# Efficient Soft-Constrained Clustering for Group-Based Labeling

Ryoma Bise[1], Kentaro Abe[1], Hideaki Hayashi[1], Kiyohito Tanaka[2], and Seiichi Uchida[1]

[1] Kyushu University, Fukuoka City, Japan. bise@ait.kyushu-u.ac.jp
[2] Kyoto Second Red Cross Hospital, Kyoto, Japan.

**Abstract.** We propose a soft-constrained clustering method for group-based labeling of medical images. Since the idea of group-based labeling is to attach the label to a group of samples at once, we need to have groups (i.e., clusters) with high purity. The proposed method is formulated to achieve high purity even for difficult clustering tasks such as medical image clustering, where image samples of the same class are often very distant in their feature space. In fact, those images degrade the performance of conventional constrained clustering methods. Experiments with an endoscopy image dataset demonstrated that our method outperformed various state-of-the-art methods.

## 1 Introduction

Collecting a large number of labeled images as training data is required in machine learning classification tasks because deep neural networks require sufficient training data to learn discriminative features for robust classification. In general object image recognition tasks, crowdsourcing services (e.g., Amazon MechanicalTurk) have been widely used for labeling images efficiently and quickly by using workers from all over the world. However, we cannot use such services for medical images because most annotators of the services cannot attach appropriate labels due to a lack of biomedical knowledge. Considering the large diversity of recent classification tasks of medical images, there is a huge demand for systems that reduce the labeling effort.

To label data efficiently, several *group-based labeling* methods [1–3] have been proposed, where the data is first clustered and then the images of each cluster are labeled by an expert, as shown in Fig. 1(a). If the purity of each cluster is high enough (i.e., if each cluster is mostly comprised of the images from the same class), group-based labeling is far more effective than instance-based (i.e., one-by-one) labeling. Imagine a situation that all images of a cluster are listed in a file viewer and an expert observes them. If the expert finds that all of them belong to the same class, she/he can attach the same label to them at once. Even when a cluster contains a small number of images from different classes, it is still easy to find and exclude them before labeling; this is because those images will be *visually salient* in the list.

In order to increase the purity of each cluster for higher efficiency of group-based labeling, *constrained clustering*, such as [4, 5], is a reasonable choice. As shown in Fig. 1 (b), an expert selects a small number of samples from the dataset and put links to them before clustering. If a pair of two samples must (cannot) belong to the same

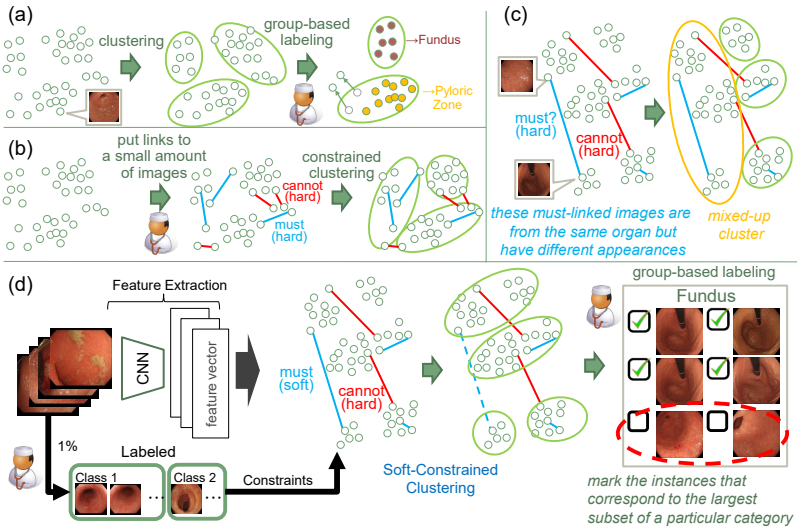


**Fig. 1.** (a) Group-based labeling with clustering. (b) Hard-constrained clustering. (c) An unexpected result by hard-constrained clustering; the middle of the mixed-up cluster is better to belong to another cluster. (d) The proposed soft-constrained clustering for group-based labeling.

cluster, they are linked by a must-link (cannot-link). Then a clustering is performed while satisfying the constraints indicated by the links. A small effort of putting the links will greatly help to increase the purity of the resulting clusters.

If a class is unimodal and thus the samples belong to the class can be grouped by a single cluster, the conventional constrained clustering will work well without any side-effect; however, if a class is not unimodal, the must-link constraints may adversely affect clustering. As shown in Fig. 1 (c), a must-link for distant samples causes a mixed-up cluster in order to satisfy the constraint by the must-link. Such an undesired must-link is not avoidable, especially for medical image applications. The appearance of a medical image often changes even in the same class due to huge appearance variations, different shooting angles, etc., and the class distribution often becomes multimodal whose components are distant to each other.

In this paper, we propose a novel soft-constrained clustering method for medical image labeling tasks. Different from the conventional constrained clustering methods, we allow the violation of must-link constraints to deal with the classes with multimodal (scattered) distributions, as shown in Fig. 1 (d). Specifically, must-links are evaluated as penalties while cannot-links are still treated as hard constraints. In addition, the proposed method is formulated as a single optimization problem, whereas the conventional soft-constrained clustering methods first solve the hard-constrained problem and then modify the cluster assignment. We have evaluated the performance of the proposed method in the task to put 20 labels to about 12,000 endoscopic images collected from several hospitals. Specifically, the proposed method achieved higher performance measurements (purity and recall of the clusters) than state-of-the-art clustering methods under different numbers of clusters and rates of links (i.e, constraints).

## 2 Related work

The most famous constrained clustering method is COP-Kmeans [4], which modifies the assignment step of K-means to satisfy must/cannot-link constraints. It performs hard-constrained clustering and thus is not suitable for medical images. Soft-constraint clustering methods, such as CVQE [6] and LCVQE [7] have been proposed to relax the hard constraints to a penalty evaluation. For example, CVQE penalizes violation of must-link constraints by adding a penalty of the distance between the two nearest cluster centers of these two points. LCVQE improves the computational costs by modifying the penalty term in the objective function of CVQE. Similar to these methods, PCK-means [8] and MPCK-means [9] methods design the penalty function as 0/1-Loss. Those soft-constrained clustering algorithms first solve the hard-constrained problem and then update the cluster assignment of each sample. This two-step organization is reasonable for discarding a small number of erroneous constraints but not for dealing with multimodal distributions; this is because as shown in Fig. 1(c) the must-links under multimodal distributions may disturb the first step. Recently, Le et al. [5] proposed a Binary Optimization for Constrained K-means (BOCK) where the optimization problem formulated as a single binary linear programming problem. Although it uses not a two-step formulation, it performs hard-constrained clustering and thus is not suitable for medical images.

Our soft-constrained clustering method is formulated as a single optimization problem while dealing with the must-constraints in a penalty term. As shown in the later experimental result, this formulation is very suitable for medical image labeling tasks because the multimodal distribution of medical images will give many undesired must-links and thus the typical two-step optimization will fail. To the authors' best knowledge, any soft-constrained clustering method like ours has been neither proposed nor applied to medical images.

## 3 Efficient labeling

Fig. 1(d) shows the overview of our labeling scheme including the proposed clustering method. In this scheme, features are first extracted for each image by DenseNet [10] pre-trained by ImageNet.[3] An expert attaches labels to a very few samples ($1\%$). From the attached label, must-link and cannot-link constraints are defined; must-link (cannot-link) for a pair of samples from the same class (different classes). Next, the proposed soft-constrained clustering method is applied by using the must-link as a penalty and the cannot-link as a hard constraint. Then, the images in the "prominent" cluster, which contains the most labeled samples among all clusters, is shown to an expert and the expert attaches the labels to all samples. If the cluster only contains samples from the same class, the expert can attach the same class label to them at once. Even if the cluster contains several samples from different classes, the expert can do the same just after discarded those samples. The remaining samples (the discarded samples and the samples in other clusters) are fed to the clustering at the next round. Finally, this process is repeated until all samples have been labeled.

[3] We selected one of the most famous networks as a feature representation network.

# 4 Soft-constrained clustering

## 4.1 Problem setting

The formulation of our soft-constrained clustering method can be started from that of the K-means clustering. Given $N$ samples $\mathcal{X} = \{\boldsymbol{x}_j \in \mathbb{R}^D\}_{j=1}^N$, where $D$ is the dimension of the feature vector, the goal of K-Means is to find a set $\mathcal{C}$ containing $K$ cluster's centroids $\mathcal{C} = \{\boldsymbol{c}_i \in \mathbb{R}^D\}_{i=1}^K$ and assign each sample $\boldsymbol{x}_j$ to one of non-overlapping clusters. The optimal clustering minimizes the within cluster sum of squares (WCSS). According to [5], it is formulated as:

$$\min_{\mathbf{C},\mathbf{S}} \|\mathbf{X} - \mathbf{SC}\|_F^2, \quad \text{s.t. } s_{ij} \in \{0,1\} \ \forall i,j, \quad \sum_{i=1}^{K} s_{ij} = 1 \ \forall j = 1,\ldots,N, \quad (1)$$

where $\mathbf{X} \in \mathbb{R}^{D\times N}$ is the matrix whose the $j$-th column corresponds to $\boldsymbol{x}_j \in \mathbb{R}^D$, and $\mathbf{C} \in \mathbb{R}^{D\times K}$ is a cluster centroid matrix whose $i$-th column corresponds to the centroid $\boldsymbol{c}_i$. $\mathbf{S} \in \mathbb{R}^{K\times N}$ is a cluster assignment matrix, where its $(i,j)$-th element $s_{ij}$ has the value of 1 if the sample $\boldsymbol{x}_i$ is assigned to the $j$-th cluster and 0 otherwise. The first constraint restricts $\mathbf{S}$ to be a binary assignment matrix and the second constraint represents each sample to be assigned to only one cluster. The $j$-th column of $\mathbf{SC}$ indicates the centroid of the cluster to which the $j$-th sample $\boldsymbol{x}_j$ is assigned. $|\cdot|_F$ denotes the Frobenius norm of a matrix.

In the constrained clustering task, we need to specify must-links and cannot-links. A must-link will be attached to a sample pair that need to belong to the same cluster. A cannot-link will be attached to a pair that cannot belong to the same cluster. Here, we assume that an expert attaches correct labels to a small number of samples for each class ($i = 1,\ldots,N_c$), where $N_c$ is the number of classes. Denoting $\mathcal{L}_i$ as the set of labeled samples for class $i$, $\boldsymbol{x}_j \in \mathcal{L}_i$ indicates that the $j$-th sample belongs to the $i$-th class. Using the labeled samples, we then generate the must-link and cannot-link constraints. A pair of samples is registered to the must-link set $\mathcal{M}$ if these labels are the same, and to the cannot-link set $\mathcal{D}$ if different.

The proposed soft-constrained clustering method also tries to optimize the set of $K$ cluster centroids $\mathcal{C}$ and the cluster assignments that minimize the sum of the WCSS distortion. The main difference from (1) is that the cannot-links $\mathcal{D}$ are imposed as hard constraints. Another, more important difference is that the must-links are evaluated as penalties; must-link constraints can be violated but penalized. This non-consistent treatment for the cannot-link and must-link comes from the property of medical images; as we noted, must-links are often too hard to satisfy for the class with a multimodal (scattered) distribution.

Consequently, the proposed method is formulated as:

$$\min_{\mathbf{C},\mathbf{S}} \ \|\mathbf{X} - \mathbf{SC}\|_F^2 + \omega \sum_{(\boldsymbol{x}_p,\boldsymbol{x}_q)\in\mathcal{M}} \sum_{i=1}^{K} |s_{ip} - s_{iq}|_1, \quad (2)$$

$$s.t. \ \ s_{ij} \in \{0,1\} \ \ \forall i,j, \quad \sum_{i=1}^{K} s_{ij} = 1 \ \ \forall j = 1,\ldots,N, \quad (3)$$

$$s_{ip} + s_{iq} \leq 1, \ \ \forall i = 1,\ldots,K \quad \forall (\boldsymbol{x}_p,\boldsymbol{x}_q) \in \mathcal{D}, \quad (4)$$

where (4) represents the cannot-link constraints. If two samples $\boldsymbol{x}_p$ and $\boldsymbol{x}_q$ are paired by a cannot-link (i..e, $(\boldsymbol{x}_p, \boldsymbol{x}_q) \in \mathcal{D}$) and they try to belong to the same cluster $i$, $s_{ip} + s_{iq} = 1 + 1 \nleq 1$; consequently, this situation violates the constraint and is not allowed. The second term of the objective function (2) represents soft-constraints of the must-link. If two samples $\boldsymbol{x}_p$ and $\boldsymbol{x}_q$ are paired by a must-link (i.e., $(\boldsymbol{x}_p, \boldsymbol{x}_q) \in \mathcal{M}$) and they try to belong to different clusters, it becomes $s_{ip} - s_{iq} = 1 > 0$ and thus penalizes the objective function by $\omega$, which is a positive constant.

### 4.2 Optimization for soft-constraints clustering

To minimize (2) with constraints (3) and (4), we take an alternating optimization approach that alternatively updates the cluster centroid matrix $\mathbf{C}$ and the assignment matrix $\mathbf{S}$ until convergence, which is similar to the original K-means optimization approach (EM algorithm). In the update step for $\mathbf{C}$ by fixing $\mathbf{S}$, we can ignore the constraints and the second penalty term of the objective function. We update $\mathbf{C}$ by solving the regularized least square problem to avoid numerical issues for large-scale problems [11]:

$$\min_{\mathbf{C}} \|\mathbf{X} - \mathbf{S}\mathbf{C}\|_F^2 + \lambda \|\mathbf{C}\|_F^2, \tag{5}$$

where $\lambda$ is the regularization parameter and we set $\lambda$ to be $10^{-4}$ in our experiments. The problem is a convex quadratic optimization problem and it can be solved in closed-form.

$$\mathbf{C} = \mathbf{X}\mathbf{S}^T(\mathbf{S}\mathbf{S} + \lambda \mathbf{I})^{-1}, \tag{6}$$

where $(\mathbf{S}\mathbf{S} + \lambda \mathbf{I})$ is guaranteed to be full-rank.

Next, we update $\mathbf{S}$ by fixing $\mathbf{C}$. Let $\mathbf{Y} \in \mathbb{R}^{K \times N}$ be a matrix whose $(i, j)$-th element $y_{ij}$ is the squared distance from $\boldsymbol{x}_j$ to its centroid $\boldsymbol{c}_i$; namely, $y_{ij} = \|\boldsymbol{c}_i - \boldsymbol{x}_j\|_2^2$. By using $\mathbf{Y}$, the first term of (2) can be written as $\langle \mathbf{Y}, \mathbf{S} \rangle$, which represents Frobenius inner product of $\mathbf{Y}$ and $\mathbf{S}$.

By rewriting the second term with $KM$ variables $\gamma_{i,(p,q)}$, which is defined for each pair of $p$ and $q$ which satisfies $(\boldsymbol{x}_p, \boldsymbol{x}_q) \in \mathcal{M}$, the updating problem of $\mathbf{S}$ becomes:

$$\min_{\mathbf{s},\boldsymbol{\gamma}} \ \langle \mathbf{Y}, \mathbf{S} \rangle + \omega \sum_{(\boldsymbol{x}_p, \boldsymbol{x}_q) \in \mathcal{M}} \sum_{i=1}^{K} \gamma_{i,(p,q)}, \tag{7}$$

$$\text{s.t.} \ \ s_{ij} \in \{0, 1\} \ \ \forall i, j, \quad \sum_{i=1}^{K} s_{ij} = 1 \ \ \forall j = 1, \ldots, N, \tag{8}$$

$$s_{ip} + s_{iq} \leq 1, \ \ \forall i = 1, \ldots, K \ \ \ \forall (\boldsymbol{x}_p, \boldsymbol{x}_q) \in \mathcal{D}, \tag{9}$$

$$s_{ip} - s_{iq} \leq \gamma_{i,(p,q)}, \ \ -s_{ip} + s_{iq} \leq \gamma_{i,(p,q)}, \ \ \forall i \ \ \forall (\boldsymbol{x}_p, \boldsymbol{x}_q) \in \mathcal{M}, \tag{10}$$

where the set $\boldsymbol{\gamma} = \{\gamma_{i,(p,q)}\}$. This problem is a mixed binary linear programming; that is, the combination of the binary programming problem for $\mathbf{S}$ and the linear programming problem for $\boldsymbol{\gamma}$. We solve this problem by the branch-and-bound optimization technique. In our experiments using over 10,000 images, the convergence time is about 30 seconds for this step. These two steps for updating $\mathbf{C}$ and $\mathbf{S}$ are alternatively iterated until convergence.

**Table 1.** 20 classes defined for stomach endoscope images. BD, UP, MD, LO, LD, and LU stand for (stomach) body, upper, middle, lower, (camera) look-down, and look-up, respectively.

| | | | |
|---|---|---|---|
| Fundus | Fundus on UP BD/ LU | UP BD/ LU | UP BD/ LD |
| UP-MD BD/ LU | UP-MD BD/ LD | MD BD/ LU | MD BD/ LD |
| MD-LO BD/ LU | MD-LO BD/ LD | LO BD/ LD | Angular Incisure LO BD/ LD |
| AntralZone on LO BD/ LD | Angular Incisure | Angular Incisure-Antral Zone | Antral Zone |
| Pyloric Antral | Pyloric Zone | Pylorus | Junction |

**Table 2.** Quantitative performance evaluation by purity and recall. Larger is better.

| Metric | Conditions | KM | BOCK[5] | MPCK[9] | LCVQE[7] | Proposed w/o must link | Proposed |
|---|---|---|---|---|---|---|---|
| purity | K=100, R=1% | 0.602 | 0.627 | 0.421 | 0.573 | 0.658 | **0.715** |
| | K=100, R=3% | 0.726 | 0.673 | 0.544 | 0.660 | 0.712 | **0.789** |
| | K=100, R=5% | 0.681 | 0.716 | 0.494 | 0.587 | **0.747** | **0.747** |
| | K=50, R=1% | 0.623 | 0.549 | 0.369 | 0.616 | 0.517 | **0.660** |
| | K=50, R=3% | 0.668 | 0.648 | 0.521 | 0.636 | 0.651 | **0.727** |
| | K=50, R=5% | 0.637 | 0.694 | 0.498 | 0.643 | 0.662 | **0.702** |
| recall | K=100, R=1% | 0.132 | 0.138 | 0.109 | 0.122 | 0.147 | **0.159** |
| | K=100, R=3% | 0.156 | 0.167 | 0.147 | 0.155 | 0.158 | **0.179** |
| | K=100, R=5% | 0.162 | 0.169 | 0.170 | 0.159 | **0.174** | **0.174** |
| | K=50, R=1% | 0.254 | 0.230 | 0.170 | 0.230 | 0.218 | **0.270** |
| | K=50, R=3% | 0.287 | 0.281 | 0.263 | 0.269 | 0.273 | **0.310** |
| | K=50, R=5% | 0.262 | 0.296 | 0.258 | 0.279 | 0.295 | **0.297** |

## 5 Experimental results

In this section, we conducted experiments to evaluate the capability of the proposed method to support efficient ground-truth annotation for endoscopic image clustering. The endoscopy images were collected from several hospitals, including 11,599 stomach images captured by endoscopy. In the clustering task, these stomach images captured were classified into 20 classes listed in Table 1, according to the part in stomach and two camera capturing angles (look-down/look-up) for several parts.

We designed the performance metrics that fit the situation where clustering is used for the actual labeling process. In the evaluation scenario, we randomly select a certain amount of the samples along with their correct label and have cannot-links and must-links by using them. Then, we perform our soft-constrained clustering and get $K$ clusters. Among them, as described in Section 3, we pick up the "prominent" cluster for each class. The prominent cluster contains the most samples labeled by the expert with a certain class and thus a reliable cluster. We thus show the images belonging to the prominent cluster to the expert for finalizing the labeling of the samples. The expert will observe the images and discard the images from different classes and put the same class label to all the remaining images. Consequently, we need to have high purity and high recall at the prominent cluster to reduce this finalizing operation. The purity for each class is the number of true positives divided by the number of unlabeled samples in the prominent cluster. The recall for each class is the number of true positives in the

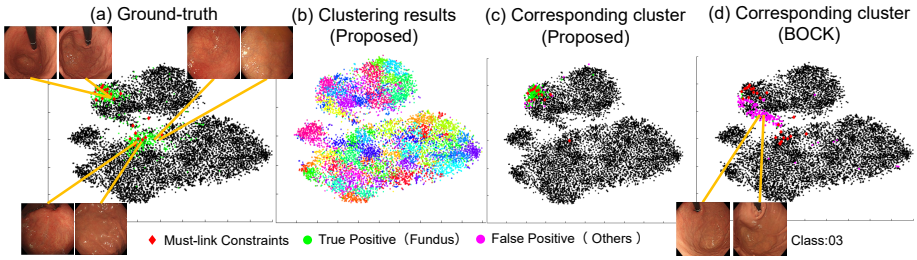


**Fig. 2.** Clustering result examples at $K = 50$ and $R = 0.03$. tSNE is used for these two-dimensional visualizations. (a) Distribution of samples whose true label is "Fundus." Green dots indicate unlabeled samples and red indicate labeled samples that were used as must-link constraints. (b) Clustering results from our method. (c) Results of proposed methods and (d) results of BOCK [5]. In (c) and (d), magenta dots indicate samples from other classes.

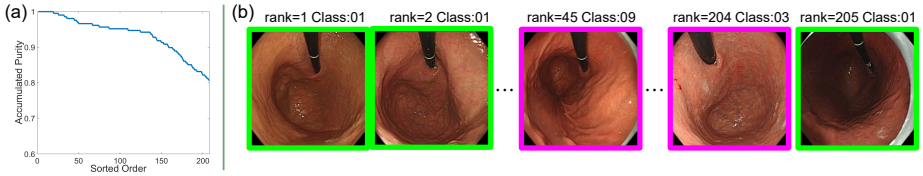


**Fig. 3.** (a) Accumulated purity curve with distance from the centroid of the cluster, where the cluster corresponds with the one in Fig. 2(c). The vertical axis indicates accumulated purity, and horizontal axis indicates sorted order by distance from centroid. (b) Images in cluster. Images were sorted by distance from centroid from left to right. "rank" shows the order by the distance in the cluster and rank=1 is the closest. "Class" is a true label of the image (e.g., Class:01 is "Fundus.")

prominent cluster divided by the total number of the samples of the class. In the later result, these measurements are averaged over all classes.

The performance of the proposed method was compared with several clustering methods: K-Means (KM) that is non-constrained clustering and the Binary Optimization approach for Constrained K-means (BOCK) that is hard-constrained clustering. Metric-based Pairwise Constrained K-means (MPCK) [9] and Linear-time Constrained Vector Quantization Error (LCVQE) [7] that are state-of-the-art soft-constrained clustering[4]. As discussed in Section 2, both soft-constrained clustering methods were designed to discard a small number of erroneous constraints. In addition, we evaluated the proposed method without using must link as an ablation study. The parameter $\omega$ was set to 50 for all experiments. To demonstrate the robustness for the rate of labeled samples over the total data $R$ and the number of clusters $K$, we evaluated the performance under several conditions: $K = 50, 100$ and $R = 1\%, 3\%, 5\%$, where the labeled samples were randomly picked up, and then the must-link and cannot-link constraints were generated

[4] A most popular hard-constrained clustering, COP-Kmeans [4] has not been compared; it did not work due to its heavy computational complexity for our large dataset.

on the basis of the labeled data. When $R = 1\%$, the average number of labeled data in each class is about 29.

As shown in Table 2, our method achieved the best performance under all conditions. When $K = 100$, the purity of the proposed method was over 0.7. Here, we note that the purity was computed using only the unlabeled data, and thus it could occur that the purity with $R = 5\%$ is better than that with $R = 10\%$. In the annotation task for separating a set of images to a particular class and others, this purity value is high enough to improve the efficiency of the annotation task compared with the individual annotation process. In particular, in the case where the number of labeled data was low $(1\%)$, the purity of the proposed method is the highest and shows $8\%(= 0.715 - 0.627)$ improvement compared with the second best, BOCK. Our method also achieved the best recall.

Fig. 3 shows the image examples in a cluster corresponding with class "Fundus." To analyze the relationship the distance from the centroid and the class label of the sample, we sorted the data by distances from the cluster centroid. Fig. 3(a) shows the accumulated purity curve. In this plot, the samples near the centroid tend to have the same label as the labeled samples in the cluster.

The average running time for clustering over 10,000 images was 14s, 1119s, 10s, 1824s, and 130s by KM, BOCK, MPCK, LCVQE, and the proposed method, respectively. Our method was much faster than the other state-of-the-arts, BOCK and LCVQE, and its speed is acceptable for the group-based labeling scheme.

## 6 Conclusion

We proposed a soft-constrained clustering suitable for group-based labeling. The proposed method performs clustering by using must-links for penalties instead of hard-constraints by considering that the class distribution of medical images is often not unimodal but multimodal whose components are distant to each other. The advantage of our method is that we formulate the soft-constrained clustering problem as a single optimization problem rather than a typical two-step optimization where the problem is solved as a hard-constrained problem and then update the result. Our method achieved higher purity and recall than several state-of-the-art clustering methods under various conditions with different numbers of clusters and rates of the must/cannot-links.

**Acknowledgments.** This work was supported by JSPS KAKENHI Grant Number JP19K22895 and AMED Grant Number JP18lk1010028.

## References

1. Biswas, A. and Jacobs, D. Active image clustering: Seeking constraints from humans to complement algorithms. *Proc. CVPR*, pp.2152-2159, (2012).
2. Galleguillos, C., McFee, B., and Lanckriet, G. Iterative category discovery via multiple kernel metric learning. *IJCV*, 108(1-2), pp.115-132, (2014).
3. Bruce, M.W., Draper, A. and Beveridge, J.R. Efficient Label Collection for Unlabeled Image Datasets. *Proc. CVPR*, pp.4594-4602, (2015).

4. Wagstaff, K., Cardie, C., Rogers, S., and Schrodl, S. Constrained k-means clustering with background knowledge. *Proc. ICML*, pp.577-584, (2001).
5. Le, H.M., Eriksson, A., Do, T.T., and Milford, M. A binary optimization approach for constrained K-Means clustering. *Proc. ACCV*, (2018).
6. Davidson, I., Ravi, S.S. Clustering with constraints: Feasibility issues and the k-means algorithm. *Proc. on SIAM Data Mining*, (2005).
7. Pelleg, D., Baras, D., K-means with large and noisy constraint sets. *Proc. ECML*, pp.674-682, (2007).
8. Basu, S., Banerjee, A., and Mooney R.J. Active semi-supervision for pairwise constrained clustering. *Proc. SIAM on Data Mining* pp.333-344, (2004).
9. Bilenko, M., Basu, S., Mooney, R.J. Integrating constraints and metric learning in semi-supervised clustering. *Proc. ICML*, pp.81-88, (2004).
10. Huang, G., Liu, Z., Maaten, L., and Weinberger, K.Q. Densely connected convolutional networks. *Proc. CVPR*, (2017).
11. Rifkin, R.M. and Lippert, R.A. Notes on regularized least squares *MIT-CSAIL Technical Report*, (2007).