

# Maximum Margin Clustering on Data Manifolds

Fei Wang, Xin Wang, Tao Li

School of Computing and Information Sciences, Florida International University, Miami, FL 33199  
 {feiwang,xwang009,taoli}@cs.fiu.edu

**Abstract**—Clustering is one of the most fundamental and important problems in computer vision and pattern recognition communities. Maximum Margin Clustering(MMC) is a recently proposed clustering technique which has shown promising experimental results. The main theme behind MMC is to extend the standard maximum margin principle in Support Vector Machine (SVM) to the unsupervised scenario. This paper will consider the problem of maximum margin clustering on data manifolds. Specifically, we propose an approach called Manifold Regularized Maximum Margin Clustering (MRMMC) which combines both the maximum margin data discrimination and data manifold information in a unified clustering objective and propose an efficient algorithm to solve it. Finally the experimental results on several real world data sets are presented to show the effectiveness of our method.

## I. INTRODUCTION

*Maximum Margin Clustering (MMC)* is a recently proposed clustering method and it extends the standard theory in *Support Vector Machine (SVM)* to the unsupervised scenario [14][16]. The *maximum margin* principle behind *SVM* is a general rule for discriminating the data from different classes and it has been used widely in computer vision (e.g., object tracking [23] and stereo vision [9]) and pattern recognition (e.g., visual category recognition [19] and face recognition [5]) communities. Most of the maximum margin methods are supervised because they can be formulated as convex problems and solve efficiently. However, in *MMC*, since we should solve the clustering hyperplanes as well as the cluster assignments, the resultant problem is usually nonconvex and hard to solve. Originally Xu *et al.* [16] proposed to first relax the *MMC* problem to a convex one and then apply *Semi-Definite Programming* to solve it. They showed promising empirical results, but the required computational complexity is too high to handle large scale data sets.

To make the *MMC* problem more practical, many efficient implementations are proposed [14][20][21][22], among which *Cutting Plane Maximum Margin Clustering (CPMMC)* [21][22] has shown state-of-the-art performances, in both accuracy and efficiency. However, the *CPMMC* algorithm directly apply the *Cutting Plane* approach [7], which is designed for convex nonsmooth problems, to solve the nonconvex *MMC* problem. Thus the convergence and optimality of the final solution may not be guaranteed.

In this paper, we propose a novel algorithm called *Manifold Regularized Maximum Margin Clustering (MRMMC)*.

Compared with existing *MMC* and conventional clustering approaches, our method has the following strengths:

(1). As indicated by [13][10], many real world data sets (especially the data sets used in pattern recognition and computer vision like faces, hand written digits and images) demonstrate some low dimensional manifold structures. However, the traditional *MMC* approaches only consider how to discriminate different clusters and leave those manifold information away. Our *MRMMC* method combines the maximum margin discrimination and data manifold information in a principled way by incorporating the cluster assignment smoothness as a prior regularization term, so that the cluster results take into account both the discriminative and geometric information contained in the data sets.

(2). Different from *CPMMC* [21][22], we propose a theoretically more elegant approach to solve the *MRMMC* problem. It first apply the *Constrained Convex Concave Procedure (CCCP)* [12] to decompose the original nonconvex problem into a series of convex problem, and then apply the *cutting plane method* to solve each of them. Thus the final solution is guaranteed to converge to a local optimum. The idea is borrowed from [6].

(3). Similar to existing *MMC* approaches, *MRMMC* can obtain the cluster assignments as well as the cluster hyperplanes, which is easy to extend to out-of-sample data points.

The rest of this paper is organized as follows. Section 2 will briefly review the basic knowledge of *MMC* and *CCCP*. The detailed procedure of *MRMMC* will be introduced in section 3. Section 4 will present the experimental results, followed by the conclusions in Section 5.

## II. PRELIMINARIES AND BACKGROUND KNOWLEDGE

### A. Maximum Margin Clustering

Mathematically, for two-class problems, given a point set  $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}^1$  and their labels  $\mathbf{y} = (y_1, \dots, y_n) \in \{-1, +1\}^n$ , *SVM* seeks a hyperplane with normal vector  $\mathbf{w}^2$

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} \quad (1)$$

<sup>1</sup>In the following we will use  $\mathcal{X}$  to denote the data set and data space interchangeably without any confusion.

<sup>2</sup>Two remarks: (1) We do not consider the offset  $b$  of the hyperplane since it can be easily incorporated into our framework by augmenting  $\tilde{\mathbf{x}}_i = [\mathbf{x}_i^T, 1]^T$  and  $\tilde{\mathbf{w}} = [\mathbf{w}^T, b]^T$ ; (2) We do not consider the kernel trick here. For nonlinear cases,  $\mathbf{x}_i$  can be viewed as the low dimensional embedding of the  $i$ -th data point by performing *Kernel PCA* [11][3].

by solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{w}, \xi_i} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{n} \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & i = 1, \dots, n, \xi_i \geq 0, y_i f(\mathbf{x}_i) \geq 1 - \xi_i \end{aligned} \quad (2)$$

where  $\{\xi_i\}$  are slack variables and  $C > 0$  is a constant.

Following standard  *SVM*, for two-class problems, *MMC* aims to find the optimal cluster hyperplane together with the cluster assignments by solving [16]:

$$\begin{aligned} \min_{\mathbf{y} \in \{-1, +1\}^n} \min_{\mathbf{w}, b, \xi_i} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{n} \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & y_i f(\mathbf{x}_i) \geq 1 - \xi_i \\ & \xi_i \geq 0 \quad i = 1, \dots, n \\ & -l \leq \mathbf{e}^T \mathbf{y} \leq l \end{aligned} \quad (3)$$

where  $l \geq 0$  is a constant controlling the cluster imbalance and  $\mathbf{e}$  is the all-one vector.

Zhao *et al.* [21] recently proposed to compute the cluster assignment of  $\mathbf{x}_i$  by

$$y_i = \text{sign}(\mathbf{w}^T \mathbf{x}_i)$$

where  $\text{sign}(\cdot)$  is a sign function. Then, they formulate the two-class *MMC* problem as follows

$$\begin{aligned} \min_{\mathbf{w}, \xi_i} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{n} \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & i = 1, \dots, n, \xi_i \geq 0, |f(\mathbf{x}_i)| \geq 1 - \xi_i \\ & -l \leq \sum_{i=1}^n f(\mathbf{x}_i) \leq l. \end{aligned} \quad (4)$$

They proved that without the cluster-balance constraint, the solution to problem (4) is identical to the solution to problem (3) and made use of the *cutting plane* method [7] to solve the problem. However, it can be easily observed that the first constraint in problem (4) is nonconvex with respect to  $\mathbf{w}$ , and the cutting plane algorithm [7] is originally designed for convex problems, thus the convergence and optimality of directly applying cutting plane to solve problem (4) may not be guaranteed.

### B. Concave-Convex Procedure

The *concave-convex procedure* (CCP) [18] is a method for solving non-convex optimization problem whose objective function can be expressed as a difference of convex functions. While [18] only considered linear constraints, [12] proposed the *constrained concave-convex procedure* (CCCP), which can solve the optimization problems with a concave-convex objective function under concave-convex constraints. Mathematically, assume we want to solve the following optimization problem [12]

$$\begin{aligned} \min_{\mathbf{z}} \quad & f_0(\mathbf{z}) - g_0(\mathbf{z}) \\ \text{s.t.} \quad & f_i(\mathbf{z}) - g_i(\mathbf{z}) \leq c_i \quad i = 1, \dots, n \end{aligned} \quad (5)$$

where  $f_i$  and  $g_i$  are real-valued convex functions on a vector space  $\mathcal{Z}$  and  $c_i \in \mathbb{R}$  for all  $i = 1, \dots, n$ . Denote by  $T_1\{f, \mathbf{z}\}(\mathbf{z}')$  the first order *Taylor expansion* of  $f$  at location  $\mathbf{z}$ , that is  $T_1\{f, \mathbf{z}\}(\mathbf{z}') = f(\mathbf{z}) + \partial_{\mathbf{z}} f(\mathbf{z})(\mathbf{z}' - \mathbf{z})$ , where  $\partial_{\mathbf{z}} f(\mathbf{z})$  is the gradient of the function  $f$  at  $\mathbf{z}$ . Given an initial point  $\mathbf{z}_0$ , the *CCCP* computes  $\mathbf{z}_{t+1}$  from  $\mathbf{z}_t$  by replacing  $g_i(\mathbf{z})$  with its first-order Taylor expansion at  $\mathbf{z}_t$ , i.e.,  $T_1\{g_i, \mathbf{z}_t\}(\mathbf{z})$ , and setting  $\mathbf{z}_{t+1}$  to be the solution of the following relaxed optimization problem

$$\begin{aligned} \min_{\mathbf{z}} \quad & f_0(\mathbf{z}) - T_1\{g_0, \mathbf{z}_t\}(\mathbf{z}) \\ \text{s.t.} \quad & f_i(\mathbf{z}) - T_1\{g_i, \mathbf{z}_t\}(\mathbf{z}) \leq c_i \quad i = 1, \dots, n \end{aligned} \quad (6)$$

The above procedure will continue until  $\mathbf{z}_t$  converges, and [12] proved that the *CCCP* will finally converge to a local minimum of problem (5).

### III. MANIFOLD REGULARIZED MAXIMUM MARGIN CLUSTERING (MRMMC)

In this section we introduce our *Manifold Regularized Maximum Margin Clustering* (MRMMC) algorithm in detail.

#### A. Motivation and Formulation

Although Equations (3) and (4) provide us an elegant way of incorporating the maximum margin principle into unsupervised learning, it does not explore any prior knowledge in the data space. Specifically, many previous research show that the data in computer vision and pattern recognition usually form some low-dimensional manifolds [13][10]. In such cases, how to take into account the geometric information contained in the data sets would be of crucial importance to the performances of the algorithms.

Consider the statistical framework of learning from examples, where there is a joint probability distribution  $P$  on the product space  $\mathcal{X} \times \mathbb{R}$ , such that the unlabeled examples  $\mathbf{x} \in \mathcal{X}$  are sampled from the marginal distribution  $P_{\mathcal{X}}$  of  $P$  by integrating the label space out. Then our goal is to learn the conditional  $P(y|\mathbf{x})$ . In general, we can make a specific assumption that there is a strong connection between the marginal  $P_{\mathcal{X}}$  and the conditional  $P(y|\mathbf{x})$  such that we can explore some prior knowledge on  $P_{\mathcal{X}}$  for better function learning. More concretely, we incorporate the *smoothness* assumption [1] in semi-supervised learning field into the *MMC* framework, which says that

**(Smoothness Assumption)**[1][2]. *If two points  $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$  are close in the intrinsic geometry of  $P_{\mathcal{X}}$ , then the conditional distributions  $P(y|\mathbf{x}_i)$  and  $P(y|\mathbf{x}_j)$  should be similar.*

An equivalent statement of the above smoothness assumption is that the prediction function  $f$  should be sufficiently smooth with respect to the intrinsic geometry of  $P_{\mathcal{X}}$ , therefore the question is how to estimate the intrinsic geometry

<sup>3</sup>For non-smooth functions, the gradient  $\partial_{\mathbf{z}} f(\mathbf{z})$  can be replaced by the subgradient [4].

of  $P_{\mathcal{X}}$ . As suggested by [1], we are particularly interested in the case when the support of  $P_{\mathcal{X}}$  is a compact submanifold  $\mathcal{M} \in \mathcal{X}$ , in which case we can estimate the smoothness of  $f$  over  $P_{\mathcal{X}}$  by

$$\|f\|_I^2 = \int_{\mathcal{M}} \langle \nabla_{\mathcal{M}} f, \nabla_{\mathcal{M}} f \rangle \quad (7)$$

where  $\nabla_{\mathcal{M}}$  is the manifold gradient. The smaller  $\|f\|_I^2$  is, the smoother  $f$  will be. [1] also pointed out that  $\|f\|_I^2$  can be approximated on the basis of the data points in  $\mathcal{X}$  as

$$\widetilde{\|f\|_I^2} = \frac{1}{n} \mathbf{f}^T \mathbf{L} \mathbf{f} \quad (8)$$

where  $\mathbf{f} = [f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)]^T$ , and  $\mathbf{L} = \mathbf{D} - \mathbf{W} \in \mathbb{R}^{n \times n}$  is the graph Laplacian defined on the data adjacency graph where  $W_{ij}$  represents the weight on the edge connecting  $\mathbf{x}_i$  and  $\mathbf{x}_j$ , and  $\mathbf{D}$  is a diagonal matrix with  $D_{ii} = \sum_j W_{ij}$ . Combining Eq.(4) and Eq.(8) together, we can derive our *MRMMC* problem as

$$\begin{aligned} \min_{\mathbf{w}, \xi_i} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C_1}{n} \mathbf{f}^T \mathbf{L} \mathbf{f} + \frac{C_2}{n} \sum_{i=1}^n \xi_i \quad (9) \\ \text{s.t.} \quad & i = 1, \dots, n, \xi_i \geq 0, |f(\mathbf{x}_i)| \geq 1 - \xi_i \\ & -l \leq \sum_{i=1}^n f(\mathbf{x}_i) \leq l \end{aligned}$$

In the following section we will introduce an efficient algorithm to solve problem (9).

### B. Problem Solving

In problem (9), we have  $n$  slack variables  $\{\xi_i\}$ . To solve it efficiently, we first derive the 1-slack form of problem (9) as in [21]. Specifically, we introduce a single slack variable  $\xi \geq 0$  and rewrite problem (9) as

$$\begin{aligned} \min_{\mathbf{w}, \xi \geq 0} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C_1}{n} \mathbf{f}^T \mathbf{L} \mathbf{f} + C_2 \xi \quad (10) \\ \text{s.t.} \quad & \forall c_i \in \{0, 1\} : \frac{1}{n} \sum_{i=1}^n c_i |f(\mathbf{x}_i)| \geq \frac{1}{n} \sum_{i=1}^n c_i - \xi \\ & -l \leq \sum_{i=1}^n f(\mathbf{x}_i) \leq l \end{aligned}$$

It can be proved that the solution to problem (10) is identical to problem (9) with  $\xi = \frac{1}{n} \sum_{i=1}^n \xi_i$  (similar to [21]). Clearly, the first constraint in problem (10) is nonconvex in  $\mathbf{w}$  and it is a difference of two convex functions, therefore we can resort to *CCCP* to solve it.

1) *CCCP Decomposition*: Given an initial point  $(\mathbf{w}^{(0)}, \xi^{(0)})$ , *CCCP* computes  $(\mathbf{w}^{(t+1)}, \xi^{(t+1)})$  from  $(\mathbf{w}^{(t)}, \xi^{(t)})$  by replacing  $|f(\mathbf{x}_i)|$  with its first order Taylor expansion at  $\mathbf{w}^{(t)}$  and optimizes the resulting convex problem. Since  $|f(\mathbf{x}_i)|$  is nonsmooth at  $\mathbf{w}^{(t)}$ , we should replace its gradient with *subgradient* when computing its tangent in *CCCP*. By Eq.(1), we can compute the tangent of  $|f(\mathbf{x}_i)|$  at  $\mathbf{w}^{(t)}$  as

$$\partial_{\mathbf{w}} (|f(\mathbf{x}_i)|) |_{(\mathbf{w}^{(t)})} = \text{sign}(f^{(t)}(\mathbf{x}_i)) (\mathbf{w}^T \mathbf{x}_i) \quad (11)$$

where  $\text{sign}(\cdot)$  is the sign function. Thus by replacing  $|f(\mathbf{x}_i)|$  in problem (10) with Eq.(11), we have the following relaxed convex optimization problem for each *CCCP* iteration

$$\begin{aligned} \min_{\mathbf{w}, \xi \geq 0} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C_1}{n} \mathbf{f}^T \mathbf{L} \mathbf{f} + C_2 \xi \quad (12) \\ \text{s.t.} \quad & \forall \mathbf{c} \in \Omega : \\ & \frac{1}{n} \sum_{i=1}^n c_i \text{sign}(f^{(t)}(\mathbf{x}_i)) f(\mathbf{x}_i) \geq \frac{1}{n} \sum_{i=1}^n c_i - \xi \\ & -l \leq \sum_{i=1}^n f(\mathbf{x}_i) \leq l \end{aligned}$$

where  $\mathbf{c} = [c_1, c_2, \dots, c_n]$  and  $\Omega = \{0, 1\}^n$ . The above problem is a standard *quadratic programming* problem which can be solved in standard ways. After obtaining the solution  $\mathbf{w}$  of the above problem, we use it as  $\mathbf{w}^{(t+1)}$  and continue the iterations until convergence.

2) *Cutting Plane Method*: Now the problem becomes how to solve Eq.(12) efficiently, which is convex and has exponential number of constraints. In the following we will employ an adaptation of the *cutting plane* algorithm [7] to solve problem (12), which targets to find a small subset of constraints from the whole set of constraints in problem (12) that ensures a sufficiently accurate solution. Using such an algorithm, we construct a nested sequence of successively tighter relaxations of problem (12). Similar to [21], we can generally find a polynomially sized subset of constraints, with which the solution of the relaxed problem fulfills all constraints from problem (12) up to a precision  $\epsilon$ , i.e., for  $\forall c_i \in \{0, 1\}$ :

$$\frac{1}{n} \sum_{i=1}^n c_i \text{sign}(f^{(t)}(\mathbf{x}_i)) f(\mathbf{x}_i) \geq \frac{1}{n} \sum_{i=1}^n c_i - (\xi + \epsilon) \quad (13)$$

That is, the remaining exponential number of constraints are guaranteed to be violated by no more than  $\epsilon$ , without the need for explicitly adding them to the optimization problem.

Specifically, our algorithm starts with only the cluster balance constraint subset and solves the following problem.

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C_1}{n} \mathbf{f}^T \mathbf{L} \mathbf{f} \quad (14) \\ \text{s.t.} \quad & -l \leq \sum_{i=1}^n f(\mathbf{x}_i) \leq l \end{aligned}$$

After getting the solution  $\mathbf{w}^{t_0}$  to the above problem<sup>4</sup>, it computes the *most violated constraint* as

$$c_i^{t_0} = \begin{cases} 1, & \text{if } \text{sign}(f^{(t)}(\mathbf{x}_i)) f^{t_0}(\mathbf{x}_i) < 1 \\ 0, & \text{otherwise} \end{cases} \quad (15)$$

where  $f^{t_0}(\mathbf{x}_i) = (\mathbf{w}^{t_0})^T \mathbf{x}_i$ . Then the algorithm adds such

<sup>4</sup>Here we use the superscript  $t_i$  to denote that this is the  $i$ -th iteration of the cutting plane algorithm for solving the problem derived from the  $t$ -th iteration of *CCCP*.

constraint to problem (14) and solve

$$\begin{aligned} \min_{\mathbf{w}, \xi \geq 0} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C_1}{n} \mathbf{f}^T \mathbf{L} \mathbf{f} + C_2 \xi \quad (16) \\ \text{s.t.} \quad & \frac{1}{n} \sum_{i=1}^n c_i^{t_0} \text{sign}(f^{(t)}(\mathbf{x}_i)) f(\mathbf{x}_i) \geq \frac{1}{n} \sum_{i=1}^n c_i^{t_0} - \xi \\ & -l \leq \sum_{i=1}^n f(\mathbf{x}_i) \leq l \end{aligned}$$

Then the algorithm will get the solution  $\mathbf{w}^{t_1}$  and computes the most violated constraint  $\{c_i^{t_1}\}$  similarly as in Eq.(15) with  $f^{t_0}$  replaced by  $f^{t_1}$ , and this constraint will be added to problem (16). This procedure will be repeated until no constraint is violated by more than  $\epsilon$ . In this way, we construct a successive strengthening approximation series of the problem (12) by a series of cutting planes that cut off the current optimal solution from the feasible set [7].

### C. Multi-Class MRMMC

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

$$\begin{aligned} \min_{\{\mathbf{w}_p\}, \xi_i \geq 0} \quad & \frac{1}{2} \sum_{p=1}^k \|\mathbf{w}_p\|^2 + \frac{C_1}{nk} \sum_{p=1}^k \mathbf{f}_p^T \mathbf{L} \mathbf{f}_p + \frac{C_2}{n} \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & \forall i = 1, \dots, n, r = 1, \dots, k : \\ & \mathbf{w}_{y_i}^T \mathbf{x}_i + \delta_{y_i, r} - \mathbf{w}_r^T \mathbf{x}_i \geq 1 - \xi_i, \quad (17) \end{aligned}$$

where we assume the data set  $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  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  $r$ .  $\mathbf{f}_p = [f_p(\mathbf{x}_1), \dots, f_p(\mathbf{x}_n)]^T$  with  $f_p(\mathbf{x}_i) = \mathbf{w}_p^T \mathbf{x}_i$ .  $y_i = \arg \max_r \mathbf{w}_r^T \mathbf{x}_i$  is the cluster membership of  $\mathbf{x}_i$ .  $\delta_{uv} = 1$  if  $u = v$  and 0 otherwise.

To further simplify problem (17), we propose to absorb  $\delta_{y_i, r}$  into  $\xi_i$  and use a separate variable  $\xi_i^r$  for each constraint. Then problem (17) can be relaxed to

$$\begin{aligned} \min_{\{\mathbf{w}_p\}, \xi_i^r \geq 0} \quad & \frac{1}{2} \sum_{p=1}^k \|\mathbf{w}_p\|^2 + \frac{C_1}{nk} \sum_{p=1}^k \mathbf{f}_p^T \mathbf{L} \mathbf{f}_p + \frac{C_2}{nk} \sum_{i=1}^n \xi_i^r \\ \text{s.t.} \quad & \forall i = 1, \dots, n, r = 1, \dots, k : \\ & \max_{p \in \{1, 2, \dots, k\}} \mathbf{w}_p^T \mathbf{x}_i - \mathbf{w}_r^T \mathbf{x}_i \geq 1 - \xi_i^r, \end{aligned}$$

In order to avoid trivial solutions, we can also enforce the class balance constraint in [22] as

$$\forall p, q \in \{1, \dots, k\} : -l \leq \sum_{i=1}^n \mathbf{w}_p^T \mathbf{x}_i - \sum_{i=1}^n \mathbf{w}_q^T \mathbf{x}_i \leq l$$

Similar to two-class MRMMC, we may find the constraint in problem (18) is nonconvex and thus we can also resort to CCCP to solve it. We first rewrite problem (18) in a *1-slack*

variable formulation as

$$\begin{aligned} \min_{\{\mathbf{w}_p\}, \xi_i^r \geq 0} \quad & \frac{1}{2} \sum_{p=1}^k \|\mathbf{w}_p\|^2 + \frac{C_1}{nk} \sum_{p=1}^k \mathbf{f}_p^T \mathbf{L} \mathbf{f}_p + C_2 \xi \quad (18) \\ \text{s.t.} \quad & \forall i = 1, \dots, n, r = 1, \dots, k, c_i^r \in \{0, 1\} : \\ & \frac{1}{nk} \sum_{i,r} c_i^r \left( \max_{p \in \{1, 2, \dots, k\}} \mathbf{w}_p^T \mathbf{x}_i - \mathbf{w}_r^T \mathbf{x}_i \right) \geq \frac{1}{nk} \sum_{i,r} c_i^r - \xi, \\ & \forall p, q \in \{1, \dots, k\} : -l \leq \sum_{i=1}^n \mathbf{w}_p^T \mathbf{x}_i - \sum_{i=1}^n \mathbf{w}_q^T \mathbf{x}_i \leq l \end{aligned}$$

Now we introduce two *concatenated* vectors as

$$\tilde{\mathbf{w}} = [\mathbf{w}_1^T, \mathbf{w}_2^T, \dots, \mathbf{w}_p^T, \dots, \mathbf{w}_k^T]^T \quad (19)$$

$$\tilde{\mathbf{x}}_{ip} = [\mathbf{0}, \mathbf{0}, \dots, \mathbf{x}_i^T, \dots, \mathbf{0}]^T \quad (20)$$

where  $\mathbf{0}$  is a  $1 \times d$  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 (18) can be reformulated as

$$\begin{aligned} \min_{\tilde{\mathbf{w}}, \xi_i^r \geq 0} \quad & \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + \frac{C_1}{nk} \tilde{\mathbf{f}}^T \tilde{\mathbf{L}} \tilde{\mathbf{f}} + C_2 \xi \quad (21) \\ \text{s.t.} \quad & \forall i = 1, \dots, n, r = 1, \dots, k, c_i^r \in \{0, 1\} : \\ & \frac{1}{nk} \sum_{i,r} c_i^r \left( \max_{p \in \{1, 2, \dots, k\}} \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{ip} - \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{ir} \right) \geq \frac{1}{nk} \sum_{i,r} c_i^r - \xi, \\ & \forall p, q \in \{1, \dots, k\} : -l \leq \sum_{i=1}^n \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{ip} - \sum_{i=1}^n \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{iq} \leq l \end{aligned}$$

where  $\tilde{\mathbf{f}} = [\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{11}, \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{21}, \dots, \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{12}, \dots, \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{nk}]^T$ , and  $\tilde{\mathbf{L}} = \mathbf{L} \otimes \mathbf{I}_k$  with  $\mathbf{I}_k$  being the identity matrix of size  $k \times k$  and  $\otimes$  is the *Kronecker product*. Next, in order to apply CCCP, we should compute the subgradient of  $\max_{p \in \{1, \dots, k\}} \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{ip}$  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*<sup>5</sup>, *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

$$\partial \max_{p \in \{1, \dots, k\}} \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{ip} = \left\{ \sum_{r=1}^k \beta_{ir} \tilde{\mathbf{x}}_{ir} \mid \sum_r \beta_{ir} = 1 \right\} \quad (22)$$

where

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

In multiclass clustering, we usually expect that we assign the data into one unique cluster (we don't consider the multi-label case in this paper), *i.e.*, we expect there is only one

<sup>5</sup>See <http://www.ee.ucla.edu/ee236b/lectures/sg.pdf>, page 8.

active function when computing the subgradient in Eq.(22). So there is a unique  $p_*$  satisfying<sup>6</sup>

$$\tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{ip_*} = \max_{p \in \{1, \dots, k\}} \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{ip} \quad (23)$$

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

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

where

$$p_*^{(t)} = \arg \max_{p \in \{1, 2, \dots, k\}} \left( \tilde{\mathbf{w}}^{(t)} \right)^T \tilde{\mathbf{x}}_{ip} \quad (24)$$

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

$$\min_{\tilde{\mathbf{w}}, \xi, \xi_r \geq 0} \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + \frac{C_1}{nk} \tilde{\mathbf{f}}^T \tilde{\mathbf{L}} \tilde{\mathbf{f}} + C_2 \xi \quad (25)$$

$$s.t. \quad \forall i = 1, \dots, n, r = 1, \dots, k, c_i^r \in \{0, 1\} :$$

$$\frac{1}{nk} \sum_{i,r} c_i^r \left( \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{ip_*^{(t)}} - \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{ir} \right) \geq \frac{1}{nk} \sum_{i,r} c_i^r - \xi,$$

$$\forall p, q \in \{1, \dots, k\} : -l \leq \sum_{i=1}^n \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{ip} - \sum_{i=1}^n \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{iq} \leq l$$

Then similar to two-class *MRMMC*, we can apply the *cutting plane* method to solve problem (25), where at the  $s$ -th step, we can compute the most violated constraint  $C_i^{ts}$  by

$$\frac{1}{nk} \sum_{i,r} c_i^{rts} \left( \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{ip_*^{(t)}} - \tilde{\mathbf{w}}^T \tilde{\mathbf{x}}_{ir} \right) \geq \frac{1}{nk} \sum_{i,r} c_i^{rts} - (\xi + \epsilon) \quad (26)$$

where

$$c_i^{rs} = \begin{cases} 1, & \text{if } \tilde{\mathbf{x}}_{ip_*^{(t)}} - \tilde{\mathbf{x}}_{ir} < 1 \\ 0, & \text{otherwise} \end{cases} \quad (27)$$

#### IV. EXPERIMENTS

In this section we will present a set of experiments to validate the effectiveness of our *MRMMC* method.

##### A. Datasets

We use three categories of data sets in our experiments, which are selected to cover a wide range of properties. Specifically, those data sets include:

(1). **UCI data**. We perform experiments on four UCI data sets: **ionosphere**, **digits**, **letter** and **satellite**<sup>7</sup>. For the **digits** data, we follow the experimental setup of [20] and focus on those pairs (3 vs 8, 1 vs 7, 2 vs 7, and 8 vs 9) that are difficult to differentiate. For the **letter** and **satellite** data sets, we use their first two classes only [20].

(2). **Text Data**. We perform experiments on four text data sets: **20-newsgroup**<sup>8</sup>, and **RCVI** [8]. For **20-newsgroup**, we choose the topic *rec* which contains *autos*, *motorcycles*,

*baseball* and *hockey* from the version 20-news-18828. For **RCVI**, we just use the data samples with the highest four topic codes (CCAT, ECAT, GCAT, and MCAT) in the ‘‘Topic Codes’’ hierarchy in the training set.

(3). **Digits Data**. We use images of digits 1, 2, 3 and 4 in the USPS<sup>9</sup> handwritten  $16 \times 16$  digits image set, which contain 1005, 731, 658 and 652 samples in each class, with a total of 3046 samples. We also perform experiments on the **MNIST** digits data sets, we give a more thorough comparison by considering all 45 pairs of digits 0 – 9.

Table I  
CLUSTERING ACCURACY(%) COMPARISONS FOR TWO-CLASS PROBLEMS.

Data	MMC	GMC	SVR	CPMMC	MRMMC
Ionosphere	78.75	76.50	77.70	72.36	<b>79.03</b>
Letter	-	-	92.80	94.47	<b>95.56</b>
Satellite	-	-	96.82	98.48	<b>99.79</b>
Text-1	-	-	96.82	95.00	<b>98.34</b>
Text-2	-	-	93.99	96.28	<b>98.21</b>
Digits	-	-	98.18	99.38	<b>100.00</b>
MNIST	-	-	92.41	95.71	<b>97.43</b>
USPS	-	-	-	94.12	<b>95.32</b>
20Newsgroup	-	-	-	70.63	<b>73.45</b>
RCVI	-	-	-	61.97	<b>65.02</b>

##### B. Experimental Setups and Comparisons

We have conducted comprehensive performance evaluations by testing our method and comparing it with 5 other representative data clustering methods using the same data corpora. The algorithms that we evaluated are listed below.

(1). **Maximum Margin Clustering (MMC)** [16]. The implementation is the same as in [16]. For multi-class *MMC*, we just follow the experimental settings in [17]. The width of the Gaussian kernel is also set by grid search from  $\{0.1\sigma_0, 0.2\sigma_0, \dots, \sigma_0\}$  with  $\sigma_0$  being the range of distance between any two data points in the data set.

(2). **Generalized Maximum Margin Clustering (GMMC)** [14]. The implementation is the same as in [14].

(3). **Iterative Support Vector Regression (IterSVR)**<sup>10</sup> [20]. The initialization is based on *k-means* with randomly selected initial data centers, and the width of the Gaussian kernel is set in the same way as in *MMC*.

(5). **Cutting Plane Maximum Margin Clustering (CPMMC)**<sup>11</sup>[21][22]. The implementation is the same as in [21] and [22].

For our *MRMMC* algorithm, we set  $\epsilon = 0.01$ ,  $\alpha = 0.01$  in our experiments, and  $\mathbf{w}_0$  is randomly initialized. The class imbalance parameter  $l$  is set by grid search from the grid  $[0, 20]$  with granularity 1. The parameter  $C_1$ ,  $C_2$

<sup>9</sup>Available from <http://www.kernel-machines.org/data.html>

<sup>10</sup>The implementation code is downloaded from [http://www.cse.ust.hk/~twinsen/itMMC\\_code.zip](http://www.cse.ust.hk/~twinsen/itMMC_code.zip).

<sup>11</sup>The implementation code is downloaded from [http://binzhao02.googlepages.com/Code\\_MMC\\_v1.rar](http://binzhao02.googlepages.com/Code_MMC_v1.rar)

<sup>6</sup>If there is multiple  $p_*$  satisfying Eq.(23), we just randomly select one.

<sup>7</sup><http://mllearn.ics.uci.edu/MLRepository.html>

<sup>8</sup><http://people.csail.mit.edu/jrennie/20Newsgroups/>.

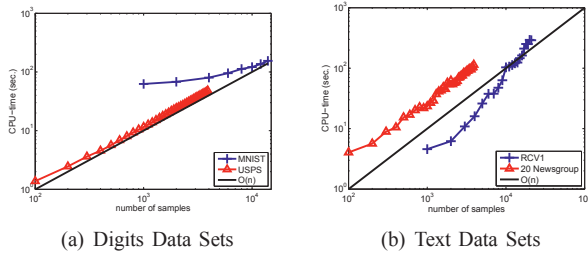


Figure 1. Times vs. Data Set Size Plots.

are searched from the exponential grid  $2^{[-8:1:6]}$ . The graph Laplacian is constructed in its standard way as in [15] using the Gaussian similarities. Note that the *GMMC* and *IterSVR* methods can only handle two-class problems.

### C. Clustering Results

The same as in previous papers [14][16][21][22], we also use *clustering accuracy* to evaluate the final clustering performance. The clustering results of all the algorithms are summarized in table I, where the results of the *k-means* algorithm are averaged over 10 independent runs with random initializations. The “-” in table I means that wither the algorithm is unable to handle multi-class cases (e.g., *GMC* and *SVR* for digits and text data sets, or the data set is too large for the algorithm to work out. From the table we can clearly see that our algorithm works better than other methods.

Besides, we also test the speed of our algorithm with respect to the scale of the data sets<sup>12</sup>. Fig.1 shows the log-log plots of speed vs. data set size on relatively large scale data sets. From the figure we can see that the computational time of our algorithm scales approximately linear with respect to the data set size, which is similar to *CPMMC* but much faster than traditional methods.

## V. CONCLUSIONS

In this paper we propose a novel *manifold regularized maximum margin clustering (MRMMC)* method, which (1) extends the classical *MMC* framework by incorporating the manifold information; (2) employs an improved solution method with better theoretical guarantee. The experimental results show that our algorithm can get better performances without the loss in speed.

### ACKNOWLEDGEMENT

The work is partially supported by NSF grants NSF grants DMS-0844513 and CCF-0830659.

<sup>12</sup>All experiments are implemented on MATLAB under MS Windows with CPU 2.2 GHz and 4G RAM

## REFERENCES

- [1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.
- [2] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.
- [3] O. Chapelle and A. Zien. Semi-supervised classification by low density separation. In *Proceedings of AISTATS*, 2005.
- [4] P. M. Cheung and J. T. Kowk. A regularization framework for multiple-instance learning. In *Proceedings of ICML*, 2006.
- [5] B. Heisele, P. Ho, and T. Poggio. Face recognition with support vector machines: Global versus component-based approach. pages 688–694, 2001.
- [6] Y. Hu, J. Wang, N. Yu, and X.-S. Hua. Maximum margin clustering with pairwise constraints. In *Proceedings of ICDM*, 2008.
- [7] J. E. Kelley. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial Applied Mathematics*, 8:703–712, 1960.
- [8] D. D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.
- [9] Y. Li and D. Huttenlocher. Learning for stereo vision using the structured support vector machine. In *Proceedings of CVPR*, 2008.
- [10] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.
- [11] B. Schölkopf, A. J. Smola, and K. R. Müller. Kernel principal component analysis. *Advances in kernel methods: support vector learning*, pages 327–352, 1999.
- [12] A. J. Smola, S. Vishwanathan, and T. Hofmann. Kernel methods for missing variables. In *Proceedings of AISTATS*, 2005.
- [13] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.
- [14] H. Valizadegan and R. Jin. Generalized maximum margin clustering and unsupervised kernel learning. In *Proceedings of NIPS*, pages 1417–1424, 2007.
- [15] F. Wang and C. Zhang. Label propagation through linear neighborhoods. In *Proceedings of ICML*, 2006.
- [16] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In *Proceedings of NIPS*, 2004.
- [17] L. Xu and D. Schuurmans. Unsupervised and semi-supervised multi-class support vector machines. In *National Conference on Artificial Intelligence (AAAI)*, 2005.
- [18] A. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15:915–936, 2003.
- [19] H. Zhang, A. C. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *Proceedings of CVPR*, pages 2126–2136, 2006.
- [20] K. Zhang, I. W. Tsang, and J. T. Kowk. Maximum margin clustering made practical. In *Proceedings of ICML*, pages 1119–1126, 2007.
- [21] B. Zhao, F. Wang, and C. Zhang. Efficient maximum margin clustering via cutting plane algorithm. In *Proceedings of The 8th SDM*, pages 751–762, 2008.
- [22] B. Zhao, F. Wang, and C. Zhang. Efficient multiclass maximum margin clustering. In *Proceedings of ICML*, pages 751–762, 2008.
- [23] W. Zhu, S. Wang, R.-S. Lin, and S. Levinson. Tracking of object with svm regression. In *Proceedings of CVPR*, pages 240–245, 2001.