IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 6, 2019, accepted May 2, 2019, date of publication May 14, 2019, date of current version May 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2916724

Improved Maximum Margin Clustering via the Bundle Method

JIANQIANG LI[1,2], (Senior Member, IEEE), JINGCHAO SUN[1], LU LIU[1], BO LIU[1], (Senior Member, IEEE), CAO XIAO[3], AND FEI WANG[4], (Senior Member, IEEE)

[1]School of Software Engineering, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[2]Beijing Engineering Research Center for IoT Software and Systems, Beijing 100124, China
[3]Center for Computational Health, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA
[4]Department of Healthcare Policy and Research, Cornell University, New York, NY 14853, USA

Corresponding author: Bo Liu (boliu@bjut.edu.cn)

This work was supported by the National Key R&D Program of China under Project 2017YFB1400105.

ABSTRACT Maximum margin clustering (MMC) is an effective clustering algorithm, which first extends a large margin principle into unsupervised learning. This paper revisits the MMC problem and points out the potential problems encountered by a cutting plane approach. We propose an improved MMC algorithm via the bundle method (BMMC). Specifically, the constrained convex–concave procedure algorithm is first applied to decompose the MMC problem into a series of convex sub-problems, and then, the bundle method is adopted to efficiently solve each sub-problem. Moreover, a simpler formulation for the multi-class MMC is presented. In addition to clustering problems, the BMMC is also extended to the semi-supervised case by incorporating the pairwise constraints, which reveals its high scalability. Compared with the previous works, the proposed solution is much simpler and faster. The experiments on several data sets are conducted to demonstrate the effectiveness of our proposed algorithm.

INDEX TERMS Bundle method, constrained convex-concave procedure, maximum margin clustering, unsupervised learning, semi-supervised learning.

I. INTRODUCTION

Clustering algorithms are often used to analyze unlabeled data. They divide data into multiple groups based on a certain optimization objective. The data items in the same group are more similar compared with those in other groups. Therefore, clustering analysis works as an effective tool of knowledge discovery in many practical problems [1]–[9], such as medical data analysis and feature extraction. Many classical clustering algorithms have been proposed, such as K-means and Spectral Clustering (SC) [10]. Motivated by the theory of support vector machine, *Maximum Margin Clustering* (*MMC*) was proposed for clustering analysis [11]. Therefore, like SVM, the main principle behind MMC is large margin principle (LMP) [12]. MMC is the first algorithm that extends LMP into unsupervised learning [13]. Different from SVM, MMC partitions unlabeled data into multiple clusters by maximizing the minimum margin in the data. It implies LMP works directly on the data in MMC, which leads to a good insight into the internal structure of the data. In supervised learning, LMP based methods are trained to find the maximum-margin hyperplane in the training data. The trained models are used to classify testing datasets and the objective is to obtain lower generalization error on the testing datasets. Consequently, roughly speaking, the application of LMP in MMC is similar to using the trained models to classify training data; and the error of training models on training data is very low. Therefore, referring to the success of SVM, MMC should have good performance in clustering tasks. Its experimental performance has been demonstrated in many researches [11], [14]–[18], which was superior than conventional clustering algorithms.

Recently, Zeng and Cheung [19] presented a pairwise constrained algorithm based on MMC. Wang and Chen [20] proposed a Soft Large Margin Clustering (SMLC), which combined advantages of MMC and the soft clustering methods. Niu *et al.* [13] proposed an maximum volume clustering (MVC) method based on large volume

The associate editor coordinating the review of this manuscript and approving it for publication was Kuo-Ching Ying.

principle (LVP) [12]. LVP is an alternative strategy for hyperplanes, which trends to certain large-volume equivalence classes [13]. MVC is presented as a binary clustering method, in which the best clustering is the partition lying in the equivalence class with the maximum volume. Zhu *et al.* [21] used a multiclass clustering algorithm based on basis of MMC and immune evolutionary method for diagnosis of electrocardiogram arrhythmias. Saradhi and Abraham [22] proposed an incremental method of MMC. Wan *et al.* [23] proposed a local graph embedding method based on LMP and fuzzy set for the dimensional reduction of face images. Zhang and Zhou [24] proposed an optimal margin distribution machine for clustering (ODMC), which could cluster data and obtain the optimal margin distribution (OMD). From this perspective, OMD can be deemed to be another statistical learning theory for clustering methods, which is like LMP and LVP. ODMC is motivated by a recent theoretical idea that maximizing the minimum margin may not achieve lower generalization error on empirical datasets in boosting-style algorithms, and instead, it is crucial to optimize the margin distribution [25], i.e., the margin mean and variance. However, it should be noted that the idea was only verified in supervised learning. For clustering problems, LMP still is an effective theory. ODMC contributes an alternative principle OMD to unsupervised learning. In this work, we focus on improving MMC based on previous works [14], [16].

For binary clustering problems, let $DS = [\mathbf{x_1}, ..., \mathbf{x_n}]$ be the dataset, where $\mathbf{x_i} \in \mathbf{X}$ for $i = 1, ..., n$ and $\mathbf{X}$ is a vector space, $\mathbf{X} \in \mathbf{R^v}$ for some positive integer v; denote $\mathbf{y} = [y_1, ..., y_n]$ as the corresponding unknown label vector, where $\mathbf{y} \in \{-1, 1\}^n$. *MMC* aims to find not only the optimal hyperplane $(\mathbf{w}^*, b^*)$, but also the optimal labeling vector $\mathbf{y}^*$ on *DS* [11]:

$$\begin{aligned} \min_{\mathbf{y}\in\{-1,+1\}^n} \min_{\mathbf{w},b,\xi_i\geq 0} \; & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum\nolimits_{i=1}^{n}\xi_i \\ s.t.\; & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad 1 \leq i \leq n \\ & -l \leq \mathbf{e}^T\mathbf{y} \leq l \end{aligned} \tag{1}$$

where $\sum_{i=1}^{n}\xi_i$ is divided by n to better capture how C scales with the data set size, $l \geq 0$ is a constant controlling the class imbalance and $\mathbf{e}$ is the all-one vector. To simplify the notations, we define

$$f(\mathbf{x}_i) = \mathbf{w}^T\mathbf{x}_i + b_i \tag{2}$$

Zhao *et al.* [14], [16] recently proposed to formulate the *MMC* problem as follows

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i\geq 0} \; & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum\nolimits_{i=1}^{n}\xi_i \\ s.t.\; & |f(\mathbf{x}_i)| \geq 1 - \xi_i, \quad -l \leq \sum\nolimits_{i=1}^{n} f(\mathbf{x}_i) \leq l \end{aligned} \tag{3}$$

and they made use of the cutting plane method [26] to solve the problem. Their cutting plane MMC algorithm (*CPMMC*) constructs a nested sequence of successively tighter relaxations of the original problem to obtain a satisfactory solution. However, in the analysis of section II, we will show that the convergence in such an algorithm may not be guaranteed because of the nonconvexity of the empirical loss

$$\ell(\mathbf{x}_i, \mathbf{w}, b) = \max(0, 1 - |f(\mathbf{x}_i)|) \tag{4}$$

In addition, in the researches [14], [16], the resulting successive QP sub-problems derived from the origin MMC problem were solved by traditional methods, such as active set and interior point methods [27]. In general, the computation effort for solving the successive QP sub-problems is high [28]. Moreover, the composite empirical loss functions of the successive QP sub-problems are convex but non-smooth, which is presented in Section III-B.

Consequently, we propose an improved MMC algorithm via the bundle method (BMMC), which is a QP-free type algorithm. Specifically, the constrained convex-concave procedure (CCCP) algorithm is first applied to decompose the MMC problem into a series of convex sub-problems, and then the bundle method (which can be viewed as a generalization of the cutting plane method) is adopted to efficiently solve each sub-problem, such that the convergence and optimality of the final solution is guaranteed. We also propose a new formulation for multi-class MMC which is much simpler than the one in [16]. Meanwhile, BMMC is extended to the semi-supervised case when pairwise constraints are incorporated, which indicates the high scalability of BMMC.

The main contributions of this paper are listed as follows:

1) We propose an improved MMC algorithm via the bundle method, which is faster and simpler than CPMMC [14], [16];
2) BMMC is also extended to the semi-supervised case, which reveals its high scalability;
3) A simpler formulation for multi-class MMC is presented;
4) Finally, the experimental results on several data sets are presented to show the effectiveness of our method.

The rest of the paper is organized as follows. Section II discusses the potential problems with CPMMC. The improved MMC algorithm is presented in section III. Section IV shows the extended semi-supervised MMC with Pairwise Constraints in detail. Theoretical analysis is conducted in Section V. Experimental results on several datasets are presented in section VI. Finally, the conclusions are given in Section VII.

II. POTENTIAL PROBLEMS WITH *CPMMC*

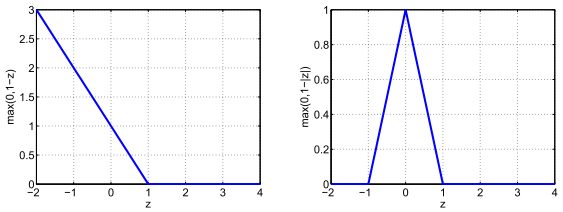
In problem (3), the empirical prediction loss on $\mathbf{x}_i$ is measured by ξ_i as

$$R_{emp}(\mathbf{x}_i) = \max(0, 1 - |f(\mathbf{x}_i)|) \tag{5}$$

In this paper, we call the loss

$$\ell(z) = \max(0, 1 - |z|) \tag{6}$$

as *symmetric hinge loss*, which is illustrated in Fig.1(b), and we also depict the traditional hinge loss for comparison.



(a) (b)

FIGURE 1. A figure illustrates the convexity of the hinge loss $\ell(z) = \max(0, 1 - z)$ and the nonconvexity of $\ell(z) = \max(0, 1 - |z|)$. (a) Hinge loss. (b) Symmetric Hinge loss.

As pointed out by [14], the *n-slack* problem (3) can be reformulated as the following *1-slack* formulation

$$\begin{aligned} &\min_{\mathbf{w},b,\xi\geq 0} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\xi \\ &s.t.\ \forall c_i \in \{0,1\}: \quad \frac{1}{n}\sum_{i=1}^{n} c_i |f(\mathbf{x}_i)| \geq \frac{1}{n}\sum\nolimits_{i=1}^{n} c_i - \xi \\ &\quad -l \leq \sum\nolimits_{i=1}^{n} f(\mathbf{x}_i) \leq l \end{aligned} \tag{7}$$

and they solve it by *cutting plane* algorithm [26].

Given a *convex* function $g(\mathbf{w})$, it is always lower-bounded by its first-order Taylor approximation, *i.e.*,

$$g(\mathbf{w}) \geq g(\mathbf{w}_0) + \langle \mathbf{w} - \mathbf{w}_0, \partial_{\mathbf{w}} g(\mathbf{w}_0) \rangle \tag{8}$$

The general principal behind the *cutting plane* algorithm is that instead of minimizing $g(\mathbf{w})$ directly, minimizing it approximately by iteratively solving a linear program arising from its lower bound. Let

$$g(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + CR_{emp}$$

where the *composite empirical loss* is

$$R_{emp} = \max\left(0, \frac{1}{n}\sum\nolimits_{i=1}^{n} c_i(1 - |f(\mathbf{x}_i)|\right)$$

Then we can derive the *CPMMC* algorithm in [14]. However, since R_{emp} is not convex in $\mathbf{w}$, then the convergence of *CPMMC* cannot be always guaranteed since Eq.(8) cannot be always satisfied.

III. IMPROVED MAXIMUM MARGIN CLUSTERING

In order to solve the MMC algorithm efficiently with guaranteed convergence, we propose to first apply the *constrained convex-concave procedure* (*CCCP*) algorithm [29] to decompose the optimization problem (7) into a series of convex problems. For each problem, we then apply the *bundle method* [30] to solve it.

A. THE CONSTRAINED CONCAVE CONVEX PROCEDURE

The concave-convex procedure (CCP) is an optimization algorithm, which solves a non-convex objective function by a sum of a convex function and a concave function [31]. However, it fails to handle optimization problems with constraints. Smola *et al.* [29] extended it to the Constrained CCP.

TABLE 1. Constrained concave convex procedure.

Randomly initialize x_0
Repeat
Find x_{E+1} as the solution of the optimization problem
$\min f_0(x) - TE_1\{g_0, x_E\}(x)$
$s.t. f_i(x) - TE_1\{g_i, x_E\}(x) \leq m_i, \forall i \in R$
Until convergence x_E

A constrained optimization problem could be described as the formulation 9:

$$\begin{aligned} &\min\ f_0(x) - g_0(x) \\ &s.t.\ f_i(x) - g_i(x) \leq m_i, \quad \forall i \in \{1, \ldots, n\} \end{aligned} \tag{9}$$

where $f_i(x)$ and $g_i(x)$ are differential convex functions on a vector space $X, \forall i \in \{0, \ldots, n\}$; $m_i \in R, \forall i \in \{1, \ldots, n\}$, R is the real number space.

For the constrained CCP, the first-Order Taylor expansion of g_i at location x_E is applied to the problem, denoted by $TE_1\{g_i, x_E\}(x) = g_i(x_E) + < x - x_E, g_i'(x_E) >$. Accordingly, for a convex function, the first-order Taylor expansion is a lower bound. Thus, for all $x_E, x \in X, f_i(x) - TE_1\{g_i, x_e\}(x) \geq f_i(x) - g_i(x)$, in which equality is maintained at $x = x_E$. Instead of minimizing the original optimization problem, the optimal solution is easy to obtained by solving the resulting convex problem. Therefore, constrained CCP can work in a simple and effective way to solve even constrained non-convex problems: linearize constraints into the conjunction of non-convex constraints and convex constraints at every step and perform optimization in the resulting problem. The details of constrained CCP are presented in Table 1.

B. CCCP DECOMPOSITION

As analyzed in [14], although the constraint

$$\frac{1}{n}\sum\nolimits_{i=1}^{n} c_i |f(\mathbf{x}_i)| \geq \frac{1}{n}\sum\nolimits_{i=1}^{n} c_i - \xi$$

in problem (7) is nonconvex, it is a difference of two convex functions. Therefore, we could resort to CCCP to solve it. Given an initial point $(\mathbf{w}^{(0)}, b^{(0)}, \xi^{(0)})$, CCCP computes $(\mathbf{w}^{(t+1)}, b^{(t+1)}, \xi^{(t+1)})$ from $(\mathbf{w}^{(t)}, b^{(t)}, \xi^{(t)})$ by replacing $|f(\mathbf{x}_i)|$ with its first order Taylor expansion at $(\mathbf{w}^{(t)}, b^{(t)})$ and optimizes the resulting convex problem. Since $|f(\mathbf{x}_i)|$ is nonsmooth at $(\mathbf{w}^{(t)}, b^{(t)})$, we should replace its gradient with *subgradient* when computing its tangent in CCCP. By Eq.(2), we can compute the tangent of $|f(\mathbf{x}_i)|$ at $(\mathbf{w}^{(t)}, b^{(t)})$ as

$$\partial_{\mathbf{w},b}(|f(\mathbf{x}_i)|)|_{(\mathbf{w}^{(t)},b^{(t)})} = \mathrm{sgn}(f^{(t)}(\mathbf{x}_i))[\mathbf{x}_i^T\mathbf{w} + b] \tag{10}$$

where $\mathrm{sgn}(\cdot)$ is the sign function. Therefore, by replacing $|f(\mathbf{x}_i)|$ in problem (7) with Eq.(10), we have the following relaxed convex optimization problem for each CCCP iteration

$$\begin{aligned} &\min_{\mathbf{w},b,\xi\geq 0} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\xi \\ &s.t.\ \forall \mathbf{c} \in \{0,1\}: \quad \frac{1}{n}\sum_{i=1}^{n} c_i \mathrm{sgn}(f^{(t)}(\mathbf{x}_i)) f(\mathbf{x}_i) \\ &\quad \geq \frac{1}{n}\sum_{i=1}^{n} c_i - \xi \\ &\quad -l \leq \sum\nolimits_{i=1}^{n} f(\mathbf{x}_i) \leq l \end{aligned} \tag{11}$$

The above problem is a standard *quadratic programming* problem which can be solved in standard ways. After obtaining the solution $(\mathbf{w}, b)$ of problem (11), we use it as $(\mathbf{w}^{(t+1)}, b^{(t+1)})$ and continue the iterations until convergence. The composite empirical loss of problem (11) is

$$R'_{emp} = \max\left(0, \frac{1}{n}\sum\nolimits_{i=1}^{n} c_i\left(1 - \text{sgn}[f^{(t)}(\mathbf{x}_i)]f(\mathbf{x}_i)\right)\right)$$

which is convex but nonsmooth. Therefore, it is natural to resort to the bundle method [30] to solve it. In the following section, we will introduce how to apply the bundle method to solve problem (11) in detail.

C. BUNDLE METHOD

Assume we are given some empirical loss R_{emp} together with a regularization term $\Omega(\mathbf{w})$, then the bundle method works in the way as shown in Table 2, where $\mathbf{w}$ is the variable to be solved, and

$$\mathbf{a}_{s+1} = \partial_{\mathbf{w}} R_{emp}(\mathbf{w}_s)$$

is the gradient vector, s is the current iteration step,

$$o_{s+1} = R_{emp}(\mathbf{w}_s) - \langle \mathbf{a}_{s+1}, \mathbf{w}_s \rangle$$

is the offset. The objective

$$J_s(\mathbf{w}) = \lambda\Omega(\mathbf{w}) + R_s(\mathbf{w}) \tag{12}$$

and

$$R_s(\mathbf{w}) = \max_{s' \le s} \left(\langle \mathbf{a}_{s'}, \mathbf{w}\rangle + o_{s'}\right) \tag{13}$$

Returning to problem (11), at iteration $s+1$,

$$\mathbf{a}_{s+1} = \partial_{(\mathbf{w},b)} R'_{emp}([\mathbf{w}_s, b_s]) = \frac{1}{n}\sum_{i=1}^{n} -c_i^s \text{sgn}[f^{(t)}(\mathbf{x}_i)]\bar{\mathbf{x}}_i \tag{14}$$

where c_i^s is computed as

$$c_i^s = \begin{cases} 1, & \text{if } \text{sgn}[f^{(t)}(\mathbf{x}_i)]f_s(\mathbf{x}_i) < 1 \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

and $\bar{\mathbf{x}}_i = [\mathbf{x}_i^T, \ 1]^T$ is the *augmented* data vector of $\mathbf{x}_i$. Then

$$\begin{aligned} \langle \mathbf{a}_{s+1}, [\mathbf{w}_s, b_s]\rangle &= \frac{1}{n}\sum\nolimits_{i=1}^{n} -c_i^s \text{sgn}[f^{(t)}(\mathbf{x}_i)](\mathbf{w}_s^T\mathbf{x}_i + b_s) \\ &= \frac{1}{n}\sum\nolimits_{i=1}^{n} -c_i^s \text{sgn}[f^{(t)}(\mathbf{x}_i)]f_s(\mathbf{x}_i) \qquad (16) \\ o_{s+1} &= R_{emp}([\mathbf{w}_s, b_s]) - \langle a_{s+1}, [\mathbf{w}_s, b_s]\rangle \\ &= \frac{1}{n}\sum_{i=1}^{n} c_i^s \qquad (17) \end{aligned}$$

$$\mathbf{a}_{s+1}^T[\mathbf{w}, b] + o_{s+1} = \frac{1}{n}\sum_{i=1}^{n} c_i^s - \frac{1}{n}\sum_{i=1}^{n} c_i^s \text{sgn}[f^{(t)}(\mathbf{x}_i)]f(\mathbf{x}_i) \tag{18}$$

Combining Eq.(12),(13),(14),(17) and (18), we can naturally adapt the bundle method in Table 2 to solve problem (11).

TABLE 2. Bundle method for regularized loss.

Initialize: $s = 0,\ \mathbf{w}_0 = \mathbf{0},\ \mathbf{a}_0 = 0,\ o_0 = 0$
$J_0(\mathbf{w}) = \lambda\Omega(\mathbf{w})$
Repeat
Find minimizer $\mathbf{w}_s = argmin_{\mathbf{w}} J_s(\mathbf{w})$
Compute gradient $\mathbf{a}_{s+1}$ and offset o_{s+1}
Increment $t \longleftarrow t+1$
Until $\epsilon_s \le \epsilon$

D. MULTICLASS PROBLEM

Following [17], we can formulate the *multi-class maximum margin clustering* problem as:

$$\begin{aligned} \min_{\mathbf{w}_1,\ldots,\mathbf{w}_k,\xi_i\ge 0} \ & \frac{1}{2}\sum\nolimits_{p=1}^{k} ||\mathbf{w}_p||^2 + \frac{C}{n}\sum\nolimits_{i=1}^{n}\xi_i \\ s.t.\ \forall i = 1,\ldots,n, \quad & r = 1,\ldots,k: \\ & \mathbf{w}_{y_i}^T\mathbf{x}_i + \delta_{y_i,r} - \mathbf{w}_r^T\mathbf{x}_i \ge 1 - \xi_i, \end{aligned} \tag{19}$$

where we assume the data set $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ comes from k clusters, and a separate weight vector $\mathbf{w}_r$ is defined for each cluster r such that $\mathbf{w}_r^T\mathbf{x}_i$ returns the confidence that $\mathbf{x}_i$ belongs to cluster r. $y_i = arg\max_r \mathbf{w}_r^T\mathbf{x}_i$ is the cluster membership of $\mathbf{x}_i$. $\delta_{uv} = 1$ if $u = v$ and 0 otherwise.

To further simplify problem (19), we propose to absorb $\delta_{y_i,r}$ into ξ_i and use a separate variable ξ_i^r for each constraint. Then problem (19) can be relaxed to

$$\begin{aligned} \min_{\mathbf{w}_1,\ldots,\mathbf{w}_k,\xi_i^r\ge 0} \ & \frac{1}{2}\sum\nolimits_{p=1}^{k} ||\mathbf{w}_p||^2 + \frac{C}{nk}\sum\nolimits_{i=1}^{n}\xi_i^r \\ s.t.\ \forall i = 1,\ldots,n, & \ r = 1,\ldots,k: \\ & \max_{p\in\{1,2,\cdots,k\}} \mathbf{w}_p^T\mathbf{x}_i - \mathbf{w}_r^T\mathbf{x}_i \ge 1 - \xi_i^r, \end{aligned} \tag{20}$$

In order to avoid trivial solutions, we can also enforce the class balance constraint in [16] as

$$\forall p, q \in \{1, 2, \cdots, k\}, \quad -l \le \sum_{i=1}^{n}\mathbf{w}_p^T\mathbf{x}_i - \sum_{i=1}^{n}\mathbf{w}_q^T\mathbf{x}_i \le l \tag{21}$$

Similar to Eq.(6), we may find the loss

$$\ell = 1 - \left(\max\nolimits_{p\in\{1,2,\cdots,k\}} \mathbf{w}_p^T\mathbf{x}_i - \mathbf{w}_r^T\mathbf{x}_i\right) \tag{22}$$

nonconvex. Then we can apply the same technique as shown in the binary case to solve the multiclass problem. We first rewrite problem (20) in a *1-slack variable* formulation as [16]

$$\begin{aligned} \min_{\mathbf{w}_1,\ldots,\mathbf{w}_k,\xi_i^r\ge 0} \ & \frac{1}{2}\sum_{p=1}^{k} ||\mathbf{w}_p||^2 + C\xi \\ s.t.\ \forall i = 1,\ldots,n, \quad & r = 1,\ldots,k,\ c_i^r \in \{0,1\}: \\ & \frac{1}{nk}\sum_{i,r} c_i^r\left(\max_{p\in\{1,2,\cdots,k\}} \mathbf{w}_p^T\mathbf{x}_i - \mathbf{w}_r^T\mathbf{x}_i\right) \\ & \ge \frac{1}{nk}\sum_{i,r} c_i^r - \xi, \\ & \forall p,\ q \in \{1,\cdots,k\}: -l \le \sum_{i=1}^{n}\mathbf{w}_p^T\mathbf{x}_i \\ & - \sum_{i=1}^{n}\mathbf{w}_q^T\mathbf{x}_i \le l \end{aligned} \tag{23}$$

Before we describe the details on how to solve the above problem, we first introduce the following two *concatenated* vectors to

$$\tilde{\mathbf{w}} = [\mathbf{w}_1^T, \mathbf{w}_2^T, \cdots, \mathbf{w}_p^T, \cdots, \mathbf{w}_k^T]^T \tag{24}$$

$$\tilde{\mathbf{x}}_{ip} = [\mathbf{0}, \mathbf{0}, \cdots, \mathbf{x}_i^T, \cdots, \mathbf{0}]^T \tag{25}$$

where $\mathbf{0}$ is an $1 \times d$ all-zero vector with d being the dimension of $\mathbf{x}_i$, *i.e.*, only the $(p-1)d$ to pd-th elements are nonzero (equals $\mathbf{x}_i$) in $\tilde{\mathbf{x}}_{ip}$. Then we have $\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip} = \mathbf{w}_p^T\mathbf{x}_i$, and problem (23) can be reformulated as

$$\begin{aligned}
&\min_{\tilde{\mathbf{w}}, \xi_i^r \geq 0} \frac{1}{2}||\tilde{\mathbf{w}}||^2 + C\xi \\
&s.t.\ \forall i = 1, \ldots, n, r = 1, \ldots, k, c_i^r \in \{0, 1\} : \\
&\quad \frac{1}{nk}\sum_{i,r} c_i^r \left(\max_{p \in \{1,2,\cdots,k\}} \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip} - \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ir} \right) \\
&\quad \geq \frac{1}{nk}\sum_{i,r} c_i^r - \xi, \\
&\quad \forall p, q \in \{1, \cdots, k\} : -l \leq \sum_{i=1}^{n} \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip} \\
&\quad - \sum_{i=1}^{n} \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{iq} \leq l
\end{aligned} \tag{26}$$

Since the first constraint is nonconvex, we can also resort to CCCP to decompose it to a series of convex problems, which can be solved via the bundle method. Specifically, in order to apply CCCP, we should compute the subgradient of $\max_{p \in \{1,\cdots,k\}} \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip}$ first. For finite pointwise maximum $f(\mathbf{x}) = \max_{p=1}^{k} f_p(\mathbf{x})$, its subdifferential is just the convex hull of the unions of *active* functions,[1] *i.e.*,

$$\partial f(\mathbf{x}) = \text{conv}\{\partial f_p(\mathbf{x}) | f_p(\mathbf{x}) = f(\mathbf{x})\}$$

Note that in our case, $f_p = \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip}$ and the variable to be solved is $\tilde{\mathbf{w}}$, therefore

$$\partial \max_{p \in \{1,\cdots,k\}} \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip} = \left\{ \sum\nolimits_{r=1}^{k} \beta_{ir}\tilde{\mathbf{x}}_{ir} \middle| \sum\nolimits_{r} \beta_{ir} = 1 \right\} \tag{27}$$

where

$$\beta_{ir} \begin{cases} = 0, & \text{if } \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ir} \neq \max_{p \in \{1,\cdots,k\}} \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip} \\ \geq 0, & \text{otherwise} \end{cases}$$

In multiclass clustering, we usually expect that we assign the data into on unique cluster (we don't consider the multi-label case in this paper), *i.e.*, we expect there is only one active function when computing the subgradient in Eq.(27). So there is a unique p_* satisfying[2]

$$\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip_*} = \max_{p \in \{1,\cdots,k\}} \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip} \tag{28}$$

[1] See http://www.ee.ucla.edu/ee236b/lectures/sg.pdf , page 8.

[2] If there is multiple p_* satisfying Eq.(28), we just randomly select one.

Then at the t-th iteration of CCCP, we can compute the first order Taylor expansion of $\max_{p \in \{1,\cdots,k\}} \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip}$ at $\tilde{\mathbf{w}}^{(t)}$ as

$$\max_{p \in \{1,\cdots,k\}} (\tilde{\mathbf{w}}^{(t)})^T\tilde{\mathbf{x}}_{ip} + \sum_{p=1}^{k} \beta_{ip}^{(t)}(\tilde{\mathbf{w}} - \tilde{\mathbf{w}}^{(t)})^T\tilde{\mathbf{x}}_{ip} = \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip_*^{(t)}}$$

where

$$p_*^{(t)} = arg\max_{p \in \{1,2,\cdots,k\}} \left(\tilde{\mathbf{w}}^{(t)}\right)^T \tilde{\mathbf{x}}_{ip} \tag{29}$$

Correspondingly we will solve the following problem at the t-th iteration of CCCP

$$\begin{aligned}
&\min_{\tilde{\mathbf{w}}, \xi_i^r \geq 0} \frac{1}{2}||\tilde{\mathbf{w}}||^2 + C\xi \\
&s.t.\ \forall i = 1, \ldots, n, \quad r = 1, \ldots, k,\ c_i^r \in \{0, 1\} : \\
&\quad \frac{1}{nk}\sum\nolimits_{i,r} c_i^r \left(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip_*^{(t)}} - \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ir} \right) \geq \frac{1}{nk}\sum\nolimits_{i,r} c_i^r - \xi, \\
&\quad \forall p, \quad q \in \{1, \cdots, k\} : -l \leq \sum_{i=1}^{n} \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip} \\
&\quad - \sum_{i=1}^{n} \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{iq} \leq l
\end{aligned} \tag{30}$$

Clearly, problem (30) is a convex optimization problem but the composite empirical loss

$$R_{emp}^m = \max\left(0, \frac{1}{nk}\sum_{i,r} c_i^r \left(1 - \left(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip_*^{(t)}} - \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ir}\right)\right)\right)$$

is non-smooth. Therefore, we can resort to bundle methods. Particularly, at iteration $s+1$ of the bundle method, we can compute that

$$\mathbf{a}_{s+1} = \partial_{\tilde{\mathbf{w}}} R_{emp}^m(\tilde{\mathbf{w}}_s) = \frac{1}{nk}\sum_{i=1}^{n} -c_i^{rs}\left(\tilde{\mathbf{x}}_{ip_*^{(t)}} - \tilde{\mathbf{x}}_{ir}\right) \tag{31}$$

where

$$c_i^{rs} = \begin{cases} 1, & \text{if } \tilde{\mathbf{x}}_{ip_*^{(t)}} - \tilde{\mathbf{x}}_{ir} < 1 \\ 0, & \text{otherwise} \end{cases} \tag{32}$$

Then

$$\begin{aligned}
\langle \mathbf{a}_{s+1}, \tilde{\mathbf{w}}_s \rangle &= \frac{1}{nk}\sum_{i=1}^{n} -c_i^{rs}\left(\tilde{\mathbf{w}}_s^T\tilde{\mathbf{x}}_{ip_*^{(t)}} - \tilde{\mathbf{w}}_s^T\tilde{\mathbf{x}}_{ir}\right) \\
o_{s+1} &= R_{emp}^m(\mathbf{w}_s) - \langle a_{s+1}, \mathbf{w}_s \rangle = \frac{1}{nk}\sum_{i=1}^{n} c_i^{rs}
\end{aligned} \tag{33}$$

$$\mathbf{a}_{s+1}^T\tilde{\mathbf{w}} + o_{s+1} = \frac{1}{nk}\sum_{i=1}^{n} c_i^{rs} - \frac{1}{n}\sum_{i=1}^{n} c_i^{rs}\left(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip_*^{(t)}} - \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ir}\right) \tag{34}$$

Combining Eq.(12),(13),(31),(33) and Eq.(34), we can adapt the bundle method in Table 2 to solve problem (30).

IV. SEMI-SUPERVISED MMC WITH PAIRWISE CONSTRAINTS

In this section we consider the problem on how to extend our maximum margin algorithm to the case of semi-supervised clustering, where we are given a set of pairwise *must-link* constraints $\mathcal{M}$ and *cannot-link* constraints $\mathcal{C}$. If two points $\{\mathbf{x}_u, \mathbf{x}_v\} \in \mathcal{M}$, then $\mathbf{x}_u$ and $\mathbf{x}_v$ should belong to the same cluster, otherwise if $\{\mathbf{x}_u, \mathbf{x}_v\} \in \mathcal{C}$, then they should belong to different clusters. In the following derivations, we drop the cluster balance constraints because the must-link and cannot-link constraints can help to enforce the cluster balance.

A. BINARY SEMI-SUPERVISED MMC

Inspired by [32], [33], we propose to apply the following objective to measure the prediction loss on the pairwise points $\{\mathbf{x}_u, \mathbf{x}_v\}$ in $\mathcal{M}$ or $\mathcal{C}$:

$$\ell' = |f(\mathbf{x}_u) - z_{uv} f(\mathbf{x}_v)| \tag{35}$$

where

$$z_{uv} = \begin{cases} 1, & \text{if } \{\mathbf{x}_u, \mathbf{x}_v\} \in \mathcal{M} \\ -1, & \text{if } \{\mathbf{x}_u, \mathbf{x}_v\} \in \mathcal{C} \end{cases} \tag{36}$$

We can see that this loss is similar to the *Laplacian loss*. Other loss forms are not applied here because we want to make the successive derivations simpler.

By introducing an additional set of slack variables $\{\xi'_k\}_{k=1}^m$, where $m = |\mathcal{M}| + |\mathcal{C}|$ is the total number of pairwise constraints, we can formulate the binary semi-supervised MMC (*SSMMC*) problem as follows

$$\begin{aligned} \min_{\mathbf{w},b,\xi\geq 0} \ & \frac{1}{2}\|\mathbf{w}\|^2 + C\xi + \frac{C'}{m}\sum_{l=1}^{m}\xi'_l \\ s.t.\ \forall c_i \in \{0,1\}: \ & \frac{1}{n}\sum\nolimits_{i=1}^{n} c_i |f(\mathbf{x}_i)| \geq \frac{1}{n}\sum\nolimits_{i=1}^{n} c_i - \xi \\ & \xi'_l \geq f(\mathbf{x}_u) - z_{uv} f(\mathbf{x}_v), \quad \xi'_l \geq -f(\mathbf{x}_u) + z_{uv} f(\mathbf{x}_v) \end{aligned} \tag{37}$$

Note that the above formulation has $m+1$ slack variables. Using the same trick as presented in Theorem 3.2 in [14], we can introduce m variables $\{t_{uv}\}$ to formulate Eq.(37) in an *extended 1-slack variable* version as

$$\begin{aligned} \min_{\mathbf{w},b,\xi\geq 0} \ & \frac{1}{2}\|\mathbf{w}\|^2 + C\xi + C'\xi' \\ s.t.\ \forall c_i \in \{0,1\}: \ & \frac{1}{n}\sum\nolimits_{i=1}^{n} c_i |f(\mathbf{x}_i)| \geq \frac{1}{n}\sum\nolimits_{i=1}^{n} c_i - \xi \\ \forall t_{uv} \in \{-1,+1\}: \ & \xi' \geq \frac{1}{m}\sum t_{uv}\,(f(\mathbf{x}_u) - z_{uv} f(\mathbf{x}_v)) \end{aligned} \tag{38}$$

where the sum in the last constraint is over all constrained pairs $(\mathbf{x}_u, \mathbf{x}_v)$.[3] Then we can apply the same technique as described in section III, where CCCP is first used to decompose problem (38) to a series of optimization problems as

$$\begin{aligned} \min_{\mathbf{w},b,\xi\geq 0} \ & \frac{1}{2}\|\mathbf{w}\|^2 + C\xi + C'\xi' \\ s.t.\ \forall c_i \in \{0,1\}: \ & \frac{1}{n}\sum_{i=1}^{n} c_i \mathrm{sgn}(f^{(t)}(\mathbf{x}_i)) f(\mathbf{x}_i) \\ & \geq \frac{1}{n}\sum_{i=1}^{n} c_i - \xi \\ \forall t_{uv} \in \{-1,+1\}: \ & \xi' \geq \frac{1}{m}\sum t_{uv}\,(f(\mathbf{x}_u) - z_{uv} f(\mathbf{x}_v)) \end{aligned} \tag{39}$$

For solving the above problem, we define the *extended constraint loss* as

$$\begin{aligned} R''_{emp} = & \frac{1}{n}\sum\nolimits_{i=1}^{n} c_i \left(1 - \mathrm{sgn}[f^{(t)}(\mathbf{x}_i)] f(\mathbf{x}_i)\right) \\ & + \frac{C'}{Cm}\sum t^*_{uv}\,(f(\mathbf{x}_u) - z_{uv} f(\mathbf{x}_v)) \end{aligned} \tag{40}$$

where

$$t^*_{uv} = \max\nolimits_{t_{uv}\in\{-1,+1\}} t_{uv}\,(f(\mathbf{x}_u) - z_{uv} f(\mathbf{x}_v)) \tag{41}$$

So we can also apply the bundle method in Table 2 to solve the above problem, such that at iteration $s+1$,

$$\begin{aligned} \mathbf{a}_{s+1} &= \partial_{(\mathbf{w},b)} R''_{emp}([\mathbf{w}_s, b_s]) \\ &= \frac{1}{n}\sum_{i=1}^{n} -c_i^s \mathrm{sgn}[f^{(t)}(\mathbf{x}_i)]\bar{\mathbf{x}}_i + \frac{C'}{Cm}\sum t^{s*}_{uv}\,(\bar{\mathbf{x}}_u - z_{uv}\bar{\mathbf{x}}_v) \end{aligned} \tag{42}$$

where c_i^s is computed as shown in Eq.(15), and

$$t^{s*}_{uv} = \max\nolimits_{t\in\{-1,+1\}} t_{uv}\left(f^s(\mathbf{x}_u) - z_{uv} f^s(\mathbf{x}_v)\right) \tag{43}$$

Then

$$\begin{aligned} \langle \mathbf{a}_{s+1}, [\mathbf{w}_s, b_s]\rangle = & \frac{1}{n}\sum\nolimits_{i=1}^{n} c_i^s \left(1 - \mathrm{sgn}[f^{(t)}(\mathbf{x}_i)] f^s(\mathbf{x}_i)\right) \\ & + \frac{C'}{Cm}\sum t^{s*}_{pq}\left(f^s(\mathbf{x}_u) - z_{uv} f^s(\mathbf{x}_v)\right) \end{aligned}$$

The offset

$$o_{s+1} = R_{emp}([\mathbf{w}_s, b_s]) - \langle a_{s+1}, [\mathbf{w}_s, b_s]\rangle = \frac{1}{n}\sum\nolimits_{i=1}^{n} c_i^s \tag{44}$$

Hence

$$\begin{aligned} \mathbf{a}_{s+1}^T [\mathbf{w}, b] + o_{s+1} = & \frac{1}{n}\sum_{i=1}^{n} c_i^s - \frac{1}{n}\sum_{i=1}^{n} c_i^s \mathrm{sgn}[f^{(t)}(\mathbf{x}_i)] f(\mathbf{x}_i) \\ & + \frac{C'}{Cm}\sum t^{s*}_{pq}\,(f(\mathbf{x}_u) - z_{uv} f(\mathbf{x}_v)) \end{aligned} \tag{45}$$

Combining Eq.(12),(13),(42),(44) and (45), we can solve problem (39) efficiently using the bundle method in Table 2.

[3] We say $(\mathbf{x}_u, \mathbf{x}_v)$ is a constraint pair if there is either a must-link constraint or cannot-link constraint between them.

B. MULTI-CLASS SEMI-SUPERVISED MMC

Following the multi-class maximum margin clustering formulation Eq.(19) and the pairwise loss Eq.(35), we can formulate the multiclass semi-supervised maximum margin clustering problem as

$$\begin{aligned}
&\min_{\tilde{\mathbf{w}},\xi_i^r\geq 0}\ \frac{1}{2}||\tilde{\mathbf{w}}||^2+C\xi+C'\xi'\\
&s.t.\ \forall i=1,\ldots,n,\quad r=1,\ldots,k,\ c_i^r\in\{0,1\}:\\
&\quad\frac{1}{nk}\sum_{i,r}c_i^r\left(\max_{p\in\{1,2,\cdots,k\}}\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip}-\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ir}\right)\\
&\quad\geq\frac{1}{nk}\sum_{i,r}c_i^r-\xi,\\
&\quad\forall t_{uv}^r\in\{-1,+1\}:\\
&\quad\xi'\geq\frac{1}{mk}\sum_{(uv),r}t_{uv}^r\left(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ur}-z_{uv}\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{vr}\right)
\end{aligned}\tag{46}$$

where the sum in the last constraint is over all $1\leq r\leq k$ and constrained pairs. Following Eq.(30), we can apply CCCP to decompose the above problem into a series of convex optimization problems such that in the t-th iteration, we solve

$$\begin{aligned}
&\min_{\tilde{\mathbf{w}},\xi_i^r\geq 0}\ \frac{1}{2}||\tilde{\mathbf{w}}||^2+C\xi+C'\xi'\\
&s.t.\ \forall i=1,\ldots,n,\quad r=1,\ldots,k,\ c_i^r\in\{0,1\}:\\
&\quad\frac{1}{nk}\sum_{i,r}c_i^r\left(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip_*^{(t)}}-\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ir}\right)\\
&\quad\geq\frac{1}{nk}\sum_{i,r}c_i^r-\xi,\\
&\quad\forall t_{uv}^r\in\{-1,+1\}:\\
&\quad\xi'\geq\frac{1}{mk}\sum_{(uv),r}t_{uv}^r\left(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ur}-z_{uv}\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{vr}\right)
\end{aligned}\tag{47}$$

where

$$\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip_*}=\max_{p\in\{1,\cdots,k\}}\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip}\tag{48}$$

Clearly, problem (47) is convex with respect to $\tilde{\mathbf{w}}$ and we can resort to the bundle method to solve it. Specifically, we can define the extended constraint loss as

$$\begin{aligned}
R_{emp}^{m'}=&\frac{1}{nk}\sum_{i,r}c_i^r\left(1-\left(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip_*^{(t)}}-\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ir}\right)\right)\\
&+\frac{C'}{Cmk}\sum t_{uv}^{r*}\left(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ur}-z_{uv}\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{vr}\right)
\end{aligned}\tag{49}$$

where

$$\tilde{\mathbf{w}}_s^T\tilde{\mathbf{x}}_{ip_*^{(t)}}=\max_{p\in\{1,\cdots,k\}}\tilde{\mathbf{w}}_s^T\tilde{\mathbf{x}}_{ip}\tag{50}$$

and

$$t_{uv}^{r*}=\max_{t_{uv}^r\in\{-1,+1\}}t_{uv}^r\left(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ur}-z_{uv}\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{vr}\right)\tag{51}$$

Then at the $s+1$-th step of the bundle method, we compute

$$\begin{aligned}
\mathbf{a}_{s+1}=\partial_{\tilde{\mathbf{w}}}R_{emp}^{m'}(\tilde{\mathbf{w}}_s)=&\frac{1}{nk}\sum_{i=1}^{n}-c_i^{rs}\left(\tilde{\mathbf{x}}_{ip_*^{(t)}}-\tilde{\mathbf{x}}_{ir}\right)\\
&+\frac{C'}{Cmk}\sum t_{uv}^{rs*}\left(\tilde{\mathbf{x}}_{ur}-z_{uv}\tilde{\mathbf{x}}_{vr}\right)
\end{aligned}\tag{52}$$

where c_i^{rs} is computed as shown in Eq.(32), and

$$t_{uv}^{rs*}=\max_{t_{uv}^r\in\{-1,+1\}}t_{uv}^r\left(\tilde{\mathbf{w}}_s^T\tilde{\mathbf{x}}_{ur}-z_{uv}\tilde{\mathbf{w}}_s^T\tilde{\mathbf{x}}_{vr}\right)\tag{53}$$

Then

$$\begin{aligned}
\langle\mathbf{a}_{s+1},\tilde{\mathbf{w}}_s\rangle=&\frac{1}{nk}\sum_{i=1}^{n}-c_i^{rs}\left(\tilde{\mathbf{w}}_s^T\tilde{\mathbf{x}}_{ip_*^{(t)}}-\tilde{\mathbf{w}}_s^T\tilde{\mathbf{x}}_{ir}\right)\\
&+\frac{C'}{Cmk}\sum t_{uv}^{rs*}\left(\tilde{\mathbf{w}}_s^T\tilde{\mathbf{x}}_{ur}-z_{uv}\tilde{\mathbf{w}}_s^T\tilde{\mathbf{x}}_{vr}\right)\\
o_{s+1}=R_{emp}^m(\mathbf{w}_s)-\langle a_{s+1},\mathbf{w}_s\rangle=&\frac{1}{nk}\sum_{i=1}^{n}c_i^{rs}
\end{aligned}\tag{54}$$

$$\begin{aligned}
\mathbf{a}_{s+1}^T\tilde{\mathbf{w}}+o_{s+1}=&\frac{1}{nk}\sum_{i=1}^{n}c_i^{rs}-\frac{1}{n}\sum_{i=1}^{n}c_i^{rs}\left(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ip_*^{(t)}}-\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ir}\right)\\
&+\frac{C'}{Cmk}\sum t_{uv}^{rs*}\left(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{ur}-z_{uv}\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_{vr}\right)
\end{aligned}\tag{55}$$

Combining Eq.(12),(13),(52),(54) and (55), we can solve problem (47) efficiently using the bundle method in Table 2.

V. THEORETICAL ANALYSIS

The convergence and time complicity of BMMC are presented in this section.

In this work, we define the shape of a dataset is represented by $n\times v$; n and v are the number of samples and features, respectively. In some special fields like bioinformatics, the datasets are sparse and high-dimensional, such as genomics data. We assume that there are $s\ll v$ non-zero features in each sample of these datasets, which indicates the sparsity. For dense datasets, s is equal to v.

A. TIME COMPLICITY

For binary BMMC, each iteration in CCCP takes time $O(ns)$. According to the formulas (2) and (18), the inner product of weights and features is needed to be computed for a sample. Thus each inner product takes time $O(s)$ and the computing time of n samples is $O(ns)$ in each CCCP iteration. For multi-class BMMC, each iteration in CCCP takes time $O(kns)$, where k is the number of classes. Base on the formulas (28) and (34), we find that multi-class decision function $f(x)$ is related to k and v, which decides the computation time $O(kns)$ in each iteration. For binary and multi-class SSBMMC, we can infer the time complicity from formulas (45) and (55), respectively. In each CCCP iteration, binary SSMMC takes time of $O(ns)$ and multi-class SSMMC takes time of $O(kns)$. In conclusion, for binary clustering tasks, the time complexity in each iteration of BMMC is $O(ns)$ and for multi-class clustering tasks, the time complexity in each iteration is $O(kns)$. Both $O(ns)$ and $O(kns)$ scale linearly with n.

B. CONVERGENCE

In our work, the CCCP is exploited to decompose MMC problem into a series of convex problems, and then the bundle

method is utilized to solve each sub-problem. The bundle method is a globally convergent algorithm for regularized risk minimization problems [30], which makes it suitable for the sub-problems. It is a QP-free algorithm and could converge at most in $O(log(1/\varepsilon))$ steps for each convex sub-problems [30]. Consequently, combination of the CCCP and bundle method can guarantee convergence and optimality of the final solution. Moreover, the proposed solution is simpler and faster than CPMMC [14], [16]. CPMMC utilizes QP to solve each CCCP sub-problem and then updates a working constraint set Ω until convergence. We assume that q_m represents the number of iterations for convergence of cutting plane algorithm (CP) in CPMMC and the required total CCCP number is q_c. For binary clustering problems, CPMMC takes time $O(|\Omega|^2 ns)$ ($O(ns)$) in each CCCP (CP) iteration. It is inferred that CPMMC roughly takes time $O(nsq_m + |\Omega|^2 nsq_c)$ to converge. BMMC approximately takes time $O(nsq)$ to converge, where q is the count of CCCP iterations in BMMC. According to [14], q is less than q_c. Intuitively, the proposed method works faster for binary clustering tasks. Likewise, for Multi-class clustering problems, CPMMC converges in $O(knsq_m + |\Omega|^2 knsq_c)$, which is slower than BMMC (converging in $O(knsq)$). In addition, the CPMMC has three stopping criteria, i.e., convergence conditions for QP, CCCP and CP, which makes parameters tuning a difficult task. However, BMMC provides a better solution for the MMC problem with only two stopping criteria, which is faster and simpler. It is easier to obtain reasonable model weights for better clustering performance with less time cost.

VI. EXPERIMENTS

A. EXPERIMENTAL SETUP

This section gives the details about datasets, parameters settings and experimental environment. We conduct a set of experiments to validate the effectiveness of the proposed method.

In the first part of experiments, CPMMC and other conventional clustering methods are implemented as baselines in comparison. The data sets we used include the **ionosphere**, **letter** and **Satellite** data sets from UCI repository[4]; a subset of **20Newsgroup** data set where we choose the topic *rec* which contains *autos*, *motorcycles*, *baseball* and *hockey* (labeled as 0,1,2 and 3) from the version 20-news-18828; and a subset of **USPS** data set[5] containing digits of 1,2,3 and 4.

In the second part of experiments, we test our BMMC on two image processing tasks, i.e., image classification and image segmentation. Three datasets are utilized to verify BMMC: a subset of handwritten digit including ten classes from **MNIST** handwritten digit database[6] and a subset of face images of two people (an2i and bpm) from **CMU** Face Images Data Set[5] and a subset from Multimodal Brain Tumor Image Segmentation (**BRATS**) challenge 2015 [34]. We evaluate the performance of our method to solve binary and multi-class image classification problems on the face dataset and handwritten digit dataset, respectively. BMMC is tested on the **BRATS** 2015 subset for image segmentation. For the handwritten subset, we randomly extract 5000 samples from the **MNIST** handwritten digit database. For face subset, we use the face images with a size of 128×120. To better capture the features of faces, we locate the heads by grabbing 69×86 sub-images from the original images. The starting point coordinates,based on the MATLAB Image Coordinate System [35], on the original images are (40, 13), (25, 13), (30, 13) and (30, 13) for the left, right, straight and up directions. **BRATS** 2015 database includes a training dataset, a testing dataset and a leaderboard dataset. In the three datasets, researchers can have access to the labels of training dataset, and the other two datasets are available only for participants. The training dataset consists of 220 cases of high grade gliomas and 54 cases of low grade gliomas. For the **BRATS** 2015 subset, we randomly extract 50 cases from the training dataset. In each case, the image data of the brain tumor is obtained by Magnetic Resonance Imaging (MRI), which is a widely used medical imaging technology and brain tumors are captured by four MRI modalities, i.e., T1-weighted (T1), T1 with gadolinium enhancing contrast (T1c), T2-weighted (T2) and FLAIR sequences. Every MRI modality has 155 axial slices with a size of 240 × 240. An example of four MRI slices is presented in Figure 2. We focus on the binary segmentation task, i.e., automatic segmentation of whole brain tumors from healthy brain tissues, which is also a main task in the related researches [36]–[38]. Since training data of **BRATS** 2013 is contained in **BRATS** 2015 dataset [39], we compare BMMC with some recently proposed segmentation methods with the data of

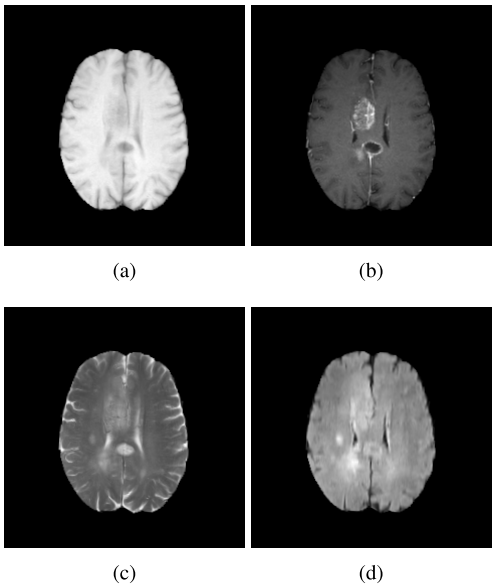


FIGURE 2. Four MRI examples with T1, T1c, T2 and FLAIR contrasts. (a) T1. (b) T1c. (c) T2. (d) Flair.

[4]http://mlearn.ics.uci.edu/MLRepository.html
[5]http://www.kernel-machines.org/data.html
[6]http://yann.lecun.com/exdb/mnist/

TABLE 3. Description of the data sets.

Data	Size	Dimensions	class
Ionoshphere	351	34	2
Letter	1555	16	2
Satellite	2236	36	2
20Newsgroup	3970	8014	4
USPS	3046	256	4
Faces	64	5934	2
Digit	5000	784	10

BRATS 2013 or 2015. Table 3 summarizes the basic properties of those data sets. Notably, **BRATS** 2015 subset has a size of 50 (cases) $\times$ 4 (MRI squeneces) $\times$ 155 (slices with a shape of 240 $\times$ 240).

For clustering and image classification tasks, Clustering accuracy (CA) is used as the performance measure in our experiments. CA is defined based on the strategy [11]: we first run clustering methods on M samples obtain the clustering results N (N represents the number of clusters); then according to the origin labels, we use the majority class in each cluster to label the samples; clustering accuracy is computed by $right\_predictions/M$, where $right\_predictions$ represents the total number of samples with correct predicted classes. Two widely used metrics are exploited to access brain tumor segmentation, i.e., DICE and Sensitivity [40]. The definitions are listed as follows:

$$DICE = \frac{2TP}{2TP + FP + FN} \tag{56}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{57}$$

where TP (TN) denotes the number of voxels with correctly predicted classes for (non) tumor regions in the segmented image; FN is the number of voxels with incorrect labels for non-tumor regions.

For our BMMC, we adopt the *linear* kernel and set $\epsilon = 0.01$ in our experiments. The class imbalance parameter l is set by grid search from the grid $[0, 20]$ with granularity 1. The parameter C is searched from the exponential grid $2^{[-8:1:6]}$. We perform experiments with MATLAB R2012b on a 2.50GHz Intel Core(TM) i7 PC running Windows 7 with 8GB main memory.

B. CLUSTERING RESULTS

In the first part of experiments, we test the effectiveness of the proposed improved maximum margin clustering methods on the first five datasets in Table 3. For comparison, we conduct a sets of experiments using the related methods, which include **K-means**,[7] **Normalized Cut** (***NC***)[8] [10], **Maximum Margin Clustering** (***MMC***),[9] **Generalized Maximum Margin Clustering** (***GMMC***) [15], **Iterative Support Vector Regression** (***IterSVR***)[10] [18], and **Cutting Plane Maximum Margin Clustering** (***CPMMC***)[11] [14], [16]. Note that *GMMC* and *IterSVR* can only handle two-class problems. The best clustering results are reported in the first experiments.

In the second part of experiments, BMMC is assessed with regard to image classification and segmentation. For image classification, we mainly compare the performance of our algorithm with the recently proposed SMLC, MVC and ODMC. In addition, the two classical methods, K-means and Spectral Clustering (SC)[12] [41] are also compared. The experiments are independently conducted twenty times for the mean clustering accuracies. For image segmentation, the segmented results of brain tumors from K-means, Fuzzy C-mean Clustering (FCM) [42] and three recently proposed methods are reported [36]–[38]. Shanker and Bhattacharya [36] combined K-means with FCM algorithm for the segmentation of MRI brain tumors. A Picture Fuzzy Clustering method was proposed for brain tumor segmentation in [37], which was based on the generalization of the traditional fuzzy set and intuitionistic fuzzy set. Shreyas and Pankajakshan [38] designed a deep learning architecture to segment brain tumors in MRI images. Notably, the datasets used in the three methods were extracted from **BRAST** 2013 or 2015 dataset. Therefore, the data distribution is same, which makes the performance comparison meaningful.

The final clustering results of the first part of experiments are summarized in Table 4. In all the experimental results, "-" indicates the corresponding algorithm do not provide such an evaluation value and the reasons may be that either the algorithm is unable to cluster data sets (*e.g.*, GMMC and IterSVR for multi-class problems) or datasets are too large for the corresponding algorithm to work out. We observe that the performance of BMMC is clearly better than that of the traditional methods.

From the results in Table 5, we can see that our methods outperform the traditional K-means and SC on the two datasets. Comparing with SLMC, BMMC obtains a more significant advantage on the two datasets. BMMC has a better performance than ODMC and is similar to MVC in CA on the **CMU** face dataset. We find that BMMC works effectively on image classification.

For brain tumor segmentation, Table 6 provides the segmented results of BMMC and the five related methods

[7]The cluster centers are initialized randomly, and the performances reported are summarized over 50 independent runs. The implementation code is downloaded from http://pwp.etb.net.co/famcastillo/codigo_spectral_clustering/kMeansCluster.html

[8]The width of the Gaussian kernel is set by grid search from $\{0.1\sigma_0, 0.2\sigma_0, \cdots, \sigma_0\}$, where σ_0 is the range of distance between any two data points in the data set.

[9]The implementation is the same as presented in [11] and [17]. The width of the Gaussian kernel is also set by grid search from $\{0.1\sigma_0, 0.2\sigma_0, \cdots, \sigma_0\}$ with σ_0 being the range of distance between any two data points in the data set.

[10]The implementation code is downloaded from http://www.cse.ust.hk/*sim*twinsen/itMMC_code.zip

[11]The implementation code is downloaded from http://binzhao02.googlepages.com/Code_MMC_v1.rar

[12]The implementation code is downloaded from http://scikit-learn.org/stable/modules/clustering.html

TABLE 4. The best clustering accuracy(%) comparisons for different conventional clustering algorithms.

Data	K-means	NC	MMC	GMMC	IterSVR	CPMMC	BMMC
Ionosphere	54.28	75.00	78.75	76.50	77.70	72.36	**78.80**
Letter	82.06	76.80	-	-	92.80	94.47	**95.02**
Satellite	95.93	95.79	-	-	96.82	98.48	**98.57**
20Newsgroup	35.27	41.89	-	-	-	70.63	**72.15**
USPS	92.15	92.81	-	-	-	94.12	**95.23**

TABLE 5. Performance comparison (mean clustering accuracies ± standard deviations %) of our methods and the other three clustering algorithms.

Data	SC	K-means	SLMC	MVC	ODMC	BMMC
Digit	37.21±0.64	61.06±1.07	67.17±2.60	-	-	**82.34 ±2.31**
Faces	50.24±0.00	92.19±12.35	94.76±0.00	97.53 ±4.90	96.32±3.80	**97.14±1.90**

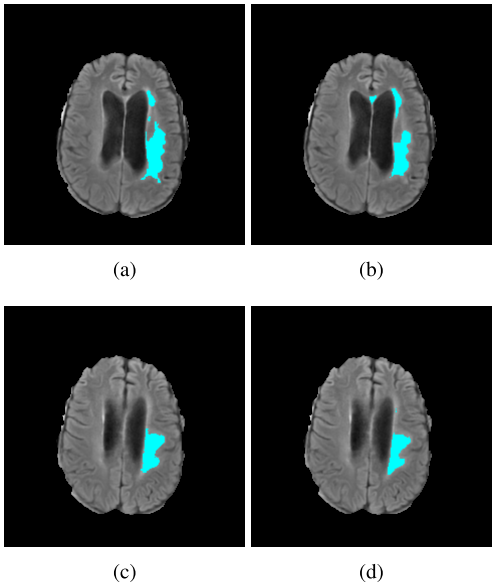


FIGURE 3. Two resulting segmentations overlayed on MRI images. (a), (c) denote the original labeled images; (b), (d) are the segmentation results.

mentioned above on MRI images. According to the table, BMMC yields better segmentations than k-means, FCM and the method in [36]. The DICE value of BMMC is relatively lower than the methods in [37], [38]. The reason is that the method in [37] could consider refusal degree to converge to a desirable brain tumor regions. The deep learning based method [38] achieved the best performance due to its supervised learning and advantages of deep architecture. Two examples of T2 MRI images with ground truth labels and the resulting images obtained from BMMC are shown in Figure 3. It can be seen that almost all the brain tumor regions are identified by BMMC according to the ground truth images. It is visually demonstrated that BMMC works on the **BRAST** 2015 dataset. In conclusion, we find that BMMC is also a competitive tool for image segmentation tasks. It should be noted that BMMC is not specifically designed for brain tumors segmentation. Therefore, we can improve BMMC to further eliminate the error prediction for (b) and (d) in Figure 3. We guess that it may obtain better

TABLE 6. Performance comparison with the five related segmentation methods on MRI images.

Mthods	DICE	Sensitivity
Proposed	**0.81**	**0.79**
K-means	0.71	0.67
FCM	0.74	0.72
Kumar et al.	0.83	-
Shanker et al.	0.77	0.73
Shreyas et al.	0.83	0.89

segmentations by using handcraft context features or post processing, such as integrating conditional random fields (a structured output method).

The experiments above indicate our BMMC method is robust on different tasks. In the first experiments, BMMC performs better than CPMMC on the five datasets. In the second experiments, our method could become a good choice for image classification and segmentation tasks. The empirical results demonstrate the better theoretical foundation of BMMC.

C. SEMI-SUPERVISED CLUSTERING RESULTS

In the second part of experiments, we test the effectiveness of our **Semi-Supervised Bundle Maximum Margin Clustering** (**SSBMMC**) method proposed in Section IV compared with some traditional methods including **MPCKmeans**[13] [43] and **Constrained EM**[14] [44] algorithms. For our SSBMMC method, we also adopt the linear kernel and the parameter C and C' are searched from the exponential grid $2^{[-8:1:6]}$. The precision ε is set to 0.01. In all the algorithms, we just set the number of clusters to be the true number of classes contained in the data set.

In our experiments, we first generate a set of constraints randomly, feed the generated constraints to all the algorithms and compute the clustering accuracies. Such procedure will be implemented 50 times independently and we report the

[13]The implementation is based on the code downloaded from http://www.cs.utexas.edu/users/ml/risc/code/

[14]The implementation code is downloaded from http://www.cs.huji.ac.il/~tomboy/code/ConstrainedEM_plusBNT.zip

TABLE 7. CPU-time (seconds) comparisons of the related clustering algorithms.

Data	K-means	NC	MMC	GMMC	IterSVR	CPMMC	BMMC
Ionosphere	0.047	0.702	654.245	8.354	1.132	0.052	**0.021**
Letter	0.031	3.713	-	-	15.765	0.077	**0.033**
Satellite	0.109	6.599	-	-	37.954	0.128	**0.040**
20Newsgroup	9.781	140.356	-	-	-	80.740	**14.992**
USPS	0.967	74.865	-	-	-	25.319	**0.571**

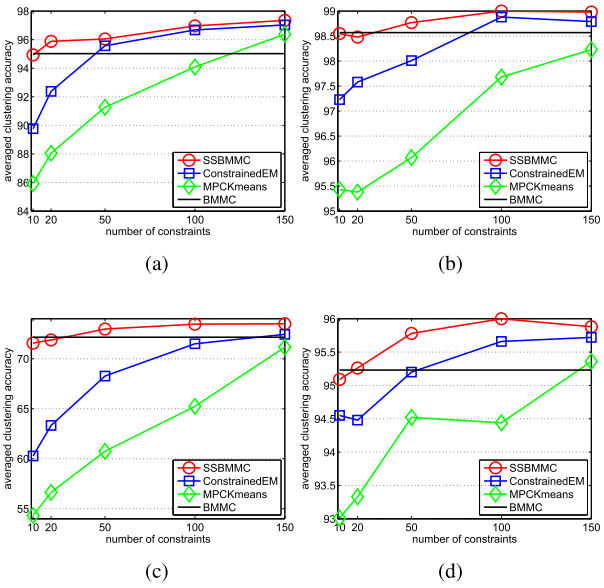


FIGURE 4. Clustering results of different semi-supervised clustering methods. (a) Letter. (b) Satellite. (c) 20Newsgroup. (d) USPS.

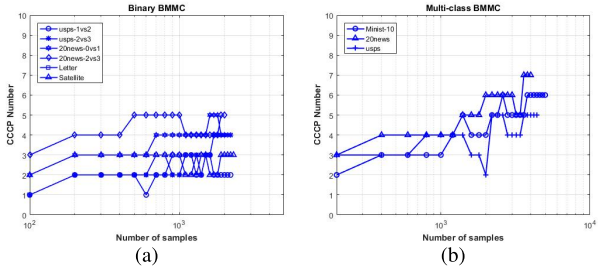


FIGURE 5. Iteration number of CCCP convergence in BMMC with different sizes of datasets.

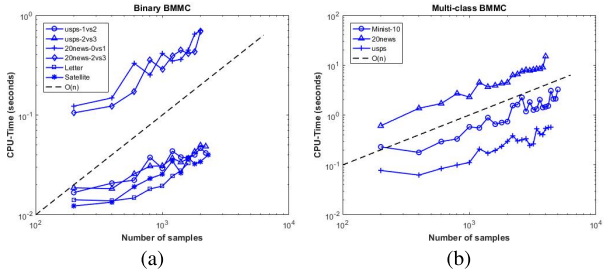


FIGURE 6. CPU-Times (seconds) of BMMC with different sizes of datasets.

average clustering accuracies in Fig.4, where different figures correspond to the results on different data sets. For all the figures, the x-axis corresponds to the number of randomly generated constraints, and the y-axis corresponds to the averaged clustering accuracies. We also plot the results of BMMC as the base line. From the figures, we find that with the increase of constraints, the clustering performance of the proposed method becomes better. Compared with other traditional semi-supervised methods, our SSBMMC method achieves superior accuracy.

D. CPU-TIME OF BMMC

The CPU-times of BMMC with the related clustering algorithms on five datasets are reported in table 7. In this table, on binary clustering tasks, BMMC is at least 2.3 times faster than CPMMC and 54 times faster than IterSVR. BMMC can work over 397 times faster than GMMC and MMC. For multi-class clustering problems, BMMC is at least 5.4 times faster than CPMMC. In addition, with the increase of dataset size, BMMC has a slower growth in CPU-time than the other MMC based methods on binary and multi-class clustering problems. This conclusion implies that BMMC has a better flexibility and scaling property with the feature size and sample size. Finally, comparing with the two traditional K-means and NC, BMMC is still competitive in CPU-time. It could obtain an appealing speed with K-means and obviously work faster than NC on the two clustering tasks. As for the MMC and GMMC, since the computation is time-consuming on the other four datasets, the CPU-times are not shown in table 7.

E. CCCP CONVERGENCE WITH DATASET SIZE

In the section, iteration number of CCCP convergence in BMMC with different sizes of datasets is reported in Figure 5. Obviously, iterations of CCCP convergence are less than 5 (7) for binary (multi-class) clustering problems. Based on CCCP, the number of iterations is irrelevant with dataset size and feature size. From the Figure 5, it can be seen that the curves of CCCP iterations do not drastically change even for large-size datasets. Therefore, immune to dataset size and feature size, the convergence of BMMC needs at most 7 CCCP iterations for clustering problems.

F. SPEED OF BMMC WITH DATASET SIZE

According to the theoretical analysis, we indicate that the computational time complexity of BMMC is linear correlation with the dataset size. In this section, two log-log plots on computational time of BMMC for binary and multi-class

clustering problems are resented in Figure 6. The plots are the related computational time of BMMC with the growth of sizes of multiple datasets. In Figure 6, we find that the lines in the two plots are correspond to polynomial growth of $O(n^h)$, in which h represents the line slope. Moreover, the speed of BMMC scales roughly O(M), which demonstrates the statement in the theoretical analysis section.

VII. CONCLUSIONS

In this paper, we propose an improved MMC algorithm based on the CCCP and bundle method, which is faster and simpler than CPMMC [14], [16]. Moreover, a new formulation for multi-class MMC is proposed. To demonstrate the effectiveness of the proposed method, we conduct two groups of experiments. In the first experiments, we validate the proposed method on several datasets and the results present the superiority of our method over CPMMC and other traditional methods. In the second experiments, we compare our method with the recently proposed SLMC, MVC and ODMC for image classification. BMMC also can yield precise segmentations for MRI brain tumors. All the results indicate BMMC is effective on image classification and segmentation tasks. We also generalize our method to the semi-supervised case by incorporating the pairwise constraints. The experimental results present that SSBMMC performs better than the traditional semi-supervised methods, which reveals the high scalability of BMMC. Based on the two group of experiments, we conclude that the proposed method is a better solution for MMC due to the better theoretical foundation.

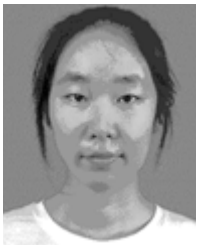
REFERENCES

[1] M. Wan, M. Li, G. Yang, S. Gai, and Z. Jin, "Feature extraction using two-dimensional maximum embedding difference," *Inf. Sci.*, vol. 274, pp. 55–69, Aug. 2014.
[2] M. Wan, G. Yang, S. Gai, and Z. Yang, "Two-dimensional discriminant locality preserving projections (2DDLPP) and its application to feature extraction via fuzzy set," *Multimedia Tools Appl.*, vol. 76, no. 1, pp. 355–371, 2017.
[3] Y. Zhang, S. Ye, and W. Ding, "Based on rough set and fuzzy clustering of MRI brain segmentation," *Int. J. Biomath.*, vol. 10, no. 2, 2017, Art. no. 1750026.
[4] T. Nguyentrang and T. Vovan, "Fuzzy clustering of probability density functions," *J. Appl. Statist.*, vol. 44, no. 4, pp. 583–601, 2017.
[5] M. Jumelle and T. Sakmeche. (2018). "Speaker clustering with neural networks and audio processing." [Online]. Available: https://arxiv.org/abs/1803.08276
[6] A. Tehreem, S. G. Khawaja, A. M. Khan, M. U. Akram, and S. A. Khan, "Multiprocessor architecture for real-time applications using mean shift clustering," *J. Real-Time Image Process.*, no. 3, pp. 1–14, 2017.
[7] H. Andrat and N. Ansari, "Analyzing game stickiness using clustering techniques," in *Advances in Computer and Computational Sciences*. Singapore: Springer, 2018, pp. 645–654.
[8] K. Chowdhury, D. Chaudhuri, and A. K. Pal, "A novel objective function based clustering with optimal number of clusters," in *Methodologies and Application Issues of Contemporary Computing Framework*. Singapore: Springer, 2018, pp. 23–32.
[9] S. Aslan, C. Yozgatligil, and C. Iyigun, "Temporal clustering of time series via threshold autoregressive models: Application to commodity prices," *Ann. Oper. Res.*, vol. 260, nos. 1–2, pp. 51–77, 2018.
[10] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[11] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, "Maximum margin clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1537–1544.
[12] V. Vapnik, "Estimation of dependences based on empirical data," *J. Roy. Stat. Soc.*, vol. 41, no. 3, pp. 462–465, 2006.
[13] G. Niu, B. Dai, L. Shang, and M. Sugiyama, "Maximum volume clustering: A new discriminative clustering approach," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 2641–2687, 2013.
[14] B. Zhao, F. Wang, and C. Zhang, "Efficient maximum margin clustering via cutting plane algorithm," in *Proc. 8th SIAM Int. Conf. Data Mining*, 2009, pp. 751–762.
[15] H. Valizadegan and R. Jin, "Generalized maximum margin clustering and unsupervised kernel learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, no. 5, 2007, pp. 1417–1424.
[16] B. Zhao, F. Wang, and C. Zhang, "Efficient multiclass maximum margin clustering," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 1248–1255.
[17] L. Xu and D. Schuurmans, "Unsupervised and semi-supervised multi-class support vector machines," in *Proc. 12th Nat. Conf. Artif. Intell. 17th Innov. Appl. Artif. Intell. Conf.*, Pittsburgh, PA, USA, Jul. 2005, pp. 904–910.
[18] K. Zhang, I. W. Tsang, and J. T. Kwok, "Maximum margin clustering made practical," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 1119–1126.
[19] H. Zeng and Y.-M. Cheung, "Semi-supervised maximum margin clustering with pairwise constraints," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 926–939, May 2012.
[20] Y. Wang and S. Chen, "Soft large margin clustering," *Inf. Sci.*, vol. 232, pp. 116–129, May 2013.
[21] B. Zhu, Y. Ding, and K. Hao, "Multiclass maximum margin clustering via immune evolutionary algorithm for automatic diagnosis of electrocardiogram arrhythmias," *Appl. Math. Comput.*, vol. 227, pp. 428–436, Jan. 2014.
[22] V. V. Saradhi and P. C. Abraham, "Incremental maximum margin clustering," *Pattern Anal. Appl.*, vol. 19, no. 4, pp. 1057–1067, 2016.
[23] M. Wan, Z. Lai, G. Yang, Z. Yang, F. Zhang, and H. Zheng, "Local graph embedding based on maximum margin criterion via fuzzy set," *Fuzzy Sets Syst.*, vol. 318, pp. 120–131, Jul. 2017.
[24] T. Zhang and Z.-H. Zhou, "Optimal margin distribution clustering," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 4474–4481.
[25] W. Gao and Z.-H. Zhou, "On the doubt about margin explanation of boosting," *Artif. Intell.*, vol. 203, no. 5, pp. 1–18, 2013.
[26] J. E. Kelley, Jr., "The cutting-plane method for solving convex programs," *J. Soc. Ind. Appl. Math.*, vol. 8, no. 4, pp. 703–712, 1960.
[27] A. Potschka, "A parametric active set method for QP solution," in *A Direct Method for Parabolic PDE Constrained Optimization Problems*. Wiesbaden, Germany: Springer, 2014.
[28] W. Hua, "A kind of nonmonotone QP-free method for constrained optimization," *Fuzzy Inf. Eng.*, vol. 2, pp. 1237–1247, Sep. 2009.
[29] A. J. Smola, S. Vishwanathan, and T. Hofmann, "Kernel methods for missing variables," in *Proc. AISTATS*, 2005, pp. 325–332.
[30] A. J. Smola, S. V. N. Vishwanathan, and Q. V. Le, "Bundle methods for machine learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2007, pp. 1377–1384.
[31] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Comput.*, vol. 15, no. 4, pp. 915–936, 2003.
[32] A. B. Goldberg, X. Zhu, and S. Wright, "Dissimilarity in graph-based semi-supervised classification," *J. Mach. Learn. Res.*, vol. 2, no. 2, pp. 155–162, 2007.
[33] R. Yan, J. Zhang, J. Yang, and A. G. Hauptmann, "A discriminative learning framework with pairwise constraints for video object classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 578–593, Apr. 2006.
[34] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.
[35] C. Solomon and T. Breckon, *Fundamentals of Digital Image Processing: A Practical Approach With Examples in MATLAB*. Hoboken, NJ, USA: Wiley, 2011.
[36] R. Shanker and M. Bhattacharya, "Brain tumor segmentation of normal and pathological tissues using K-mean clustering with fuzzy C-mean clustering," in *Proc. Eur. Congr. Comput. Methods Appl. Sci. Eng.*, 2017, pp. 286–296.
[37] S. V. A. Kumar, B. S. Harish, and V. N. M. Aradhya, "A picture fuzzy clustering approach for brain tumor segmentation," in *Proc. 2nd Int. Conf. Cogn. Comput. Inf. Process.*, Aug. 2016, pp. 1–6.
[38] V. Shreyas and V. Pankajakshan, "A deep learning architecture for brain tumor segmentation in MRI images," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, Oct. 2017, pp. 1–6.

[39] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, Jan. 2017.
[40] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Med. Imag.*, vol. 15, no. 1, p. 29, 2015.
[41] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.
[42] H.-C. Liu, J.-M. Yih, D.-B. Wu, and S.-W. Liu, "Fuzzy C-mean clustering algorithms based on picard iteration and particle swarm optimization," in *Proc. Int. Workshop Educ. Technol. Training*, Dec. 2008, pp. 838–842.
[43] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, p. 11.
[44] N. Shental, A. Bar-Hillel, T. Hertz, and D. Weinshall, "Computing Gaussian mixture models with EM using equivalence constraints," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 465–472.

JIANQIANG LI received the B.S. degree in mechatronics from the Beijing Institute of Technology, Beijing, China, in 1996, and the M.S. and Ph.D. degrees in control science and engineering from Tsinghua University, Beijing, China, in 2001 and 2004, respectively. He was a Researcher with the Digital Enterprise Research Institute, National University of Ireland, Galway, from 2004 to 2005. From 2005 to 2013, he was with NEC Laboratories, China, as a Researcher. He was with the Department of Computer Science, Stanford University, as a Visiting Scholar, from 2009 to 2010. He joined the Beijing University of Technology, Beijing, China, in 2013, as a Beijing Distinguished Professor. He has over 40 publications and 37 international patent applications (19 of them have been granted in China, the U.S., or Japan). His research interests include Petri nets, enterprise information systems, business processes, data mining, information retrieval, semantic web, privacy protection, and big data. He served as a PC Member for multiple international conferences and organized the IEEE Workshop on Medical Computing.

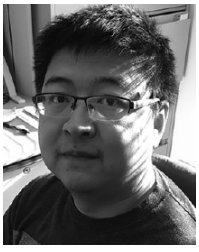
JINGCHAO SUN received the B.S. degree from Xinjiang University, in 2014. He is currently pursuing the Ph.D. degree with the School of Software Engineering, Beijing University of Technology. His research interests include data mining, machine learning, and big data.



LU LIU received the M.S. degree from Datong University, in 2012. She is currently pursuing the Ph.D. degree with the School of Software Engineering, Beijing University of Technology. Her research interests include data mining, information retrieval, privacy protection, and big data.

BO LIU (M'14–SM'17) received the B.S. degree from the Department of Automation, Beijing Institute of Technology, Beijing, China, and the M.S. and Ph.D. degrees from the Department of Automation, System Integration Institute, Tsinghua University, Beijing, in 2003 and 2008, respectively. She was with the NEC Laboratory, China, as a Researcher, from 2008 to 2010 and from 2013 to 2015. She was a Research Professional with the Computation Institute, The University of Chicago, Chicago, IL, USA, and the Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL, USA, from 2011 to 2012. She joined as an Associate Professor with the Beijing University of Technology, in 2015. She has authored over 50 articles and 40 inventions. Her current research interests include big data, data mining, machine learning, cloud computing, scientific workflow, semantic web, and ontology reasoning.

CAO XIAO received the Ph.D. degree from the University of Washington, Seattle, WA, USA, in 2016. She is currently a Research Staff Member with the Center for Computational Health, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA. Her research interests include developing novel machine learning and data mining models to solve real world healthcare challenges. Particularly, she is interested in deep computational phenotyping and risk prediction, deep representation learning for graphs, causal inference from observational data, tensor decomposition for integrating multiple medical data sources, and translational informatics research.



FEI WANG is currently an Assistant Professor with the Division of Health Informatics, Department of Healthcare Policy and Research, Cornell University. He has published more than 150 papers on top data mining and medical informatics venues. His papers have received more than 3300 citations so far. His major research interests include data analytics and its applications in health informatics. He won the Best Student Paper for ICDM 2015, the Best Research Paper Nomination for ICDM 2010, the Marco Romani Best Paper Nomination in AMIA TBI 2014, and his paper was selected as the Best Paper Finalist in SDM 2011 and 2015. He is the Vice Chair of the KDD Working Group in AMIA.