability. For many real applications, hard overlapping clustering results carry more interpretability. For example, overlapping clustering techniques have been employed to discover roles to implement a role based access control mechanism [5]. In the setting of role based access control, a user either assumes a role, or not. The second reason is that the k-extended technique represents a data point by the mean of the cluster representatives to which the data point belongs. While some overlapping clustering techniques, e.g. [1], [6], represent a data point by the sum of the cluster representatives to which the data point belongs, we think the mean is more appropriate in representing the relationship between a data point and its associated cluster representatives.

Like many other overlapping clustering techniques, the k-extended technique simply assumes that a data point can belong to any number of clusters without imposing any constraint on cluster assignments. This way of modeling might be able to obtain an overlapping clustering result, which describes the dataset very well. However, the clustering result could be far away from the ground truth.

To overcome the limitation, we propose sparsity constrained overlapping clustering, which is able to incorporate prior knowledge on cluster memberships into the overlapping clustering process. Technically, our overlapping clustering technique is to decompose a data matrix into two matrices, where one matrix consists of cluster representatives and the other is a binary coefficient matrix showing cluster memberships, while the form of the binary coefficient matrix is regulated according to available prior knowledge. Mathematically our problem can boil down to a constrained optimization problem. As the decomposed coefficient matrix is binary, due to the combinatorial nature, this problem is very difficult to solve. So we propose an alternating minimization solution, which minimizes the objective function by fixing one of the decomposed matrices and proceeds in an alternative fashion. The derived subproblem of minimizing the objective function while fixing the cluster representative matrix is proven to be NP-hard. To solve it, we propose a branch-and-bound exact algorithm, which is suited for small size problems, and a simulated annealing algorithm, which is applicable to large size problems. To evaluate our technique of overlapping clustering with sparseness constraints, extensive experiments on both synthetic and real datasets are conducted.

III. Prior Work

Overlapping clustering has recently attracted much attention from both the data mining and computational biology fields. However, little awareness of sparsity constraints in overlapping clustering has been observed. The modified nonnegative sparse coding model proposed by [7] is closely related to our work. Mathematically, it is formulated as the problem of minimizing $\frac{1}{2}||A-XC||_F^2+\lambda\sum_{ij}X_{ij}$, where A and λ are given, C is restricted to be nonnegative, and rows of X are forced to unit norm. The main difference in our work is that X is converted from the binary memberships of S, such that $X_{ij} = S_{ij}/\sum_j S_{ij}$. Therefore, in addition to the constraint of unit norm for rows of X, positive elements in each row must be the same, which coincides with most of the existing overlapping clustering models including [3], [8], [2], [6]. Note that some of them are stated as probabilistic clustering approaches. However they boil down to matrix decomposition problems eventually. Another work closely related to our work is the model proposed by Zhu [9]. It is based on the plaid model, which is a co-clustering model and attempts to approximate a data matrix with the sum of k submatrices. On the basis of the plaid model, Zhu's model minimizes the approximation error along with the size of submatrices in hope to find some cohesive submatrices.

Imposing sparseness constraints in data analysis tasks in attempt to discover real data patterns or relationships is not a new idea. One of the most important works is the Lasso model, a shrinkage and selection method for linear regression, proposed by Tibshrani [10], which has been widely used in many fields. The power of sparseness constraints has also been well-appreciated by the machine learning community. One important work is the model proposed by Hoyer [11], which incorporates spareness constraints in non-negative matrix factorization. Heiler et al. [12] even proposed a sequential cone programming approach to this sparsity constrained non-negative matrix factorization problem.

IV. Problem Definitions

In this section, we will present the formal definition of our sparsity constrained overlapping clustering model. Before doing that, we would like to first introduce the k-extended overlapping clustering technique, as our model is built on it.

Definition 1 (k-Extended [2]): Given m observations $\{A_1, ..., A_m\} \in \mathcal{R}^n$, discover k clusters $\{\mathcal{S}_1, ..., \mathcal{S}_k\}$ with respective representatives $\{C_1, ..., C_k\} \in \mathcal{R}^n$ such that

- An observation can belong to multiple clusters;
- The sum of distances between each observation and the mean of its assigned cluster representatives is minimized.

Like many other overlapping clustering models, the k-extended technique can be described as a matrix decomposition problem: Decompose a matrix $A_{m\times n}$ into a binary matrix $S_{m\times k}$, where $S_{ij} = 1$ means data point i belongs to cluster j, and a real matrix $C_{k\times n}$, where row i is the representative of cluster i. As the goal is to discover the decomposition solution which can best describe the observed data, so the k-extended technique can be formulated as the following optimization problem.

$$\begin{aligned} min \quad & ||A_{m\times n} - X_{m\times k} \times C_{k\times n}||_2^2 \\ s.t. & \begin{cases} X_{ij} = \frac{S_{ij}}{\sum_j S_{ij}} \\ S_{ij} \in \{0,1\} \end{cases} \end{aligned} \tag{1}$$

If we include the constraints into the objective function, the above optimization problem can be reformatted as the following unconstrained programming problem.

$$min\ f_0 = \sum_i ||A_i - \frac{\sum_j S_{ij}C_j}{\sum_j S_{ij}}||_2,\ s.t.\ S_{ij} \in \{0,1\}. \tag{2}$$

The k-extended technique is essentially an extension of the well known k-means clustering technique, which assigns each observation to only one cluster, which can be enforced by adding a constraint of $\sum_j S_{ij} = 1$ to Equation 2. By allowing a data point to belong to multiple clusters, the data description accuracy can be significantly improved indeed. However, this might cause the problem of overfitting the data. To address it, one straightforward solution is to take advantage of available prior knowledge on cluster memberships and regulate the form of the coefficient matrix S.

In reality, some applications may have explicit prior knowledge about the maximum clusters a point can belong to, while others may not. To reflect real situations, we present two overlapping clustering techniques. They are explicit sparsity constrained overlapping clustering and implicit sparsity constrained overlapping clustering.

Problem 1: (Explicit Sparsity Constrained Overlapping Clustering)

$$\begin{aligned} min \quad f_1 = \sum_i ||A_i - \frac{\sum_j S_{ij} C_j}{\sum_j S_{ij}}||_2^2 \\ s.t. \begin{cases} \sum_j S_{ij} \leq \delta^* \\ S_{ij} \in \{0, 1\} \end{cases} . \end{aligned} \tag{3}$$

In explicit sparsity constrained overlapping clustering, there is an explicit constraint on the maximum clusters that a point can belong to, which is enforced by $\sum_j S_{ij} \leq \delta^*$. It is possible that different data points may have different limits on the maximum clusters that they can be assigned to. However, given the optimization model as Equation 3, it is not difficult to extend to the personalized case. So in this paper we only consider the case that all data points have the same limit.

Problem 2: (Implicit Sparsity Constrained Overlapping Clustering)

$$\begin{aligned} min \quad f_2 = \sum_i ||A_i - \frac{\sum_j S_{ij} C_j}{\sum_j S_{ij}}||_2^2 + \lambda ||S||_1 \\ s.t. \ S_{ij} \in \{0, 1\}. \end{aligned} \tag{4}$$

As its name implies, the implicit sparsity constrained overlapping clustering has no explicit restriction on the maximum number of clusters that a point can belong to. Instead, there is a penalty on the L_1 norm of the coefficient matrix $\lambda \sum_{ij} S_{ij}$, where λ is a tuning parameter controlling the penalty level. In cases where no explicit prior knowledge is available, one can repeatedly adjust the tuning parameter λ and choose the one which gives a satisfactory clustering result. Both enforcing the constraint of $\sum_j S_{ij} \leq \delta^*$ and adding a penalty of $\lambda ||S||_1$ in the objective function would limit the number of 1's elements in S. Therefore our technique is called sparsity constrained overlapping clustering.

V. Alternating Minimization Algorithms

In this section, we will present alternating minimization algorithms for our sparsity constrained overlapping clustering problems. Alternating minimization is a method widely used to solve difficult problems in data mining and machine learning.

The sparsity constrained overlapping clustering problems consist of two groups of variables, S_{ij} and C_j. It is difficult to optimize over all variables, while it is not difficult to optimize the problem when either S_{ij} or C_j is fixed. So we present an alternating minimization algorithm, which starts with a set of initial cluster representatives $\{C_1, ..., C_k\}$ and then repeats the following two-step procedure:

- Assignment Step: Minimize the objective function f_1/f_2 by fixing $\{C_1, ..., C_k\}$ and obtain cluster memberships S_{ij};
- Update step: Minimize the objective function f_1/f_2 by fixing S_{ij} and obtain updated $\{C_1, ..., C_k\}$.

VI. Update Step

In terms of the update step, the explicit and implicit sparsity constrained overlapping clustering problems are the same. The update step is given cluster memberships to update cluster representatives. The explicit sparsity constrained overlapping clustering problem has a constraint of $\sum_j S_{ij} \leq \delta^*$. When cluster memberships S_{ij} are fixed, the problems reduces to minimizing $\sum_i ||A_i - \frac{\sum_j S_{ij} C_j}{\sum_j S_{ij}}||_2$. For the implicit sparsity constrained overlapping clustering problem, when S_{ij} is fixed, a part of its objective function (Equation 4) becomes a constant and the problem reduces to minimizing $\sum_i ||A_i - \frac{\sum_j S_{ij} C_j}{\sum_j S_{ij}}||_2$ as well.

For the update step, we only need to look at the problem of minimizing $\sum_i ||A_i - \frac{\sum_j S_{ij} C_j}{\sum_j S_{ij}}||_2$, which can be solved through linear least squares. To do that, we first replace S_{ij} by X_{ij}, such that $X_{ij} = \frac{S_{ij}}{\sum_j S_{ij}}$. Thus the problem becomes to minimize $||A - X \times C||_2^2$, which is equal to $\sum_i ||A_i - X \times C_i||_2^2$.

Minimizing $||A_i - X \times C_i||_2^2$ is a typical linear regression problem, where (X, A_i) can be viewed as observations and C_i are unknown parameters to be determined. $||A_i - X \times C_i||_2^2$ can be expanded as the following:

$$A_i^T X^T X A_i - 2 C_i^T X A_i + C_i^T C_i.$$

Since this is a quadratic expression, the global minimum can be found by differentiating it with respect to C_i. Thus we have

$$C_i = (X^T X)^{-1} X A_i.$$

Therefore, at each update step, we need to update each cluster representative to be $(X^T X)^{-1} X A_i$ given new cluster memberships.

VII. Assignment Step

The assignment step is to assign observations to given cluster representatives while minimizing errors. In this section, we will study the complexity of the cluster assignment problems and propose both exact and heuristic algorithms.

A. Complexity Analysis

The assignment step in both explicit and implicit sparsity constraint overlapping clustering is NP-hard, which can be proved by a reduction to a known NP-hard problem, the subset sum problem. [13], which is described as the follows.

Definition 2 (Subset Sum Problem [13]): Given a set of integers $\{I_1, ..., I_n\}$, does the sum of some non-empty subset equal exactly zero?

Theorem 1: The cluster assignment problem in explicit sparsity constrained overlapping clustering is NP-hard.

Proof. The cluster assignment problem is a minimization problem. Its decision instance can be described as: Given a vector set $\{C_1, ..., C_k \in R^m$, a point $x \in R^m$, a cluster assignment threshold δ, and a real number b, is there some vector subset $\mathcal{S}$, such that $||x - \frac{\sum_{C_i \in \mathcal{S}} C_i}{|\mathcal{S}|}||_2^2 \leq b$, where $|\mathcal{S}|$ is the number of vectors in $\mathcal{S}$?

The cluster assignment problem belongs to P, because for any instance, it is easy to check if a solution is true. Next we will show that every cluster assignment instance is polynomially reducible to a subset sum instance. For any subset sum instance of $\{n_1, ..., n_t\}$, we can construct a corresponding cluster assignment instance such that:

- All data points are 1-dimensional;
- Cluster representatives are $\{n_1, ..., n_t\}$;
- δ is t;
- b is 0.

Such a constructed cluster assignment instance is equivalent to finding a subset of $\{n_1, ..., n_t\}$, such that the sum of contained numbers is equal to zero. Clearly the constructed cluster assignment instance is true if and only if the subset sum instance is true. So the theorem is proven. □

Theorem 2: The assignment step in implicit sparsity constrained overlapping clustering is NP-hard.

Proof. It is not difficult to see that the assignment step belongs to P. A decision cluster assignment instance in implicit sparsity constrained overlapping coursing is as: Given a vector set $\{C_1, ..., C_k\}$, a data point x, a penalty parameter λ, and a real number b, is there some vector subset $\mathcal{S}$, such that $||x - \frac{\sum_{C_i \in \mathcal{S}} C_i}{|\mathcal{S}|}||_2^2 + \lambda|\mathcal{S}| \leq b$, where $|\mathcal{S}|$ is the number of vectors in $\mathcal{S}$?

For each subset sum instance $\{I_1, ..., I_n\}$, we can find an assignment step instance such that:

- All data points are 1-dimensional;
- Cluster representatives are $\{n_1, ..., n_t\}$;
- λ=0;
- b is 0.

Clearly the constructed cluster assignment instance is true if and only if the subset sum instance is true. So the theorem is proven. □

B. Exact Algorithm

Although the assignment step in both explicit and implicit sparsity constrained overlapping clustering are NP-hard, it is still possible to apply an exact search algorithm in many real applications, because unlike the number of data points, the number of clusters usually is not too large. But an exhaustive search algorithm is too computationally expensive in any case. In this section, we will propose an efficient branch-and-bound (B&B) exact algorithm. B&B algorithms have been widely used in finding optimal solutions of various optimization problems, especially in discrete and combinatorial optimization. It consists of a systematic enumeration of all candidate solutions, where large subsets of candidate solutions are discarded by using upper and lower estimated bounds of the quantity being optimized.

1) Explicit Sparsity Constrained Overlapping Clustering: The cluster assignment subproblem in explicit sparsity constrained overlapping clustering can be formulated as an optimization problem as the follows.

$$\begin{aligned} &min \quad f_1 = ||A - \frac{\sum_i S_i C_i}{\sum_i S_i}||_2 \\ &s.t. \begin{cases} \sum_i S_i \leq \delta^* \\ S_i \in \{0, 1\}. \end{cases} \end{aligned} \tag{5}$$

For simplicity, we consider A to be a vector. In other words, we want to assign a data point A to given cluster representatives $\{C_1, ..., C_k\}$ appropriately. A straightforward approach for such a combinatorial problem is to search through the whole solution space $S \in \{0, 1\}^m$. The computational time would be $O(2^m)$, which is inhibitive for large values of m. B&B is a strategy that reduces computational time by avoiding the search in some solution subspaces where the optimal solution is guaranteed not to exist.

The outline of the B&B algorithms proposed for the cluster assignment problem in explicit sparsity constrained overlapping clustering is given in the Algorithm 1.

The explicit explanation of Algorithm 1 is given as follows:

- Lines 1-3 indicate that the recursive algorithm terminates when the all variables have been branched; in other words, the whole solution space has been searched.
- In lines 4-6 if the currently visited solution $\{s_1 = s'_1, ..., s_{\ell-1} = s'_{\ell-1}, s_\ell = 0, ..., s_k = 0\}$ is better than all previously visited solutions, update the best visited solution S^* and the lowest objective value z accordingly.
- Line 8 gives an estimate for the lower bound of $minimum(f_1)$ in the solution subspace of $s_j \in \{0, 1\}$

Algorithm 1 Branch-and-Bound (ℓ) for Explicit Sparsity Constrained Overlapping Clustering

```
Input: (i) $\ell$ is the index of the next variable to branch at.
    (ii) The current solution is $s_j = s'_j$ for $j = 1, ..., \ell - 1$.
    (iii) The current solution subspace is $s_j \in \{0, 1\}$ for $j = \ell, ..., k$.
    (iv) The best solution visited so far is $S^*$ and the current lowest objective value is $z$.
1: if $\ell > n$ then
2:     Return;
3: else
4:     if $f_1(s_1 = s'_1, ..., s_{\ell-1} = s'_{\ell-1}, s_\ell = 0, ..., s_k = 0) < z$ then
5:         $z = f_1(s_1 = s'_1, ..., s_j = s'_{\ell-1}, s_\ell = 0, ..., s_k = 0)$;
6:         $S^* = \{s_1 = s'_1, ..., s_j = s'_{\ell-1}, s_\ell = 0, ..., s_k = 0\}$;
7:     end if
8:     Estimate a lower bound $LB$ for the minimum of $f_1$ given $s_j = s'_j$ for $j = 1, ..., \ell - 1$.
9:     if $LB < z$ and $\sum_{i=1}^{\ell-1} s'_j < \delta^*$ then
10:        $S_\ell = 1$, Branch-and-Bound($\ell + 1$);
11:        $S_\ell = 0$, Branch-and-Bound($\ell + 1$);
12:    end if
13: end if
```

for $j = \ell, ..., k$ with $s_j = s'_j$ for $j = 1, ..., \ell - 1$, which will be searched later. We defer our discussion of how to obtain the estimated lower bound until after the explanation of the algorithm.

- Lines 9-12 are the essence of this branch-and-bound algorithm. Without them, the algorithm is just an ordinary brute force algorithm. There are two conditions used to determine whether or not keeping searching along the current branch. $LB < z$ means there might be some solution in this branch that outperforms the current best solution. $\sum_{i=1}^{\ell-1} s'_j < \delta^*$ means the current solution does not violate the cluster assignment constraint. So if both constraints are satisfied, there might be a feasible solution outperforming the current best solution and the algorithm should proceed.

To further explain the branch-and-bound algorithm, consider the illustration in Figure 1. The figure gives a tree-like representation for the whole solution space. Assume that at the beginning we have an initial solution S^* and its corresponding objective value z. The algorithm first proceeds to $\{1, ...\}$, which is the solution subspace with S_1 fixed to be 1 and the other components in S to be binary. The algorithm will estimate a lower bound of f_1 in that solution subspace. If we ascertain that there is no solution better than S^*, the current best solution, it is clearly unnecessary to proceed any further in that branch. The branch is then discarded and the algorithm moves to other branches.

The essence of a branch-and-bound algorithm is to save computational time by avoiding searching some unnecessary branches by intelligently employing a upper bound (the best objective value visited so far) and a lower bound (the estimated best objective value in the current solution subspace to be searched). In the B&B algorithm, at each branching point, we need to estimate a lower bound of the objective value in the current solution subspace. An accurate estimated lower bound would improve the algorithm performance significantly.

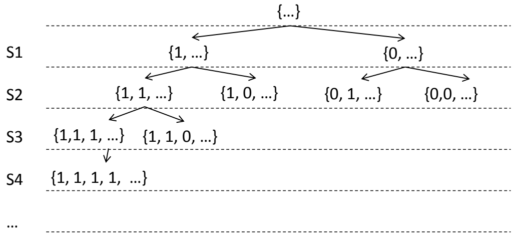

Figure 1: Branch and Bound Illustration

Consider Equation 5 of the cluster assignment problem. Suppose that the value of a portion of S, $\{S_1, ..., S_{\ell-1}\}$, has been determined. For notational convenience, we assume that among $\{S_1, ..., S_{\ell-1}\}$, $\{S_1, ..., S_{\ell'}\}$ are 1 and the remaining $\{S_{\ell'+1}, ..., S_{\ell-1}\}$ are 0. To obtain the exact lower bound of f_1, we need to solve the minimization function of f_1: $min \quad f_1 = ||\frac{\sum_i S_i c_i}{\sum_i S_i} - x||_2, \quad s.t. \quad S_i \in \{0, 1\}$ where $\{S_1, ..., S_{\ell-1}\}$ have been determined.

This problem could be as hard as the original cluster assignment problem. Let us take a deeper look at f_1. As a part of S has been determined, so f_1 can be reorganized as follows:

$$\begin{aligned} f_1 &= ||\frac{\sum_i S_i C_i}{\sum_i S_i} - x||_2 \\ &= ||\sum_{i=1}^{\ell'} \frac{1}{\ell' + \sum_{i=\ell}^{k} S_i} C_i + \sum_{i=\ell}^{k} \frac{S_i}{\ell' + \sum_{i=\ell}^{k} S_i} C_i - x||_2 \\ &= ||\sum_{i=\ell}^{k} \frac{S_i}{\ell' + \sum_{i=\ell}^{k} S_i} C_i - (x - \sum_{i=1}^{\ell'} \frac{1}{\ell' + \sum_{i=\ell}^{k} S_i} C_i)||_2 \end{aligned}$$

The function f_1 can then be viewed as discovering a linear combination of vectors $(C_\ell, ..., C_k)$ with coefficients of $\{\sum_{i=\ell}^{k} \frac{S_i}{\ell' + \sum_{i=\ell}^{k} S_i}\}$ to approximate the target $(x - \sum_{i=1}^{\ell'} \frac{1}{\ell' + \sum_{i=\ell}^{k} S_i} C_i)$. For convenience, in the following we denote $(x - \sum_{i=1}^{\ell'} \frac{1}{\ell' + \sum_{i=\ell}^{k} S_i} C_i)$ by x^*.

S_i being binary makes the problem difficult to solve. To estimate the lower bound of f_1, we relax the binary S_i to be the real value y_i and then solve the following problem:

$$min \; ||\sum_{i=\ell}^{k} y_i C_i - x^*||_2^2$$

where y_i are real variables. Certainly, the optimal objective value of such a relaxed problem gives a lower bound to the minimum of f_1.

The above relaxation problem is also a typical linear least squares problem. For notational convenience, we rewrite the problem as $min\ ||CY - x^*||_2^2$, where the matrix C is $\{C_\ell, ..., C_k\}$, Y is $(y_\ell, ..., y_k)^T$, and x^* is $(x - \sum_{i=1}^{\ell'} \frac{1}{\ell' + \sum_{i=\ell}^{k} S_i} C_i)$. Then the optimal solution of Y is $(C^TC)^{-1}C^Tx^*$. So the estimated lower bound for f_1 is:

$$||C(C^TC)^{-1}C^Tx^* - x^*||_2^2. \tag{6}$$

At step 8 in the Algorithm 1, we derive the estimated lower bound $||C(C^TC)^{-1}C^Tx^* - x^*||_2^2$ and use it to determine whether or not to continue searching the current branching space.

2) Implicit Sparsity Constrained Overlapping Clustering: Now we look at the cluster assignment problem in implicit sparsity constrained overlapping clustering. Unlike the cluster assignment in the explicit case, the objective function in the implicit case minimizes both approximation error and the L_1-norm of the assignments. For simplicity, we consider one data point A. The problem of assigning A to given cluster representatives $\{C_1, ..., C_k\}$ in the implicit case can be formulated as the following optimization problem:

$$min \quad f_2 = ||A - \frac{\sum_j S_j C_j}{\sum_j S_j}||_2^2 + \lambda ||S||_1 \tag{7}$$

where λ is given, and S_j, which denotes cluster membership, is to be determined.

A B&B algorithm for the cluster assignment subproblem of the implicit sparsity constrainedly overlapping clustering, is provided in Algorithm 2.

Algorithm 2 Branch-and-Bound (ℓ) for Implicit Sparsity Constrained Overlapping Clustering

```
Input: (i) $\ell$ is the index of the next variable to branch at.
    (ii) The current solution is $s_j = s'_j$ for $j = 1, ..., \ell - 1$.
    (iii) The current solution subspace is $s_j \in \{0, 1\}$ for $j = \ell, ..., k$.
    (iv) The best solution visited so far is $S^*$ and the current lowest objective value is $z$.
 1: if $\ell > n$ then
 2:     Return;
 3: else
 4:     if $f_2(s_1 = s'_1, ..., s_{\ell-1} = s'_{\ell-1}, s_\ell = 0, ..., s_k = 0) < z$ then
 5:         $z = f_2(s_1 = s'_1, ..., s_j = s'_{\ell-1}, s_\ell = 0, ..., s_k = 0)$;
 6:         $S^* = \{s_1 = s'_1, ..., s_j = s'_{\ell-1}, s_\ell = 0, ..., s_k = 0\}$;
 7:     end if
 8:     Estimate a lower bound $LB$ for the minimum of $f_2$ given $s_j = s'_j$ for $j = 1, ..., \ell - 1$.
 9:     if $LB < z$ then
10:         $S_\ell = 1$, Branch-and-Bound($\ell + 1$);
11:         $S_\ell = 0$, Branch-and-Bound($\ell + 1$);
12:     end if
13: end if
```

Since Algorithm 2 is similar to Algorithm 1, we will skip the explanation of the main body of the algorithm. Instead, we point out the differences in Algorithm 2:

- In lines 4 and 5, the objective function is f_2, which is $||A - \frac{\sum_j S_j C_j}{\sum_j S_j}||_2^2 + \lambda||S||_1$;
- In line 8, the way of estimating the lower bound of f_2 is different, since the objective function is different;
- In line 9, the condition is $LB \leq z$ only, since there is no explicit constraint on the maximum cluster assignments in this case.

At step 8, we need to estimate the lower bound of f_2 at each branching point. Here we will present an estimation method.

The objective function f_2 consists of two parts, $||A - \frac{\sum_j S_j C_j}{\sum_j S_j}||_2^2$, which is f_1, and $\lambda||S||_1$. Suppose that at the branching pint of ℓ, where $\{S_1, ..., S_{\ell-1}\}$ have been determined, $\{S_1, ..., S_{\ell'}\}$ are 1 and the remaining $\{S_{\ell'+1}, ..., S_{\ell-1}\}$ are 0. According to the estimated lower bound for f_1 in Equation 6, we clearly have $||A - \frac{\sum_j S_j C_j}{\sum_j S_j}||_2^2 \leq ||C(C^TC)^{-1}C^Tx^* - x^*||_2^2$, where C is $\{C_\ell, ..., C_k\}$ and x^* is $(x - \sum_{i=1}^{\ell'} \frac{1}{\ell' + \sum_{i=\ell}^{k} S_i} C_i)$. We also have $\lambda||S||_1 \leq \lambda \sum_{i=1}^{\ell'} S_i$. Therefore an estimated lower bound for f_2 is

$$||C(C^TC)^{-1}C^Tx^* - x^*||_2^2 + \lambda \sum_{i=1}^{\ell'} S_i. \tag{8}$$

C. Heuristic

As the cluster assignment problem in both explicit and implicit cases are NP-hard, exact search algorithms are not appropriate when the number of clusters is large. So in this section, we will present efficient simulated annealing (SA) heuristics, which usually run fast and produce satisfactory results.

Let us look at the cluster assignment problem in explicit sparseness constrained overlapping clustering first. It is given an input vector X and cluster representatives $C_1, ..., C_k$ to assign X to clusters appropriately such that the mean of the assigned cluster representatives is the closest to x. The solution space of the cluster assignment S is $\{0, 1\}^m$. Our simulated annealing heuristic designed for the cluster assignment problem is described as follows:

- Firstly, we find the cluster representative C_i closest to X and initialize the value of S by letting its ith component be 1 and the others be 0.
- A next candidate solution is found by randomly selecting one component of the current solution S and flipping its value from 0 to 1 or from 1 to 0. Repeat it, if there are more than δ^* elements with the value of 1 in S.
- If the new solution is closer to the target X, update the current solution to be the new solution. Even if the new solution is not better, with a certain probability less than 1, the current solution is still updated to be

Algorithm 3 Simulated Annealing for Explicit Sparseness Constrained Overlapping Clustering

```
Input: $x$, $\{C_1, ..., C_k\} \in \{0,1\}^{m \times 1}$, and $count^*$;
Output: $S \in \{0,1\}^{k \times 1}$;
 1: $i = arg: \ min_j \ ||C_j - x||_2$;
 2: $S(i) = 1$ and $S(j) = 0$, $\forall j \neq i$;
 3: $count = 1$;
 4: while $count \leq count^*$ do
 5:    Generate a random number $t$ in $\{1, ..., k\}$;
 6:    $S' = S$ and $S'(t) = 1 - S'(t)$;
 7:    if $f_1(S') < f_1(S)$ and $\sum_i S'(i) \leq \delta^*$ then
 8:       $S = S'$;
 9:    else
10:       Generate a random number $r$ in $[0, 1]$;
11:       if $r < exp[-log(count + 1)(f_1(S') - f_1(S))]$ then
12:          $S' = S$;
13:       end if
14:       $count = count + 1$;
15:    end if
16: end while
```

the new solution. This property reduces the chance of being stuck at a local optimum.

- Repeat the previous two steps until some terminating condition is satisfied, such as the maximum iteration steps are reached or the objective value is not improved for a certain number of iterations. At the end, choose the best solution that has been visited to be the final solution.

As mentioned in the above steps, when the candidate solution is worse than the current solution, there is a certain probability of moving to that inferior solution anyway. We adopt the transition probability formula proposed by Besag [14], which is

$$exp[-log(n+1) \times max(0, f_1(S') - f_1(S))] \qquad (9)$$

where S is the current solution, S' is the candidate solution, and n is the number of current iterations.

The complete pseudocode of the simulated annealing algorithm is provided in Algorithm 3.

A simulated annealing algorithm for the cluster assignment problem in implicit sparseness constrained overlapping clustering can be easily obtained by making a few changes to Algorithm 3:

- At line 7, the terminating condition is changed to $f_2(S') < f_2(S)$;
- At line 11, the condition is changed to $r < exp[-log(count + 1)(f_2(S') - f_2(S))]$.

VIII. Experimental Study

In this section, three experiments are designed to study our proposed explicit and implicit sparsity constrained overlapping clustering models. All experiments are implemented in Matlab and run on a Dell desktop with Intel Core 2 Duo CPU E8400 @ 3.00GHz and 2.96 GB of RAM.

Experiment 1. The first experiment is to evaluate the proposed exact B&B algorithm and SA heuristic. For simplicity, we consider explicit sparsity constrained overlapping clustering. So we study Algorithm 1 and Algorithm 3.

As both of these algorithms are designed for the cluster assignment step instead of the whole overlapping clustering problem, we compare the alternating minimization algorithm coupled with the exact B&B algorithm, and the same alternating minimization algorithm coupled with the SA heuristic.

Notice that although the B&B algorithm gives an optimal solution for the cluster assignment problem, the alternating minimization algorithm coupled with the exact B&B algorithm is not guaranteed to produce an optimal solution for an explicit sparsity constrained overlapping clustering problem.

The experiment is conducted on seven synthetic datasets including a large size dataset. The detailed data generation procedure is: (i) First, randomly generate k representative vectors of d attributes with element values ranging from 1 to 50; (ii) Second, randomly generate a binary cluster membership matrix S with each row consisting of no more than δ^* elements with the value of 1; (iii) Finally, construct a data matrix based on the representative vectors and the cluster membership matrix S. Those seven datasets are generated in different parameter settings. *Dataset 1*: $n = 20$, $d = 5$, $k = 4$, and $\delta^* = 2$; *Dataset 2*: $n = 40$, $d = 10$, $k = 6$, and $\delta^* = 3$; *Dataset 3*: $n = 60$, $d = 15$, $k = 8$, and $\delta^* = 4$; *Dataset 4*: $n = 80$, $d = 20$, $k = 10$, and $\delta^* = 5$; *Dataset 5*: $n = 100$, $d = 25$, $k = 12$, and $\delta^* = 6$; *Dataset 6*: $n = 120$, $d = 30$, $k = 14$, and $\delta^* = 7$; *Dataset 7*: $n = 10,000$, $d = 100$, $k = 20$, and $\delta^* = 10$. *Dataset 7* is of large size, which is used to test the scalability of our proposed algorithms.

We compare the alternating minimization algorithm coupled with the B&B algorithm and the same algorithm coupled with the SA heuristic in terms of approximation error and computational time. For both algorithms, we assume δ^* is known and use it to regulate the form of the binary coefficient decomposed matrix. For the SA heuristic, the maximum number of iterations is k^2.

The measure of approximation error is defined as the following: $error = \frac{||A - A'||_2}{||A||_2}$ where A is the original data matrix and A' is the reconstructed data matrix.

Results are plotted in Figures 2 and 4. As the alternating minimization algorithm coupled with the B&B algorithm cannot return a result for *Dataset 6* in a limited time, in Figure 2 only the comparison on *Datasets 1-6* is provided.

We observe that the alternating minimization algorithm coupled with the B&B algorithm does outperform the alternating minimization algorithm coupled with the SA heuristic. However, the performance of the alternating minimization algorithm coupled with the SA heuristic is satisfactory. In many cases, the approximation error is even less than 0.05.

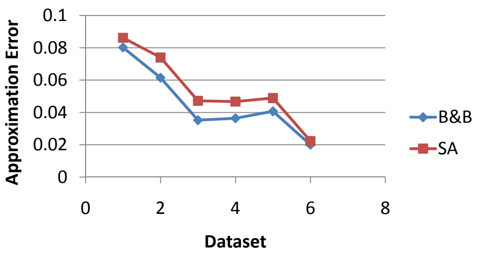

Figure 2: Comparison w.r.t. Approximation Error

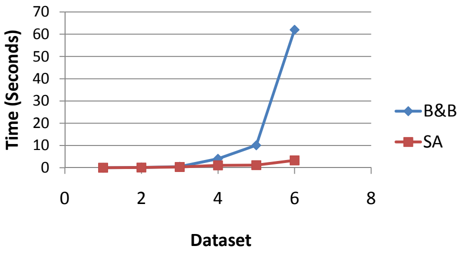

Figure 3: Comparison w.r.t. Computational Time

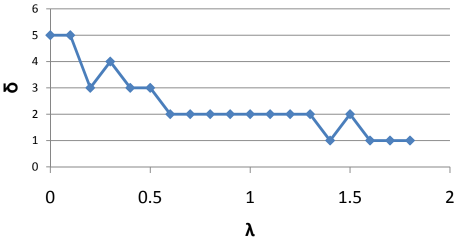

Figure 4: Relation Between λ and δ

Figure 3 provides the comparison result on computational time for all datasets except for *Dataset 7*. The alternating minimization algorithm coupled with the SA heuristic *Dataset 7* takes 1,849 seconds to cluster *Dataset 7* of 10,000 records and the resulting approximation error is only 0.0214. The result validates the scalbility of our proposed alternating minimization algorithm coupled with the SA heuristic. We also observe that when the data size is small, two algorithms are comparable. However, the required computational time for the alternating minimization algorithm coupled with the B&B algorithm grows exponentially with the data size. For a data matrix with 120 records and 30 attributes, it takes about 60 seconds. When the alternating minimization algorithm coupled with the SA heuristic, the required computational time grows slowly with the data size. The underlying reason is that the SA heuristic runs in polynomial time, while the B&B algorithm is an exact algorithm. In the worst case, it can take as much as an exhaustive search algorithm. The Figure 2 and Figure 3 suggest that the B&B algorithm is suited for small scale problems and the SA heuristic is good for large scale problems.

Experiment 2. The second experiment is to investigate the relation between explicit and implicit sparsity constrained overlapping clustering models. The experiment is conducted on a synthetic dataset, generated in the same way as employed in the first experiment. The parameter setting is $n = 120$, $d = 30$, $k = 3$, and $\delta^* = 7$. We run the alternating minimization algorithm for the implicit clustering, coupled with the SA heuristic for each value of λ ranging from 0 to 1.8. The maximum number of iterations is set to be k^2. For each λ value, we find the maximum number of cluster assignments, δ, which is not the real cluster assignment limit δ^*, in the clustering result. The results are plotted in the Figure 4. There are two observations. First, δ decreases when the value of λ increases. The reason is that more penalty are imposed on the total number of cluster assignments when the value of λ increases. The second observation is that for most values of λ ranging from 0.2 to 0.5, δ is 3, which is the true cluster assignment limit. δ goes to 4 when λ is 0.3 since the alternating minimization algorithm does not necessarily find the global optimum.

Experiment 3. The third experiment is to evaluate the soundness of our sparsity constrained overlapping clustering approach. Specifically we compare our alternating minimization algorithm coupled with the SA heuristic for the explicit sparsity constrained overlapping clustering model with the k-extended algorithm for the conventional overlapping clustering model. We assume the true cluster assignment limit is known to our algorithm. In the SA heuristic, the maximum number of iterations is set to be the square of the cluster assignment limit.

The experiment is conducted on both synthetic datasets and a real dataset, the MovieLens dataset[1]. The synthetic dataset generation procedure is the same as before. The specific parameter settings are as follows. (1) *small-synthetic*: a dataset with $n = 75$, $d = 30$, $k = 10$, and $\delta^* = 3$; (2) *medium-synthetic*: a dataset with $n = 200$, $d = 50$, $k = 10$, and $\delta^* = 5$, (3) *large-synthetic*: a dataset with $n = 1000$, $d = 150$, $k = 30$, and $\delta^* = 10$. The MovieLens dataset consists of ratings and tags for movies by users. We generate three rating matrices. (1) *small-real*: 100 movies and 38 users; (2) *medium-real*: 150 movies and 12 users; (3) *large-real*: 200 movies and 7 users. Because most of users rate a small portion of movies, when many movies are considered, we can only find a very few users, who rate all of the selected movies. The tags list genres of every movie. According to the real data, a movie can belong to up to six genres. We use six as the cluster assignment limit for our explicit sparsity constrained overlapping clustering model.

We adopt the comparison measure employed in [6]. To evaluate the clustering results, precision, recall, and F-

[1] http://www.grouplens.org

Data	F-measure		Precision		Recall	
	Sparse OC	K-Extended	Sparse OC	K-Extended	Sparse OC	K-Extended
small-synthetic	**0.4804**	0.4684	**0.3562**	0.3536	**0.7377**	0.6933
medium-synthetic	**0.6587**	0.6424	**0.5556**	0.5514	**0.8088**	0.7692
large-synthetic	**0.6703**	0.5732	**0.5783**	0.5538	**0.7972**	0.5941
small-real	**0.4345**	0.2889	**0.6668**	0.6655	**0.3220**	0.1845
medium-real	**0.5255**	0.3503	0.5921	**0.5982**	**0.4724**	0.2477
large-real	**0.5446**	0.4082	0.5867	**0.6014**	**0.5082**	0.3089

Figure 5: Comparison of results on all datasets

measure were calculated over pairs of points. For each pair of points that share at least one cluster in the overlapping clustering results, these measures try to estimate whether the prediction of this pair as being in the same cluster was correct with respect to the underlying true categories in the data. Precision is calculated as the fraction of pairs correctly put in the same cluster, recall is the fraction of actual pairs that were identified, and F-measure is the harmonic mean of precision and recall.

Comparison of results is provided in Figure 5. For synthetic datasets, the explicit sparsity constrained overlapping clustering model outperforms the k-extended model with respect to any clustering comparison measure. For the real datasets, the performance of our model is significantly better than the k-extended model with respect to the measures of F-measure and recall and is comparable to the k-extended model with respect to precision. The experimental results validates the soundness of our sparsity constrained overlapping clustering model.

IX. Conclusion

This paper studies the problem of overlapping clustering with sparseness constraints. Specifically, this paper proposes two new methods, explicit sparsity constrained overlapping clustering and implicit sparsity constrained overlapping clustering, which respectively incorporate explicit and implicit sparseness constraints into overlapping clustering. In addition, we propose alternating minimization algorithms to solve these two problems. Furthermore, as the cluster assignment step in both of these algorithms is NP-hard, we propose an efficient branch-and-bound exact algorithm and a simulated annealing heuristic. Experimental results show that our methods perform better than the existing overlapping clustering method.

References

[1] E. Segal, A. Battle, and D. Koller, "Decomposing gene expression into cellular processes," in *In Proc. of 8th Pacific Symposium on Biocomputing (PSB)*, pp. 89–100, 2003.

[2] G. Cleuziou, "An extended version of the k-means method for overlapping clustering," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pp. 1 –4, 2008.

[3] L. Lazzeroni and A. Owen, "Plaid models for gene expression data," *Statistica Sinica*, vol. 12, pp. 61–86, 2000.

[4] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 1981.

[5] H. Lu, J. Vaidya, and V. Atluri, "Optimal boolean matrix decomposition: Application to role engineering," in *ICDE '08: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, (Washington, DC, USA), pp. 297–306, IEEE Computer Society, 2008.

[6] A. Banerjee, C. Krumpelman, and J. Ghosh, "Model-based overlapping clustering," in *In KDD*, pp. 532–537, ACM Press, 2005.

[7] L. Badea, D. Tilivea, L. Badea, and D. Tilivea, "Sparse factorizations of gene expression data guided by binding data," in *Pacific Symposium on Biocomputing*, pp. 447–458, 2005.

[8] Q. Fu and A. Banerjee, "Multiplicative mixture models for overlapping clustering," in *ICDM*, pp. 791–796, 2008.

[9] H. Zhu, G. Mateos, G. Giannakis, N. Sidiropoulos, and A. Banerjee, "Sparsity-cognizant overlapping co-clustering for behavior inference in social networks," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 3534 –3537, march 2010.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[11] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, December 2004.

[12] M. Heiler, C. Schnorr, P. Bennett, and E. Parrado-hernandez, "Learning sparse representations by non-negative matrix factorization and sequential cone programming," *Journal of Machine Learning Research*, vol. 7, p. 2006, 2006.

[13] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. MIT Press and McGraw-Hill, 2001.

[14] J. B. Peter, P. Green, D. Higdon, and K. Mengersen, "Bayesian computation and stochastic systems," *Statistical Science*, vol. 10, pp. 3–67, 1995.