Knowl Inf Syst (2013) 36:489–515
DOI 10.1007/s10115-012-0592-8

REGULAR PAPER

Improving clustering with pairwise constraints: a discriminative approach

Hong Zeng · Aiguo Song · Yiu Ming Cheung

Received: 23 December 2011 / Revised: 17 August 2012 / Accepted: 3 December 2012 /
Published online: 22 December 2012

Abstract To obtain a user-desired and accurate clustering result in practical applications, one way is to utilize additional pairwise constraints that indicate the relationship between two samples, that is, whether these samples belong to the same cluster or not. In this paper, we put forward a discriminative learning approach which can incorporate pairwise constraints into the recently proposed two-class maximum margin clustering framework. In particular, a set of pairwise loss functions is proposed, which features robust detection and penalization for violating the pairwise constraints. Consequently, the proposed method is able to directly find the partitioning hyperplane, which can separate the data into two groups and satisfy the given pairwise constraints as much as possible. In this way, it makes fewer assumptions on the distance metric or similarity matrix for the data, which may be complicated in practice, than existing popular constrained clustering algorithms. Finally, an iterative updating algorithm is proposed for the resulting optimization problem. The experiments on a number of real-world data sets demonstrate that the proposed pairwise constrained two-class clustering algorithm outperforms several representative pairwise constrained clustering counterparts in the literature.

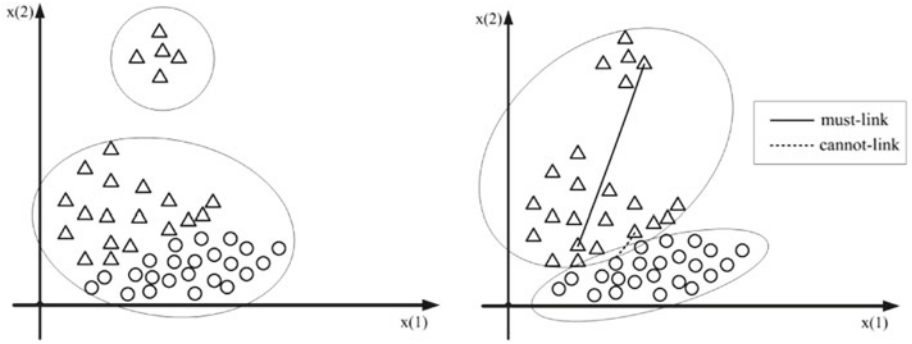
Keywords Maximum margin clustering · Pairwise constraints · Discriminative approach · Robust pairwise loss function

1 Introduction

The goal of clustering is to split a collection of data into meaningful groups or clusters. It has been extremely valuable for data analysis and plays a key role in various application domains

H. Zeng (✉) · A. Song
School of Instrument Science and Engineering, Southeast University,
Nanjing, China
e-mail: hzeng@seu.edu.cn

Y. M. Cheung
Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China



(**a**)The clustering result without pairwise constraints guided (**b**)The clustering result with pairwise constraints guided

Fig. 1 An example for clustering guided with pairwise constraints. Samples from different true categories are represented by *circle* and *triangle*, respectively. The *ellipses* denote the clusters. The result in (**a**) does not correspond to the true semantic categories of data, while the result in (**b**) has been greatly improved with the guidance of pairwise constraints

ranging from engineering, bioinformatics, earth science, medical science, social science to economics [1–5]. In practice, it is, therefore, of great importance to obtain a user-desired and accurate grouping of data, that is, the clusters obtained by a clustering algorithm correspond to the true semantic categories of the data [6]. However, due to its unsupervised nature, a traditional unsupervised clustering algorithm may fail to provide the desired partitions, for example, several documents that talk about different topics but with many words in common may be grouped into the same cluster. One solution for this problem is to introduce prior information that may guide the partitioning of the data. In this paper, we are particularly interested in how to incorporate pairwise constraints to improve the clustering performance. Specifically, two kinds of pairwise constraints are additionally provided for the clustering task at hand: One is the *must-link* constraint describing that the pair of instances should belong to the *same* cluster; the other is the *cannot-link* constraint specifying that the pair of instances should be assigned to *different* clusters. Although pairwise constraints provide weak supervision information as they only indicate the relationship between the splits rather than the splits themselves, it is generally believed that they can boost the clustering performance [7,8]. A simple example for pairwise constrained clustering is illustrated in Fig. 1.

In practice, pairwise constraints can be often obtained conveniently. First, in many application areas, pairwise constraints can be collected automatically along with the unlabeled data. For example, the protein co-occurring information in the Database of Interacting Proteins (DIP) data set can be used as the must-link constraints when performing gene clustering [9]. In video surveillance applications, any two faces from the same frame must come from two different persons, thus forming a cannot-link constraint, whereas those from temporally adjacent frames with spatial overlap can be assumed to be must-link [10]. Second, pairwise constraints can also be collected by human specification with little efforts. Unlike labels that would require users to have prior knowledge or experience with a data set, it is less difficult for a user, who is even not an expert in a domain, to make a judgement whether two objects are similar or not than to provide them with the exact class labels.

Since the last decade, a number of approaches have been developed, which have achieved reasonable performance improvement in clustering tasks using pairwise constraints. The most predominant strategy of these methods is to learn a better distance metric or similarity matrix by leveraging pairwise constraints first and then to apply standard clustering algorithms like K-means or spectral methods with the resulting metric or similarity matrix to determine the cluster assignments [9,11–20]. Despite the success in their application domains, several existing works [9,13–16] simply utilize the limited number of prespecified pairwise constrained samples in the metric learning step, while the abundant unconstrained ones are not taken into account. As a result, the obtained metric or similarity may not be reliable for the whole data set [21]. Thereby, it is more rational to make use of valuable intrinsic information with the unconstrained data in addition to the given pairwise constraints, in order to boost the effect of such scarce supervision information. Nevertheless, since the task of clustering is to split a collection of data into meaningful clusters, it may make more sense for pairwise constraints to directly guide the learning for the partitioning boundaries with the largest margins between categories, than to help in modeling the metrics of data with which to perform the clustering afterwards. The plausible reason is that the partitioning boundaries might be simpler to estimate by such margin-based approaches, whereas the ground-truth distance metrics might be considerably complex or against the model assumption [10,22]. Thus, by extending maximum margin clustering (MMC) algorithm [23] that directly finds the hyperplanes that well separate different groups of data by a generalized large margin criterion, Hu et al. have proposed a pairwise constrained two-class MMC and a pairwise constrained multi-class MMC in [25]. Moreover, they have shown that it is much easier for the pairwise constrained variants of MMC to obtain the cluster memberships for out-of-sample data points, because the discriminative partitioning boundaries can be established using the available data [25]. Such a merit is important for generalization, predictive purposes and large-scale data. By contrast, it is generally nontrivial for some existing semi-supervised clustering methods (e.g., the pairwise constrained variants of spectral clustering approaches [11,12,16,18–20]) to determine the cluster memberships for new samples, as no partitioning model is built in such methods. Nevertheless, our previous work [26] finds that the multi-class constrained clustering method in [25] may not be able to robustly detect whether the cannot-link constraints have been violated, particularly when the cannot-link pair of samples actually belong to different semantic categories yet with very similar appearance. Furthermore, we find that the constrained two-class clustering algorithm in [25] also suffers from such a drawback. Besides, we have further observed that it may generate unexpected high penalty even though some given must-link constraints are not violated, especially for the case where the must-link pair of samples are from far away subgroups in a semantic class.

In this paper, aimed at improving the clustering performance on the two-class data set, we propose a novel approach to adapt the unsupervised MMC for its pairwise constrained variant. In particular, to overcome drawbacks of the preliminary pairwise constrained two-class MMC approach [25], a new set of pairwise loss functions is proposed based on a two-class discriminative classification model. It is sensitive to detect the violations of pairwise constraints and is able to provide robust penalization to such violations based on the margin principle. Meanwhile, it would not produce unexpected high penalty if the pairwise constraints are satisfied. In addition, as the extension of MMC, there are also three other advantages of the proposed method over some pairwise constrained clustering counterparts in the literature. Firstly, it makes fewer assumptions on the distance metric or similarity which might be complicated in practice. Secondly, it is able to exploit the limited number of pairwise constraints together with abundant unconstrained samples in a unified discriminative learning framework. Thirdly, it is much easier for the proposed method to determine the cluster memberships for out-of-sample examples than its constrained spectral clustering counterparts. For the resulting optimization problem, we have developed an effective and efficient iterative algorithm to solve it. Experimental results evaluated on extensive benchmark real-world data sets demonstrate that the proposed pairwise constrained clustering method outperforms state-of-the-art constrained clustering counterparts for two-class clustering tasks.

The remainder of the paper is organized as follows. In Sect. 2, we present related work on pairwise constrained clustering. We formulate the basic pairwise constrained two-class maximum margin clustering problem in Sect. 3. In Sect. 4, the proposed constrained two-class MMC approach is given in detail. Experimental results are given in Sect. 5. Finally, some concluding remarks are given in Sect. 6.

2 Related work

A lot of efforts have been made in pairwise constrained classification [10,33–36] and pairwise constrained clustering [8,9,11,13–21,25–29,31,32] for improving the performance of the learning algorithms. In the following, we will only focus on the work utilizing the pairwise constraints for clustering. Generally, the current pairwise constrained clustering methods can be divided into four categories: *constraint-based* ones, *metric-based* ones, *similarity-based* ones and *penalty-based* ones.

The *constraint-based* methods: The seminal effort on *constraint-based* semi-supervised clustering, that is, the COP-K-means algorithm, has been made by Wagstaff et al. in [8]. It follows the same clustering procedure of K-means while respecting the pairwise constraints, requiring that none of the constraints is violated in the clustering process. Nevertheless, such a *strict* enforcement of constraints often suffers from great performance degradation in case of possibly noisy constraints.

The *metric-based* methods: In these approaches, a distance metric is firstly trained to satisfy the prespecified pairwise constraints as much as possible, and then K-means clustering is performed with the obtained metric. The algorithm in [13] learns a Mahalanobis distance metric which corresponds to a linear transformation of the data features, by minimizing the distance between must-link data points while keeping cannot-link data points apart. Nevertheless, it is computationally too expensive to be applied to practical applications. Bar-Hillel et al. thus proposed the relevant components analysis (RCA) [14], which is efficient and achieves comparable results as the algorithm in [13]. The main drawback with RCA is that it is unable to handle the cannot-link constraints. Such a problem has been addressed by discriminative component analysis (DCA) in [15], which can make use of both the must-link and cannot-link constraints to learn the optimized transformation. Most recently, Wu et. al assert that the Mahalanobis distance metric model is unable to effectively handle binary-class data with multi-modal distribution, which is, however, common in practice, because it simply assumes a fixed metric for the entire input space [22]. Furthermore, these metric-based pairwise constrained clustering algorithms exclude unconstrained data from the metric training step, therefore leading to less reliable metric for the whole data set [17,31,32].

The *similarity-based* methods: These approaches resort to seeking an appropriate similarity matrix with the assistance of pairwise constraints for conducting the spectral clustering. With a given affinity matrix, the methods of [9,16] simply modify the similarities between constrained objects only. Not surprisingly, such an ad-hoc modification usually does not work well in practice due to insufficient pairwise constraints, and the obtained similarity matrix generally may not be a valid kernel matrix. The methods in [18,19] resolve this problem by introducing the manifold regularizer when optimizing the similarity matrix with respect to the pairwise constraints. However, they require semi-definite programming (SDP) whose computation cost is considerably intensive and prohibits such approaches applicable to real applications, even for data sets of moderate size. In a way consistent with the given affinities, a much more efficient work in [20] propagates the pairwise constraints from a few specified sample pairs to the rest of the entries in the similarity matrix using Gaussian process.

The *penalty-based* methods: These approaches augment penalty functions for violating the pairwise constraints with the original clustering objective function; therefore, they can utilize the unconstrained samples and the constrained ones in a single framework. The proposed method also follows this line. In this way, the paper [21] and references in [27–30] adapt the K-means and the nonnegative matrix factorization (NMF) clustering to their constrained variants, respectively, but with no extension to new points. Most recently, preliminary efforts on constrained MMC have been made in [25,26] along this direction. Despite their success, we find that the loss function for violating the pairwise constraints in [25] may not be able to detect and discourage the violation of given constraints robustly. We further find that the performance of the pairwise constrained multi-class clustering algorithm in [26] may degrade for two-class clustering tasks, when the number of must-link constraints and that of cannot-link constraints are imbalanced.

3 Problem formulation

In this paper, only the two-class clustering problem is considered where we group the data into two clusters. Motivated by the success of large margin methods in supervised learning, MMC [23] aims at extending such methods to unsupervised learning. Specifically, for a given set of data points, it tries to find the maximum margin hyperplane that can separate the data into two different clusters among all possible binary labels. Formally, let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{m} (\mathbf{x}_i \in \mathbb{R}^d)$ denote the set of data points. Suppose the hyperplane is parameterized by the weight vector $\mathbf{w} \in \mathbb{R}^{d\times 1}$ and the bias term $b \in \mathbb{R}$, that is, the discriminative function can be written as $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}+b$, and then the MMC problem is cast into the following optimization problem [38], that is,

$$\min_{f} \left\{ \frac{1}{2}\|f\|_{\mathcal{I}}^2 + \lambda \sum_{i=1}^{m} L(|f(\mathbf{x}_i)|) \right\},$$
$$subject\ to:\ -s \le \sum_{i=1}^{m} f(\mathbf{x}_i) \le s, \tag{1}$$

where $\|f\|_{\mathcal{I}}^2$ is the regularization term controlling the complexity of the hypothesis space, and $\|\mathbf{w}\|^2$ is often adopted for $\|f\|_{\mathcal{I}}^2$. The item λ is a positive trade-off constant. The item $L(|f(\mathbf{x}_i)|)$ is the "generalized" margin constraint violation cost function with $L(z) = \max(1-z, 0)$. Such a cost function is similar to that for the transductive support vector machine (SVM) [24], forcing the unlabeled samples far away from the hyperplane. The constraint in (1) is a cluster balance constraint introduced to avoid the trivially optimal solutions because a large margin value can be always achieved by assigning all the samples to a single cluster [37]. The constant $s \ge 0$ controls the cluster imbalance. Ultimately, MMC learns the hypothesis f via solving (1), and finally, the cluster assignment for $\mathbf{x}_i$ can be determined as follows: It belongs to group 1 if $f(\mathbf{x}_i) \ge 0$, or group 2 otherwise. Moreover, the clustering memberships of the unseen data can be obtained similarly with the obtained partitioning hypothesis. Recently, several efforts [39–43] have been made to enhance the efficiency of the basic MMC [37], while some extend the basic MMC with subspace learning [44], feature selection [45] and multiple kernel learning [46] for improving its effectiveness. Nevertheless, such extensions are all unsupervised.

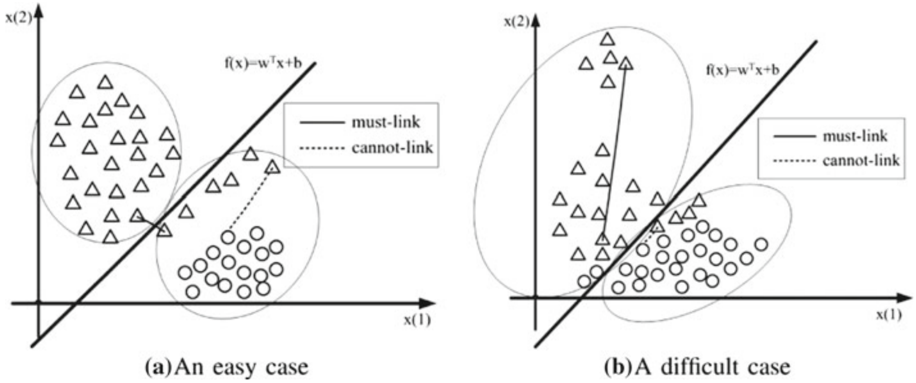


Fig. 2 Two intermediate clustering results with premature hyperplanes. Samples from different true categories are represented by *circle* and *triangle*, respectively. The premature hyperplane for each clustering task is depicted with a *straight line* across the two-dimensional space. The *ellipses* indicate the clusters determined by the hyperplane. As can be seen, the clustering result has violated the prespecified constraints. The pairwise loss function in [25] is able to provide appropriate penalization to the constraint violation for the case shown in (**a**), but may fail for the case in (**b**)

In this paper, to further improve the performance of MMC, we focus on introducing more informative supervision for adapting it to its semi-supervised variant. Specifically, a set of pairwise constraints is provided, that is, $\mathcal{C} = \{C_j = (\mathbf{x}_{j1}, \mathbf{x}_{j_2}, \ell_j)\}_{j=1}^{n}$, where $\ell_j = 1$ indicates that the pair of samples $(\mathbf{x}_{j_1}, \mathbf{x}_{j_2})$ must link, while $\ell_j = -1$ means that the pair of samples $(\mathbf{x}_{j_1}, \mathbf{x}_{j_2})$ cannot link. We denote the sets of prescribed must-link constraints and cannot-link constraints by $\mathcal{ML}$ and $\mathcal{CL}$, respectively, with $\mathcal{C} = \mathcal{ML} \cup \mathcal{CL}$. To derive the constrained MMC, the key lies in the way that the pairwise constraints are being incorporated into MMC. Under the same discriminative learning framework, they can be introduced, as another set of cost functions which attempts to penalize the violation of the given constraints, into the original MMC objective function, that is,

$$\min_{f} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + \lambda \sum_{i=1}^{m} L(|f(\mathbf{x}_i)|) + \gamma \sum_{j=1}^{n} L'(f(\mathbf{x}_{j_1}), f(\mathbf{x}_{j_2}), \ell_j) \right\}, \tag{2}$$

where $\mathbf{x}_i$ represents the sample that does not involve in the constraints set, the item $L'(f(\mathbf{x}_{j_1}), f(\mathbf{x}_{j_2}), \ell_j)$ denotes the cost function for violating the pairwise constraint C_j, and the coefficient γ is the another positive trade-off constant. Note that the cluster imbalance constraint is excluded in (2) because the cannot-link constraints are able to prevent the trivially optimal solutions [25]. From now on, we will use the shorthands f_i, f_{j_1} and f_{j_2} for $f(\mathbf{x}_i)$, $f(\mathbf{x}_{j_1})$ and $f(\mathbf{x}_{j_2})$, respectively.

A preliminary work on constrained two-class MMC has been proposed in [25], in which $|f_1 - \ell f_2|$ is adopted for the pairwise loss function, that is,

$$L'(f_1, f_2, \ell) = |f_1 - \ell f_2|. \tag{3}$$

Although the improvement over unsupervised MMC has been observed in [25], we find that this work with such a pairwise loss function may not work robustly in practice. Its robustness will be investigated through two cases as shown in Fig. 2.

For the case described in Fig. 2a, the premature hyperplane has resulted in the must-link pair $(\ell_j = 1)$ of samples, which are close to each other, residing on its two sides with opposite signs. Therefore, it will incur a large cost for penalizing such an violation, according to the pairwise loss function $|f_{j_1} - f_{j_2}|$. The same figure depicts a constraint ($\ell_j = -1$) that is posed on two far apart samples from different ground-truth classes. Such a pair of samples have been incorrectly placed on the same side of the hyperplane with the hyperplane with the same sign (i.e., the constraint is not fulfilled). Moreover, they are both far off the hyperplane. Therefore, the significant cost $|f_{j_1} + f_{j_2}|$ will effectively drive the algorithm to find a better hyperplane.

However, in practical applications, users are more inclined to specify some highly informative pairwise constraints for dealing with cases in which samples are much more difficult to be partitioned correctly. For instance, samples from different classes usually show serious overlap in real problems, that is, there are many samples belonging to different categories but seeming to be very similar (e.g., some face images may have similar appearance but belong to different persons, several documents may get many words in common but discuss different topics, etc.). Therefore, the cannot-link constraints are particularly provided to guide the partitioning in such difficult scenarios. As shown in Fig. 2b, a cannot-link constraint is specified in the severely overlapping region and the premature hyperplane passes through such a region. Consequently, the cannot-link pair of samples will also be close to the premature hyperplane in this case, for example, see Fig. 2b. As a result, the values of the decision function f for both samples, that is, f_{j_1} and f_{j_2}, will be near zero, and then $|f_{j_1} + f_{j_2}|$ would naturally reduce to zero as if the constraint was not violated. That is to say, even though the current partition does not fulfill such a cannot-link constraint in the case where the constraint is specified from a severe overlapping regions as shown in Fig. 2b, the pairwise loss function of [25] would not be able to discourage such a violation effectively.

Meanwhile, we often expect that two points belonging to the same true class but with a large distance (e.g., experimental samples from two subtypes of lung cancer [47]) could still be grouped in the same cluster. Thus, a must-link constraint is usually provided for such a pair of samples. In Fig. 2b, the must-link pair of samples happens to be correctly put into the same partition by the hyperplane. In other words, the must-link constraint is not violated. Nevertheless, because these two points are very different from each other, the form of $|f_{j_1} - f_{j_2}|$ might trigger an unexpected high cost, leading to a possibly poor hyperplane instead.

On the other hand, for the difficult case in Fig. 2b, the paper [22] argues that a fixed metric for the entire input space (e.g., the Mahalanobis metric) may not be able to effectively model the desired distance metric, with which originally far away samples from subgroups of a single semantic class (see the data points with the must-link constraints in Fig. 2b) are near to each other, while the samples belonging to different categories but with similar appearance (see the data points with the cannot-link constraints in Fig. 2b) become far apart. By contrast, it can be observed from Fig. 2 that it may be much easier to find a hyperplane which can well separate the samples into two clusters and fulfill the given constraints by the constrained MMC.

To this end, we will propose a new discriminative pairwise constrained two-class MMC algorithm. In particular, we put forward a set of pairwise loss functions, which could overcome the drawbacks of the work in [25] and features robust penalization to the violation of constraints.

4 The proposed discriminative pairwise constrained maximum margin clustering approach

Recall that the unsupervised two-class MMC is to find a binary labeling for the samples, so that the obtained hyperplane under such a labeling result has the largest margin from both the two labeled clusters [23]. With the given pairwise constraints, the main idea of our proposed pairwise constrained two-class clustering approach is as follows: The labeling for a pair of constrained samples should fulfill the must-link/cannot-link constraint on the cluster memberships of them; meanwhile, the corresponding hyperplane with such a labeling should possess the maximum margin from the constrained samples. Base on such an idea, the proposed new pairwise loss functions are presented first.

4.1 The proposed pairwise loss functions

Concerning the definitions for must-link and cannot-link constraints in the context of binary maximum margin clustering, we propose the following pairwise loss function $H_j(f)$ for $L'(f_{j1}, f_{j2}, \ell_j)$:

$$H_j(f) \triangleq \min\{L(f_{j1}) + L(\ell_j f_{j2}), L(-f_{j1}) + L(-\ell_j f_{j2})\}. \tag{4}$$

It essentially requires that if $\mathbf{x}_{j1}$ is labeled with 1 (or -1), then $\mathbf{x}_{j2}$ ought be labeled with ℓ_j (or $-\ell_j$) correspondingly; furthermore, $\mathbf{x}_{j1}$ and $\mathbf{x}_{j2}$ should be far away from the hyperplane at the same time. To evaluate the robustness of the proposed pairwise loss functions, the following two propositions are presented.

Proposition 1 *If C_j is violated, the pairwise loss for violating such a constraint $H_j(f)$ is always not less than 1.*

Proof Suppose C_j is a must-link constraint ($\ell_j = 1$) and is violated, recall that a sample $\mathbf{x}_i$ belongs to cluster 1 if $f_i \geq 0$, or cluster 2 otherwise; therefore, either f_{j1} or f_{j2} should be negative. Because $L(z) = \max(1-z, 0)$ is a nonnegative and monotonic decreasing function, we have

$$\begin{cases} L(f_{j1}) + L(f_{j2}) \geq L(0) = 1, \\ L(-f_{j1}) + L(-f_{j2}) \geq L(0) = 1. \end{cases}$$

Thereby, $H_j(f) \geq 1$ holds if a must-link constraint C_j is violated. In a similar way, we can prove that $H_j(f) \geq 1$ also holds if a cannot-link constraint C_j is violated. □

Proposition 2 *If C_j is not violated, the pairwise loss $H_j(f)$ is always less than 2.*

Proof Suppose C_j is a must-link constraint ($\ell_j = 1$) and is not violated. Without loss of the generality, we assume $f_{j1} \geq 0$ and $f_{j2} \geq 0$, and then we have $H_j(f) = \min\{L(f_{j1}) + L(f_{j2}), L(-f_{j1}) + L(-f_{j2})\} = L(f_{j1}) + L(f_{j2})$. Furthermore, recall that $L(z) = \max(1 - z, 0)$; we, therefore, obtain that $L(f_{j1}) \leq 1$ and $L(f_{j2}) \leq 1$. Consequently, $L(f_{j1}) + L(f_{j2}) \leq 2$ holds. In a similar way, we can prove that $H_j(f) \leq 2$ also holds if a cannot-link constraint C_j is not violated. □

According to Proposition 1 and Proposition 2, several remarks can be made as follows:

- When C_j is violated, the pairwise loss for violating such a constraint $H_j(f)$ is always not less than 1. Furthermore, $H_j(f)$ will produce a greater value than 1 when the hyperplane is close to either $\mathbf{x}_{j1}$ or $\mathbf{x}_{j2}$ (i.e., $|f_{j1}| < 1$ or $|f_{j2}| < 1$). Therefore, in comparison with (3) which may fail to effectively penalize the violation of cannot-link constraints specified in seriously overlapped region as depicted in Fig. 2b, $H_j(f)$ is expected to be able to overcome such a drawback in this case.

- The value of $H_j(f)$ is upper bounded if C_j is not violated. Moreover, when C_j is satisfied, no matter how different $\mathbf{x}_{j_1}$ and $\mathbf{x}_{j_2}$ may be, as both of them are far away from the hyperplane (i.e., $|f_{j_1}| > 1$ and $|f_{j_2}| > 1$), $H_j(f)$ will be near 0. Thereby, compared to (3), it would not trigger an undesired significant cost for the aforementioned case in Fig. 2b, where a must-link pair of samples stay in the same partition yet have a large distance.

4.2 Pairwise constrained maximum margin clustering algorithm

Subsequently, the corresponding constrained two-class MMC optimization problem can be written as follows:

$$Q(f) \triangleq \frac{1}{2}\|\mathbf{w}\|^2 + \lambda \sum_{i=1}^{m} L(|f_i|) + \gamma \sum_{j=1}^{n} H_j(f). \tag{5}$$

Note that pairwise constraints have already been proposed to be incorporated into SVMs [10,33]. However, the classification problem is the main focus in [10,33], where the labels for some examples are known during training. By contrast, we handle the clustering problem where no labels are provided in (5).

Although $H_j(f)$ may be a preferred choice compared to the loss function (3), it is generally intractable to optimize. Therefore, we will introduce several auxiliary functions as follows:

$$H_j^{(1)}(f) \triangleq L(f_{j_1}) + L(\ell_j f_{j_2}),$$
$$H_j^{(2)}(f) \triangleq L(-f_{j_1}) + L(-\ell_j f_{j_2}).$$

Further, we define

$$\widetilde{H}_j(f, p) = p \cdot H_j^{(1)}(f) + (1 - p) \cdot H_j^{(2)}(f), \tag{6}$$

where $p \in [0, 1]$. We have

$$H_j(f) = \min_{p \in [0,1]} \widetilde{H}_j(f, p). \tag{7}$$

Moreover, it can be verified that

$$p^* \triangleq \arg \min_{p \in [0,1]} \widetilde{H}_j(f, p), \tag{8}$$

should be either 0 or 1, which can thus be viewed as a selector for $H_j(f)$ from $\{H_j^{(1)}(f), H_j^{(2)}(f)\}$. For the proposed constrained MMC objective function in (5) and the following formulation

$$\widetilde{Q}(f, \{p_j\}_{j=1}^n) \triangleq \frac{1}{2}\|\mathbf{w}\|^2 + \lambda \sum_{i=1}^{m} L(|f_i|) + \gamma \sum_{j=1}^{n} \widetilde{H}_j(f, p), \tag{9}$$

with (6)–(8), we have

$$Q(f) = \min_{\{p_j\}_{j=1}^n \in [0,1]^n} \widetilde{Q}(f, \{p_j\}_{j=1}^n). \tag{10}$$

Consequently, the problem of minimizing $Q(f)$ with respect to f can be equivalently cast into the one of the minimizing $\widetilde{Q}(f, \{p_j\}_{j=1}^n)$ with respect to f and $\{p_j\}_{j=1}^n \in [0, 1]^n$, that is,

$$\min_f Q(f) = \min_{f, \{p_j\}_{j=1}^n} \widetilde{Q}(f, \{p_j\}_{j=1}^n). \tag{11}$$

Under the circumstances, we have developed an iterative optimization algorithm for the proposed pairwise constrained MMC problem. In each iteration, it minimizes $\widetilde{Q}(f, \{p_j\}_{j=1}^n)$ by iteratively keeping either f or $\{p_j\}_{j=1}^n$ fixed and optimizing $\widetilde{Q}(\cdot)$ with respect to the other, that is,

- Fixing f at $f^{(t-1)}$, we will solve the following problem for $\{p_j^{(t)}\}_{j=1}^n$, that is,

$$\{p_j^{(t)}\}_{j=1}^n = \arg \min_{\{p_j\}_{j=1}^n} \widetilde{Q}(f^{(t-1)}, \{p_j\}_{j=1}^n). \tag{12}$$

- Fixing $\{p_j\}_{j=1}^n$ at $\{p_j^{(t)}\}_{j=1}^n$, we then optimize the following problem for $f^{(t)}$, that is,

$$f^{(t)} = \arg \min_f \widetilde{Q}(f, \{p_j^{(t)}\}_{j=1}^n). \tag{13}$$

Based on (7), it can be seen that the problem (12) can be directly solved by

$$p_j^{(t)} = \begin{cases} 1, & \text{if } H_j^{(1)}(f^{(t-1)}) < H_j^{(2)}(f^{(t-1)}); \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

In the following, we will develop the detailed procedure for (13). Specifically, the corresponding optimization problem can be written below:

$$\min_f \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + \lambda \sum_{i=1}^m L(|f_i|) + \gamma \sum_{j=1}^n \widetilde{H}_j(f, p^{(t)}) \right\}. \tag{15}$$

By introducing slack variables as in the support vector machine algorithm, (15) can be equivalently formulated as follows:

$$\min_{f,\boldsymbol{\xi}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + \lambda \sum_{i=1}^m \xi_i + \gamma \sum_{j=1}^n \widetilde{H}_j(f, p^{(t)}) \right\}$$
$$\textit{subject to}: \ |f(\mathbf{x}_i)| \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, m, \tag{16}$$

where $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_m]^T$. From Eq. (6), it is clear that $\widetilde{H}_j(f, p^{(t)})$ is just the linear combination of the convex loss functions with nonnegative coefficients; thus, it is still convex. Consequently, the objective function in (16) is convex. However, the inequality constraints are not convex, but can be viewed as a difference of two convex functions. Fortunately, such a nonconvex optimization problem in (16) can be efficiently solved by resorting to the constrained concave–convex procedure (CCCP) [48,49], which will be described in the following subsection.

4.2.1 The constrained concave–convex procedure for the optimization

CCCP is an effective technique proposed recently to solve problems with concave–convex objective function under the concave–convex constraints in the following form [38,48,49]:

$$\min_{\mathbf{z}} \left\{ h_0(\mathbf{z}) - g_0(\mathbf{z}) \right\}$$
$$\text{subject to}: \ h_i(\mathbf{z}) - g_i(\mathbf{z}) \leq c_i, i = 1, \ldots, q, \tag{17}$$

where h_0, g_0, h_i, g_i $(i = 1, \ldots, q)$ are convex and differentiable functions, and $c_i \in \mathbb{R}$. Given an initial guess on $\mathbf{z}_0$, in the $(r+1)$-th iteration, CCCP first replaces $g_0(\mathbf{z}), g_i(\mathbf{z})$ in (17) with its linearization at $\mathbf{z}_r$ and then solves the resulting convex problem for $\mathbf{z}_{r+1}$:

$$\begin{aligned} &\min_{\mathbf{z}} \left\{ h_0(\mathbf{z}) - \left[g_0(\mathbf{z}_r) + \nabla g_0(\mathbf{z}_r)^T (\mathbf{z} - \mathbf{z}_r) \right] \right\} \\ &\quad \text{subject to : } h_i(\mathbf{z}) - \left[g_i(\mathbf{z}_r) + \nabla g_i(\mathbf{z}_r)^T (\mathbf{z} - \mathbf{z}_r) \right] \le c_i, i = 1, \ldots, q, \end{aligned} \tag{18}$$

where $\nabla g_0\ (\mathbf{z}_r), \nabla g_i\ (\mathbf{z}_r)(i = 1, \ldots, q)$ are the gradients of $g_0(.), g_i(.)$ at $\mathbf{z}_r$, respectively. It can be shown that the objective (18) in each CCCP iteration decreases monotonically [48]. When CCCP converges, it will arrive at a local minimum of (17) [49].

For our problem, as $|\mathbf{w}^T \mathbf{x}_i + b|$ is nonsmooth, the gradient in (18) should then be replaced by the subgradient[1]. Let $f_r(\mathbf{x}_i) = \mathbf{w}_r^T \mathbf{x}_i + b_r$; we have

$$\begin{aligned} \frac{\partial}{\partial \mathbf{w}} |\mathbf{w}^T \mathbf{x}_i + b| \Big|_{\mathbf{w}_r, b_r} &= sgn(\mathbf{w}_r^T \mathbf{x}_i + b_r)\mathbf{x}_i, \\ \frac{\partial}{\partial b} |\mathbf{w}^T \mathbf{x}_i + b| \Big|_{\mathbf{w}_r, b_r} &= sgn(\mathbf{w}_r^T \mathbf{x}_i + b_r), \end{aligned}$$

where $sgn(\,)$ denotes the sign function. Following CCCP, the $|f(\mathbf{x}_i)|$ term in (16) is replaced by its first-order Taylor expansion

$$\begin{aligned} &|\mathbf{w}_r^T \mathbf{x}_i + b_r| + sgn(\mathbf{w}_r^T \mathbf{x}_i + b_r)[(\mathbf{w} - \mathbf{w}_r)^T \mathbf{x}_i + (b - b_r)] \\ &\quad = sgn(\mathbf{w}_r^T \mathbf{x}_i + b_r)(\mathbf{w}^T \mathbf{x}_i + b) = sgn(f_r(\mathbf{x}_i)) f(\mathbf{x}_i). \end{aligned}$$

Then, we obtain the following convex optimization problem for each iteration of CCCP:

$$\begin{aligned} &\min_{f, \boldsymbol{\xi}} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + \lambda \sum_{i=1}^{m} \xi_i + \gamma \sum_{j=1}^{n} \widetilde{H}_j(f, p^{(t)}) \right\} \\ &\quad subject\ to : .\ sgn(f_r(\mathbf{x}_i)) f(\mathbf{x}_i) \ge 1 - \xi_i,\ \xi_i \ge 0,\ i = 1, \ldots, m, \end{aligned} \tag{19}$$

which can be reformulated into

$$\min_{f} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + \lambda \sum_{i=1}^{m} L(sgn(f_r(\mathbf{x}_i)) f(\mathbf{x}_i)) + \gamma \sum_{j=1}^{n} \widetilde{H}_j(f, p^{(t)}) \right\}. \tag{20}$$

Subsequently, starting with f_0, we will solve (20) for f_{r+1}. Suppose it converges in the R-th CCCP iteration, the solution to (15), that is, $f^{(t)}$, is then obtained by $f^{(t)} = f_R$.

For solving (20), let

$$\hat{y}_i \triangleq \begin{cases} 1 & \text{if } f_r(\mathbf{x}_i) \ge 0; \\ -1 & \text{otherwise.} \end{cases} \tag{21}$$

$$\hat{y}_{j_1} \triangleq \begin{cases} 1 & \text{if } C_j \in \mathcal{ML} \text{ and } p_j^{(t)} = 1; \\ -1 & \text{if } C_j \in \mathcal{ML} \text{ and } p_j^{(t)} = 0; \\ 1 & \text{if } C_j \in \mathcal{CL} \text{ and } p_j^{(t)} = 1; \\ -1 & \text{if } C_j \in \mathcal{CL} \text{ and } p_j^{(t)} = 0. \end{cases} \tag{22}$$

[1] It can be shown that the CCCP remains valid when using any subgradient of the concave function [50]. A subgradient of f at $\mathbf{x}$ is any vector $\mathbf{g}$ that satisfies the inequality $f(\mathbf{y}) \le f(\mathbf{x}) + \mathbf{g}'(\mathbf{y} - \mathbf{x})$ for all $\mathbf{y}$ [51].

$$\hat{y}_{j2} \triangleq \begin{cases} 1 & \text{if } C_j \in \mathcal{ML} \text{ and } p_j^{(t)} = 1; \\ -1 & \text{if } C_j \in \mathcal{ML} \text{ and } p_j^{(t)} = 0; \\ -1 & \text{if } C_j \in \mathcal{CL} \text{ and } p_j^{(t)} = 1; \\ 1 & \text{if } C_j \in \mathcal{CL} \text{ and } p_j^{(t)} = 0. \end{cases} \tag{23}$$

We further let γ equal to λ, and then it can be observed that (20) is just a standard SVM problem, with $\{(\mathbf{x}_i, \hat{y}_i)\}_{i=1}^m$, $\{(\mathbf{x}_{j1}, \hat{y}_{j1})\}_{j=1}^n$ and $\{(\mathbf{x}_{j2}, \hat{y}_{j2})\}_{j=1}^n$ being the training examples. The problem in (20) can be efficiently solved by directly applying an off-the-shelf SVM solver, for example, SVM-perf[2] which is used in this paper. Therefore, subjecting to pairwise constraints, we finally reformulate the two-class clustering problem into a series of supervised binary learning problems and then make use of an effective and efficient SVM solver for optimizing such supervised problems. This makes our proposed method theoretically attractive. Its effectiveness in practice is demonstrated through experiments in Sect. 5.

The detailed procedure for solving (13) using CCCP is summarized in Algorithm 1. We stop the algorithm when the *relative* difference between objective values of the convex problem in (20) is less than ϵ_1 in two successive iterations, which means the current objective value of (20) is only larger than $(1 - \epsilon_1)$ of the objective value in the last iteration. In this paper, ϵ_1 is set at 0.01.

Algorithm 1: The procedure for (13) using CCCP.

```
input : $\{\mathbf{x}_i\}_{i=1}^m$, $\{(\mathbf{x}_{j_1}, \mathbf{x}_{j_2}, \ell_j)\}_{j=1}^n$, $p_j^{(t)}$, $\lambda$, $f_0$
Determine $\hat{y}_{j_1}, \hat{y}_{j_2}$ by (22)-(23), $j = 1, \ldots, n$;
Determine $\hat{y}_i$ by (21) with $r = 0$, $i = 1, \ldots, m$;
repeat
    Solve (20) for $f_{r+1}$ by an SVM solver, with $\{(\mathbf{x}_i, \hat{y}_i)\}_{i=1}^m$, $\{(\mathbf{x}_{j_1}, \hat{y}_{j_1})\}_{j=1}^n$ and $\{(\mathbf{x}_{j_2}, \hat{y}_{j_2})\}_{j=1}^n$
    being the training examples;
    $f_r \leftarrow f_{r+1}$;
    Determine $\hat{y}_i$ by (21), $i = 1, \ldots, m$;
until stopping criterion is satisfied ;
output: $f_r$. Then we use $f_r$ for $f^{(t)}$.
```

4.2.2 The complete algorithm

A proper initialization for f is critical because the original problem in (5) is nonconvex. We, therefore, provide a heuristic method for the initialization. Since MMC projects each sample to a one-dimensional space with the projection vector $[\mathbf{w}^T \; b]^T$, it is expected that the cannot-link pairs of samples should be as far away as possible, while the must-link pairs should be as close as possible. To this end, we try to find a projection vector $\mathbf{u} \in \mathbb{R}^{d+1}$ to map the data, formed by appending 1 to each instance, into such a space by solving

$$\max_{\mathbf{u}} \left\{ \frac{1}{N_c} \sum_{j:\ell_j=-1} \left(\left(\mathbf{x}_{j1}^T \; 1 \right)^T \mathbf{u} - \left(\mathbf{x}_{j2}^T \; 1 \right)^T \mathbf{u} \right)^2 = \mathbf{u}^T \mathbf{S}_c \mathbf{u} \right\},$$

$$\textit{subject to}:$$

$$\frac{1}{N_m} \sum_{j:\ell_j=1} \left(\left(\mathbf{x}_{j1}^T \; 1 \right)^T \mathbf{u} - \left(\mathbf{x}_{j2}^T \; 1 \right)^T \mathbf{u} \right)^2 = \mathbf{u}^T \mathbf{S}_m \mathbf{u} = 1, \tag{24}$$

[2] http://svmlight.joachims.org/svm_perf.html.

where N_c and N_m denote the total number of must-link and cannot-link constraints, respectively, and

$$\mathbf{S}_c = \frac{1}{N_c} \sum_{j:\ell_j=-1} \left[\left(\mathbf{x}_{j1}^T\ 1\right) - \left(\mathbf{x}_{j2}^T\ 1\right) \right]^T \left[\left(\mathbf{x}_{j1}^T\ 1\right) - \left(\mathbf{x}_{j2}^T\ 1\right) \right], \tag{25}$$

$$\mathbf{S}_m = \frac{1}{N_m} \sum_{j:\ell_j=1} \left[\left(\mathbf{x}_{j1}^T\ 1\right) - \left(\mathbf{x}_{j2}^T\ 1\right) \right]^T \left[\left(\mathbf{x}_{j1}^T, 1\right) - \left(\mathbf{x}_{j2}^T\ 1\right) \right]. \tag{26}$$

By the Lagrange method, the variable vector $\mathbf{u}$ can be easily solved, that is, the optimal $\mathbf{u}$ should be the first eigenvector of $(\mathbf{S}_m)^{-1}\mathbf{S}_c$ (it can be obtained by the "eigs( )" function in Matlab). Ultimately, the solution for $\mathbf{u}$ is used to initialize f.

The complete algorithm for the proposed pairwise constrained MMC is summarized in Algorithm 2. We stop the algorithm when the *relative* difference between objective values of the problem in (9) is less than ϵ_2 in two successive iterations. In our later experiments, ϵ_2 is set at 0.01.

Algorithm 2: The proposed pairwise constrained MMC algorithm.

```
input : $\{\mathrm{x}_i\}_{i=1}^m$, $\{(\mathrm{x}_{j_1}, \mathrm{x}_{j_2}, \ell_j)\}_{j=1}^n$, $\lambda(\gamma)$
Initialize $f^{(0)}$ by solving (24);
repeat
    Compute $p_j^{(t)}$ by (14);
    Compute $f^{(t)}$ by Algorithm 1;
until stopping criterion is satisfied (suppose it stops in the $\bar{T}$-th iteration) ;
output: $f^{(t)}$. $f^{(t)}$ is then denoted as $\hat{f}$. $\mathrm{x}_i$ belongs to group 1 if $\hat{f}(\mathrm{x}_i) \geq 0$, or group 2 otherwise.
```

Proposition 3 *The proposed pairwise constrained two-class MMC algorithm in Algorithm 2 always converges.*

Proof By (14), we have

$$\widetilde{Q}\left(f^{(t+1)}, \left\{p_j^{(t)}\right\}_{j=1}^n\right) \geq \widetilde{Q}\left(f^{(t+1)}, \left\{p_j^{(t+1)}\right\}_{j=1}^n\right).$$

Moreover, by the CCCP in Algorithm 1 which solves (13), we have

$$\widetilde{Q}\left(f^{(t)}, \left\{p_j^{(t)}\right\}_{j=1}^n\right) \geq \widetilde{Q}\left(f^{(t+1)}, \left\{p_j^{(t)}\right\}_{j=1}^n\right).$$

We, therefore, have

$$\begin{aligned} Q(f^{(t)}) = \widetilde{Q}\left(f^{(t)}, \left\{p_j^{(t)}\right\}_{j=1}^n\right) &\geq \widetilde{Q}\left(f^{(t+1)}, \left\{p_j^{(t)}\right\}_{j=1}^n\right) \geq \widetilde{Q}\left(f^{(t+1)}, \left\{p_j^{(t+1)}\right\}_{j=1}^n\right) \\ &= Q(f^{(t+1)}). \end{aligned}$$

Hence, the objective function $Q(f^{(t)})$ does not increase throughout the iterations in Algorithm 2. Furthermore, note that the objective function for the proposed method in (5) is nonnegative. In other words, its lower bound is 0. Then, we can conclude that the Algorithm 2 shall converge to a local optimum according to the Gauchys criterion for convergence [52]. □

4.2.3 Time complexity analysis

This section analyzes the time complexity of the proposed pairwise constrained MMC algorithm. For the initialization, it takes $O(nd^2)$ time to build $\mathbf{S}_c$ and $\mathbf{S}_m$, and the time complexity for solving the eigen problem is $O(d^2)$ by the eigs() function in Matlab. Therefore, the initialization by finding the projection vector generally takes $O(nd^2)$ time. In each iteration of the Algorithm 2, the computational cost of calculating $\{p_j^{(t)}\}_{j=1}^n$ by (14) is $O(nd)$. Algorithm 1 actually performs a sequence of SVM training with the linear kernel. Thanks to the cutting plane optimization techniques utilized by the off-the-shelf SVM solver SVM-Perf [53], the SVM training scales with time $O(d(m+2n))$. Hence, the time complexity of Algorithm 1 can be denoted as $O(d(m+2n))$. Eventually, the overall complexity of the proposed pairwise constrained binary MMC in Algorithm 2 is $O(nd^2) + O(\bar{T}d(m+2n))$, where $\bar{T}$ is the number of iterations for Algorithm 2 to converge.

4.3 Discussions

In our previous work [26], we have proposed a pairwise constrained multi-class MMC algorithm, whose possible drawback will be discussed when applied for the two-class clustering task in this section. Regarding to a two-class clustering task, [26] reduces to optimize the following problem:

$$\min_{\mathbf{w}_1,\mathbf{w}_2} \left\{ \frac{\beta}{2}\sum_{r=1}^{2}\|\mathbf{w}_r\|^2 + \sum_{i=1}^{m}\sum_{r=1}^{2} L\left(\max_{p\in\{1,2\}} \mathbf{w}_p^T\mathbf{x}_i - \mathbf{w}_r^T\mathbf{x}_i\right) \right.$$
$$\left. + \sum_{j=1}^{n} L\left(\ell_j\left[\max_{r_1=r_2}\left(\mathbf{w}_{r_1}^T\mathbf{x}_{j_1} + \mathbf{w}_{r_2}^T\mathbf{x}_{j_2}\right) - \max_{r_1\neq r_2}\left(\mathbf{w}_{r_1}^T\mathbf{x}_{j_1} + \mathbf{w}_{r_2}^T\mathbf{x}_{j_2}\right)\right]\right)\right\}, \tag{27}$$

where β is a positive constant, $r_1, r_2 \in \{1, 2\}$, and $\mathbf{w}_r$ is the weight vector for cluster r ($r \in \{1, 2\}$). The value $\mathbf{w}_r^T\mathbf{x}_i$ is the score for the sample $\mathbf{x}_i$ associated with the cluster r, and $\mathbf{w}_{r_1}^T\mathbf{x}_{j_1} + \mathbf{w}_{r_2}^T\mathbf{x}_{j_2}$ with $r_1 = r_2$ denotes the score for the pair of samples $\mathbf{x}_{j_1}, \mathbf{x}_{j_2}$ assigned to the same cluster. $\mathbf{w}_{r_1}^T\mathbf{x}_{j_1} + \mathbf{w}_{r_2}^T\mathbf{x}_{j_2}$ with $r_1 \neq r_2$ represents the score for the pair of samples $\mathbf{x}_{j_1}, \mathbf{x}_{j_2}$ assigned to different clusters. Optimizing such an objective function of the algorithm in [26] essentially requires that the score for assigning a sample to some cluster to which it is most likely to belong should be greater than the score for assigning it to other cluster by at least a margin, and the score for the most possible assigning scheme satisfying the specified pairwise constraint should be always greater than that for any assigning scheme which violates the constraint by at least a margin. It has been demonstrated that the algorithm in [26] features robust penalty for violating the pairwise constraints.

To solve the nonconvex optimization problem in (27), the CCCP is applied to decompose (27) into a series of convex problems [26]. In addition, according to [26], we have $\mathbf{w}_1 = -\mathbf{w}_2$ in such convex problems for the two-class clustering case. Therefore, the convex optimization problem in the t-th CCCP iteration can be rewritten as follows:

$$\min_{\mathbf{w}_1,\mathbf{w}_2} \left\{ \frac{\beta}{2}\sum_{r=1}^{2}\|\mathbf{w}_r\|^2 + \sum_{i=1}^{m} L\left(2\mathbf{w}_{p^*}^T\mathbf{x}_i\right) + \sum_{j=1}^{n} L\left(\ell_j\left[\mathbf{w}_{u^*}^T(\mathbf{x}_{j_1}+\mathbf{x}_{j_2}) - \mathbf{w}_{v^*}^T(\mathbf{x}_{j_1}-\mathbf{x}_{j_2})\right]\right)\right\}, \tag{28}$$

Table 1 The possible values for $\left[\mathbf{w}_{u^*}^T(\mathbf{x}_{j_1}+\mathbf{x}_{j_2})-\mathbf{w}_{v^*}^T(\mathbf{x}_{j_1}-\mathbf{x}_{j_2})\right]$

	$u^*=1$	$u^*=2$
$v^*=1$	$2\mathbf{w}_1^T\mathbf{x}_{j_2}$	$-2\mathbf{w}_1^T\mathbf{x}_{j_1}$
$v^*=2$	$2\mathbf{w}_1^T\mathbf{x}_{j_1}$	$-2\mathbf{w}_1^T\mathbf{x}_{j_2}$

where

$$\begin{aligned} q^* &= \arg\max_{r\in\{1,2\}} \mathbf{w}_r^{(t-1)T}\mathbf{x}_i, \\ u^* &= \arg\max_{r\in\{1,2\}} \mathbf{w}_r^{(t-1)T}(\mathbf{x}_{j_1}+\mathbf{x}_{j_2}), \\ v^* &= \arg\max_{r\in\{1,2\}} \mathbf{w}_r^{(t-1)T}(\mathbf{x}_{j_1}-\mathbf{x}_{j_2}), \end{aligned}$$

and $\mathbf{w}_r^{(t-1)}$ denotes the solution of $\mathbf{w}_r$ in the $(t-1)$-th iteration of CCCP. Considering $\mathbf{w}_1=-\mathbf{w}_2$, it is easy to verify that the item $\left[\mathbf{w}_{u^*}^T(\mathbf{x}_{j_1}+\mathbf{x}_{j_2})-\mathbf{w}_{v^*}^T(\mathbf{x}_{j_1}-\mathbf{x}_{j_2})\right]$ takes one of the four values listed in Table 1.

Then (28) can be reformulated as:

$$\min_{\mathbf{w}_1}\left\{\beta\|\mathbf{w}_1\|^2+\sum_{i=1}^m L\left(\widehat{\ell_i}2\mathbf{w}_1^T\mathbf{x}_i\right)+\sum_{j=1}^n L\left(\ell_j 2\mathbf{w}_1^T\tilde{\mathbf{x}}_j\right)\right\}, \tag{29}$$

where

$$\widehat{\ell_i} \triangleq \begin{cases} 1 & \text{if } q^*=1; \\ -1 & \text{if } q^*=2, \end{cases}$$

and $\tilde{\mathbf{x}}_j$ can be one of the samples from the set $\{\mathbf{x}_{j_1},\mathbf{x}_{j_2},-\mathbf{x}_{j_1},-\mathbf{x}_{j_2}\}$ depending on $\mathbf{w}_r^{(t-1)T}$ (see Table 1). Subsequently, let $\mathbf{w}=2\mathbf{w}_1$, (29) is equivalent to:

$$\min_{\mathbf{w}}\left\{\frac{\beta}{2}\|\mathbf{w}\|^2+2\sum_{i=1}^m L\left(\widehat{\ell_i}\mathbf{w}^T\mathbf{x}_i\right)+2\sum_{j=1}^n L\left(\ell_j\mathbf{w}^T\tilde{\mathbf{x}}_j\right)\right\}. \tag{30}$$

Hence, in essence, for the case of two-class pairwise constrained MMC, optimizing the problem in (30) requires that the sample $\mathbf{x}_i$ $(i=1,\ldots,m)$ and one of the samples from $\{\mathbf{x}_{j_1},\mathbf{x}_{j_2},-\mathbf{x}_{j_1},-\mathbf{x}_{j_2}\}$ $(j=1,\ldots,n)$ should be as far away from the hyperplane as possible.

In [26], in order to find a good initial model for performing the CCCP, it utilizes only the more reliable supervisory information (i.e., the pairwise constraints) other than the unconstrained samples in the first few iterations of CCCP, that is, it solves the following problem for initializing $\widehat{\ell_i}$'s $(i=1,\ldots,m)$ in (30):

$$\min_{\mathbf{w}}\left\{\frac{\beta}{2}\|\mathbf{w}\|^2+2\sum_{j=1}^n L(\ell_j\mathbf{w}^T\tilde{\mathbf{x}}_j)\right\}, \tag{31}$$

which is a binary SVM problem but without the bias term b, and ℓ_j's $(j=1,\ldots,n)$ can be viewed as the binary labels for $\tilde{\mathbf{x}}_j$'s $(j=1,\ldots,n)$. However, it is known that the bias term plays a crucial role when the distribution of the labels is uneven [54–57]. Thereby, in a common case, for example, when the positive ℓ_j's greatly outnumber the negative ones, it may result in an improper hyperplane by the unbiased SVM in (31). Consequently, the so-obtained hyperplane may not be a good initialization for the problem (30). Furthermore, since the original optimization problem of [26] in (27) is nonconvex and CCCP only converges to a local minimum [49], such an initialization may lead to poor solutions eventually. In other words, for the two-class pairwise constrained clustering, when the number of must-link constraints greatly deviates from the number of cannot-link ones, the performance of our previous work in [26] may degrade.

To this end, this paper particularly has designed a new pairwise constrained two-class MMC method. The proposed one and our previous work [26] when applied in two-class clustering differ from each other in two aspects. One is that they use different pairwise loss functions, which lead to different objective functions in the CCCP iterations. Specifically, for a pair of constrained samples, there is only one sample or its negative involved in the objective function in (20), while both of the two samples from the constraint are included in (30). The other is that the proposed new method does include the bias term b in the hyperplane model. Therefore, for the case where the number of must-link constraints and the one of cannot-link constraints are imbalanced, the proposed method is expected to be more robust than our previous work [26] which is, however, particularly developed for handling the multi-class clustering tasks.

5 Experiments

In this section, we evaluated the accuracy and efficiency of our proposed algorithm on a couple of real-world binary-class data sets. Moreover, the generalization capability of the proposed approach, that is, the performance on the out-of-sample data points, was investigated as well. All the experiments were conducted with Matlab 7.0 on a 2.62GHz Intel Dual-Core PC running Windows 7 with 2GB main memory.

5.1 Evaluation metrics

To evaluate the effectiveness of clustering algorithms, we used the clustering accuracy (ACC) and the normalized mutual information (NMI) in this paper. The efficiency of each clustering algorithm was assessed by the central processing unit (CPU) time in seconds during the execution.

Following the strategy in [11,25,41], we firstly take a set of labeled data, remove the labels and run the clustering algorithms. Then, we label each of the resulting clusters with the majority class according to the original training labels. Finally, the clustering accuracy (ACC) is defined as the matching degree between the obtained labels and the original true labels [11,25,41]:

$$ACC = \frac{\sum_{i=1}^{N} I(\hat{t}_i = t_i)}{N}, \tag{32}$$

where N is the number of samples in the data set, $\hat{t}_i$ is the label obtained by the above steps for $\mathbf{x}_i$, and t_i is the ground-truth label. The *NMI* is defined by [58]:

$$NMI = \frac{\sum_{i=1}^{C}\sum_{j=1}^{C} n_{i,j} \log\left(\frac{N\cdot n_{i,j}}{n_i \tilde{n}_j}\right)}{\sqrt{\left(\sum_{i=1}^{C} n_i \log\frac{n_i}{N}\right)\left(\sum_{j=1}^{C} \tilde{n}_j \log\frac{\tilde{n}_j}{N}\right)}} \tag{33}$$

where n_i denotes the number of data contained in the cluster $\mathcal{C}_i (1 \le i \le C)$, $\tilde{n}_j$ is the number of data belonging to the ground-truth class $\mathcal{C}_j^{'}$ $(1 \le j \le C)$, and $n_{i,j}$ denotes the number of data that is in the intersection between the cluster $\mathcal{C}_i$ and the class $\mathcal{C}_j^{'}$. The *NMI* ranges from 0 to 1. The larger the value it is, the more similar the groupings by clustering and those by the true class labels.

5.2 Competing algorithms

We compared the proposed algorithm, denoted as **TwoClaCMMC**, with the following competing clustering algorithms.

1) **K-means** [59]: This is the traditional unsupervised K-means algorithm without any pairwise constraints incorporated, and it served as the baseline.
2) **DCA+K-means** [15]: The discriminative component analysis (DCA) firstly learns a distance metric based on the pairwise constraints.[3] Then, K-means is performed with this metric. This method is a typical *metric-based* one.
3) **AffPropag** [20]: AffPropag[4] is a representative pairwise constrained spectral clustering algorithm, which belongs to the *similarity-based* methods.
4) **CPCMMC** [25]: This is a *penalty-based* approach, and it is a preliminary work on pairwise constrained two-class maximum margin clustering.
5) **MultiClaCMMC** [26]: This is our previous approach, which is also a *penalty-based* one. It is proposed in order to overcome the drawbacks of multi-class clustering algorithm in [25].

5.3 Experiments for evaluating the effectiveness and the efficiency

This set of experiments compare the effectiveness and the efficiency of TwoClaCMMC and its competing algorithms.

5.3.1 Evaluation data sets

We used four categories of data sets in this set of experiments for evaluating the effectiveness and the efficiency of all the algorithms. Specifically, those data sets are as follows:

- **UCI data**: `live-disorder, pima, sonar` are from the UCI repository.[5]
- **Image data**: `COIL2`[6] is a subset from the Columbia object image library (COIL-100), including color images of objects taken from different angles. It is known that there are subgroups within a single semantic class for this data set [60].
- **Gene expression data**: We performed experiments on two gene expression data sets, and they have high dimensionalities, with many features (i.e., genes) in common for samples from different classes. `colon cancer` data set consists of samples from two classes of patients: the normal healthy ones and the tumor ones [61]. We preprocessed the raw data by carrying out a base 10 logarithmic transformation. `leukemia` data set contains samples from two classes of leukemia: the acute lymphoblastic leukemia (ALL) and the acute myeloid leukemia (AML). It has been reported that there are subgroups in the class ALL [62]. The raw leukemia data set was preprocessed following the protocol in [63].

[3] http://www.cs.cmu.edu/~liuy/dca.zip.

[4] http://faculty.ucmerced.edu/mcarreira-perpinan/research/software/AffProp.zip.

[5] http://www.ics.uci.edu/~mlearn/MLRepository.html.

[6] http://olivier.chapelle.cc/ssl-book/benchmarks.html.

Table 2 The data sets used in experiments for evaluating the effectiveness and the efficiency

Dataset	#Dimension	#Sample (#+/#-)
`live-disorder`	6	345 (145/200)
`pima`	8	768 (268/500)
`sonar`	60	208 (97/111)
`COIL2`	241	1,500 (750/750)
`coloncancer`	2,000	62 (22/40)
`leukemia`	3,567	72 (47/25)
`20NewsGroup-similar2`	1,864	197 (98/99)
`20NewsGroup-related2`	3,225	190 (93/97)
`WebKBPage`	3,000	1,051 (230/820)

- **Text data**: `20NewsGroup-Similar2, 20NewsGroup-Related2` are from the 20 newsgroup document databases,[7] and `WebKBPage` is from [64]. `20NewsGroup-Similar2` data set consists of 2 newsgroups on similar topics (comp.os.ms-windows, comp.windows.x) with significant cluster overlap. `20NewsGroup-Related2` contains articles from two related topics (talk.politics.misc, talk.politics.guns). The `WebKB Page` data set is a subset of web documents of the computer science departments of four universities. The two categories are course or noncourse. Each document is represented by the textual content of the webpage. It is known that the articles from these two classes often have a lot of words in common [64].

In a word, most of the data sets used in this set of experiment may feature the multi-modal data distribution within a single semantic class and severe overlap among samples from different classes, thus are quite appropriate for evaluating the robustness of the pairwise constrained clustering algorithms. A summary of all the binary-class data sets used in this set of experiments is shown in Table 2. Note that the values in the parentheses denote sample numbers of the two classes.

5.3.2 Evaluation settings

For our method (TwoClaCMMC) and CPCMMC, γ was set to be equal to λ; thus, they both have only one free parameter. Then, λ was searched from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}\}$, and the best result was reported on each data set for TwoClaCMMC and CPCMMC. The only one free parameter β in MultiClaCMMC was set by searching from the grid $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}\}$. For AffPropag, to build the basic similarity matrix, the width of the Gaussian kernel was set by exhaustive search from the grid $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$. The K-means and DCA+K-means do not have free parameters.

Moreover, all the data sets were preprocessed so that their features have zero mean and unit standard deviation, except the three text data sets whose features are already sparse enough. In the experiments, we let the number of clusters be the true number of classes (i.e., 2) in each data set for all the algorithms, although the selection of the number of clusters is a crucial issue, which is, however, beyond the scope of this paper. For each data set, we evaluated the performance with different numbers of pairwise constraints. As in [20,25,26],

[7] http://people.csail.mit.edu/jrennie/20Newsgroups/.

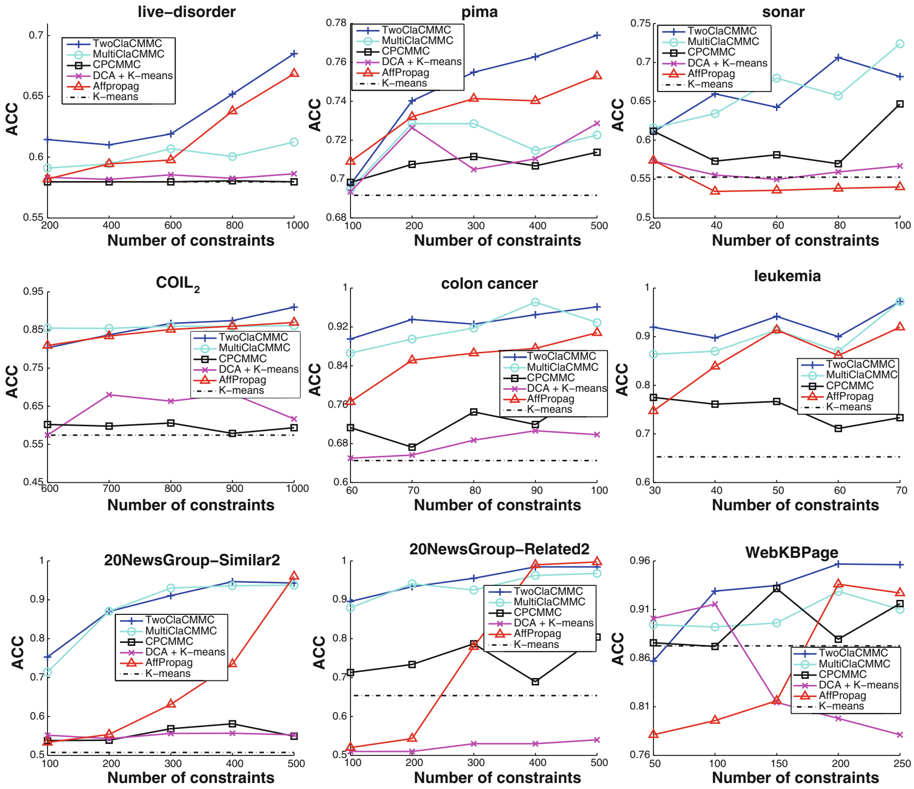


Fig. 3 Comparison of clustering accuracy over the different number of pairwise constraints

each constraint was generated by randomly selecting a pair of samples. If the samples belong to the same class, a must-link constraint was formed. Otherwise, a cannot-link constraint was formed. Then, the same constraints are applied to different algorithms. For a fixed number of pairwise constraints, the results were averaged over 10 realizations of different pairwise constraints.

5.3.3 Evaluation results on the effectiveness and the efficiency

To compare the clustering effectiveness, the *ACC* and *NMI* curves for the clustering results of all the algorithms are shown in Figs. 3 and 4.[8]

It can be observed that TwoClaCMMC indeed can improve the performance of the baseline K-means, even though the maximum number of constraints provided for each data set in our experiment is only an extremely small fraction of the total number of constraints that can be generated (e.g., 0.12 % for the `COIL2` data set for which only up to 1,000 pairwise constraints are provided, and 0.04 % for the `WebKBPage` data set for which up to 250 pairwise constraints are provided). Therefore, such a result indicates that the proposed method may be competent in the practical scenarios where the nonexpert users generally label only a small portion of sample pairs for guiding the clustering.

[8] Since DCA+K-means has memory overflow problem on the `leukemia` data set whose dimensionality is high, we do not include it for comparison on this data set.

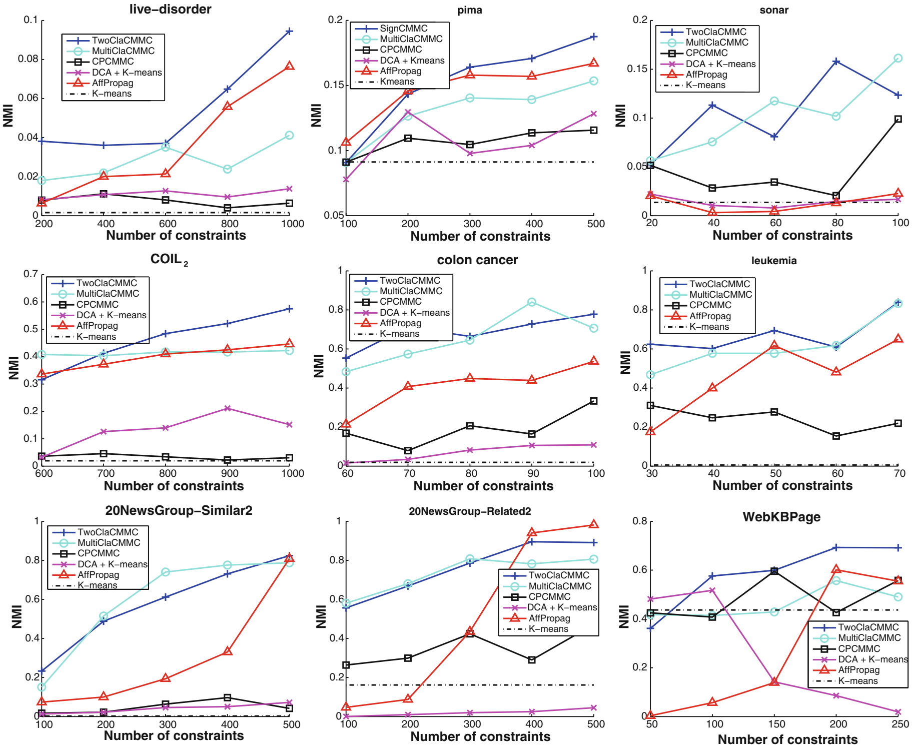


Fig. 4 Comparison of normalized mutual information over the different number of pairwise constraints

In contrast, the *ACC*/*NMI* curves for DCA+K-means often show a slow rise on some data sets with the increase in the amounts of pairwise constraints, or even fall behind those for the baseline. On the one hand, such phenomena may be due to the fact that the unconstrained samples are not involved in the DCA metric learning algorithm. As a result, useful supervision information contained in the insufficient pairwise constraints may not be effectively propagated to the whole data set. On the other hand, the degrading performance of DCA+K-means suggests that the ground-truth metrics for the data used in this set of experiments might be considerably complex in general. Furthermore, one can also observe that the clustering performance of AffPropag is always inferior to that of the proposed algorithm. These observations corroborate the advantage of the proposed discriminative approach, which directly searches the possibly much simpler partitioning boundary and thus makes fewer assumptions about the possibly complicated underlying metric or similarity for the data.

It is also observed that the improvement made by CPCMMC on most data sets is often less significant than those by TwoClaCMMC and MultiClaCMMC. This phenomenon may be due to the reason that there is generally severe overlap in most used data sets and the ground-truth categories may contain subgroups as well, but the loss function for violating the cannot-link constraints in CPCMMC [25] may not be able to effectively discourage the violation of constraints in such scenarios, and the pairwise loss function for the must-link constraints may produce unexpected high penalty even when the must-link constraints are not violated.

Besides, from Figs. 3 and 4, one can observe that although MultiClaCMMC is comparable to TwoClaCMMC on the `sonar`, `COIL2`, `20NewsGroup-Similar2`,

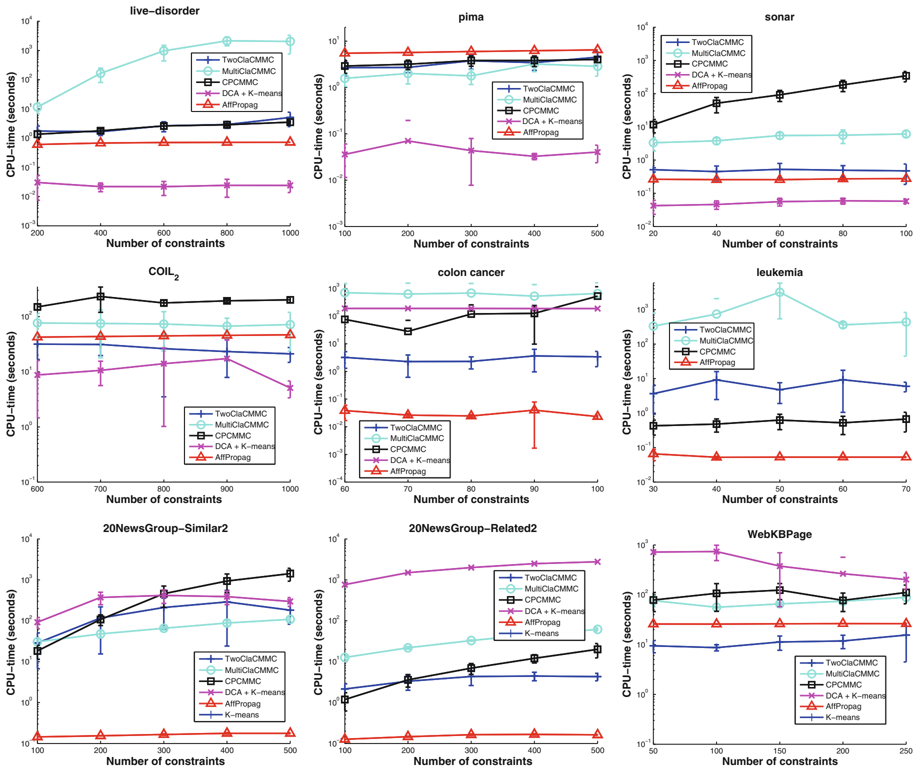


Fig. 5 Comparison of execution time for the pairwise constrained clustering algorithms

20NewsGroup-Related2 data sets, its performance is inferior to that of the newly proposed TwoClaCMMC on the remaining data sets, especially when the number of pairwise constraints is small. Recall in our experiment setups a pairwise constraint was generated as follows: randomly selecting a pair of samples and then forming a must-link or cannot-link constraint based on whether the ground-truth labels of such a pair of samples are the same or not. From Table 2, it can be seen that the numbers of positive and negative samples are imbalanced in the live-disorder, pima, colon cancer, leukemia and WebKBPage data sets. Then, for such data sets, according to the permutation and combination theories, such a scheme for producing the pairwise constraints will be prone to result in significantly less cannot-link constraints than must-link constraints (e.g., the average number of must-link constraints is about two times of that of cannot-link constraints, when 30 and 100 constraints were generated for the leukemia and WebKBPage data set, respectively). Then, as we have discussed in Sect. 4.3, the performance of MultiClaCMMC is likely to degrade under such circumstances.

To compare the efficiency, the average and standard deviation of the CPU time consumption for the 10 trials of each constrained clustering algorithm are demonstrated in Fig. 5. It can be seen that the proposed method TwoClaCMMC is generally time efficient. For instance, in this set of experiments, it takes about 30 s to partition the COIL2 data set with the largest number of sample size, and around 10 s to handle the leukemia data set with the largest dimensionality. Although it is not always the fastest algorithm on all the data sets, it is much

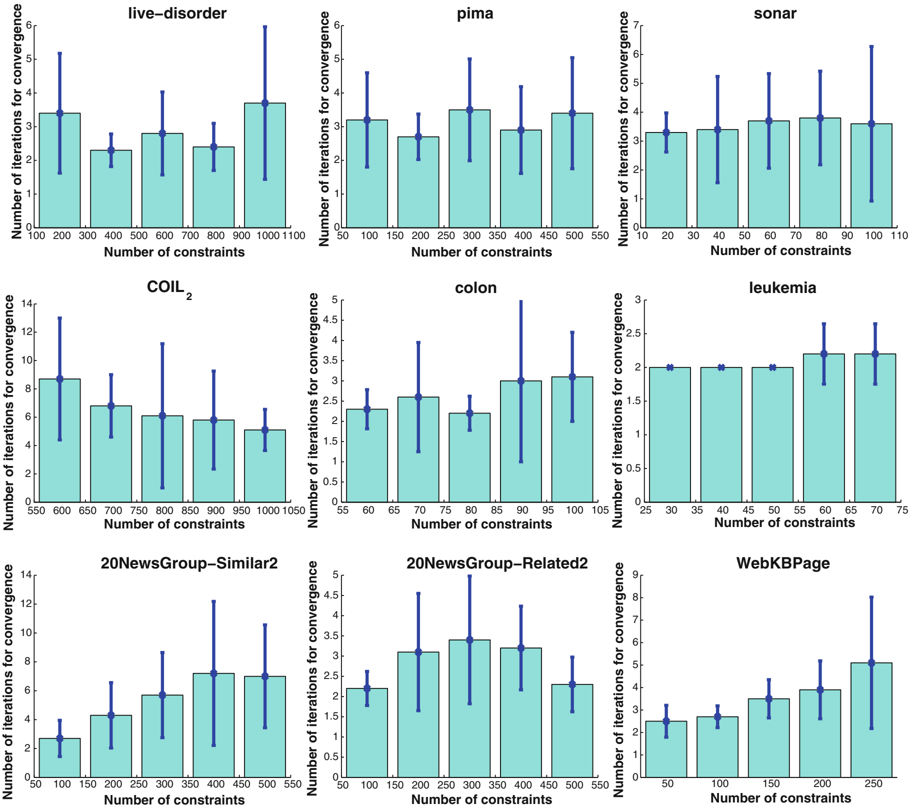


Fig. 6 Number of iterations for convergence of the proposed algorithm

more advantageous than its counterparts in terms of the accuracy. Moreover, the average and standard deviation of the number of iterations for the proposed approach to converge are shown in Fig. 6. As can be seen from Fig. 6, the proposed method usually converges fast, and the number of iterations for the proposed approach to converge never exceeded 20 on all the data sets tried so far in this set of experiments.

5.4 Experiments for evaluating the generalization performance

In practice, it is often desirable for a clustering algorithm to use the partitioning model or metric, which is learned with a small amount of currently available data, for determining the cluster memberships for newly observed samples. Such a merit is particularly useful for handling the large-scale data, because many clustering algorithms cannot be run on the whole large data sets due to memory problem. The proposed constrained clustering method TwoClaCMMC is essentially based on SVM model, which can achieve small generation error on new data; thus, it is expected to naturally inherit such a merit. In this subsection, a set of experiments compares the generalization performance for some competing algorithms.

5.4.1 Evaluation data sets

We performed experiments on three data sets from UCI repository. Two confusing classes ("I" vs. "J", "3" vs. "8") were chosen from the letter and pendigit databases to construct the

Table 3 The data sets used in experiments for evaluating the generalization performance

Dataset	#Dimension	#Sample
letterIJ	16	1,502
pendigit38	16	2,110
magic	10	19,020

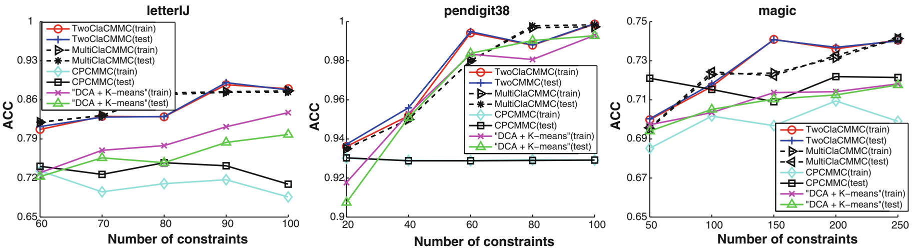


Fig. 7 Comparison of generalization performance over the different number of pairwise constraints in the 10-fold runs

letterIJ data set and pendigit38 data set, respectively. In addition, magic was used, which is a large-scale data set. The basic information about those data sets is summarized in Table 3.

5.4.2 Evaluation settings

For each data set, it was first divided into 10 nonoverlapping folds. Then, we sequentially picked onefold of samples as the out-of-sample points for testing the generalization performance, and the samples in the remaining folds were used as the training set to learn the partitioning model or metric.

With regard to TwoClaCMMC, MultiClaCMMC and CPCMMC, we used the discriminative model learned on the training data to directly cluster the testing data. For DCA+K-means, the metric obtained on the training data was used to perform K-means clustering on the testing data. Note that AffPropag does not provide such a generalization ability; even worse, the large-scale data set magic cannot be run by AffPropag on our machine due to the out-of-memory problem.

5.4.3 Evaluation results on the generalization performance

Figure 7 shows the average results on the previously available data (marked by "train") and the out-of-sample data (marked by "test") in the 10-fold runs. As can been seen from Fig. 7, TwoClaCMMC and MultiClaCMMC again outperform CPCMMC as well as DCA+K-means on the out-of-sample data points. Moreover, for TwoClaCMMC and MultiClaCMMC, the clustering performance on out-of-sample examples is comparable with that on the training subset. In other words, for handling large-scale data sets from a practical viewpoint, we may firstly perform TwoClaCMMC on a small subset and then use the learned model to cluster the remaining points.

6 Conclusion

We have presented a discriminative pairwise constrained maximum margin clustering approach. In particular, a new set of pairwise loss functions featuring robust detection and penalization for the violation to the constraints have been proposed. In this way, without explicitly estimating the possibly complex underlying distance metric or similarity matrix, the proposed method directly models the partitioning boundary that separates the data into two groups and fulfills the pairwise constraints as much as possible. An iterative updating algorithm has been developed for the resulting optimization problem. Extensive experiments on real-world data sets demonstrate that the proposed pairwise constrained two-class clustering algorithm could considerably improve the baseline unsupervised clustering. It also outperforms the state-of-the-art pairwise constrained clustering counterparts with the same number of constraints. In addition, the experimental results show that the proposed method generalizes well on the out-of-sample data points.

There are several potential directions for future research. Firstly, we mainly focus on developing a pairwise constrained two-class maximum margin clustering algorithm in this paper; however, it is more common to perform multi-class clustering tasks in real applications. We plan to derive the multi-class extension of the proposed method in the future research. Secondly, the time complexity of the initialization procedure for the proposed method is $O(nd^2)$, which may be time consuming for very high-dimensional data. More efficient initialization scheme should be developed. Thirdly, it would be also interesting to study how to actively identify the most informative pairwise constraints during the clustering. Lastly, instead of prespecifying the constraints in batch, we will explore the situation where the constraints are incrementally given. We plan to consider the problem of efficiently updating a clustering to satisfy the new and old constraints rather than re-clustering the entire data set.

Acknowledgments The authors would like to thank the anonymous reviewers for their valuable comments and helpful suggestions. This work was supported by the National Natural Science Foundation of China (No. 61105048, 60972165, 61104206, 51175080), the Doctoral Fund of Ministry of Education of China (No. 20100092120012, 20110092120034), the Natural Science Foundation of Jiangsu Province (No. BK2010240, BK2010423), the Technology Foundation for Selected Overseas Chinese Scholar, Ministry of Human Resources and Social Security of China (No. 6722000008), and the Open Fund of Jiangsu Province Key Laboratory for Remote Measuring and Control (No. YCCK201005).

References

1. Peng W, Li T (2011) Temporal relation co-clustering on directional social network and author-topic evolution. Knowl Inf Syst 26:467–486
2. Tang M, Zhou Y, Li J, Wang W et al (2011) Exploring the wild birds imigration data for the disease spread study of H5N1: a clustering and association approach. Knowl Inf Syst 27:227–251
3. Baralis E, Bruno G, Fiori A (2011) Measuring gene similarity by means of the classification distance. Knowl Inf Syst 29:81–101
4. Zhao W, He Q, Ma H, Shi Z (2011) Effective semi-supervised document clustering via active learning with instance-level constraints. Knowl Inf Syst 30:569–587
5. Kalogeratos A, Likas A (2011) Text document clustering using global term context vectors. Knowl Inf Syst. doi:10.1007/s10115-011-0412-6
6. Li Z, Liu J (2009) Constrained clustering by spectral kernel learning. Proceedings of the 12th IEEE international conference on computer vision, pp 421–427
7. Basu S, Davidson I, Wagstaff K (2008) Constrained clustering: advances in algorithms, applications and theory. CRC Press, Boca Raton
8. Wagstaff K, Cardie C, Schroedl S (2001) Constrained k-means clustering with background knowledge. Proceedings of the 18th international conference on, machine learning, pp 577–584

9. Kulis B, Basu S, Dhillon I, Mooney R (2005) Semi-supervised graph glustering: a kernel approach. Proceedings of the 22th international conference on, machine learning, pp 457–464
10. Yan R, Zhang J, Yang J, Hauptmann A (2006) A discriminative learning framework with pairwise constraints for video object classification. IEEE Trans Pattern Anal Mach Intell 28(4):578–593
11. Domeniconi C, Peng J, Yan B (2011) Composite kernels for semi-supervised clustering. Knowl Inf Syst 28:99–116
12. Wang F, Li P, König AC, Wan M (2011) Improving clustering by learning a bi-stochastic data similarity matrix. Knowl Inf Syst. doi:10.1007/s10115-011-0433-1
13. Xing EP, Ng AY, Jordan MI, Russell S (2003) Distance metric learning with application to clustering with side-information. Adv Neural Inf Process Syst 15:521–528
14. Bar-Hillel A, Hertz T, Shental N, Weinshall D (2003) Learning distance functions using equivalence relations. Proceedings of the 20th international conference on, machine learning, pp 11–18
15. Hoi SCH, Liu W, Lyu MR, Ma WY (2006) Learning distance metrics with contextual constraints for image retrieval. Proceedings of the 9th international conference on computer vision and, pattern recognition, pp 2072–2078
16. Kamvar SD, Klein D, Manning C (2003) Spectral learning. Proceedings of the 18th international joint conference on, artificial intelligence, pp 561–566
17. Davis JV, Kulis B, Jain P, Sra S, Dhillon IS (2007) Information-theoretic metric learning. Proceedings of the 24th international conference on, machine learning, pp 209–216
18. Li ZG, Liu J, Tang X (2008) Pairwise constraint propagation by semidefinite programming for semi-supervised classification. Proceedings of the 25th international conference on, machine learning, pp 576–583
19. Hoi SCH, Jin R, Lyu MR (2007) Learning nonparametric kernel matrices from pairwise constraints. Proceedings of the 24th international conference on, machine learning, pp 361–368
20. Lu Z, Carreira-Perpinan MA (2008) Constrained spectral clustering through affinity propagation. Proceedings of the 11th IEEE international conference on computer vision and, pattern recognition, pp 1–8
21. Bilenko M, Basu S, Mooney RJ (2004) Integrating constraints and metric learning in semi-supervised clustering. Proceedings of the 21st international conference on, machine learning, pp 81–88
22. Wu L, Jin R, Hoi SCH, Zhu J, Yu N (2009) Learning bregman distance functions and its application for semi-supervised clustering. Adv Neural Inf Process Syst 22:2089–2097
23. Xu L, Neufeld J, Larson B, Schuurmans D (2005) Maximum margin clustering. Adv Neural Inf Process Syst 17:1537–1544
24. Collobert R, Sinz F, Weston J, Bottou L (2006) Large scale transductive svms. J Mach Learn Res 7:1687–1712
25. Hu Y, Wang J, Yu N, Hua XS (2008) Maximum margin clustering with pairwise constraints. Proceedings of the 8th IEEE international conference on data mining, pp 253–262
26. Zeng H, Cheung YM (2012) Semi-supervised maxmum margin clustering with pairwise constraints. IEEE Trans Knowl Data Eng 24(5):926–939
27. Chen Y, Rege M, Dong M, Hua J (2007) Incorporating user provided constraints into document clustering. Proceedings of the 7th IEEE international conference on data mining, pp 103–112
28. Wang F, Li T, Zhang CS (2008) Semi-supervised clustering via matrix factorization. Proceedings of the 8th SIAM international conference on data mining, pp 1–12
29. Li T, Ding C, Jordan MI (2007) Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. Proceedings of the 7th IEEE international conference on data mining, pp 577–582
30. Chen Y, Rege M, Dong M, Hua J (2008) Non-negative matrix factorization for semi-supervised data clustering. Knowl Inf Syst 17:355–379
31. Hoi SCH, Liu W, Chang SF (2008) Semi-supervised distance metric learning for collaborative image retrieval. Proceedings of the 11th IEEE international conference on computer vision and, pattern recognition, pp 1–7
32. Zhang DQ, Zhou ZH, Chen SC (2007) Semi-supervised dimensionality reduction. Proceedings of the 7th SIAM international conference on data mining, pp 629–634
33. Nguyen N, Caruana R (2008) Improving classification with pairwise constraints: a margin-based approach. Proceedings of the 19th European conference on machine learning and knowledge discovery in databases, pp 113–124
34. Goldberg A, Zhu X, Wright S (2007) Dissimilarity in graph-based semi-supervised classification. Proceedings of the 12th international conference on artificial intelligence and, statistics, pp 155–162
35. Tong W, Jin R (2007) Semi-supervised learning by mixed label propagation. Proceedings of the 22nd national conference on, artificial intelligence, pp 651–656
36. Zhang C, Cai Q, Song Y (2010) Boosting with pairwise constraints. Neurocomputing 73(4–6):908–919

37. Xu L, Schuurmans D (2005) Unsupervised and semi-supervised multi-class support vector machines. Proceedings of the 20th national conference on, artificial intelligence, pp 904–910
38. Zhang K, Tsang IW, Kwok JT (2009) Maximum margin clustering made practical. IEEE Trans Neural Netw 20(4):583–596
39. Valizadegan H, Jin R (2007) Generalized maximum margin clustering and unsupervised kernel learning. Adv Neural Inf Process Syst 19:1417–1424
40. Zhang K, Tsang IW, Kwok JT (2007) Maximum margin clustering made practical. Proceedings of the 24th international conference on, machine learning, pp 1119–1126
41. Zhao B, Wang F, Zhang C (2008) Efficient multiclass maximum margin clustering. Proceedings of the 25th international conference on, machine learning, pp 1248–1255
42. Li YF, Tsang IW, Kwok JT, Zhou ZH (2009) Tighter and convex maximum margin clustering. Proceedings of the 12th international conference on artificial intelligence and, statistics, pp 344–351
43. Wang F, Zhao B, Zhang C (2010) Linear time maximum margin clustering. IEEE Trans Neural Netw 21(2):319–332
44. Gu Q, Zhou J (2009) Subspace maximum margin clustering. Proceedings of the 18th ACM conference on information and, knowledge management, pp 1337–1346
45. Zhao B, Kwok J, Wang F, Zhang C (2009) Unsupervised maximum margin feature selection with manifold regularization. Proceedings of the 12th IEEE conference on computer vision and, pattern recognition, pp 888–895
46. Zhao B, Kwok JT, Zhang C (2009) Multiple kernel clustering. Proceedings of the 9th SIAM international conference on data mining, pp 638–649
47. Shen RL, Olshen AB, Ladanyi M (2010) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics 26:292–293
48. Yuille AL, Rangarajan A (2003) The concave-convex procedure. Neural Comput 15(4):915–936
49. Smola AJ, Vishwanathan SVN, Hofmann T (2005) Kernel methods for missing variables. Proceedings of the 20th international workshop on artificial intelligence and, statistics, pp 325–332
50. Collobert R, Sinz F, Weston J et al (2006) Large scale transductive SVMs. J Mach Learn Res 7:1687–1712
51. Bonnans JF, Gilbert JC, Lemaréchal C et al (2003) Numerical optimization. Springer, Berlin, Germany
52. Rudin W (1978) Principles of mathematical analysis, 3rd edn. McGray-Hill, New York
53. Joachims T (2006) Training linear SVMs in linear time. Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pp 217–226
54. Li Y, Bontcheva K, Cunningham H (2009) Adapting svm for data sparseness and imbalance: a case study in information extraction. Nat Lang Eng 15:241–271
55. He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284
56. Shalev-Shwartz S, Singer Y, Srebro N (2007) Pegasos: primal estimated sub-gradient solver for SVM. Proceedings of the 24th international conference on, machine learning, pp 807–814
57. Núñez Castro H, González Abril L, Angulo Bahón C (2011) A post-processing strategy for SVM learning from unbalanced data. Proceedings of the 15th European symposium on artificial, neural networks, pp 195–200
58. Strehl A, Ghosh J (2003) Cluster ensembles-a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 3:583–617
59. Duda RO, Hart PE, Stork DG (2001) Pattern classification. Wiley, New York
60. Chapelle O, Schölkopf B, Zien A (2006) Semi-supervised learning. MIT Press, Cambridge, MA
61. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci 96:6745–6750
62. Golub TR, Slonim DK, Tamayo P et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–537
63. Dudoit S, Fridlyand J, Speed TP (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 97:77–87
64. Sindhwani V, Niyogi P, Belkin M (2005) Beyond the point cloud: from transductive to semi-supervised learning. Proceedings of the 22nd international conference on, machine learning, pp 824–831

Author Biographies

Hong Zeng received the Ph.D. degree in Computer Science from Hong Kong Baptist University, in 2010. He is currently with the Robot Sensor and Control Laboratory (RSCL) in the School of Instrument Science and Engineering, Southeast University, China. His research interests are in the areas of signal processing, pattern recognition, machine learning, data mining and their applications for bioelectrical signals.

Aiguo Song received the Ph.D. degree in Measurement and Control from Southeast University, Nanjing, P. R. China in 1996. From 1996 to 1998, he was an Associate Researcher with the Intelligent Information Processing Laboratory, Southeast University, P. R. China. From 1998 to 2000, he was an associate Professor with the Department of Instrument Science and Engineering, Southeast University, P. R. China. From 2000 to 2003, he was the Director of the Robot Sensor and Control Lab, Southeast University, P. R. China. From April, 2003 to April, 2004, he was a visiting scientist in the Lab for Intelligent Mechanical Systems (LIMS), Northwestern University, Evanston, USA. He is currently a Professor in the School of Instrument Science and Engineering, Southeast University, P. R. China. His current interests concentrate on teleoperation, haptic display, rehabilitation robot, signal processing.

Yiu-ming Cheung received the Ph.D. degree at Department of Computer Science and Engineering from the Chinese University of Hong Kong in 2000. Currently, he is a full professor at the Department of Computer Science in Hong Kong Baptist University. His research interests include machine learning, information security, signal processing, pattern recognition and data mining. He is the founding chair of Computational Intelligence Chapter of IEEE Hong Kong Section. Also, he is a senior member of IEEE and ACM. More details can be found at: http://www.comp.hkbu.edu.hk/~ymc.