

Maximum Margin Multiple Instance Clustering with its applications to Image and Text Clustering

Dan Zhang, Fei Wang, Luo Si, and Tao Li

Abstract—In multiple instance learning problems, patterns are often given as *bags* and each bag consists of some *instances*. Most of existing research in the area focuses on multiple instance classification and multiple instance regression, while very limited work has been conducted for multiple instance clustering. This paper formulates a novel framework as *Maximum Margin Multiple Instance Clustering (M³IC)* for multiple instance clustering. However, it is impractical to directly solve the optimization problem of M³IC. Therefore, M³IC is relaxed in the paper to enable an efficient optimization solution with a combination of the *Constrained Concave-Convex Procedure (CCCP)* and the *Cutting Plane* method. Furthermore, this paper presents some important properties of the proposed method and discusses the relationship between the proposed method and some other related ones. An extensive set of empirical results are shown to demonstrate the advantages of the proposed method against existing research for both effectiveness and efficiency.

Index Terms—Multiple Instance Clustering, Maximum Margin, Constrained Concave-Convex Procedure (CCCP), Cutting Plane

1 INTRODUCTION

In a traditional supervised learning problem, given a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathcal{X}$ is an instance, and $y_i \in \mathcal{Y}$ is the corresponding label, the main objective of most related algorithms is to learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ such that the prediction error rate on the unlabeled examples can be minimized.

Multiple instance learning (MIL) can be viewed as a variation of the traditional learning methods for problems with incomplete knowledge on the examples (or *instances*). In the MIL setting, patterns are given as *bags*, and each bag consists of some instances. Most of the current MIL research focuses on solving binary classification [2][7][23] and regression problems [1][24][40]. In a binary multiple instance classification problem, the labels are assigned to bags, rather than instances. A typical assumption for binary multiple instance classification is that a bag should be labeled as positive if at least one of its instances is positive; and negative if all of its instances are negative. Given a binary multiple instance classification problem, the main objective of the MIL learning algorithm is to learn a classifier based on the labeled bags, and then use this classifier to determine the labels for unlabeled bags. MIL has been widely employed in areas such as text mining [2], drug design [7], Localized Content Based Image Retrieval (LCBIR) [23], etc.

As another branch of machine learning, clustering [12] is a fundamental research topic in both data mining and machine learning, and plays an important role in applications such

as text mining, web analysis, marketing, etc [11]. It aims at dividing the whole dataset into several clusters of similar objects, or the hidden patterns in an unsupervised way, where these different patterns may correspond to different underlying data concepts.

However, almost all of the current clustering methods are designed for the traditional learning problems, while in many cases they should be better formulated as MIL ones.

Similar to LCBIR [23], in image clustering, there is natural ambiguity as to what portion of the image contains the clustering concept, where the concept can be a tiger, an elephant, etc, while most portion of the image may be irrelevant. In this case, we can treat each image as a bag, and each instance in this bag corresponds to a region in this image. It is clear that this application requires the solution of *Multiple Instance Clustering (MIC)* to help users to partition these bags.

As a popular topic in text mining, almost all of the previous text clustering methods [18] are designed for the traditional single instance learning problems, i.e., each document is treated as an instance. However, it is very likely that not all parts of a document are related to a specific topic. So, text clustering can be better formulated as a multiple-instance problem, where different articles are represented as different collections of overlapping text passages [2]. Then, each document can be regarded as a bag, and the passages in it are instances in this bag. To cluster these documents, represented as bags, we need the solution of MIC.

Recently, very limited research has been conducted for multiple instance clustering. In [39], the authors regard bags as atomic data items and use some forms of distance metric to measure the distances between bags. Then they adapt the K-medoids algorithm to cluster these bags based on the defined distances. Their method is efficient in some applications such as drug discovery. But, as claimed by [23], defining distances between bags in an unsupervised way may not be able to reflect their actual content differences. For example, two

- Dan Zhang is with the Department of Computer Science, Purdue University, West Lafayette, IN, 47906.
E-mail: danzhang2008@gmail.com
- Fei Wang is with Healthcare Transformation Group, IBM T. J. Watson Research Center at Hawthorne, NY, US. Luo Si is with the Department of Computer Science, Purdue University. Tao Li is with School of Computer Science & Information Sciences, Florida International University

pictures may share identical background and only differ in that one contains an elephant and the other contains a fox. By using the minimal Hausdorff distance to measure distances between bags [35], the distance between these two pictures will be very low even though their actual contents (or concepts) may differ. The calculation of the distances between bags is quite time consuming, since it needs to calculate all the distances between instances in different bags.

This paper proposes a novel *Maximum Margin Multiple Instance Clustering (M³IC)* framework. The new formulation aims at finding desired hyperplanes that maximize the margin differences on at least one instance per bag in an unsupervised way. The initial formulation of M³IC is a non-convex optimization problem, which is impossible to be solved directly. Therefore, we relax the original M³IC problem and propose a method – M³IC-MBM, which is a combination of *Constrained Concave-Convex Procedure (CCCP)* and *Cutting Plane* methods, to solve the relaxed optimization task. The two major contributions of this work are: 1. We designed a clustering method that can truly incorporate the assumption of MIL. 2. We proposed an optimization method for the task with the property of guaranteed convergence.

This paper substantially extends our preliminary work in [42]. Compared with the previous work, we derive the MIC framework from a different viewpoint, making it more clear. Moreover, in this paper, we derive the dual form of the corresponding objective function and give proofs to some important theorems. Furthermore, to validate the performance of our method, more experiments are conducted on more datasets and the analysis is more precise.

The rest of the paper is organized as follows: Section 2 introduces some related works. Section 3 formulates the novel M³IC problem and proposes an efficient method to solve it. Section 4 presents the experimental results. Section 5 points out some future directions. Section 6 concludes the whole paper.

2 RELATED WORKS

In this section, some related works are introduced. In Section 2.1, we will revisit the multiple instance classification algorithms. In Section 2.2, the only related work [39] will be reviewed. Since the proposed method is a generalization of Maximum Margin Clustering and employs the CCCP, we will introduce them separately in Section 2.3 and Section 2.4.

2.1 Multiple Instance Classification

The notion of multiple-instance learning was first introduced by Dietterich et al. [7] to deal with the drug activity prediction. After that, a lot of researchers have studied the binary multiple instance classification problem. The binary multiple instance classification algorithms can be roughly divided into three groups: the group that is specifically designed to solve multiple-instance problems, the group that tries to modify single-instance learning for MIL by introducing MIL constraints and the group that tries to convert MIL to a traditional single-instance problem and solve it using traditional learning methods.

For the first group, the first MIL method is APR [7], which represents positive instances by an axis-parallel rectangle in the feature space. Maron and Lozano-Pérez proposed a method called Diverse Density (DD) [20] [21], which intends to find a concept point in the feature space that resembles positive instance most, and classify instances according to the distances between the instances and this concept point. In [25] [40], the authors accelerated the DD method by using Expectation-Maximization (EM), and proposed EM-DD.

As for the second group. Andrews et al. [2] proposed two MIL formulations based on SVM [29], one (mi-SVM) for the instance-based classification and the other (MI-SVM) for bag-level classification. Since the MIL formulations are all non-convex, Gehler and Chapelle applied deterministic annealing to solve them and obtained a better local solution [9]. Gärtner et al. [10] proposed a kernel function directly for bags. Later, Kwok and Cheung [16] extended their work by proposing a marginalized MI kernel and converting the MIL problem from an incomplete data problem to a complete data problem. In [26], the authors make some modifications on the loss functions of single-instance SVM and focus more on the positive bags with smaller sizes.

For the third group, DD-SVM [5] selects a set of prototypes from the local solutions of DD method and then a SVM is trained based on the bag features summarized by these selected prototypes. In [6], bags are embedded into a feature space defined by instances, and a 1-norm SVM is applied to build the bag level classifiers.

Beside the binary multiple instance classification methods, the multiple-instance multiple-label problem [50] [17] [13] [41] is also a popular topic recently, where the labels are not restricted to be binary, but can be a vector. Some other special problems in binary multiple instance classification are also proposed, such as Semi-Supervised Multiple Instance Learning [23] [51] [45], Multiple Instance Active Learning [28] [44] and Multiple Instance Transfer Learning [43].

A common strait of these classification methods is that they need to utilize the supervised information of all the labeled bags to determine which part/parts are the desired object/objects. However, in MIC, such label information is not available. So, it is hard to determine the desired concepts underlying different clusters, which is the main obstacle to solve the MIC problem.

2.2 Multiple Instance Clustering

Recently, very limited research [39] addresses the task of MIC. In [39], the authors defined the distances between bags in three different ways, i.e., the maximal, minimal and the average Hausdorff distance [1][35][39]:

$$\begin{aligned} \max H(\mathbf{A}, \mathbf{B}) &= \max\{\max_{\mathbf{a} \in \mathbf{A}} \min_{\mathbf{b} \in \mathbf{B}} \|\mathbf{a} - \mathbf{b}\|, \max_{\mathbf{b} \in \mathbf{B}} \min_{\mathbf{a} \in \mathbf{A}} \|\mathbf{b} - \mathbf{a}\|\} \\ \min H(\mathbf{A}, \mathbf{B}) &= \min_{\mathbf{a} \in \mathbf{A}, \mathbf{b} \in \mathbf{B}} \|\mathbf{a} - \mathbf{b}\| \\ \text{ave}H(\mathbf{A}, \mathbf{B}) &= \frac{\sum_{\mathbf{a} \in \mathbf{A}} \min_{\mathbf{b} \in \mathbf{B}} \|\mathbf{a} - \mathbf{b}\| + \sum_{\mathbf{b} \in \mathbf{B}} \min_{\mathbf{a} \in \mathbf{A}} \|\mathbf{b} - \mathbf{a}\|}{|\mathbf{A}| + |\mathbf{B}|}, \quad (1) \end{aligned}$$

where \mathbf{A} and \mathbf{B} are two bags of instances, $\mathbf{A} = \{\mathbf{a}_1, \dots, \mathbf{a}_m\}$ and $\mathbf{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$, and $|\cdot|$ measures the cardinality of that set. The function $aveH(\cdot, \cdot)$ calculates the average distances between each instance in one bag and its nearest instance in the other one. In this way, bags can be considered as atomic data items, and the K-medoids algorithm can be employed to cluster these data items based on the distances between them.

Although these three methods define the distances between bags in different ways, they are all unsupervised methods. And without identifying the desired objects/concepts in each bag, the unsupervised distance definitions can not reveal the true distances between bags. Different from these methods, the proposed method solves this problem in a more elegant way by incorporating the cluster assignments into the MIC formulation, the detail of which will be explained in Section 3.

2.3 Maximum Margin Clustering

Maximum Margin learning is a key technique for classification and dimensionality reduction [31]. Recently, the idea of maximum margin learning has also been applied to data clustering, which is usually referred to as Maximum Margin Clustering (MMC). MMC was first proposed in [36]. In [36], the authors borrowed the idea of a standard machine learning principle - maximum margin principle, and used it for clustering. More precisely, they try to assign instances to two classes $\{-1, +1\}$ so that the separation between the two classes can be as large as possible, which can be formulated as:

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

where \mathbf{x}_i represents the i -th instance, and y_i is the associated label. \mathbf{w} is the separation hyperplane, while b represents the bias. ξ_i is the hinge loss associated with the i -th example and the corresponding \mathbf{w} and b . The loss function $\sum_{i=1}^n \xi_i$ is divided by n to better capture how the hinge loss parameter 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.

The maximum margin clustering method has a solid theoretical foundation and performs much better than the previous methods. But at first it can only deal with the two class separation problem. In the later work [37], it was extended to the multi-class case. However, one of the potential problems in these methods is that they are too time consuming. In [46], [48], [49] and [34], the authors have relaxed the original problem and solved it in an efficient way.

Although greatly improved, unfortunately, MMC can only deal with the traditional learning problem. In this paper, similar to MMC, we extend the Maximum Margin principle to MIC and propose the M³IC problem.

2.4 Constrained Concave Convex Programming

The concave-convex procedure (CCP) [38] is an optimization method for solving the non-convex problem whose objective

function can be expressed as the difference of convex functions. While [38] only considered linear constraints, [27] put forward the constrained concave-convex procedure (CCCP), that can solve the optimization problems with a concave-convex objective function under concave-convex constraints. Mathematically, suppose we want to solve the following optimization problem [27]

$$\begin{aligned} \min_{\mathbf{z}} \quad & f_0(\mathbf{z}) - g_0(\mathbf{z}) \\ \text{s.t.} \quad & f_i(\mathbf{z}) - g_i(\mathbf{z}) \leq v_i, i = 1, \dots, n, \end{aligned} \quad (3)$$

where f_i and g_i are convex functions on vector space \mathcal{Z} and $v_i \in \mathbb{R}$ 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} , which is $T_1\{f, \mathbf{z}\}(\mathbf{z}') = f(\mathbf{z}) + \partial_{\mathbf{z}}f(\mathbf{z})(\mathbf{z}' - \mathbf{z})$, and $\partial_{\mathbf{z}}f(\mathbf{z})$ is the gradient of the function f at \mathbf{z} . Given an initial point \mathbf{z}_0 , CCCP computes \mathbf{z}_{t+1} from \mathbf{z}_t by substituting $g_i(\mathbf{z})$ 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 the optimal solution of the following relaxed problem

$$\begin{aligned} \min_{\mathbf{z}} \quad & f_0(\mathbf{z}) - T_1\{g_0, \mathbf{z}_t\}(\mathbf{z}) \\ \text{s.t.} \quad & f_i(\mathbf{z}) - T_1\{g_i, \mathbf{z}_t\}(\mathbf{z}) \leq v_i, i = 1, \dots, n. \end{aligned} \quad (4)$$

The above procedure will iterate until \mathbf{z}_t converges, and [27] has proved that the CCCP will finally converge to a local minimum of problem (3).

3 THE PROPOSED METHOD

3.1 Problem Statement

Suppose a set of n bags, $\{\mathbf{B}_i, i = 1, 2, \dots, n\}$ is given, and the instances in the bag \mathbf{B}_i are denoted as $\{\mathbf{B}_{i1}, \mathbf{B}_{i2}, \dots, \mathbf{B}_{in_i}\}$, where n_i is the total number of instances in this bag. The goal of MIC is to partition this given dataset into k clusters such that the concepts in different clusters can be “distinct” from each other. A $1 \times n$ vector \mathbf{y} is used here to denote the cluster assignment array, with y_i being the cluster assignment for bag \mathbf{B}_i .

3.2 Motivation

In Section 2.1, we have reviewed some binary multiple instance classification methods. Most of them are developed based on the typical assumption that each positive bag contains at least one positive instance, while instances in negative bags are all negative. Given labels, the multiple instance classifier can be learnt by simply looking for the common concepts in positive bags, which are also far away from all of the negative instances. Since any instance in a positive bag can be either positive or negative, looking for the real positive instance is very difficult and can not be formulated as a convex optimization problem.

MIC is a much harder problem than the previous binary multiple instance classification problems. In MIC, no label information is given beforehand, and therefore it is even harder to find the different concepts representing different clusters. Another problem is that almost all of the traditional multiple instance learning methods can only classify the unlabeled bags as either positive or negative ones. This is because almost all of

the current multiple instance learning methods were designed to solve application problems where only the relevances of objects need to be determined, such as in LCBIR and drug discovery. But in MIC, we are facing a problem where the whole dataset needs to be categorized into several, probably more than two, clusters.

So, in MIC, we need to both identify the most relevant instances and cluster these instances into several different groups. Both of these two objectives are very difficult, let alone couple them together. Fortunately, the idea of MMC [36][37][48][49], which was introduced in Section 2.3, gives us some inspirations to solve this problem. The clustering principle we investigate here is to find a desired labeling so that if one were to subsequently run a multiple instance SVM [2], the margin obtained would be maximal over all possible labels. The details of the proposed algorithm will be elaborated in the following sections.

3.3 Formulation

In this subsection, the M³IC problem will be formulated. For each class $p \in \{1, \dots, k\}$, we define a weight vector \mathbf{w}_p . We wish to solve for a labeling \mathbf{y} that leads to the maximum (soft) margin. Straightforwardly, one could attempt to tackle the following optimization problem directly:

$$\begin{aligned} \min_{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k, y_1, y_2, \dots, y_n, \xi_i \geq 0} \quad & \frac{1}{2} \sum_{p=1}^k \|\mathbf{w}_p\|^2 + \frac{C}{n} \sum_{i=1}^n \xi_i \quad (5) \\ \text{s.t.} \quad & \forall i \in \{1, \dots, n\}, \forall p_i \in \{1, 2, \dots, k\} \setminus y_i \\ & \bigcup_{j=1}^{n_i} I_{(\mathbf{w}_{y_i}^T \mathbf{B}_{ij} - \mathbf{w}_{p_i}^T \mathbf{B}_{ij} \geq 1 - \xi_i)} = 1, \end{aligned}$$

where $\sum_{i=1}^n \xi_i$ is divided by n to make this formulation scalable to the dataset size and “ \setminus ” means ruling out. $I_{(\text{expression})}$ is an indication function. It is true when the “*expression*” holds. “ \bigcup ” is an union operator. \mathbf{B}_{ij} is defined in Section 3.1. It can be seen that the large margin constraint is imposed on at least one instance per bag. The reason for this is similar to the assumption in the binary multiple instance classification problems, i.e., there should be at least one common concept in each bag. But there are also some differences. In binary multiple instance classification problems, negative bags are often referred to as irrelevant information, which may not represent any interesting group (For example, in LCBIR, negative bags are always some arbitrary background pictures). However, in MIC, we need to partition the whole dataset into k groups, and therefore the common concept constraint should be imposed on all the clusters within the whole dataset.

The form of problem (5) is very complicated. However, as shown in Theorem 1, it can be transformed to a much more simplified form, in which the number of variables that need to be optimized could be reduced as well.

Theorem 1: Problem (5) is equivalent to:

$$\begin{aligned} \min_{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k, \xi_i \geq 0} \quad & \frac{1}{2} \sum_{p=1}^k \|\mathbf{w}_p\|^2 + \frac{C}{n} \sum_{i=1}^n \xi_i \quad (6) \\ \text{s.t.} \quad & i = 1, \dots, n, \\ & \max_{j \in \mathbf{B}_i} (\mathbf{w}_{u_{ij}^*}^T \mathbf{B}_{ij} - \mathbf{w}_{v_{ij}^*}^T \mathbf{B}_{ij}) \geq 1 - \xi_i, \end{aligned}$$

where, $u_{ij}^* = \arg \max_p (\mathbf{w}_p^T \mathbf{B}_{ij})$, $v_{ij}^* = \arg \max_{p \setminus u_{ij}^*} (\mathbf{w}_p^T \mathbf{B}_{ij})$ ¹ and the optimal value for y_i , which is \hat{y}_i , equals $\arg \max_{y_i \in \{1, 2, \dots, k\}} \mathbf{w}_{y_i}^T \mathbf{B}_{ij^*}$, where $j^* = \arg \max_{j \in \mathbf{B}_i} [\mathbf{w}_{u_{ij}^*}^T \mathbf{B}_{ij} - \mathbf{w}_{v_{ij}^*}^T \mathbf{B}_{ij}]$.

Proof:

It is clear that $\bigcup_{j=1}^{n_i} I_{(\mathbf{w}_{y_i}^T \mathbf{B}_{ij} - \mathbf{w}_{p_i}^T \mathbf{B}_{ij} \geq 1 - \xi_i)} = 1$ is equivalent to $I_{(\max_{j \in \mathbf{B}_i} (\mathbf{w}_{y_i}^T \mathbf{B}_{ij} - \mathbf{w}_{p_i}^T \mathbf{B}_{ij}) \geq 1 - \xi_i)} = 1$. So, the constraints in problem (5) can be transformed to:

$$\begin{aligned} \forall i \in \{1, \dots, n\}, \forall p_i \in \{1, 2, \dots, k\} \setminus y_i \\ \max_{j \in \mathbf{B}_i} (\mathbf{w}_{y_i}^T \mathbf{B}_{ij} - \mathbf{w}_{p_i}^T \mathbf{B}_{ij}) \geq 1 - \xi_i. \quad (7) \end{aligned}$$

Since the second part of the problem (5) is $\frac{C}{n} \sum_{i=1}^n \xi_i$, it is evident that the optimal value for ξ_i can be expressed as:

$$\hat{\xi}_i = \min_{y_i \in \{1, 2, \dots, k\}} \max_{p_i \in \{1, 2, \dots, k\} \setminus y_i} \max_{j \in \mathbf{B}_i} 1 - (\mathbf{w}_{y_i}^T \mathbf{B}_{ij} - \mathbf{w}_{p_i}^T \mathbf{B}_{ij}). \quad (8)$$

For any $j \in \mathbf{B}_i$, if $\mathbf{w}_{y_i}^T \mathbf{B}_{ij} < \mathbf{w}_{p_i}^T \mathbf{B}_{ij}$, then, the right handside of this equation will be larger than one. So, \hat{y}_i , which is the optimal solution of y_i , can only take values from the u_{ij}^* s that achieves the largest output within the instances in bag \mathbf{B}_i , i.e., $\hat{y}_i \in \{u_{ij}^*, j = 1, \dots, n_i\}$ and $u_{ij}^* = \arg \max_{y_{ij} \in \{1, 2, \dots, k\}} \mathbf{w}_{y_{ij}}^T \mathbf{B}_{ij}$. Similarly, the optimal value \hat{p}_i would also take values from v_{ij}^* s that satisfy $v_{ij}^* = \arg \max_{p_{ij} \setminus u_{ij}^*} \mathbf{w}_{p_{ij}}^T \mathbf{B}_{ij}$. Since the optimization of $[1 - (\mathbf{w}_{y_i}^T \mathbf{B}_{ij} - \mathbf{w}_{p_i}^T \mathbf{B}_{ij})]$ is also taken w.r.t. all the instances in \mathbf{B}_i , it is clear that $\hat{y}_i = \arg \max_{y_i \in \{1, 2, \dots, k\}} \mathbf{w}_{y_i}^T \mathbf{B}_{ij^*}$, where $j^* = \arg \max_{j \in \mathbf{B}_i} [\mathbf{w}_{u_{ij}^*}^T \mathbf{B}_{ij} - \mathbf{w}_{v_{ij}^*}^T \mathbf{B}_{ij}]$. Integrating j^* and the solution of \hat{y}_i to constraint (7), this constraint can be transformed to:

$$\begin{aligned} i = 1, \dots, n, \\ \max_{j \in \mathbf{B}_i} (\mathbf{w}_{u_{ij}^*}^T \mathbf{B}_{ij} - \mathbf{w}_{v_{ij}^*}^T \mathbf{B}_{ij}) \geq 1 - \xi_i. \end{aligned}$$

□

By employing Theorem 1, compared with the original optimization problem (5), the number of variables that need to be optimized are reduced by n . Besides the direct derivative from problem (5), this transformed problem (6) can also be perceived from the large margin perspective. And the new explanation will serve as the basis of understanding the algorithm that will be proposed later.

To give this explanation, first of all, the *Bag Margin (BM)* for a bag \mathbf{B}_i is defined as:

$$\max_{j \in \mathbf{B}_i} (\mathbf{w}_{u_{ij}^*}^T \mathbf{B}_{ij} - \mathbf{w}_{v_{ij}^*}^T \mathbf{B}_{ij}). \quad (9)$$

1. This equation can also be written as: $v_{ij}^* = \arg \max_{p \neq u_{ij}^*} (\mathbf{w}_p^T \mathbf{B}_{ij})$

Then problem (6) can be interpreted as: instead of labeling all the samples by running an SVM implicitly over all possible labels as that in MMC [36], in M^3IC , we try to find a labeling on bags that results in several classifiers that maximize margins on bags, and, for each bag, its BM is determined by its most “discriminative” instance.

Beside the large margin constraint, one has to enforce the class balance constraint to avoid the trivially “optimal” solution, where all of the bags are assigned to one cluster and achieve an unbounded margin, and to minimize the impact of some outliers, which may form some very small clusters [36]. Then, the formulation of M^3IC can be formally defined as:

$$\begin{aligned} \min_{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k, \xi_i \geq 0} \quad & \frac{1}{2} \sum_{p=1}^k \|\mathbf{w}_p\|^2 + \frac{C}{n} \sum_{i=1}^n \xi_i \quad (10) \\ \text{s.t.} \quad & i = 1, \dots, n, \\ & \max_{j \in \mathbf{B}_i} (\mathbf{w}_{u_{ij}^*}^T \mathbf{B}_{ij} - \mathbf{w}_{v_{ij}^*}^T \mathbf{B}_{ij}) \geq 1 - \xi_i \\ & \forall p, q \in \{1, 2, \dots, k\} \\ -l \leq \sum_{i=1}^n \sum_{j \in \mathbf{B}_i} I_{ij} \mathbf{w}_p^T \mathbf{B}_{ij} - \sum_{i=1}^n \sum_{j \in \mathbf{B}_i} I_{ij} \mathbf{w}_q^T \mathbf{B}_{ij} \leq l. \end{aligned}$$

Here, I_{ij^*} equals 1 if $j^* = \arg \max_{j \in \mathbf{B}_i} (\mathbf{w}_{u_{ij}^*}^T \mathbf{B}_{ij} - \mathbf{w}_{v_{ij}^*}^T \mathbf{B}_{ij})$, and otherwise 0. l is a parameter that controls the cluster balance as in [36]. It is clear that, in this formulation, these two sets of constraints are imposed only on the instances that determine the bag margins of their corresponding bags. Once these “witness” instances have been identified, the other instances become irrelevant. If we can obtain results from problem (10), the cluster assignment of a specific bag \mathbf{B}_i can be determined by $y_i = \arg \max_p \sum_{j \in \mathbf{B}_i} I_{ij} \mathbf{w}_p^T \mathbf{B}_{ij}$.

However, it is implausible to get the exact solution of problem (10) efficiently. In the following sections, we will state the reason, and suggest a method to approximate it.

3.4 M^3IC -MBM

For the first set of constraints in problem (10), i.e., $\max_{j \in \mathbf{B}_i} (\mathbf{w}_{u_{ij}^*}^T \mathbf{B}_{ij} - \mathbf{w}_{v_{ij}^*}^T \mathbf{B}_{ij}) \geq 1 - \xi_i$, the convexity of $\mathbf{w}_{u_{ij}^*}^T \mathbf{B}_{ij}$ cannot be determined and the “max” function also makes this constraint non-convex. For the second set of constraints, the indication function I_{ij} makes these constraints non-convex. Therefore, it is impractical to search for the global optimal solution of the optimization problem (10).

To solve this problem, in this section, we first relax the original problem, and then an efficient method – M^3IC -MBM is proposed to solve this resulting optimization problem.

3.4.1 Relaxation

As previously discussed, the non-convexity of problem (10) are mainly caused by the two sets of constraints. So, we consider to relax these two sets of constraints, so that a solution that is as close to that of the original problem as possible can be approximated.

First of all, its first set of constraints is relaxed. To achieve this goal, a new notion – *Modified Bag Margin (MBM)* is

introduced. For the bag \mathbf{B}_i , *MBM* is defined as:

$$\max_{j \in \mathbf{B}_i} (\max_u \mathbf{w}_u^T \mathbf{B}_{ij} - \text{mean}_{v \setminus u_{ij}^*} (\mathbf{w}_v^T \mathbf{B}_{ij})), \quad (11)$$

where, the “mean” function calculates the average value of the input function with respect to the subscript variable. As another way of defining bag margins, MBM can be considered as a relaxation of BM, which would slightly increase the feasible domain of problem (10), but avoid dealing with the intractable second maximal function v_{ij}^* .

Replacing BM with MBM, the first set of constraints in problem (10) becomes: $\max_{j \in \mathbf{B}_i} (\max_u \mathbf{w}_u^T \mathbf{B}_{ij} - \text{mean}_{v \setminus u_{ij}^*} (\mathbf{w}_v^T \mathbf{B}_{ij})) \geq 1 - \xi_i$. This is also equivalent to $\frac{k}{k-1} \max_{j \in \mathbf{B}_i} (\max_u \mathbf{w}_u^T \mathbf{B}_{ij} - \text{mean}_v (\mathbf{w}_v^T \mathbf{B}_{ij})) \geq 1 - \xi_i$.

For the second constraint in problem (10), the non-convexity is mainly brought in by the indication function I_{ij} . So, we relax it, and rewrite this constraint as follows:

$$\begin{aligned} \forall p, q \in \{1, 2, \dots, k\} \quad (12) \\ -l \leq \sum_{i=1}^n \sum_{j \in \mathbf{B}_i} \frac{1}{n_i} \mathbf{w}_p^T \mathbf{B}_{ij} - \sum_{i=1}^n \sum_{j \in \mathbf{B}_i} \frac{1}{n_i} \mathbf{w}_q^T \mathbf{B}_{ij} \leq l. \end{aligned}$$

Now, the two constraints in the original problem have already been relaxed. Furthermore, for the convenience of computation, without loss of generality, we introduce two *concatenated* vectors as:

$$\begin{aligned} \tilde{\mathbf{w}} &= [\mathbf{w}_1^T, \mathbf{w}_2^T, \dots, \mathbf{w}_p^T, \dots, \mathbf{w}_k^T]^T \quad (13) \\ \mathbf{B}_{ij(p)} &= [\mathbf{0}, \mathbf{0}, \dots, \mathbf{B}_{ij}^T, \dots, \mathbf{0}]^T, \end{aligned}$$

² where, $\mathbf{0}$ is a $1 \times d$ zero vector, and d is the dimension of \mathbf{B}_{ij} . In $\mathbf{B}_{ij(p)}$, only the $(p-1)d$ to pd -th elements are nonzero and equals \mathbf{B}_{ij} and it is clear that $\tilde{\mathbf{w}}^T \mathbf{B}_{ij(p)} = \mathbf{w}_p^T \mathbf{B}_{ij}$.

After the relaxation of the two constraints in Eq. (11), Eq. (12), and the introduction of the two *concatenated* vectors in Eq.(13), the original problem (10) can be transformed to:

$$\begin{aligned} \min_{\tilde{\mathbf{w}}, \xi_i \geq 0} \quad & \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + \frac{C}{n} \sum_i \xi_i \quad (14) \\ \text{s.t.} \quad & i = 1, \dots, n, \\ & \frac{k}{k-1} \max_{j \in \mathbf{B}_i} (\max_u \tilde{\mathbf{w}}^T \mathbf{B}_{ij(u)} - \text{mean}_v (\tilde{\mathbf{w}}^T \mathbf{B}_{ij(v)})) \\ & \geq 1 - \xi_i \\ & \forall p, q \in \{1, 2, \dots, k\} \\ -l \leq \sum_{i=1}^n \sum_{j \in \mathbf{B}_i} \frac{1}{n_i} \tilde{\mathbf{w}}^T (\mathbf{B}_{ij(p)} - \mathbf{B}_{ij(q)}) \leq l. \end{aligned}$$

In the following sections, we will propose a method – *M^3IC with Modified Bag Margin (M^3IC -MBM)*, which is characterized by an outer CCCP iteration and an inner Cutting Plane iteration, to solve this relaxed problem.

² Please note that, after this reformulation, $\tilde{\mathbf{w}}$ and $\mathbf{B}_{ij(p)}$ are still two one dimensional column vectors, rather than matrices.

3.4.2 CCCP Decomposition

Although some relaxations and simplifications have been made on the original problem, its first set of constraints is still non-convex. Fortunately, it is the difference between two convex functions. As introduced in Section 2.4, the constrained concave-convex procedure (CCCP) is just devised to solve this kind of optimization problems [27]. Next, we will show how to use CCCP to solve this problem.

To simplify the formulation, let $f(\tilde{\mathbf{w}}, i)$ be $\frac{k}{k-1} \max_{j \in \mathbf{B}_i} g(\tilde{\mathbf{w}}, i, j)$ and $g(\tilde{\mathbf{w}}, i, j)$ be $\max_u \tilde{\mathbf{w}}^T \mathbf{B}_{ij(u)} - \text{mean}_v(\tilde{\mathbf{w}}^T \mathbf{B}_{ij(v)})$. Then, the first set of constraints in problem (14) turns to: $f(\tilde{\mathbf{w}}, i) \geq 1 - \xi_i$. It is obvious that this set of constraints is, although not convex, the difference of two convex functions, which could be solved efficiently by using CCCP.

Given an initial point $\tilde{\mathbf{w}}^{(0)}$, CCCP computes $\tilde{\mathbf{w}}^{(t+1)}$ from $\tilde{\mathbf{w}}^{(t)}$ ³ iteratively by replacing $f(\tilde{\mathbf{w}}, i)$ with its first order Taylor expansions at $\tilde{\mathbf{w}}^{(t)}$, and solves the resulting quadratic programming problem, until convergence.

Therefore, to use CCCP, we should first calculate the first-order Taylor expansion of $f(\tilde{\mathbf{w}}, i)$ at $\tilde{\mathbf{w}}^{(t)}$. However, $f(\tilde{\mathbf{w}}, i)$ is a non-smooth function w.r.t. $\tilde{\mathbf{w}}$. So, we replace its gradient with the corresponding subgradient⁴ as follows:

$$\begin{aligned} & \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} \Big|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}} \\ &= \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial g(\tilde{\mathbf{w}}, i, j)} \times \frac{\partial g(\tilde{\mathbf{w}}, i, j)}{\partial \tilde{\mathbf{w}}} \Big|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}} \\ &= \sum_{j \in \mathbf{B}_i} \left(z_{ij}^{(t)} \times \frac{k}{k-1} \left(\sum_{r=1}^k \gamma_{ijr}^{(t)} \mathbf{B}_{ij(r)} - 1/k \sum_{p=1}^k \mathbf{B}_{ij(p)} \right) \right), \end{aligned} \quad (15)$$

where,

$$z_{ij}^{(t)} = \begin{cases} 1, & \text{if } j = \arg \max_{j \in \mathbf{B}_i} g(\tilde{\mathbf{w}}^{(t)}, i, j) \\ 0, & \text{otherwise} \end{cases}, \quad (16)$$

and

$$\gamma_{ijr}^{(t)} = \begin{cases} 1, & \text{if } r = \arg \max_{r \in \{1, 2, \dots, k\}} (\tilde{\mathbf{w}}^{(t)})^T \mathbf{B}_{ij(r)} \\ 0, & \text{otherwise} \end{cases}. \quad (17)$$

3. We use the superscript t to denote that the result is obtained from the t -th CCCP iteration, i.e., $\tilde{\mathbf{w}}^{(t)}$ is the optimized weight vector from the t -th CCCP iteration step.

4. Please note that for non-smooth functions, at the non-smooth points, we cannot find its gradient as we usually do on differentiable functions. So, the notion of the subgradient is introduced to solve this problem. At the non-smooth point, the subgradient takes values from an interval that satisfies some specific constraints [4]. In practical use, we can simply choose one value from this interval.

Then, we decompose $f(\tilde{\mathbf{w}}, i)$ at $\tilde{\mathbf{w}}^{(t)}$ as:

$$\begin{aligned} & f(\tilde{\mathbf{w}}, i) \\ &= f(\tilde{\mathbf{w}}^{(t)}, i) + (\tilde{\mathbf{w}} - \tilde{\mathbf{w}}^{(t)})^T \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} \Big|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}} \\ &= \tilde{\mathbf{w}}^T \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} \Big|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}} + \frac{k}{k-1} \\ & \quad \max_{j \in \mathbf{B}_i} \left(\max_u \left((\tilde{\mathbf{w}}^{(t)})^T \mathbf{B}_{ij(u)} \right) - \text{mean}_v \left((\tilde{\mathbf{w}}^{(t)})^T \mathbf{B}_{ij(v)} \right) \right) \\ & \quad - (\tilde{\mathbf{w}}^{(t)})^T \times \\ & \quad \sum_{j \in \mathbf{B}_i} \left(z_{ij}^{(t)} \times \frac{k}{k-1} \left(\sum_{r=1}^k \gamma_{ijr}^{(t)} \mathbf{B}_{ij(r)} - 1/k \sum_{p=1}^k \mathbf{B}_{ij(p)} \right) \right) \\ &= \tilde{\mathbf{w}}^T \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} \Big|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}}. \end{aligned} \quad (18)$$

So, for the t -th CCCP iteration, by replacing $f(\tilde{\mathbf{w}}, i)$ in problem (14) with Eq. (18), this problem can be transformed to:

$$\begin{aligned} & \min_{\tilde{\mathbf{w}}, \xi_i \geq 0} \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + \frac{C}{n} \sum_i \xi_i \\ & \text{s.t. } i = 1, \dots, n, \\ & \quad \tilde{\mathbf{w}}^T \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} \Big|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}} \geq 1 - \xi_i \\ & \quad \forall p, q \in \{1, 2, \dots, k\} \\ & \quad -l \leq \sum_{i=1}^n \sum_{j \in \mathbf{B}_i} \frac{1}{n_i} \tilde{\mathbf{w}}^T (\mathbf{B}_{ij(p)} - \mathbf{B}_{ij(q)}) \leq l. \end{aligned} \quad (19)$$

3.4.3 Cutting Plane

Directly solving the problem (19) as a *Quadratic Programming (QP)* problem would be a time consuming work when the number of bags is large. So, in this section, we seek a method that would greatly accelerate the optimization procedure. This method was originally proposed in [14], and was shown to be effective and efficient for solving similar tasks.

There are n slack variables ξ_i in problem (19). As in [14], to solve this problem efficiently, the 1-slack form of problem (19) is first derived. More precisely, we introduce a single slack variable ξ and rewrite the problem (19) as:

$$\begin{aligned} & \min_{\tilde{\mathbf{w}}, \xi \geq 0} \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + C\xi \\ & \text{s.t. } \forall \mathbf{c} \in \{0, 1\}^n \\ & \quad \frac{1}{n} \tilde{\mathbf{w}}^T \sum_{i=1}^n \mathbf{c}_i \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} \Big|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}} \geq \frac{1}{n} \sum_{i=1}^n \mathbf{c}_i - \xi \\ & \quad \forall p, q \in \{1, 2, \dots, k\} \\ & \quad -l \leq \sum_{i=1}^n \sum_{j \in \mathbf{B}_i} \frac{1}{n_i} \tilde{\mathbf{w}}^T (\mathbf{B}_{ij(p)} - \mathbf{B}_{ij(q)}) \leq l. \end{aligned} \quad (20)$$

It can be proved that the solution to problem (20) is identical to problem (19) with $\xi = \frac{1}{n} \sum_{i=1}^n \xi_i$ (similar to [14]).

Now the problem turns to how to solve problem (20) efficiently, which is convex, but has exponential number of constraints because of the large number of feasible \mathbf{c} . To solve this problem, we employ an adaption of the cutting plane algorithm [15], whose main objective is to find a small subset

Algorithm: M³IC-MBM
Input:
1. bags $\{\mathbf{B}_1, \dots, \mathbf{B}_n\}$
2. parameters: regularization constant C , CCCP solution precision ϵ_1 , cutting plane solution precision ϵ_2 , cluster number k , cluster size balance l
Output:
The cluster assignment \mathbf{y}
CCCP Iterations:
1. Construct $\tilde{\mathbf{B}} = \{\mathbf{B}_{ij(r)}\}$
2. Initialize $\tilde{\mathbf{w}}^0, t=0, \Delta J = 10^{-3}, J^{-1} = 10^{-3}$
3. while $\Delta J/J^{t-1} > \epsilon_1$ do
4. Derive problem (25). Set the constraint set $\Omega = \phi, \forall 1 \leq i \leq n, \mathbf{c}_{j(i)}=0, s = -1$
Cutting Plane Iterations:
5. while H^{t_s} is true do
6. $s = s + 1$
7. Get $(\tilde{\mathbf{w}}^{(t_s)}, \xi^{(t_s)})$ by solving (25) under Ω^{t_s}
8. Compute the most violated bags, i.e., $\mathbf{c}_i^{t_s}$, by
$\mathbf{c}_i^{t_s} = \begin{cases} 1, & \text{if } (\tilde{\mathbf{w}}^{(t_s)})^T \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} \Big _{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t_s)}} \leq 1 \\ 0, & \text{otherwise} \end{cases}$
and update the constraint set Ω^{t_s} by $\Omega^{t_{s+1}} = \Omega^{t_s} \cup \mathbf{c}_i^{t_s}$.
9. end while
10. $t = t + 1$
11. $\tilde{\mathbf{w}}^{(t)} = \tilde{\mathbf{w}}^{(t-1)s}$
12. $\Delta J = J^{t-1} - J^t$
13. end while
14. Cluster Assignment:
For bag $\mathbf{B}_i, y_i = \arg \max_p (\tilde{\mathbf{w}}^{(t)})^T \mathbf{B}_{ij^*(p)}$, where $j^* = \arg \max_{j \in \mathbf{B}_i} (\max_u (\tilde{\mathbf{w}}^{(t)})^T \mathbf{B}_{ij(u)} - \text{mean}_v ((\tilde{\mathbf{w}}^{(t)})^T \mathbf{B}_{ij(v)}))$

TABLE 1
Algorithm: M³IC-MBM

of constraints Ω^{t-5} from the whole set of constraints $\{0, 1\}^n$ in problem (20) that guarantees an accurate solution. By using this method, a nested sequence of tighter relaxations can be constructed. More precisely, the first constraint is replaced by:

$$\forall \mathbf{c} \in \Omega^t$$

$$\frac{1}{n} \tilde{\mathbf{w}}^T \sum_{i=1}^n \mathbf{c}_i \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} \Big|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}} \geq \frac{1}{n} \sum_{i=1}^n \mathbf{c}_i - \xi. \quad (21)$$

A subset of constraints Ω^t with polynomial size can be generally found. With the constraints in Ω^t , the solution of the relaxed problem satisfies all the constraints from problem (20) up to a precision ϵ_2 , i.e., $\forall \mathbf{c} \in \{0, 1\}^n$:

$$\frac{1}{n} \tilde{\mathbf{w}}^T \sum_{i=1}^n \mathbf{c}_i \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} \Big|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}} \geq \frac{1}{n} \sum_{i=1}^n \mathbf{c}_i - (\xi + \epsilon_2) \quad (22)$$

which means, the remaining exponential number of constraints will not be violated up to the precision ϵ_2 . Therefore, they don't need to be explicitly added to Ω^t .

The algorithm constructs Ω^t in Eq.(21) iteratively. It starts

5. t is used to denote that it is in the t -th iteration of CCCP.

with an empty set Ω^{t_0} as follows:

$$\min_{\tilde{\mathbf{w}}} \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 \quad (23)$$

$$s.t. \quad i = 1, \dots, n, \forall p, q \in \{1, 2, \dots, k\}$$

$$-l \leq \sum_{i=1}^n \sum_{j \in \mathbf{B}_i} \frac{1}{n_i} \tilde{\mathbf{w}}^T (\mathbf{B}_{ij(p)} - \mathbf{B}_{ij(q)}) \leq l.$$

After solving this problem and getting the corresponding solution $\tilde{\mathbf{w}}^{(t_0)}$, the most violated constraint can be computed as:

$$\mathbf{c}_i^{t_0} = \begin{cases} 1, & \text{if } (\tilde{\mathbf{w}}^{(t_0)})^T \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} \Big|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t_0)}} \leq 1 \\ 0, & \text{otherwise} \end{cases} \quad (24)$$

Then, this new constraint will be added to Ω^t and the optimization problem turns to:

$$\min_{\tilde{\mathbf{w}}, \xi \geq 0} \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + C\xi \quad (25)$$

$$s.t. \quad \forall \mathbf{c} \in \Omega^{t_1},$$

$$\frac{1}{n} \tilde{\mathbf{w}}^T \sum_{i=1}^n \mathbf{c}_i \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} \Big|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}} \geq \frac{1}{n} \sum_{i=1}^n \mathbf{c}_i - \xi$$

$$\forall p, q \in \{1, 2, \dots, k\}$$

$$-l \leq \sum_{i=1}^n \sum_{j \in \mathbf{B}_i} \frac{1}{n_i} \tilde{\mathbf{w}}^T (\mathbf{B}_{ij(p)} - \mathbf{B}_{ij(q)}) \leq l.$$

For the current cutting plane step, in Ω^{t_1} , there is only one element, which is obtained from Eq.(24). From this updated optimization problem, the optimal solution of $\tilde{\mathbf{w}}^{(t_1)}$ can be computed. Then, the most violated constraint vector $\mathbf{c}_i^{t_1}$ would be computed similarly as in Eq.(24), where the only difference is that the weight vector $\tilde{\mathbf{w}}^{(t_0)}$ is replaced by $\tilde{\mathbf{w}}^{(t_1)}$. This procedure is repeated until all the constraints satisfy the requirements in Eq.(22). In this way, the successive strengthening approximation series of the problem (20) is constructed by the expanding number of cutting planes that cut off the current optimal solution from the feasible set [15].

3.4.4 Dual Form

Problem (25) is a QP problem. In many cases, QP problems are solved from their dual forms. So, in this section, we derive the dual form of problem (25) with t_1 being replaced by t_s .

First of all, the following simplifications are made: we define a set of examples $\mathbf{X}^{t_s} = \{\mathbf{x}_1^{t_s}, \dots, \mathbf{x}_s^{t_s}, \mathbf{x}_{s+1}^{t_s}, \dots, \mathbf{x}_{s+2k^2}^{t_s}\}$. Here, $[\mathbf{x}_1^{t_s}, \dots, \mathbf{x}_s^{t_s}]$ are defined as: $\mathbf{x}_j^{t_s} = \frac{1}{n} \sum_{i=1}^n \mathbf{c}_i \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} \Big|_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t_s)}}, j = 1, \dots, s$, $[\mathbf{x}_{s+1}^{t_s}, \dots, \mathbf{x}_{s+2k^2}^{t_s}]$ are defined as:

$$\forall p, q \in \{1, 2, \dots, k\}$$

$$\mathbf{x}_{(s+(p-1) \times k + q)}^{t_s} = - \sum_{i=1}^n \sum_{j \in \mathbf{B}_i} \frac{1}{ln_i} (\mathbf{B}_{ij(p)} - \mathbf{B}_{ij(q)}). \quad (26)$$

6. Here, we denote t_i as the i -th iteration of the cutting plane algorithm for solving the problem from the t -th iteration of CCCP. And therefore, $\Omega^{t_0} = \emptyset, |\Omega^{t_s}| = s$, where $|\cdot|$ denote the number of vectors in Ω^{t_s} .

$[\mathbf{x}_{s+k^2+1}^{t_s}, \dots, \mathbf{x}_{s+2k^2}^{t_s}]$ are defined as:

$$\forall p, q \in \{1, 2, \dots, k\} \quad (27)$$

$$\mathbf{x}_{(s+k^2+(p-1) \times k+q)}^{t_s} = \sum_{i=1}^n \sum_{j \in \mathbf{B}_i} \frac{1}{\ln_i} (\mathbf{B}_{ij(p)} - \mathbf{B}_{ij(q)}).$$

Then problem (25) can be transformed to:

$$\begin{aligned} \min_{\tilde{\mathbf{w}}, \xi \geq 0} \quad & \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + C\xi \\ \text{s.t.} \quad & \forall j \in \{1, \dots, s\} \\ & \tilde{\mathbf{w}}^T \mathbf{x}_j^{t_s} \geq \frac{1}{n} (\mathbf{c}^{t_j})^T \mathbf{1} - \xi \\ & \forall j \in \{s+1, \dots, s+2k^2\} \\ & \tilde{\mathbf{w}}^T \mathbf{x}_j^{t_s} \geq -1. \end{aligned} \quad (28)$$

Proposition 2: The objective function of the dual problem of (28) is given by:

$$D^{t_s}(\alpha^{t_s}) = -\frac{1}{2} (\alpha^{t_s})^T \mathbf{K}^{t_s} \alpha^{t_s} + \frac{1}{n} \sum_{i=1}^s \alpha_i^{t_s} (\mathbf{c}^{t_i})^T \mathbf{1} - \sum_{i=s+1}^{s+2K^2} \alpha_i^{t_s}, \quad (29)$$

where $\mathbf{K}_{i,j}^{t_s} = \langle \mathbf{x}_{i,j}^{t_s}, \mathbf{x}_{j,i}^{t_s} \rangle$ and α^{t_s} is a s -dimensional vector. The corresponding dual optimization problem can then be formulated as:

$$\begin{aligned} \max_{\alpha^{t_s} \geq 0} \quad & D^{t_s}(\alpha^{t_s}) \\ \text{s.t.} \quad & \sum_{i=1}^s \alpha_i^{t_s} \leq C. \end{aligned} \quad (30)$$

Proof: Please refer to Appendix A.

3.4.5 Algorithm

The proposed method is characterized by an outer iteration, i.e., CCCP iteration and an inner iteration, i.e., Cutting Plane iteration. H^{t_s} is used to denote the constraint $\frac{1}{n} (\tilde{\mathbf{w}}^{(t_s)})^T \sum_{i=1}^n c_i^{t_s} \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} |_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}} \leq \frac{1}{n} \sum_{i=1}^n c_i^{t_s} - (\xi^{(t_s)} + \epsilon_2)$ and $J^t = \frac{1}{2} \|\tilde{\mathbf{w}}^{(t)}\|^2 + C\xi^{(t)}$. The whole method is summarized in Table 1. We have also elaborated how the kernel form of the proposed method can be derived in Appendix D.

3.5 Discussion

3.5.1 Properties

The outer iteration of our method is CCCP. It has already been shown that CCCP decreases the objective function monotonically and converges to a local minimum solution [27]. As for the inner iteration – the Cutting Plane iteration, we have the following three theorems. Theorem 3 justifies the Step 8 of the proposed algorithm in Table 1. Theorem 4 gives the calculation complexity for each cutting plane iteration. And Theorem 5 gives an upper bound on the convergence rate.

Theorem 3: The most violated constraint for problem (25), with t_1 being replaced by t_s , is as follows:

$$\mathbf{c}_i^{t_s} = \begin{cases} 1, & \text{if } (\tilde{\mathbf{w}}^{(t_s)})^T \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} |_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}} \leq 1 \\ 0, & \text{otherwise} \end{cases}. \quad (31)$$

Proof: The most violated constraints are the ones that give the largest ξ . The maximal of ξ can be given by:

$$\begin{aligned} \xi^* &= \max_{\mathbf{c}_i^{t_s} \in \{0,1\}} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{c}_i^{t_s} - \frac{1}{n} (\tilde{\mathbf{w}}^{t_s})^T \sum_{i=1}^n \mathbf{c}_i^{t_s} \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} |_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}} \right) \\ &= \frac{1}{n} \max_{\mathbf{c}_i^{t_s} \in \{0,1\}} \left(\sum_{i=1}^n \mathbf{c}_i^{t_s} (1 - (\tilde{\mathbf{w}}^{t_s})^T \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} |_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}}) \right) \\ &= \frac{1}{n} \sum_{i=1}^n \max_{\mathbf{c}_i^{t_s} \in \{0,1\}} \left(\mathbf{c}_i^{t_s} (1 - (\tilde{\mathbf{w}}^{t_s})^T \frac{\partial f(\tilde{\mathbf{w}}, i)}{\partial \tilde{\mathbf{w}}} |_{\tilde{\mathbf{w}}=\tilde{\mathbf{w}}^{(t)}}) \right). \end{aligned} \quad (32)$$

So, $\mathbf{c}_i^{t_s}$ should be calculated using Eq.(31). \square

Theorem 4: Each iteration from step 5 to step 9 in Table 1 takes time $O(en)$ for a constant constraint set Ω^{t_s} , where e is the average number of nonzero features of \mathbf{B}_{ij} and $e = d$ for non-sparse data.

Proof: Each inner product in Step 5 and Step 8 takes $O(ek)$ time. For each line, n inner products need to be calculated. So, in total, it takes $O(ekn)$ time. As for Step 7, we can solve it through the dual form, setting up the dual form of problem (25) is dominated by computing $(s+2k^2)^2$ inner products, and each inner product takes around $O(ekn)$ calculations after $\mathbf{x}_j^{t_s}, j \in \{s+1, \dots, s+2k^2\}$ are precomputed. In total, the cost time is $O((s+2k^2)^2ekn)$. Solving the dual requires time $O((s+2k^2)^2)$. Therefore, each CCCP iteration takes time $O(en)$ for a constant constraint set Ω^{t_s} . \square

Theorem 5: The Cutting Plane iteration in Table 1 terminates after at most $\{\frac{2}{\epsilon_2}, \frac{8CR^2}{\epsilon_2^2}\}$ steps, where $R^2 = \frac{k}{k-1} \times \max_{ij} \|\mathbf{B}_{ij}\|^2$.

Proof: Please refer to Appendix B. \square

Although the convergence of M³IC-MBM can be guaranteed, its outer iteration – CCCP iteration only converges to a local minimal solution. Therefore, in this paper, during each trial the M³IC-MBM algorithm is run several times, and the solution with the minimal J^t value is chosen as the output. We will show, in the experiments, that M³IC-MBM is very fast and with only a few repetition times, we can get good results.

3.5.2 Relationship with Prior Work

In [47], [48] and [49], the authors accelerate the MMC and Semi-Supervised SVM problems for the traditional single instance learning. They first divide the original problem into a series of non-convex sub-problems by using Cutting Plane, then solve each non-convex sub-problem using CCCP iteratively. These methods have shown state-of-the-art performances, both in accuracy and efficiency, and they look similar to M³IC-MBM in this paper. But the main common problem in their methods is that the Cutting Plane approach is designed

to solve convex problems, rather than nonconvex problems. Since, they try to solve a non-convex problem by using cutting plane, the convergence and optimality of their methods may not be guaranteed. In both [48] and [49], the authors try to prove the convergence rate from the dual form. However, the constraints of their objective problem are not convex. In this case, the optimal solution of the original problem does not equal to that of the dual problem [3]. So, the proof of the convergence of the dual problem cannot be used to justify the convergence of the primal problem in their papers.

Different from their method, in M³IC-MBM, we first apply the CCCP to decompose the original nonconvex problem into a series of convex ones, and then use the Cutting Plane method to solve each of them. In this way, the final solution can be guaranteed to converge to a local optimal value as proved in Appendix B. Therefore, M³IC-MBM is theoretically much more elegant than the previous related methods.

Dataset	Categories
Corel	Fox, Tiger, Elephant
SIVAL1	AjaxOrange, Apple, Banana, BlueScrunge, CandleWithHolde
SIVAL2	CardboardBox, CheckeredScarf, CokeCan, DataMiningBook, DirtyRunningShoe
SIVAL3	DirtyWorkGloves, FabricSoftenerBox, FeltFlowerRug, GlazedWoodPot, GoldMedal
SIVAL4	GreenTeaBox, JuliesPot, LargeSpoon, RapBook, SmileyFaceDoll
SIVAL5	SpriteCan, StripedNotebook, TranslucentBowl, WD40Can, WoodRollingPin
SIVAL6	AjaxOrange, Apple, Banana, FabricSoftenerBox, GoldMedal
Natural Scene1	desert, mountains, sea
Natural Scene2	sea, sunset, trees
Natural Scene3	desert, mountains, sunset
Natural Scene4	desert, mountains, sea, sunset, trees
Reuters1	earn, acq, money-fx
Reuters2	crude, grain, trade, interest
Reuters3	acq, money-fx, grain
Reuters4	acq, crude, trade
Reuters5	earn, money-fx, crude, trade
Reuters6	earn, acq

TABLE 2
The composition of datasets

4 EXPERIMENTS

In this section, a set of experiments are presented to show the effectiveness and the efficiency of the proposed method.

4.1 Real-World Datasets

MIC is a relatively new research area, and there is no benchmark dataset to measure its performance. Fortunately, we can employ some of the available multiple instance benchmark datasets⁷, which were originally used for measuring the performances of multiple instance classification algorithms, and make them eligible for the MIC tasks. The details of these datasets are described as follows:

7. All of the datasets used in this paper can be downloaded online, and we will specify them separately later.

Corel Pictures from three categories of the Corel dataset, namely elephant, fox, and tiger, are merged together. More specifically, we merge the positive bags from the benchmark datasets – elephant, fox, and tiger [2]⁸. The reason why the negative bags in these datasets are not used is that the main objective of clustering task is to discover the hidden concepts/patterns in a dataset. But, in these datasets, the negative bags are just some background pictures and contain no common hidden concept/pattern. The detailed descriptions of this combined dataset are summarized in Table 2 and Table 3.

SIVAL There are in total 25 categories in the SIVAL dataset [23]. For each category, there are 60 images. We randomly partition these 25 categories into 6 groups, with each group containing 5 categories. We name the six groups as SIVAL1, SIVAL2, SIVAL3, SIVAL4, SIVAL5, and SIVAL6. The descriptions of these datasets are also summarized in Table 2 and Table 3.

Natural Scene This dataset was originally used to test the performances of multiple instance multiple label algorithms⁹, where each bag can be associated with more than one labels. There are in total 2000 pictures in this dataset, associated with 5 labels. We choose, from this dataset, the 1543 pictures, which are associated with only one label. We randomly partition these pictures into 4 groups, namely Natural Scene1, Natural Scene2, Natural Scene3, and Natural Scene4. Please refer to Table 2 and Table 3 for more details.

Reuters This dataset is from Reuters-21578 collection¹⁰, and was also originally used for the multiple instance multiple label problem. In this dataset, there are in total 2000 bags and they are associated with 7 labels. We use the same strategy as that we employed on the Natural Scene dataset, and delete the bags that are associated with more than one labels. There are in total such 1,701 bags left. We further partition them into six groups, and summarize the details of these datasets in Table 2 and Table 3.

Musk The MUSK datasets, i.e., MUSK1 and MUSK2 [7], are two most popular benchmark datasets in binary multiple instance classification. However, they can **not** be used here. This is because there is only one potential concept – musk in both of these two datasets, while, to measure the performance of clustering algorithms, at least two different underlying concepts should exist within these two datasets.

4.2 Experimental Setups

We have conducted comprehensive performance evaluations by testing our method and comparing it with the only existed MIC method–BAMIC [39] and two other single instance learning methods, i.e., K-means [8] and CPM3C [48].

For BAMIC, we used the three bag distance measurement methods as in [39], i.e., minimal Hausdorff distance, maximal Hausdorff distance and average Hausdorff distance. We name

8. It can be downloaded at: <http://www.cs.columbia.edu/~andrews/mil/datasets.html>

9. It can be downloaded at <http://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/annex/miml-image-data.htm>

10. It can be downloaded at <http://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/annex/miml-text-data.htm>

Dataset	Categories	Features	Bags	Instances	Min Number of Bags per Class	Max Number of Bags per Class	Ave Number of Bags per Class
Corel	3	230	300	1953	100	100	100
SIVAL1	5	30	300	9300	60	60	60
SIVAL2	5	30	300	9300	60	60	60
SIVAL3	5	30	300	9300	60	60	60
SIVAL4	5	30	300	9300	60	60	60
SIVAL5	5	30	300	9300	60	60	60
SIVAL6	5	30	300	9300	60	60	60
Natural Scene1	3	15	949	8541	268	341	316.3
Natural Scene2	3	15	935	8415	216	378	311.7
Natural Scene3	3	15	824	7416	216	340	274.7
Natural Scene4	5	15	1543	13887	216	378	308.6
Reuters1	3	243	1316	3726	81	798	438.7
Reuters2	4	243	385	1779	73	109	96.3
Reuters3	3	243	627	2171	81	437	209
Reuters4	3	243	640	2393	94	437	213.3
Reuters5	4	243	1061	3140	73	109	89.3
Reuters6	2	243	1235	3409	437	798	617.5

TABLE 3
The detailed description of the datasets

the BAMIC methods with these three bag distance measurements as BAMIC1, BAMIC2, and BAMIC3, respectively. BAMIC is in fact based on K-medoids, but doesn't provide a way to avoid the local minimal problem. So, for each dataset, we run each of these BAMIC algorithms 10 times independently, and report *only the best performance* of these 10 independent runs.

For M³IC-MBM, we set $\epsilon_1 = 0.01$, $\epsilon_2 = 0.01$, The class imbalance parameter l is set by grid search from the grid $[0, 0.001, 0.01, 0.1, 1 : 1 : 5, 10]$ and the parameter C is searched from the exponential grid $2^{[-4:1:4]}$. $\tilde{\mathbf{w}}^0$ is randomly initialized. To avoid the local minimal problem that we have mentioned in Section 3.5, for each experiment, we run the M³IC algorithm 5 times independently and report the final result *with the minimal J^t* in Table 1.

For K-means and CPM3C, they are not designed to solve the multiple instance clustering, but can be adapted to solve it. More specifically, we first cluster all of the instances in bags by using these two methods separately (For K-means, the result is summarized over 50 independent runs). Then, for each bag, the cluster assignment is determined by the cluster assignment that appears most frequently on the instances of this bag. These two methods are used as the traditional single instance learning opponents.

CPM3C is related to the proposed method. But as mentioned in Section 3.5.2, one of the significant differences between CPM3C and the proposed method is that the outer iteration of CPM3C is the cutting plane and its inner iteration is the CCCP iteration. However, when using the cutting plane method to solve a nonconvex problem, the convergence property may not hold. So, in order to avoid the infinite cutting plane loop, the maximum number of cutting plane loops is set to 20. The set of parameters for CPM3C are exactly the same as that of the proposed method. Therefore, in this paper, we employed the same way to tune the parameters of CPM3C as the one used in the proposed method.

4.3 Evaluation

In experiments, the number of clusters k is set to the true number of classes of these datasets. To evaluate the performances of different clustering algorithms, the following criterions are employed:

4.3.1 Clustering Accuracy (Acc)

The clustering accuracy (Acc) [33][36][48][49] is used to evaluate the final clustering performance. Specifically, we first take a set of labeled bags, remove the labels of these bags and run the clustering algorithms, then we relabel these bags using the clustering assignments returned by the algorithms. Finally, we measure the percentage of correct classifications by comparing the true labels and the labels assigned by the clustering algorithms as follows:

$$Acc = \frac{1}{n} \sum_{i=1}^n \delta(y_i, \text{map}(c_i)),$$

where, $\text{map}(\cdot)$ is a function that maps each cluster index to a class label, which can be found by the Hungarian algorithm [22]. c_i and y_i are the cluster index of \mathbf{x}_i and the true class label. $\delta(a, b)$ is a function that equals 1 when a equals b , and 0 otherwise.

It is clear that Acc discovers the one-to-one relationship between clusters, and measures the extent to which cluster assignments are associated with the corresponding true categories. The greater clustering accuracy is, the better the clustering algorithm performs.

4.3.2 Normalized Mutual Information (NMI)

Another evaluation metric we employed here is the Normalized Mutual Information (NMI) [30], which was originally a symmetric measure to quantify the statistical information shared between two distributions. More precisely, for two random variables X and Y , NMI is defined as:

$$NMI(X, Y) = \frac{I(X, Y)}{\sqrt{H(X)H(Y)}}. \quad (33)$$

Here, $I(X, Y)$ is mutual information between X and Y , while $H(X)$ and $H(Y)$ are entropies of X and Y . If there is a one to one correspondence between the distribution of X and that of Y , NMI equals one. From the perspective of information theory, it means all of the information encoded in X (sender) has already been correctly delivered to Y (receiver). Otherwise, if Y is merely a uniform distribution, NMI equals zero, which means no information would be transferred from X to Y . It is clear that $NMI(X, X) = 1$, which achieves the maximal possible value of NMI. Given two clustering results, we can consider the clustering assignments as distributions of random variables, and NMI in Eq.(33) can be estimated as:

$$NMI = \frac{\sum_{p=1}^k \sum_{q=1}^k n_{p,q} \log\left(\frac{n_{p,q}}{n_p \hat{n}_q}\right)}{\sqrt{\left(\sum_{p=1}^k n_p \log \frac{n_p}{n}\right) \left(\sum_{q=1}^k \hat{n}_q \log \frac{\hat{n}_q}{n}\right)}}, \quad (34)$$

where n_p refers to the number of data contained in the p -th cluster, \hat{n}_q is the number of data belonging to the q -th cluster, and $n_{p,q}$ denotes the number of data that are in the intersection between the p -th and the q -th clusters. We give two examples here. Suppose the true clustering result is $[1, 1, 2, 2, 3, 3]$ and predicted cluster assignment is $[2, 2, 3, 3, 1, 1]$. The corresponding NMI would be 1, which means that these two clustering assignments are exactly the same. And if predicted one is $[3, 3, 3, 3, 3, 3]$, the estimated NMI would be 0, which indicates that the predicted result makes no sense. The value calculated using the above equation is used as a performance measure for the given clustering result, since some datasets used in experiments are imbalanced.

4.3.3 CPU Running Time

All of our experiments are conducted on MATLAB r2008a on a 2.5GHZ Intel CoreTM2 Duo PC running Windows Vista with 2.0GB main memory. The average CPU-times for each independent run of these algorithms are also used as evaluation.

4.4 Clustering Results

The clustering accuracies and the NMI for different algorithms are reported in Table 4 and 5, while the average CPU time of all the independent runs for these algorithms is reported in Table 6.

From Table 4 and Table 5, it is easy to tell that, in most cases, M³IC-MBM works much better than the other methods. From Table 6, it is clear that our method runs much faster than BAMIC and CPM3C. Within most of these 17 datasets, the running time of M³IC-MBM is comparable with that of K-means. However, it is also clear that among the datasets, where K-means runs faster, the clustering performance of K-means is much worse than M³IC-MBM. So, it is safe to say, in these datasets, the proposed method is both effective and efficient than the comparison methods. From Reuters3 and Reuters4 results reported in Table 4, it seems that K-means performs better than the proposed method. But as we can see from Table 5, the corresponding NMI values of K-means are all 0, which means K-means gives the same cluster assignments to all of

the bags. It sounds unbelievable! In traditional single instance clustering problems, K-means will not give the same cluster assignments to all of the instances. But in our experiments of multiple instance clustering, the cluster assignments of bags depend on that of the associated instances. In this case, it is possible that all of the bags are grouped into the same cluster by K-means, since it is performed on all of the instances, rather than directly on bags. The high accuracies of K-means on these two datasets are in fact brought in by the extremely imbalanced nature of different categories in the corresponding datasets.

The reason for the good performance of M³IC-MBM is that it can incorporate the nature of multiple instance setting into the clustering problem more naturally. So, it gives a better clustering performance than the other methods. And in our algorithm, the outer iteration-CCCP iteration, as well as the inner iteration-Cutting Plane iteration, converges very fast. So, the proposed method is very efficient.

For BAMIC, it is an adaption of K-medoids. It tries to integrate the multiple instance learning into the clustering problem. The motivation is good. But this method tries to define the distances between bags in an unsupervised way, which brings the ambiguous problem into the clustering. The most important ‘‘concept’’ in each bag is not identified. So, in these datasets, its performances are not quite promising. And since it needs to calculate the distances between instances in different bags many times, its speed will be badly affected.

As for K-means, it is relatively fast in these datasets because it is an iterative method, and during each iteration, it only needs to calculate the distances between the instances and the clustering centers. But it is not designed for multiple instance clustering, and therefore cannot take the multiple instance assumptions into account. This is the main reason why K-means performs worse than M³IC-MBM does.

CPM3C is a both effective and efficient method in some single instance learning datasets, as shown in [48]. However, it is not suitable to be adapted to solve the multiple instance clustering problems. Its potential problem, i.e., the convergence problem, is very evident in these real world multiple instance datasets, which validates our comments in Section 3.5.2 very well. And its performance in clustering accuracies and NMI are not quite promising in these datasets either.

4.5 Properties

In this section, some properties of the proposed method are analyzed through a set of experiments. SIVAL6, Natural Scene4 and Reuters6 are used in these experiments.

4.5.1 Scalability

First of all, the performance of the proposed method with different dataset sizes are studied here.

For each trial, the same amount of bags from each category are randomly chosen, and then the corresponding performances of different algorithms are measured. This procedure is repeated 100 times and the average results are reported in Fig.1, Fig.2, and Fig.3.

From these comparison results, it can be seen that with different dataset sizes, the clustering accuracies do not change

	M ³ IC-MBM	BAMIC1	BAMIC2	BAMIC3	K-means	CPM3C
Corel	54.0	40.3	47.3	36.7	52.0	49.7
SIVAL1	47.0	26.3	35.3	38.0	26.0	23.3
SIVAL2	42.0	29.0	31.7	39.3	29.0	28.3
SIVAL3	41.0	30.0	35.7	38.7	27.0	29.7
SIVAL4	39.0	26.0	32.7	30.0	25.3	26.0
SIVAL5	40.7	25.7	36.3	34.3	26.0	28.0
SIVAL6	44.7	25.7	31.7	24.7	24.0	23.3
Natural Scene1	56.6	39.7	33.5	45.0	40.2	39.4
Natural Scene2	63.2	38.1	52.5	44.9	50.5	49.4
Natural Scene3	60.4	55.1	46.2	42.1	40.4	45.6
Natural Scene4	41.7	33.4	32.4	38.9	29.3	37.8
Reuters1	77.0	66.8	49.5	62.7	60.7	75.1
Reuters2	69.9	35.1	63.6	40.8	28.3	38.4
Reuters3	61.1	55.7	38.6	58.9	69.7	69.7
Reuters4	61.9	36.4	44.5	49.4	68.2	68.2
Reuters5	54.2	42.3	32.8	47.6	30.5	31.1
Reuters6	89.7	67.9	70.3	68.9	65.7	85.9

TABLE 4

Clustering accuracy (%) comparisons (the best results in each data set are highlighted in boldface)

	M ³ IC-MBM	BAMIC1	BAMIC2	BAMIC3	K-means	CPM3C
Corel	0.159	0.027	0.092	0.024	0.154	0.012
SIVAL1	0.178	0.015	0.107	0.048	0.030	0.024
SIVAL2	0.166	0.023	0.075	0.048	0.087	0.107
SIVAL3	0.266	0.015	0.105	0.088	0.097	0.092
SIVAL4	0.137	0.006	0.079	0.027	0.023	0.028
SIVAL5	0.162	0.011	0.037	0.023	0.023	0.038
SIVAL6	0.120	0.018	0.119	0.025	0.018	0.076
Natural Scene1	0.151	0.020	0.018	0.009	0.025	0.024
Natural Scene2	0.156	0.046	0.063	0.092	0.065	0.067
Natural Scene3	0.239	0.023	0.033	0.191	0.094	0.092
Natural Scene4	0.136	0.059	0.073	0.135	0.083	0.114
Reuters1	0.302	0.115	0.226	0.248	0.239	0.385
Reuters2	0.369	0.050	0.100	0.217	0.046	0.052
Reuters3	0.199	0.011	0.116	0.189	0.000	0.000
Reuters4	0.296	0.078	0.211	0.115	0.000	0.000
Reuters5	0.301	0.173	0.192	0.198	0.228	0.245
Reuters6	0.530	0.279	0.229	0.288	0.259	0.499

TABLE 5

Normalized Mutual Information comparisons (the best results in each data set are highlighted in boldface)

	M ³ IC-MBM	BAMIC1	BAMIC2	BAMIC3	K-means	CPM3C
Corel	1.2	267.8	257.1	261.6	6.1	137.6
SIVAL1	1.8	95.5	96.1	92.5	7.0	87.6
SIVAL2	3.1	95.1	98.4	95.8	5.6	108.3
SIVAL3	2.7	93.3	100.4	95.7	5.0	93.1
SIVAL4	2.9	107.7	100.5	94.8	7.7	86.9
SIVAL5	3.2	95.1	117.2	106.0	7.1	95.7
SIVAL6	6.7	69.7	70.5	71.2	8.3	69.6
Natural Scene1	11.4	68.3	69.2	68.3	2.2	49.0
Natural Scene2	17.1	63.9	66.4	66.6	2.6	138.2
Natural Scene3	13.2	49.3	51.2	51.5	2.0	52.8
Natural Scene4	30.6	290.0	316.2	320.6	4.1	80.3
Reuters1	13.9	42.0	44.3	42.8	17.0	198.8
Reuters2	3.5	7.4	7.4	7.4	3.9	419.9
Reuters3	3.3	8.8	10.7	9.2	6.8	218.4
Reuters4	3.4	10.2	10.9	11.0	6.1	211.2
Reuters5	13.8	19.6	21.8	21.4	8.1	148.5
Reuters6	8.2	65.8	74.1	68.5	7.2	54.2

TABLE 6

CPU Running Time (in seconds, the best results in each data set are highlighted in boldface)

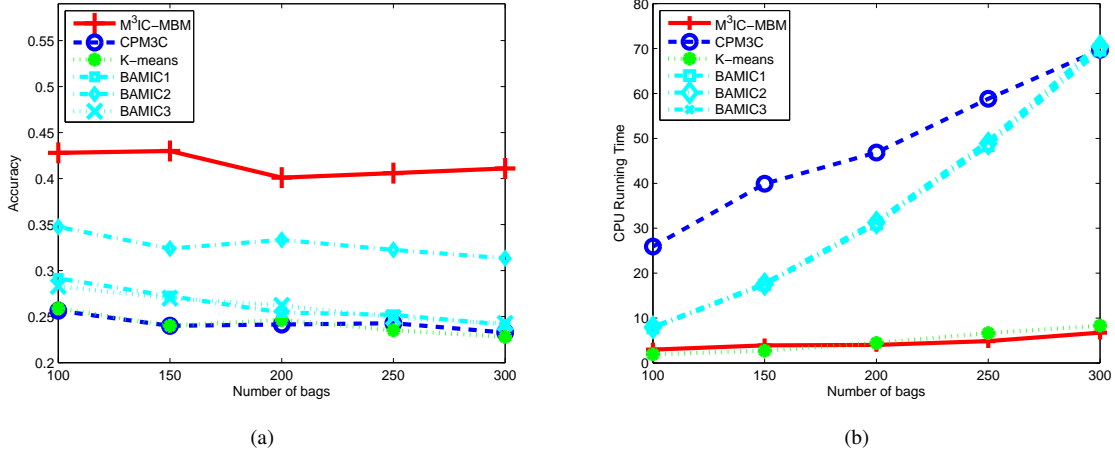


Fig. 1. The Comparison of Clustering Accuracy and the CPU Running Time (in seconds) with Varying Dataset Sizes on SIVAL

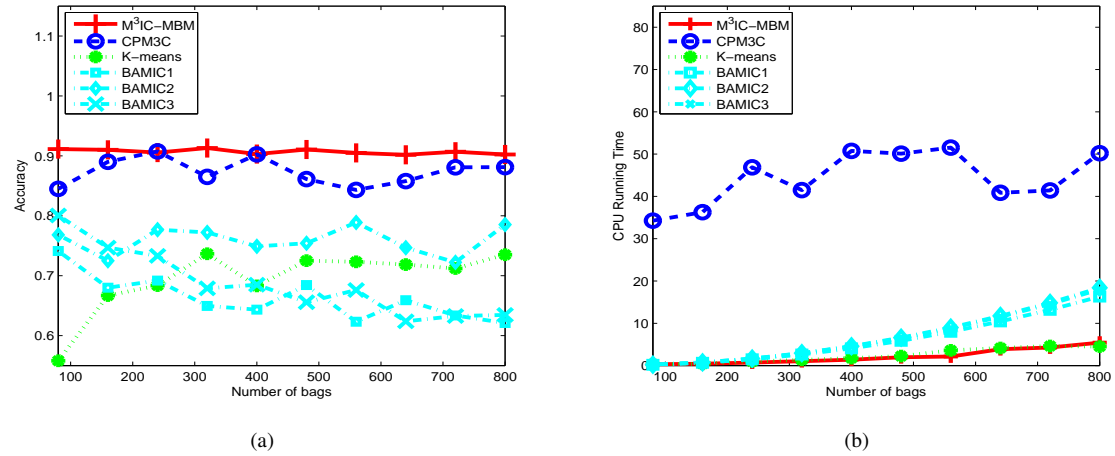


Fig. 2. The Comparison of Clustering Accuracy and the CPU Running Time (in seconds) with Varying Dataset Sizes on Reuters

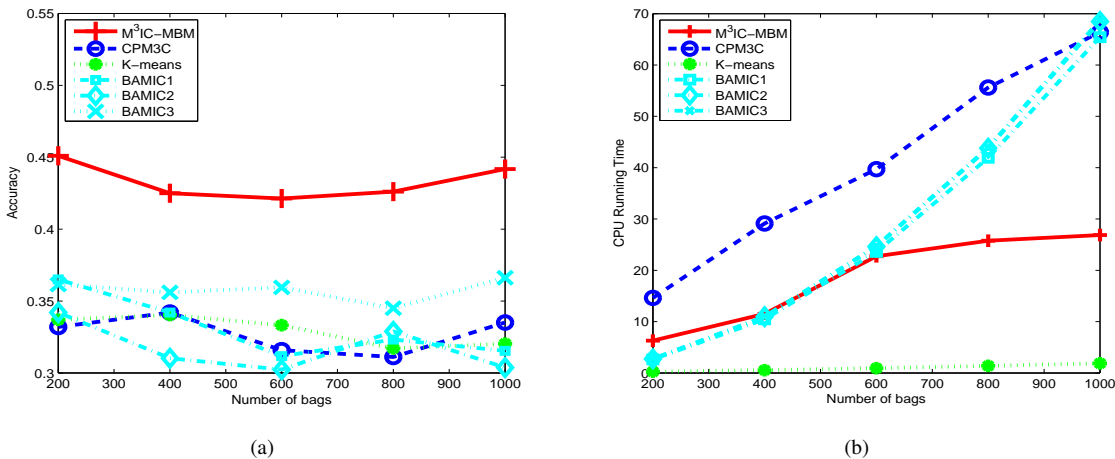


Fig. 3. The Comparison of Clustering Accuracy and the CPU Running Time (in seconds) with Varying Dataset Sizes on Natural Scene

a lot. The clustering accuracy of the proposed method is much better than that of the other ones. As for the CPU running time, with the increase of the bags, the CPU running time of BAMIC increases almost exponentially with the increase of bags, while that of M³IC-MBM and K-means are relatively stable. And in these datasets, in most cases, the CPU running time of K-means is shorter than that of M³IC-MBM. But it is evident that, the clustering accuracy performance of K-means is much worse than that of M³IC-MBM. For CPM3C, it is clear that its clustering accuracy is good compared with the other methods except M³IC-MBM. However, its CPU running time is too long which may hinder its practical use in this kind of applications. This is mainly due to the convergence issue as discussed before.

4.5.2 Sensitivity

When conducting experiments on M³IC-MBM, C and l are set by grid search. In this section, we will study the impact of these two parameters on the proposed method.

For each dataset, each time, we will only fix one parameter, while tuning the other one. In this way, the parameter sensitivity experiments of C and l are reported in Fig.4, Fig.5 and Fig.6. It is clear that with the increase of the parameter C , the performance of M³IC-MBM is relatively stable. As for the experimental results on l , in SIVAL, the performance of the proposed method deteriorates very quickly after l becomes larger than 2, while in Reuters and Natural Scene, its impacts are relatively small. So, in practice, we suggest using a relatively small l .

5 FUTURE WORKS

This section discusses some future research directions based on the proposed framework in this paper.

5.1 Multiple Instance Multiple Clustering

In many clustering algorithms, such as K-means, Spectral Clustering, Gaussian Mixture Models, etc, there is an assumption – each example can only be assigned to one cluster. They can not assign more than one clusters to each image, because they represent the whole image by using just one feature vector and the features of different concepts in this image are merged together and can hardly be separated. However, in some cases, this is problematic. Let’s consider the following problem:

Problem: Suppose we are given three groups of images: Group A contains pictures with both tigers and elephants; Group B contains pictures with both elephants and fox; Group C contains contains pictures with both foxes and tigers. It is clear that in this case, each group should belong to two different clusters.

In this case, we can not longer assume that each image belongs to only one cluster. A striking characteristic of our current M³IC formulation is that it can be extended to solve this problem, since each image is represented by a set of instances where each instance corresponds to one portion of the image and each instance can belong to one cluster. In this way, each bag can be assigned to multiple clusters. In the future, we plan to extend our current work to solve this

problem (Multiple Instance Multiple Clustering), which can be considered as an unsupervised version of multiple instance multiple label learning [13][17][41][50].

5.2 The Convergence Rate of CCCP Iteration

In this paper, we have proved the convergence rate of the inner cutting plane iteration. As for the outer CCCP iteration, when conducting the experiments, we found that it normally terminates in a few steps. But the precise form of the convergence rate of the outer CCCP iteration is still an interesting theoretical problem that can be solved in the future.

6 CONCLUSIONS

In this paper, we formulate a novel M³IC problem for the multiple instance clustering problem, which can be directly used to solve applications such as image and text clustering. In order to avoid solving a non-convex problem directly, we relax the original problem. Then, a combination of *Constrained Concave-Convex Procedure (CCCP)* and the *Cutting Plane* method – M³IC-MBM is used to solve the relaxed problem. After that, some important properties of the proposed method are also demonstrated and proved. In the experiment part, we compare the proposed method with the existed method–BAMIC, and two other single instance clustering methods on several real-world image and text datasets. The comparison results are very promising. In the last section, some future directions are outlined.

APPENDIX A

PROOF OF THE DUAL PROBLEM

Proposition 2: The dual problem of (28) is given by:

$$\begin{aligned} \max_{\alpha^{t_s} \geq 0} \quad & -\frac{1}{2} \alpha^{t_s T} \mathbf{K}^{t_s} \alpha^{t_s} + \frac{1}{n} \sum_{i=1}^s \alpha_i^{t_s} (\mathbf{c}^{t_i})^T \mathbf{1} - \sum_{i=2^{n+1}}^{s+2K^2} \alpha_i^{t_s} \\ \text{s.t.} \quad & \sum_{i=1}^s \alpha_i^{t_s} \leq C, \end{aligned} \quad (35)$$

where $\mathbf{K}_{i,j}^{t_s} = \langle \mathbf{x}_i^{t_s}, \mathbf{x}_j^{t_s} \rangle$ and α^{t_s} is a s -dimensional vector.

Proof: The Lagrangian form of problem (28) is:

$$\begin{aligned} L(\tilde{\mathbf{w}}, \xi, \eta, \alpha^{t_s}) = & \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + C\xi \\ & + \sum_{j=1}^s \alpha_j^{t_s} (-\xi - \tilde{\mathbf{w}} \mathbf{x}_j^{t_s} + \frac{1}{n} (\mathbf{c}^{t_j})^T \mathbf{1}) \\ & + \sum_{j=s+1}^{s+2K^2} \alpha_j^{t_s} (-1 - \tilde{\mathbf{w}} \mathbf{x}_j^{t_s}) - \eta\xi, \end{aligned} \quad (36)$$

where, $\alpha^{t_s} \geq 0$ and $\eta > 0$. Differentiating this Lagrangian form with respect to $\tilde{\mathbf{w}}$ and ξ , we can get:

$$\frac{\partial L(\tilde{\mathbf{w}}, \xi, \eta, \alpha^{t_s})}{\partial \tilde{\mathbf{w}}} = \tilde{\mathbf{w}} - \sum_{i=1}^{s+2k^2} \alpha_i^{t_s} \mathbf{x}_i^{t_s}. \quad (37)$$

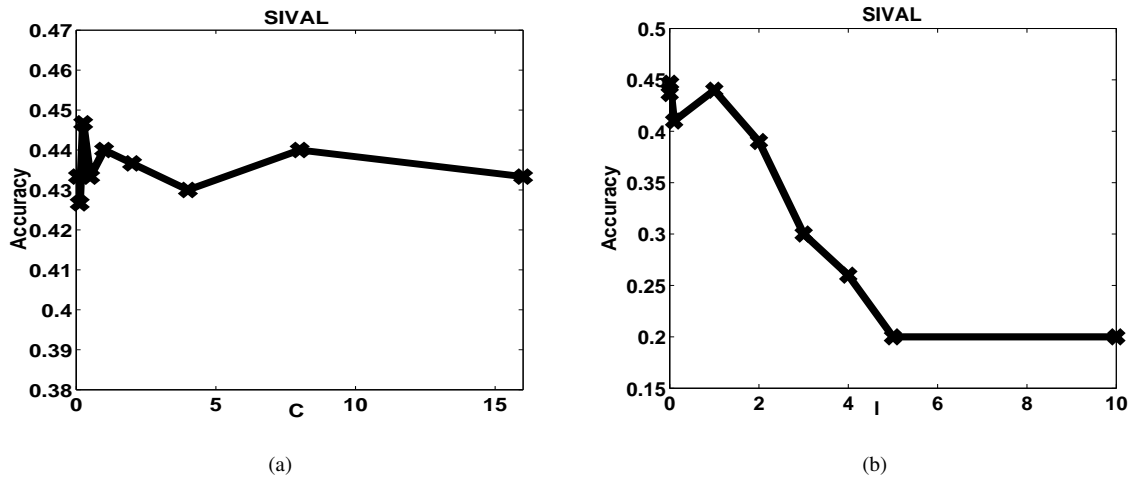


Fig. 4. The left figure measures the sensitivity of the parameter C on SIVAL, and the right figure measures the sensitivity of the parameter I on SIVAL.

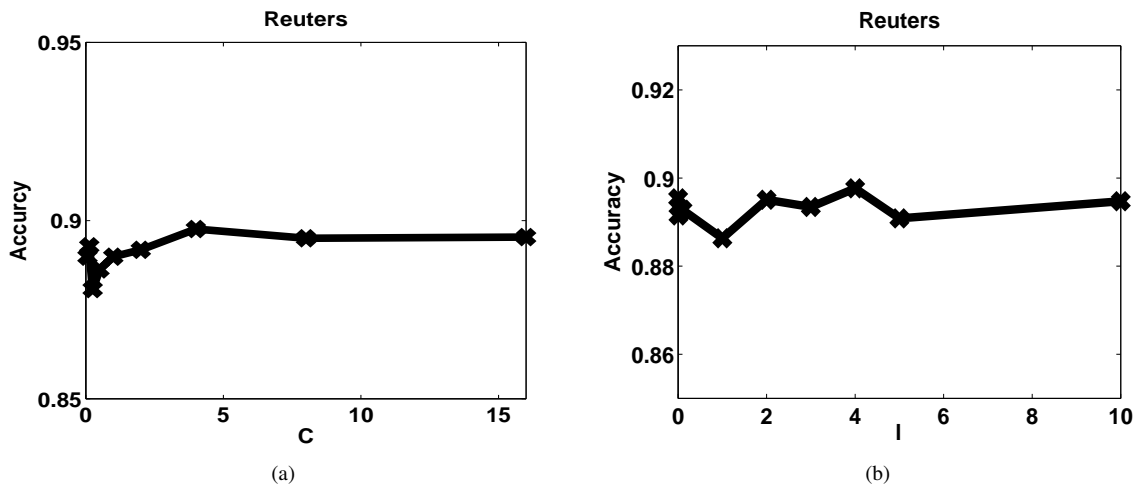


Fig. 5. The left figure measures the sensitivity of the parameter C on Natural Scene, and the right figure measures the sensitivity of the parameter I on Reuters.

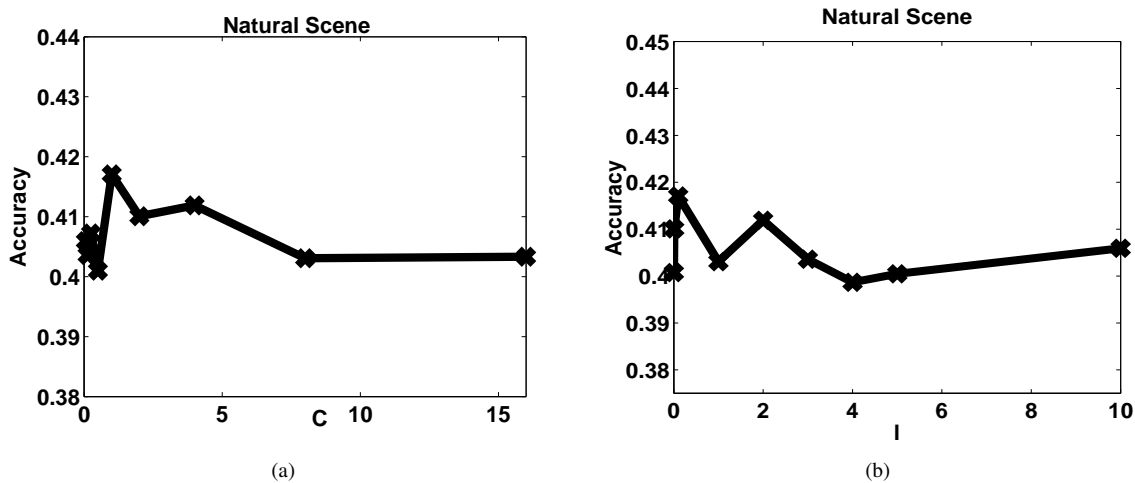


Fig. 6. The left figure measures the sensitivity of the parameter C on Natural Scene, and the right figure measures the sensitivity of the parameter I on Natural Scene.

$$\frac{\partial L(\tilde{\mathbf{w}}, \xi, \eta, \alpha^{t_s})}{\partial \xi} = C - \sum_{i=1}^s \alpha_i^{t_s} - \eta. \quad (38)$$

Setting these derivatives to zeros, the optimal solution for $\tilde{\mathbf{w}}$ and ξ are required as follows:

$$\begin{aligned} \tilde{\mathbf{w}} &= \sum_{i=1}^{s+2k^2} \alpha_i^{t_s} \mathbf{x}_i^{t_s} \\ \sum_{i=1}^s \alpha_i^{t_s} &= C - \eta. \end{aligned} \quad (39)$$

Plugging Eq.(39) into the Lagrangian, we obtain the dual problem:

$$\begin{aligned} \max_{\alpha^{t_s} \geq 0} & -\frac{1}{2} \alpha^{t_s T} \mathbf{K}^{t_s} \alpha^{t_s} + \frac{1}{n} \sum_{i=1}^s \alpha_i^{t_s} (\mathbf{c}^{t_i})^T \mathbf{1} - \sum_{i=s+1}^{s+2k^2} \alpha_i^{t_s} \\ \text{s.t.} & \sum_{i=1}^s \alpha_i^{t_s} \leq C. \end{aligned} \quad (40)$$

where $\mathbf{K}_{i,j}^{t_s} = \langle \mathbf{x}_i^{t_s}, \mathbf{x}_j^{t_s} \rangle$.

□

APPENDIX B PROOF OF THE CONVERGENCE RATE

Theorem 5: The Cutting Plane iteration in Table 1 terminates after at most $\{\frac{2}{\epsilon_2}, \frac{8CR^2}{\epsilon_2^2}\}$ steps, where $R^2 = \frac{k}{k-1} \times \max_{i,j} \|\mathbf{B}_{ij}\|^2$.

Proof: The objective function value of problem (28) is upper bounded by C , since $\mathbf{w} = 0$ and $\xi = 1$ is a feasible point. Next, we will show, during each Cutting Plane iteration, the objective function will increase by at least a constant value.

Let $\mathbf{c}^{t_{s+1}}$ be the newly added constraint. Let $\tilde{\alpha}^{t_s}$ be the optimal solution of the dual problem before this constraint is added. Suppose $\tilde{\alpha}^{t_s} = [\hat{\alpha}_1^{t_s}, \dots, \hat{\alpha}_s^{t_s}, 0, \hat{\alpha}_{s+1}^{t_s}, \dots, \hat{\alpha}_{s+2k^2}^{t_s}]$. To lower bound the progress made by each cutting plane iteration, we consider the increase in a specific direction of the dual function by line search, which can be described as follows:

$$\max_{0 \leq \beta \leq C} D^{t_{s+1}}(\tilde{\alpha}^{t_s} + \beta \eta) - D^{t_{s+1}}(\tilde{\alpha}^{t_s}) \quad (41)$$

where

$$D^{t_{s+1}}(\tilde{\alpha}) = -\frac{1}{2} \tilde{\alpha}^T \mathbf{K}^{t_{s+1}} \tilde{\alpha} + \frac{1}{n} \sum_{i=1}^{s+1} \tilde{\alpha}_i (\mathbf{c}^{t_i})^T \mathbf{1} - \sum_{i=s+1}^{s+1+2k^2} \tilde{\alpha}_i, \quad (42)$$

and the direction η is specified as:

$$\eta_j = \begin{cases} -\frac{1}{C} \tilde{\alpha}_j^{t_s}, & j = 1, \dots, s \\ 1, & j = s+1 \\ 0, & j = s+2, \dots, s+1+2k^2 \end{cases}. \quad (43)$$

Using this way to construct η , $\tilde{\alpha}^{t_s} + \beta \eta$ is guaranteed to lie in the feasible region of the dual.

Then, Theorem 6, which has been proved in [32] is employed.

In our formulation, we have:

$$\begin{aligned} j &= 1, \dots, s \\ \frac{\partial D^{t_{s+1}}(\tilde{\alpha}^{t_s})}{\partial \tilde{\alpha}_j^{t_s}} &= \frac{1}{n} (\mathbf{c}^{t_j})^T \mathbf{1} - (\tilde{\mathbf{w}}^{t_s})^T \mathbf{x}_j^{t_{s+1}} = \xi, \text{ if } \tilde{\alpha}_j^{t_s} \neq 0, \end{aligned} \quad (44)$$

and

$$\begin{aligned} j &= s+2, \dots, s+1+2k^2 \\ \frac{\partial D^{t_{s+1}}(\tilde{\alpha}^{t_s})}{\partial \tilde{\alpha}_j^{t_s}} &= -1 - (\tilde{\mathbf{w}}^{t_s})^T \mathbf{x}_j^{t_{s+1}} = 0, \text{ if } \tilde{\alpha}_j^{t_s} \neq 0. \end{aligned} \quad (45)$$

These two equations can be directly derived from the KKT conditions for solving a convex optimization problem [3].

For the newly added constraint,

$$\frac{\partial \tilde{D}^{t_{s+1}}(\tilde{\alpha}^{t_s})}{\partial \tilde{\alpha}_{s+1}^{t_s}} = \frac{1}{n} (\mathbf{c}^{t_j})^T \mathbf{1} - (\tilde{\mathbf{w}}^{t_s})^T \mathbf{x}_j^{t_{s+1}} \geq \xi + \varepsilon_2 \quad (46)$$

Then, we can get:

$$\nabla D^{t_{s+1}}(\tilde{\alpha}^{t_s})^T \eta \geq \xi + \varepsilon_2 - \sum_{j=1}^s \frac{\tilde{\alpha}_j^{t_s}}{C} \xi \geq \varepsilon_2 \geq 0, \quad (47)$$

$$\begin{aligned} \eta^T \mathbf{K}^{t_{s+1}} \eta &= (\mathbf{x}_{s+1}^{t_{s+1}})^2 - \frac{2}{C} \sum_{i=1}^s \tilde{\alpha}_i^{t_s} \mathbf{K}^{t_{s+1}}(\mathbf{x}_i^{t_{s+1}}, \mathbf{x}_{s+1}^{t_{s+1}}) \\ &\quad - 2 \sum_{i=s+2}^{s+2k^2+1} \eta_i \mathbf{K}^{t_{s+1}}(\mathbf{x}_i^{t_{s+1}}, \mathbf{x}_{s+1}^{t_{s+1}}) \\ &\quad + \frac{1}{C^2} \sum_{i=1}^s \sum_{j=1}^s \tilde{\alpha}_i^{t_s} \tilde{\alpha}_j^{t_s} \mathbf{K}^{t_{s+1}}(\mathbf{x}_i^{t_{s+1}}, \mathbf{x}_j^{t_{s+1}}) \\ &\quad - \frac{2}{C} \sum_{i=s+2}^{s+k^2+1} \sum_{j=1}^s \eta_i \tilde{\alpha}_j^{t_s} \mathbf{K}^{t_{s+1}}(\mathbf{x}_i^{t_{s+1}}, \mathbf{x}_j^{t_{s+1}}) \\ &\quad + \sum_{i=s+2}^{s+k^2+1} \sum_{i=s+2}^{s+k^2+1} \eta_i \eta_j \mathbf{K}^{t_{s+1}}(\mathbf{x}_i^{t_{s+1}}, \mathbf{x}_j^{t_{s+1}}) \\ &\leq R^2 + 2R^2 + \frac{1}{C^2} C^2 R^2 \\ &= 4R^2. \end{aligned} \quad (48)$$

By utilizing Theorem 6, the minimum increase per cutting plane iteration is:

$$\max_{0 \leq \beta \leq C} D^{t_{s+1}}(\tilde{\alpha}^{t_s} + \beta \eta) - D^{t_{s+1}}(\tilde{\alpha}^{t_s}) \geq \min\left\{\frac{C\epsilon_2}{2}, \frac{\epsilon_2^2}{8R^2}\right\}. \quad (49)$$

It is clear that this increase is independent of s and t . So, the cutting plane algorithm will terminate after at most $\max\{\frac{2}{\epsilon_2}, \frac{8CR^2}{\epsilon_2^2}\}$ iterations.

□

APPENDIX C PROOF OF THE UPPER BOUND

Theorem 6 [32]: \mathbf{J} is a symmetric, positive semi-definite matrix. We have the following objective function:

$$\Theta(\alpha) = -\frac{1}{2} \alpha^T \mathbf{J} \alpha + \langle \mathbf{h}, \alpha \rangle \quad (50)$$

α_0 is a solution of this objective function and η is a direction such that $\langle \nabla \Theta(\alpha_0), \eta \rangle > 0$. Then, the function:

$$\max_{0 < \beta \leq D} \Theta(\alpha_0 + \beta \eta) - \Theta(\alpha_0) \quad (51)$$

is upper bounded by:

$$\begin{aligned} & \max_{0 < \beta \leq D} \Theta(\alpha_0 + \beta \eta) - \Theta(\alpha_0) \\ & \geq \frac{1}{2} \min\{D, \frac{\langle \nabla \Theta(\alpha_0), \eta \rangle}{\eta^T J \eta}\} \langle \nabla \Theta(\alpha_0), \eta \rangle. \end{aligned}$$

Proof: It has been proved in [32]. \square

APPENDIX D KERNEL FORM

In many practical applications, kernel methods play important roles [19]. To derive the kernel form of the proposed method – Kernel M³IC, first of all, we need to replace all of the \mathbf{B}_{ij} in the primal form with $\phi(\mathbf{B}_{ij})$, where $\phi(\cdot)$ is a nonlinear map that maps instances in the primal feature space to another very high dimensional feature space so that the instances that are nonseparable in the primal space can be separable in the new high dimensional space. Normally, we don't really need to deal with this high dimensional space explicitly, but resort to the kernel trick¹¹. The inner product of $\phi(\mathbf{B}_{ij_1})$ and $\phi(\mathbf{B}_{ij_2})$ can be depicted as: $\langle \phi(\mathbf{B}_{ij_1}), \phi(\mathbf{B}_{ij_2}) \rangle = K(\mathbf{B}_{ij_1}, \mathbf{B}_{ij_2})$, where where K is a kernel function, such as RBF kernel, polynomial kernel, etc[29].

Then, the formulation of Kernel M³IC can be rewritten from Eq.(10) as:

$$\begin{aligned} \min_{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k, \xi_i \geq 0} & \quad \frac{1}{2} \sum_{p=1}^k \|\mathbf{w}_p\|^2 + \frac{C}{n} \sum_{i=1}^n \xi_i \quad (52) \\ \text{s.t.} & \quad i = 1, \dots, n, \\ & \quad \max_{j \in \mathbf{B}_i} (\mathbf{w}_{u^*}^T \phi(\mathbf{B}_{ij}) - \mathbf{w}_{v^*}^T \phi(\mathbf{B}_{ij})) \geq 1 - \xi_i \\ & \quad \forall p, q \in \{1, 2, \dots, k\} \\ -l \leq & \sum_{i=1}^n \sum_{j \in \mathbf{B}_i} I_{ij} \mathbf{w}_p^T \phi(\mathbf{B}_{ij}) - \sum_{i=1}^n \sum_{j \in \mathbf{B}_i} I_{ij} \mathbf{w}_q^T \phi(\mathbf{B}_{ij}) \leq l. \end{aligned}$$

After the relaxation like what we have done in Section 3.4.1, this formulation can be transformed to:

$$\begin{aligned} \min_{\tilde{\mathbf{w}}, \xi_i \geq 0} & \quad \frac{1}{2} \|\tilde{\mathbf{w}}\|^2 + \frac{C}{n} \sum_i \xi_i \quad (53) \\ \text{s.t.} & \quad i = 1, \dots, n, \\ & \quad \frac{k}{k-1} \max_{j \in \mathbf{B}_i} (\max_u \tilde{\mathbf{w}}^T \phi(\mathbf{B}_{ij(u)}) - \text{mean}_v (\tilde{\mathbf{w}}^T \phi(\mathbf{B}_{ij(v)}))) \\ & \quad \geq 1 - \xi_i \\ & \quad \forall p, q \in \{1, 2, \dots, k\} \\ -l \leq & \sum_{i=1}^n \sum_{j \in \mathbf{B}_i} \frac{1}{n_i} \tilde{\mathbf{w}}^T (\phi(\mathbf{B}_{ij(p)}) - \phi(\mathbf{B}_{ij(q)})) \leq l, \end{aligned}$$

where $\phi(\mathbf{B}_{ij(p)})$ is defined as $[\mathbf{0}, \mathbf{0}, \dots, \mathbf{B}_{ij}^T, \dots, \mathbf{0}]$, which is similar to Eq.(13), and $\tilde{\mathbf{w}}$ is also defined as $[\mathbf{w}_1^T, \mathbf{w}_2^T, \dots, \mathbf{w}_p^T, \dots, \mathbf{w}_k^T]^T$. Since both $\tilde{\mathbf{w}}$ and $\phi(\mathbf{B}_{ij(p)})$ are in very high dimensional kernel spaces, it is impractical to calculate their dot products directly. Therefore, the kernel trick is used here. According to the representer theorem [29], the final solution of $\tilde{\mathbf{w}}$ of this optimization problem should be a linear combination of $\phi(\mathbf{B}_{ij(p)})$, i.e., $\tilde{\mathbf{w}} = \sum_{i,j,p} \beta_{ijp} \phi(\mathbf{B}_{ij(p)})$, and $\tilde{\mathbf{w}}^T \phi(\mathbf{x}) = \sum_{i,j,p} \beta_{ijp} K(\mathbf{B}_{ij(p)}, \mathbf{x})$. By replacing $\tilde{\mathbf{w}}$ with $\sum_{i,j,p} \beta_{ijp} \phi(\mathbf{B}_{ij(p)})$, and substituting the inner products between high dimension vectors with their corresponding kernel outputs, Eq.(53) can be turned to an equivalent optimization problem with respect to β and ξ . In this way, all of the following optimization problems from Eq.(15) to Eq.(25) can also be transformed to and solved by their kernel forms.

REFERENCES

- [1] R. Amar, D.R. Dooley, S.A. Goldman, and Q. Zhang, Multiple-instance learning of real-valued data. Proceedings of the Eighteenth International Conference on Machine Learning, pages 3–10, 2001
- [2] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multipleinstance learning. Advances in Neural Information Processing Systems, Cambridge, MA: MIT Press., pages 561–568, 2003.
- [3] S.P. Boyd, L. Vandenberghe. Convex optimization. Cambridge university press, 2004
- [4] J.M. Borwein, A.S. Lewis, Convex analysis and nonlinear optimization: Theory and examples, Springer Verlag, 2006
- [5] Y. Chen, J.Z. Wang, Image categorization by learning and reasoning with regions, Journal of Machine Learning Research, Vol.5, pages 913–939, 2004.
- [6] Y. Chen, J. Bi and J.Z. Wang, Miles: Multiple-instance learning via embedded instance selection, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 28, pages 1931–1947, 2006.
- [7] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. In Artificial Intelligence, pages 1–8, 1998.
- [8] R.O. Duda, P.E. Hart and D.G. Stork, Pattern classification, Wiley-Interscience, 2001
- [9] P. Gehler, O. Chapelle, Deterministic annealing for multiple-instance learning, Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, pages 123–130 2007.
- [10] T. Gartner, P.A. Flach, A. Kowalczyk and A.J. Smola, Multi-instance kernels, in: International Conference on Machine Learning, pages 179–186, 2002.
- [11] Jiawei Han, Micheline Kamber. Data Mining. Morgan Kaufmann Publishers, 2001.
- [12] Anil K. Jain, Richard C. Dubes. Algorithms for clustering data. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.
- [13] R. Jin, S. Wang, and Z.H. Zhou. Learning a distance metric from multi-instance multi-label data. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 2038–2048, 2009.
- [14] T. Joachims. Training linear SVMs in linear time. International Conference on Knowledge Discovery and Data Mining, pages 217–226. 2006.
- [15] J.E. Kelley. The cutting plane method for solving convex programs. Journal of the SIAM, 8(4):703-712, 1960.
- [16] J. Kwok, P. M. Cheung, Marginalized multi-instance kernels, Proceedings of the 19th International Joint Conference on Artificial Intelligence, 2007.
- [17] Y.X. Li, S. Ji, J. Ye, S. Kumar, and Z.H. Zhou. Drosophila gene expression pattern annotation through multi-instance multi-label learning. Proceedings of the 21st International Joint Conference on Artificial Intelligence, pages 1445–1450, 2009.
- [18] C. Manning, P. Raghavan, Introduction to Information Retrieval, Cambridge University Press, 2008.
- [19] M. Girolami. Mercer kernel based clustering in feature space, IEEE Transaction on Neural Networks, 13, pages 2780- 2784, 2002.

11. For a detailed description of the kernel methods, please refer to [29]

- [20] O. Maron, T. Lozano-Prez, A framework for multiple-instance learning, in: *Advances in Neural Information Processing Systems*, Vol.10, Cambridge, pages 570–576, 1998.
- [21] O. Maron, A.L. Ratan, Multiple-instance learning for natural scene classification, *International Conference on Machine Learning'98*, Morgan Kaufmann, pages 341–349, 1998.
- [22] C.H. Papadimitriou, and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Dover Publications, 1998
- [23] R. Rahmani and S. A. Goldman. Missl: Multiple-instance semi-supervised learning. *International Conference on Machine Learning*, volume 10, pages 705–712, 2006.
- [24] S. Ray and D. Page, Multiple instance regression. *International Conference on Machine Learning*, pages 425–432, 2001
- [25] R. Rahmani, S.A. Goldman, H.Zhang, J. Krettek, J.E. Fritts, Localized content based image retrieval, *International Workshop on Multimedia Information Retrieval*, pages 227–236 2005.
- [26] C.B. Razvan, R.J. Mooney, Multiple instance learning for sparse positive bags, *International Conference on Machine Learning*, pages 112–119, 2007.
- [27] A.J. Smola, SVN Vishwanathan, and T. Hofmann. Kernel methods for missing variables. *International Workshop on Artificial Intelligence and Statistics*, pages 325–332, 2005.
- [28] B. Settles, M. Craven and S. Ray, Multiple instance active learning, *Advances in Neural Information Processing Systems*, pages 1289–1296, 2007.
- [29] B. Scholkopf, A. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002
- [30] A. Strehl, J. Ghosh, Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, Volume 3, pages 583–617, 2002
- [31] Q. Tao, D. Chu and J. Wang. Recursive Support Vector Machines for Dimensionality Reduction. *IEEE Transactions on Neural Networks*. Vol. 19, no. 1. pages 189–193. 2008.
- [32] I. Tsochantaris, T. Joachims, T. Hofmann, Y. Altun, Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, Volume 6, pages 1453–1478, 2006.
- [33] H. Valizadegan, R. Jin., Generalized Maximum Margin Clustering and Unsupervised Kernel Learning. *Advances in Neural Information Processing Systems*, pages 1417–1424, 2006.
- [34] F. Wang, B. Zhao and C. Zhang. Linear Time Maximum Margin Clustering. *IEEE Transactions on Neural Networks*. Vol. 21, no. 2. pages 319–332. 2010.
- [35] J. Wang, J.D. Zucker. Solving the Multiple-Instance Problem: A Lazy Learning Approach. *International Conference on Machine Learning*, pages 1119–1126, 2000.
- [36] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. *Advances in Neural Information Processing Systems*, pages 1537–1544, 2005.
- [37] L. Xu, D. Schuurmans. Unsupervised and semi-supervised multi-class support vector machines. *National Conference on Artificial Intelligence*, pages 904–910, 2005
- [38] A. Yuille, A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.
- [39] M.L. Zhang, Z.H. Zhou. Multi-instance clustering with applications to multi-instance prediction. *Applied Intelligence*, 31(1): 47–68, 2009.
- [40] Q. Zhang, S.A. Goldman, EM-DD: An improved multiple-instance learning technique. *Advances in Neural Information Processing Systems*, pages 1073–1080, 1998
- [41] M.L. Zhang and Z.H. Zhou. M3MIML: A maximum margin method for multi-instance multi-label learning. In: *Proceedings of the 8th IEEE International Conference on Data Mining*, pp.688–697, 2008.
- [42] D. Zhang, F. Wang, L. Si and T. Li. M³IC: Maximum Margin Multiple Instance Clustering, *International Joint Conference on Artificial Intelligence*, pages 1339–1344, 2009.
- [43] D. Zhang, L. Si, Multiple Instance Transfer Learning, *Workshop on Optimization Based Methods for Emerging Data Mining Problems*, in Conjunction with the *IEEE International Conference on Data Mining*, pages 406–411, 2009
- [44] D. Zhang, F. Wang, Z.W. Shi, C.S. Zhang. Interactive Localized Content-Based Image Retrieval with Multiple Instance Active Learning. *Pattern Recognition*, 43(2): 478–484, 2010.
- [45] D. Zhang, Z.W. Shi, Y.Q. Song and C.S. Zhang. Localized Content-Based Image Retrieval Using Semi-Supervised Multiple Instance Learning. *Asian Conference on Computer Vision*, pages 180–188, 2007
- [46] K. Zhang, I. W. Tsang, and J. T. Kwok. Maximum Margin Clustering Made Practical. *IEEE Transactions on Neural Networks*. Vol. 20, no. 4. pages 583–596. 2009.
- [47] B. Zhao, F. Wang, and C. Zhang. Cuts3vm: a fast semi-supervised svm algorithm. *International Conference on Knowledge Discovery and Data Mining*, pages 830–838, 2008.
- [48] B. Zhao, F. Wang, and C. Zhang. Efficient maximum margin clustering via cutting plane algorithm. *SIAM International Conference on Data Mining*, pages 751–762, 2008.
- [49] B. Zhao, F. Wang, and C. Zhang. Efficient multiclass maximum margin clustering. *International Conference on Machine Learning*, pages 751–762, 2008.
- [50] Z.H. Zhou, M.L. Zhang, Multi-instance multi-label learning with application to scene classification, *Advances in Neural Information Processing Systems*, pages 1609–1616, 2006.
- [51] Z.H. Zhou, J.M. Xu, On the relation between multi-instance learning and semi-supervised learning, *International Conference on Machine Learning*, pages 1167–1174, 2007



Dan Zhang is now a graduate student in Computer Science Department, Purdue University, West Lafayette, IN, USA. Before this, he got his Master's degree from Department of Automation, Tsinghua University, Beijing, China in 2008. His research interests include Information Retrieval, machine learning, data mining and pattern recognition.

Fei Wang is currently a postdoctoral research scientist in Healthcare Transformation Group, IBM T. J. Watson Research Center at Hawthorne, NY, US. Before this, he was a Postdoctoral Researcher in Department of Statistical Sciences, Cornell University, and a Postdoctoral Researcher in School of Computer Science, Florida International University, Miami, FL, USA. He got his PhD's degree from Department of Automation, Tsinghua University in 2008. His main research interests include machine learning, data mining and pattern recognition.

Luo Si is an Assistant Professor in Computer Science Department and Statistics Department (by courtesy), Purdue University. Before this, he got his PHD's degree from Computer Science Department, Carnegie Mellon University in 2006. His main research interest include: Information Retrieval, Knowledge Management, Machine Learning, Text/Data Mining for Life Science, Speech and Multimedia Processing, Natural Language Processing.

Tao Li is an Associate Professor in School of Computer Science, Florida International University, Miami, FL, USA. He received his Ph.D. in computer science from the Department of Computer Science, University of Rochester in 2004. His main research interests include machine learning, data mining and pattern recognition.