Multiple-Instance Ordinal Regression

Yanshan Xiao, Bo Liu, and Zhifeng Hao

***Abstract*— Ordinal regression (OR) is a paradigm in supervised learning, which aims at learning a prediction model for ordered classes. The existing studies mainly focus on single-instance OR, and the multi-instance OR problem has not been explicitly addressed. In many real-world applications, considering the OR problem from a multiple-instance aspect can yield better classification performance than from a single-instance aspect. For example, in image retrieval, an image may contain multiple and possibly heterogeneous objects. The user is usually interested in only a small part of the objects. If we represent the whole image as a global feature vector, the useful information from the targeted objects that the user is of interest may be overridden by the noisy information from irrelevant objects. However, this problem fits in the multiple-instance setting well. Each image is considered as a bag, and each object region is treated as an instance. The image is considered as of the user interest if it contains at least one targeted object region. In this paper, we address the multi-instance OR where the OR classifier is learned on multiple-instance data, instead of single-instance data. To solve this problem, we present a novel multiple-instance ordinal regression (MIOR) method. In MIOR, a set of parallel hyperplanes is used to separate the classes, and the label ordering information is incorporated into learning the classifier by imputing the parallel hyperplanes with an order. Moreover, considering that a bag may contain instances not belonging to its class, for each bag, the instance which is nearest to the middle of the corresponding class is selected to learn the classifier. Compared with the existing single-instance OR work, MIOR is able to learn a more accurate OR classifier on multiple-instance data where only the bag label is available and the instance label is unknown. Extensive experiments show that MIOR outperforms the existing single-instance OR methods.**

***Index Terms*— Multiple-instance data, ordinal regression (OR).**

I. Introduction

ORDINAL regression (OR) is a learning paradigm where the training data are marked by a set of ranks and there exists an ordering between different ranks [1]–[3]. For example, in image retrieval, each image is associated with one of the three ratings: *relevant*, *partial relevant*, and *irrelevant*. An image in the *partial relevant* rating has higher relevance to the user query than that in the *irrelevant* rating, and having the *relevant* rating should be more relevant than having the other ratings. It is seen that there is a natural order among the classes. For this type of problems, the misclassification costs are not the same for different classes [4]–[7]. Classifying a *relevant* image as *irrelevant* should have a higher penalty error than classifying it as *partial irrelevant*. In the literature, OR is sometimes referred to as multiclass classification with ordered labels [8]–[10]. The experimental results in the previous work [10], [11] have shown that for the multiclass classification problem with ordered labels, the OR methods [10], [11] can obtain better classification performance than the traditional classification methods, especially when there is only a small amount of labeled data available to train the classifier. To date, OR has been applied to various applications, such as disease treatment [6], facial recognition [7], and information retrieval [9].

Considerable work [12]–[14] has been done on OR. For example, [11] extends support vector machine (SVM) to solve the OR problem and puts forward a threshold OR model where a set of hyperplanes is used to separate the ordinal classes. [15] proposes two SVM-based OR models: support vector OR with explicit constraints (SVOR-EXCs) and support vector OR with implicit constraints (SVOR-IMCs). In SVOR-EXC, the order of classes is incorporated into the learning problem by explicitly imposing inequalities on the hyperplane order. In SVOR-IMC, the ordinality relation of classes is implicitly embedded into the formulation by considering the training data from all classes to determine each hyperplane. Moreover, neural networks [16], Gaussian process [17], and ensemble learning [18] are adapted to solve the OR problems.

Despite much progress in this area, most of the existing methods are proposed for single-instance OR, and the multi-instance OR problem has not been explicitly addressed. In single-instance OR, the training set contains a number of instances, and the label is associated with a single instance. In multi-instance OR, the training set consists of a number of bags and each bag has several instances. The label is associated with a bag of instances, and the specific instance label is unavailable. A bag belongs to class p if it has at least one instance of class p. That is to say, the bag of class p may contain instances not belonging to class p, in addition to at least one instance of class p [19]–[21]. The task of multi-instance OR is to build an OR classifier on the multiple-instance data where only the bag label is available and the instance label is unknown.

In real-world applications, considering the data as a single instance may lead to misclassification. Let us take image

Manuscript received November 11, 2016; revised July 9, 2017 and October 7, 2017; accepted October 16, 2017. This work was supported in part by the Natural Science Foundation of China under Grant 61472090, Grant 61672169, and Grant 61472089, in part by the NSFC-Guangdong Joint Found under Grant U1501254, in part by the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S2013050014133, in part by the Natural Science Foundation of Guangdong under Grant 2015A030313486, Grant 2014A030306004, and Grant 2014A030308008, in part by the Science and Technology Planning Project of Guangdong under Grant 2013B051000076, Grant 2015B010108006, Grant 2017A040405050, and Grant 2015B010131015, and in part by the Science and Technology Planning Project of Guangzhou under Grant 201707010492 and Grant 201604016041. *(Corresponding author: Bo Liu.)*

Y. Xiao is with the School of Computers, Guangdong University of Technology, Guangzhou 510006, China (e-mail: xiaoyanshan@189.cn).

B. Liu is with the School of Automation, Guangdong University of Technology, Guangzhou 510006, China (e-mail: csbliu@189.cn).

Z. Hao is with the School of Mathematics and Big Data, Foshan University, Foshan 528000, China (e-mail: zfhao@fosu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2017.2766164

Fig. 1. Sample images from the MSRA-MM data set. The labels are attached underneath the images. Assuming that the user is interested in the images on "soccer," (b) and (c) should have higher similarity than (a) and (b). However, (a) and (b) are expected to deliver higher similarity than (b) and (c) if the whole image is transformed into a global feature vector. Instead of transforming the whole image into a single instance, we can consider it from a multiple-instance's view. In multiple-instance settings, an image is segmented into several content regions. Each image is treated as a bag, and each region is considered as an instance. An image is classified as of the user interest if it includes at least one targeted region. From a multiple-instance's view, (b) and (c) will be classified to the "soccer" class, since they contain the regions on "soccer."

retrieval as an example. In single-instance image retrieval, the whole image is treated as a single instance, and the features characterizing the global view of an image, such as color histograms, are used to compute the similarity between images. However, the user is usually interested in only a small part of the image, rather than the whole image. If we use the features from the whole image area to represent an image, the useful information that the user is of interest may be overridden by noisy information from irrelevant regions. Consider an example in Fig. 1 which shows some images from the MSRA-MM data set. Supposing that the user is interested in the images on "soccer," Fig. 1(b) and (c) is associated with the label "soccer" and should have higher similarity than Fig. 1(a) and (b). Nevertheless, if the whole image is transformed into a global feature vector, Fig. 1(a) and (b) is expected to give higher similarity than Fig. 1(b) and (c). Instead of transforming the whole image into a single instance, we can consider it from a multiple-instance's view. In multiple-instance setting, an image is segmented into several content regions. Each image is considered as a bag, and each region is treated as an instance. A bag (image) is classified as of the user interest if it contains at least one instance (region) that the user is of interest [22]–[25]. It can be seen that the bag (image) label is determined by only the instance (region) that the user is of interest, with the rest instances (regions) being irrelevant, which can reduce the influence of noisy information from irrelevant regions. From a multiple-instance aspect, Fig. 1(b) and (c) will be classified to the "soccer" class, rather than the "rugby" class, since they contain the regions on "soccer," but do not involve the regions on "rugby." Moreover, the experimental results in the previous studies [22]–[25] have shown that transforming the data into multiple-instance data can obtain improved performance than considering it as a single instance.

In this paper, we address the multi-instance OR problem where the OR classifier is built on multiple-instance data, instead of single-instance data. Different from single-instance OR where an instance is explicitly associated with a label, in multi-instance OR, only the bag label is given and the specific instance label is unavailable. To deal with this problem, we propose a novel multiple-instance ordinal regression (MIOR) method. In MIOR, a set of parallel hyperplanes is utilized to separate the multiple-instance data, and the ordering information of labels is incorporated into learning the model by imputing the hyperplanes with an order. Moreover, considering that the bag may contain instances not belongs to its class, for each bag, the instance which is closest to the middle of the corresponding class is selected to learn the classifier. As a result, the derived optimization function is nonconvex, and the constrained concave–convex procedure (CCCP) [26]–[29] is applied to decompose it into a series of convex subproblems.

The main contributions of this paper are as follows.

1) The OR problem on multiple-instance data is introduced. To the best of our knowledge, this paper is the first attempt that addresses the multi-instance OR problem.
2) Compared with the single-instance OR methods [11], [15], [40], MIOR is able to learn the OR classifier on multiple-instance data where the label is associated with a set of instances, and the label of a single instance is unavailable.
3) Numerical studies on real-world data sets show that MIOR achieves explicitly improvements in terms of the mean zero-one errors and mean absolute errors compared with the existing single-instance OR methods.

MIOR is an SVM-based method, which can handle the multiclass classification problems when the labels are in orders and the data are in multiple-instance form. It is different from the traditional multiclass multiple-instance classification methods [31]–[35]. The traditional multiclass multiple-instance classification methods do not take the label ordering information into account, and the learned classifier is usually a group of disordered intersecting hyperplanes. By contrast, MIOR learns a set of ordered parallel hyperplanes. For the multiclass classification data where the classes are ranked in orders, when only a small amount of labeled data is available, it is more desirable to separate it using a set of ordered parallel hyperplanes, rather than a group of disordered intersecting hyperplanes. Moreover, MIOR is distinguished from the multiple-instance regression methods [36], [37]. The labels in multi-instance OR are discrete and finite, while the target values in multiple-instance regression are continuous and infinite.

The rest of this paper is organized as follows. Section II reviews the existing work on OR and multiple-instance learning. Section III presents the details of the proposed MIOR method. Experiments are conducted in Section IV. Section V concludes this paper and outlines the future work.

II. Related Work

In this paper, we extend the single-instance OR methods to multiple-instance data and propose a novel MIOR model. In the following, the previous work on single-instance OR and multiple-instance learning is reviewed. Then, the typical single-instance OR method-SVOR-EXCs is introduced.

A. Ordinal Regression

Recently, a considerable number of methods [18], [20], [38], [39] have been proposed to solve the OR problem. Reference [11] utilizes a number of hyperplanes to separate the ordinal classes and presents an SVM-based threshold model. Reference [30] proposes two variants of OR methods: fix-margin SVOR and sum-up-margin SVOR. In fix-margin SVOR, the margins of the closest neighboring classes are fixed and the same, and the margin is then maximized. In sum-up-margin SVOR, the margins of the closest neighboring classes can be different, and the sum of these margins is maximized. Reference [15] presents two SVM-based OR approaches: SVOR-EXCs and SVOR-IMCs. SVOR-EXC introduces explicit constraints on label orders. SVOR-IMC considers the training data from all classes to determine each hyperplane, and the order of labels can be implicitly satisfied. The basic idea of RED-SVM [40], [41] is similar to SVOR-IMC, and the main difference is that RED-SVM minimizes the threshold term θ_i^2. By considering $\langle \mathbf{w}, \theta_i \rangle$ as a new vector, it can be resolved by the standard SVM solvers. Besides SVM, other learning techniques, such as neural networks [16], Gaussian process [17], and ensemble learning [19], are adapted to resolve the OR problems.

Nevertheless, the existing methods are mainly proposed for single-instance OR, and the multi-instance OR problem has not been addressed. In practice, many OR problems, e.g., sentiment classification and image retrieval, can be solved from the multiple-instance aspect. Hence, we extend the single-instance OR method to multiple-instance data and present the MIOR method. Compared with the existing single-instance OR methods, MIOR is able to learn an OR classifier on multiple-instance data where the label is attached to a set of instances, and the instance label is unknown.

B. Multiple-Instance Learning

In multiple-instance learning [35], [42]–[45], a label is associated with a bag of instances and the specific instance label is unknown. Many methods have been put forward for multiple-instance learning. For example, Reference [46] represents the target concept by an axis-parallel rectangle. Reference [44] proposes the diverse density (DD) value to measure co-occurrence of similar instances from different positive bags. Reference [47] combines expectation–maximization with DD. Reference [35] adapts boosting to multiple-instance classification. Reference [42] puts forward DD-SVM through a feature mapping defined on instance prototypes learned by DD. Reference [31] presents mi-SVM for instance-level classification and MI-SVM for bag-level classification. mi-SVM initializes all instances in positive bags as positive, and MI-SVM selects one instance to replace the positive bag. The classifier is learned iteratively until each positive bag has at least one instance as positive. Reference [43] embeds the bags in the feature space and selects the important instances.

However, most of the existing multiple-instance learning methods assume that there is no order among the labels. In many real-world multiple-instance applications, there exhibits an order among the labels. For example, in image retrieval, there is a natural order among the labels: *relevant*, *partial relevant* and *irrelevant*, and misclassifying the *relevant* image as *irrelevant* should have a larger penalty error than misclassifying it as *partial relevant*. This label ordering information can be included to improve the classification performance. The existing multiple-instance learning methods do not take the scenario of label orders into account, and the classes separated by the classifier are usually disordered. In this paper, we propose a novel learning model for multi-instance OR problems. It is able to incorporate the label ordering information into improving the classification performance.

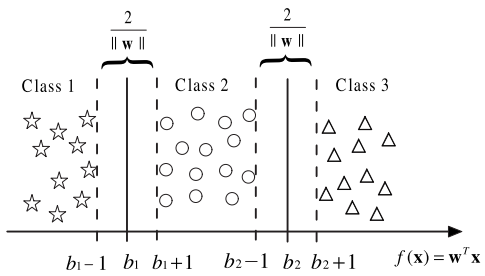

Fig. 2. Example of SVOR-EXC. Class 1: SVOR-EXC is proposed for single-instance OR problems. The training set contains a number of instances and all instance labels are available. The instances of class 1, 2, and 3 are represented as "⋆," "○," and "Δ," respectively. Class 2: SVOR-EXC seeks two parallel hyperplanes, i.e., $(\mathbf{w}, b_1)$ and $(\mathbf{w}, b_2)$, to separate the three classes. Class 3: Margin between class 1 and 2 is the same with that between class 2 and 3, being $(2/\|\mathbf{w}\|)$. The SVOR-EXC problem is formulated by maximizing the margin between the closest neighboring classes, i.e., $(2/\|\mathbf{w}\|)$.

C. SVOR-EXC

There are K classes of training instances $\{\mathbf{x}_1^p, \ldots, \mathbf{x}_{n_p}^p\}_{p=1}^K$, where n_p is the number of instances in class p. To separate the K classes, SVOR-EXC [15] seeks $K-1$ parallel hyperplanes $(\mathbf{w}, b_1), \ldots, (\mathbf{w}, b_{K-1})$, where there is one vector $\mathbf{w}$ and $K-1$ variables $b_1, \ldots, b_{K-1}$ needed to be optimized. The margins of the closest neighboring classes are the same, being $(2/\|\mathbf{w}\|)$, and the learning problem of SVOR-EXC is formulated to maximize the margin $(2/\|\mathbf{w}\|)$, as follows:

$$
\begin{aligned}
\min_{\zeta_i^p, \zeta_i^{'p1} \geq 0} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{p=2}^{K}\sum_{i=1}^{n_p}\zeta_i^p + \sum_{p=1}^{K-1}\sum_{i=1}^{n_p}\zeta_i^{'p}\right) \\
\text{s.t. } & \mathbf{w}^T\mathbf{x}_i^1 - b_1 \leq -1 + \zeta_i^{'1}, \quad i = 1, \ldots, n_1 \\
& \mathbf{w}^T\mathbf{x}_i^p - b_{p-1} \geq 1 - \zeta_i^p, \quad i = 1, \ldots, n_p \\
& p = 2, \ldots, K-1 \\
& \mathbf{w}^T\mathbf{x}_i^p - b_p \leq -1 + \zeta_i^{'p}, \quad i = 1, \ldots, n_p \\
& p = 2, \ldots, K-1 \\
& \mathbf{w}^T\mathbf{x}_i^K - b_{K-1} \geq 1 - \zeta_i^K, \quad i = 1, \ldots, n_K \\
& b_{p-1} \leq b_p, \quad p = 2, \ldots, K-1
\end{aligned} \tag{1}
$$

where C is a regularization parameter; ζ_i^p and $\zeta_i^{'p}$ are error terms. The first set of constraints indicates that the instances of class 1 should lie on the negative side of hyperplane $(\mathbf{w}, b_1)$. The second and third sets of constraints imply that the instances of class p $(p = 2, \ldots, K-1)$ need to lie between hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$. The fourth set of constraints represents that the instances of class K are located on the positive side of hyperplane $(\mathbf{w}, b_{K-1})$. For example, in Fig. 2, class 1, 2, and 3 are separated by two parallel hyperplanes $(\mathbf{w}, b_1)$ and $(\mathbf{w}, b_2)$. The instances of class 1 are located on the negative (left) side of hyperplane $(\mathbf{w}, b_1)$.

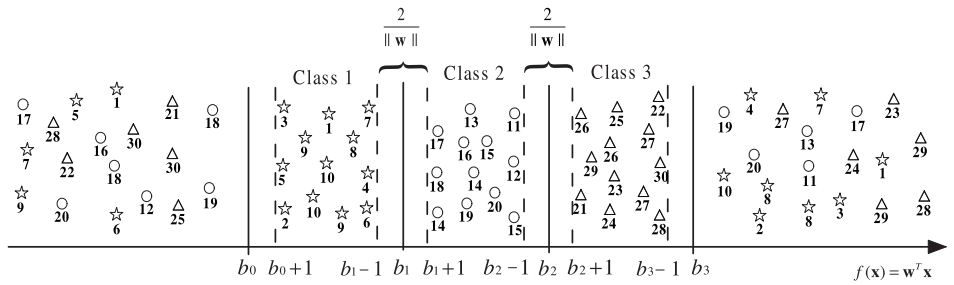

Fig. 3. Example of the proposed MIOR method. (1) MIOR is designed for multi-instance OR problems. The training set consists of three classes, and each class has a number of bags. The "⋆," "○," and "Δ" signs denote the instances in class 1, 2, and 3, respectively. The signs with the same number indicate the instances in the same bag. For example, "⋆" signs with the number "7" are instances in the 7th bag, and this bag belongs to class 1. It can be seen that each bag contains several instances. The bag label is known, but the specific instance labels are unavailable. For example, the 7th bag ("⋆" signs with the number "7") has three instances. The 7th bag belongs to class 1, but the three instances in the 7th bag do not necessarily belong to class 1. According to the description of multiple-instance learning, the bag of class p contains at least one instance belonging to class p. This is to say, the bag of class p may contain instances not belonging to class p. For the three instances in the 7th bag, one instance falls inside class 1 (i.e., inside the range of hyperplanes $(\mathbf{w}, b_0)$ and $(\mathbf{w}, b_1)$), while the other two instances lie outside the three classes (i.e., outside the range of hyperplanes $(\mathbf{w}, b_0)$ and $(\mathbf{w}, b_3)$). Since the 7th bag belongs to class 1, the instance which falls inside class 1 is more likely to belong to class 1, compared with the other two instances. In this paper, MIOR is proposed by requiring that each bag has at least one instance classified to its corresponding class. (2) Different from SVOR-EXC, MIOR seeks four hyperplanes $(\mathbf{w}, b_0)$, …, $(\mathbf{w}, b_3)$. The three classes are included between hyperplanes $(\mathbf{w}, b_0)$ and $(\mathbf{w}, b_3)$, and the classes are separated by hyperplanes $(\mathbf{w}, b_1)$ and $(\mathbf{w}, b_2)$. (3) Besides requiring that each bag has at least one instance classified to its corresponding class, MIOR is formulated by maximizing the margin $(2/\|\mathbf{w}\|)$ between the closest neighboring classes, and meanwhile minimizing the margin between the first hyperplane and the last hyperplane, i.e., $b_3 - b_0$.

Those of class 2 are between hyperplane $(\mathbf{w}, b_1)$ and $(\mathbf{w}, b_2)$. Those of class 3 lie on the positive (right) side of hyperplane $(\mathbf{w}, b_2)$.

The margins of the closest neighboring classes are the same. As shown in Fig. 2, the margin between class 1 and 2 is the same with that between class 2 and 3, being $(2/\|\mathbf{w}\|)$. The margin $(2/\|\mathbf{w}\|)$ is maximized in the learning problem (1). After the $K-1$ hyperplanes are obtained, a test instance $\mathbf{x}$ is classified to class 1 if $\mathbf{w}^T\mathbf{x} - b_1 \leq 0$ is satisfied. It is assigned to class p $(p = 2, \ldots, K-1)$ if it has $\mathbf{w}^T\mathbf{x} - b_{p-1} > 0$ and $\mathbf{w}^T\mathbf{x} - b_p \leq 0$. It is categorized to class K if $\mathbf{w}^T\mathbf{x} - b_{K-1} > 0$ is met.

SVOR-EXC is designed for single-instance OR problems where each instance has an explicit label. However, in multi-instance OR problems, a label is associated with a bag of instances, and the instance label is not given. Hence, SVOR-EXC cannot be straightforwardly applied to deal with the multi-instance OR problem. To deal with this problem, we present a novel MIOR approach, which can effectively solve the OR problem when the instance label is unavailable.

III. Proposed Approach

A. Problem Statement

Suppose that the training set consists of K classes. Class p $(p = 1, \ldots, K)$ has n_p bags: $B_1^p, \ldots, B_{n_p}^p$, where B_i^p is the ith bag in class p. Each bag has a number of instances. The jth instance in bag B_i^p is denoted as B_{ij}^p. A bag belongs to class p $(1 \leq p \leq K)$ if it contains at least one instance of class p. The key challenge of multi-instance OR is that the specific instance labels in a bag are unavailable. The bag of class p may contain instances not belonging to class p, in addition to at least one instance of class p [48], [49]. The goal of multi-instance OR is to learn an OR classifier on multiple-instance data where the bag label is given, but the instance label is unknown.

SVOR-EXC seeks $K-1$ hyperplanes to separate K classes. Slightly different from SVOR-EXC, MIOR constructs $K+1$ parallel hyperplanes $(\mathbf{w}, b_0)$, $(\mathbf{w}, b_1)$, …, $(\mathbf{w}, b_K)$, where there is one vector $\mathbf{w}$ and $K+1$ variables $b_0, \ldots, b_K$ needed to be optimized. As an example in Fig. 3, the three classes are located between the first hyperplane $(\mathbf{w}, b_0)$ and the last hyperplane $(\mathbf{w}, b_K)$, and the K classes are separated by hyperplanes $(\mathbf{w}, b_1)$, $(\mathbf{w}, b_2)$, …, $(\mathbf{w}, b_{K-1})$. The margin between the closest neighboring classes is $(2/\|\mathbf{w}\|)$, and the margin between the first hyperplane $(\mathbf{w}, b_0)$ and the last hyperplane $(\mathbf{w}, b_K)$ is $b_K - b_0$. The MIOR problem is formulated by maximizing the margin between the closest neighboring classes, i.e., $(2/\|\mathbf{w}\|)$, and minimizing the margin between the first hyperplane and the last hyperplane, i.e., $b_K - b_0$.

B. Formulation

MIOR includes the training data between hyperplanes $(\mathbf{w}, b_0)$ and $(\mathbf{w}, b_K)$, and separates the different classes using a set of parallel hyperplanes. The learning problem can be derived by minimizing the margin $b_K - b_0$ between the first hyperplane and the last hyperplane, and maximizing the margin $(2/\|\mathbf{w}\|)$ of the closest neighboring classes, as follows:

$$\begin{aligned}
\min_{\xi_{ij}^p, \xi_{ij}^{'p}, \eta \geq 0} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C_0\eta + C_1 \sum_p \sum_i \sum_j \left(\xi_{ij}^p + \xi_{ij}^{'p}\right) \\
\text{s.t.} \quad & \bigcup_{j=1}^{|B_i^p|} \left[I\left(\mathbf{w}^T B_{ij}^p - b_{p-1} \geq 1 - \xi_{ij}^p\right) \right. \\
& \left. \cap I\left(\mathbf{w}^T B_{ij}^p - b_p \leq -1 + \xi_{ij}^{'p}\right) \right] = 1 \\
& i = 1, \ldots, n_p, \quad p = 1, \ldots, K \\
& b_K - b_0 \leq \eta, \\
& b_{p-1} \leq b_p, \quad p = 1, \ldots, K
\end{aligned} \tag{2}$$

where C_0 and C_1 are regularization parameters, ξ_{ij}^p and $\xi_{ij}^{'p}$ are error terms, η is a variable to be optimized, and $|B_i^p|$ denotes the number of instances in bag B_i^p. I(expression) is an indication function. I(expression) $= 1$ holds if the expression is true. Otherwise, it has I(expression) $= 0$. "$\cup$" is a union operator and "$\cap$" is an intersection operator.

As shown in Fig. 3, $(2/\|\mathbf{w}\|)$ is the margin of the closest neighboring classes. In problem (2), minimizing $(1/2)\|\mathbf{w}\|^2$ is

equivalent to maximizing the margin of the closest neighboring classes. $b_K - b_0$ is the margin between the first hyperplane $(\mathbf{w}, b_0)$ and the last hyperplane $(\mathbf{w}, b_K)$. We let $b_K - b_0 \leq \eta$, and minimizing η implies the minimization of $b_K - b_0$. The data can be better generalized by minimizing the tube between the first hyperplane and the last hyperplane, and meanwhile maximizing the margin of the closest neighboring classes.

The first set of constraints in problem (2) requires that for each bag B_i^p of class p, there exists at least one instance B_{ij}^p which belongs to class p, lying inside the range of hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$. Specifically, when $I(\mathbf{w}^T B_{ij}^p - b_{p-1} \geq 1-\xi_{ij}^p) \cap I(\mathbf{w}^T B_{ij}^p - b_p \leq -1+\xi_{ij}^{'p}) = 1$ holds, it indicates that instance B_{ij}^p lies between hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$. Moreover, when $\cup_{j=1}^{|B_i^p|}[I(\mathbf{w}^T B_{ij}^p - b_{p-1} \geq 1-\xi_{ij}^p) \cap I(\mathbf{w}^T B_{ij}^p - b_p \leq -1+\xi_{ij}^{'p})] = 1$ holds, it implies that for all instances in bag B_i^p, there is at least one instance B_{ij}^p which falls inside the range of hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$.

In single-instance OR, as shown in problem (1), each instance is associated with a label. The instance $\mathbf{x}_i^p$ of class p $(p = 2, \ldots, K-1)$ is required to lie between hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$, and to satisfy constraints $\mathbf{w}^T \mathbf{x}_i^p - b_{p-1} \geq 1 - \zeta_i^p$ and $\mathbf{w}^T \mathbf{x}_i^p - b_p \leq -1 + \zeta_i^{'p}$. In multi-instance OR, although the instance label is unavailable, each bag contains at least one instance of its class. For bag B_i^p of class p, there must exist at least one instance which belongs to class p, lies between hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$, and satisfies constraints $\mathbf{w}^T B_{ij}^p - b_{p-1} \geq 1 - \xi_{ij}^p$ and $\mathbf{w}^T B_{ij}^p - b_p \leq -1 + \xi_{ij}^{'p}$. Hence, in problem (2), the first set of constraints $\cup_{j=1}^{|B_i^p|}[I(\mathbf{w}^T B_{ij}^p - b_{p-1} \geq 1-\xi_{ij}^p) \cap I(\mathbf{w}^T B_{ij}^p - b_p \leq -1+\xi_{ij}^{'p})] = 1$ is proposed to ensure that each bag B_i^p of class p should contain at least one instance B_{ij}^p which belongs to class p, lying inside the range of hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$.

The third set of constraints $b_{p-1} \leq b_p$ requires that the parallel hyperplanes should be ranked to represent the label ordering information. It is noted that problem (2) involves the computation of indication functions which are difficult to solve. We can transform it into a more simplified form. Let $(\mathbf{w}, ((b_{p-1} + b_p)/2))$ be the hyperplane lying in the middle of hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$.

Theorem 1: If bag B_i^p contains at least one instance lying inside the range of hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$, the instance B_{ij}^p $(j = \arg\min_{j \in B_i^p}((|\mathbf{w}^T B_{ij}^p - ((b_{p-1} + b_p)/2)|)/\|\mathbf{w}\|))$ which is nearest to the middle hyperplane $(\mathbf{w}, ((b_{p-1} + b_p)/2))$ must be located inside this range.

Proof: Consider an example in Fig. 4, where class p is bounded by hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$; bag B_i^p belongs to class p and contains three instances B_{ij}^p, B_{ig}^p, and B_{ih}^p. The distance from instance B_{ij}^p to the middle hyperplane $(\mathbf{w}, ((b_{p-1} + b_p)/2))$ is $d_j = ((|\mathbf{w}^T B_{ij}^p - ((b_{p-1} + b_p)/2)|)/\|\mathbf{w}\|)$. The distance between the middle hyperplane $(\mathbf{w}, (b_{p-1} + b_p/2))$ and the bounded hyperplane $(\mathbf{w}, b_{p-1})$ or $(\mathbf{w}, b_p)$ is $((b_p - b_{p-1})/2\|\mathbf{w}\|)$. It is seen that B_{ij}^p is the instance nearest to the middle hyperplane $(\mathbf{w}, ((b_{p-1} + b_p)/2))$.

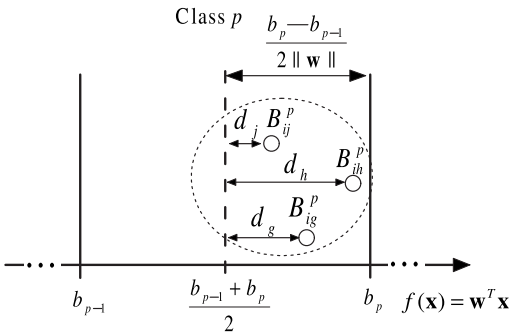

Fig. 4. Illustration of Theorem 1. To make the presentation clear, only class p is presented and the other classes are neglected. In class p, only bag B_i^p is drawn out, and the other bags are omitted. Bag B_i^p contains three instances B_{ij}^p, B_{ig}^p, and B_{ih}^p. B_{ij}^p is the instance nearest to the middle hyperplane $(\mathbf{w}, ((b_{p-1} + b_p)/2))$. If bag B_i^p has at least one instance lying inside the range of hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$, instance B_{ij}^p must be located inside this range. This is because B_{ij}^p is the instance nearest to the middle hyperplane $(\mathbf{w}, ((b_{p-1} + b_p)/2))$, and the distance d_j from B_{ij}^p to the middle hyperplane is smaller than d_g and d_h. If instance B_{ij}^p is out of the range of hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$, instances B_{ig}^p and B_{ih}^p must be outside this range.

Assume that bag B_i^p has at least one instance lying inside the range of hyper-planes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$. We can deduce that instance B_{ij}^p must be inside this range. This is because B_{ij}^p is the instance nearest to the middle hyper-plane $(\mathbf{w}, ((b_{p-1} + b_p)/2))$, and thus, it has $d_g > d_j$ and $d_h > d_j$, where d_h, d_g, and d_j are the distances from instances B_{ih}^p, B_{ig}^p, and B_{ij}^p to the middle hyper-planes $(\mathbf{w}, ((b_{p-1} + b_p)/2))$, respectively. If instance B_{ij}^p is outside the range of hyper-planes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$, the distance d_j from instance B_{ij}^p to the middle hyper-plane $(\mathbf{w}, ((b_{p-1} + b_p)/2))$ must be larger than $((b_p - b_{p-1})/2\|\mathbf{w}\|)$, i.e., $d_j > ((b_p - b_{p-1})/2\|\mathbf{w}\|)$. Since $d_g > d_j$ and $d_h > d_j$ hold, it has $d_g > d_j > ((b_p - b_{p-1})/2\|\mathbf{w}\|)$ and $d_h > d_j > ((b_p - b_{p-1})/2\|\mathbf{w}\|)$. Hence, all the instances B_{ij}^p, B_{ig}^p, and B_{ih}^p are outside the range of hyper-planes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$. Hence, if bag B_i^p has at least one instance inside the range of $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$, the instance B_{ij}^p nearest to the middle hyperplane $(\mathbf{w}, ((b_{p-1} + b_p)/2))$ must be inside this range.

Based on Theorem 1, problem (2) can be transformed into

$$
\begin{aligned}
\min_{\zeta_i^p, \zeta_i^{'p}, \eta \geq 0} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C_0\eta + C_1 \sum_p \sum_i \left(\zeta_i^p + \zeta_i^{'p}\right) \\
\text{s.t.} \quad & \mathbf{w}^T B_{ij}^p - b_{p-1} \geq 1 - \zeta_i^p \\
& \mathbf{w}^T B_{ij}^p - b_p \leq -1 + \zeta_i^{'p} \\
& j = \arg\min_{j \in B_i^p} \left| \mathbf{w}^T B_{ij}^p - \frac{b_{p-1} + b_p}{2} \right| \\
& i = 1, \ldots, n_p, \quad p = 1, \ldots, K \\
& b_K - b_0 \leq \eta \\
& b_{p-1} \leq b_p, \quad p = 1, \ldots, K.
\end{aligned} \tag{3}
$$

Compared with Theorem 1, we simplify $j = \arg\min_{j\in B_i^p} (((|\mathbf{w}^T B_{ij}^p - ((b_{p-1}+b_p)/2)|)/\|\mathbf{w}\|)$ as $j = \arg\min_{j\in B_i^p} |\mathbf{w}^T B_{ij}^p - ((b_{p-1}+b_p)/2)|$ in problem (3), since $\|\mathbf{w}\|$ is the same for all instances.

In multi-instance OR, each bag B_i^p of class p should contain at least one instance B_{ij}^p, which belongs to class p, lying inside the range of hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$. To ensure that each bag B_i^p has at least one instance located inside the range of hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$, problem (2) employs the indication functions $I(\text{expression})$. When one instance is located inside the range of hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$, it has $I(\mathbf{w}^T B_{ij}^p - b_{p-1} \geq 1 - \xi_{ij}^p) \cap I(\mathbf{w}^T B_{ij}^p - b_p \leq -1 + \xi_{ij}^{'p}) = 1$.

However, the indication functions are difficult to solve. Different from problem (2), problem (3) does not use the indication functions. In order to ensure that each bag B_i^p has at least one instance located inside the range of hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$, problem (3) requires the instance B_{ij}^p, which is nearest to the middle hyperplane $(\mathbf{w}, ((b_{p-1}+b_p)/2))$ to lie between hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$ and to satisfy constraints $\mathbf{w}^T B_{ij}^p - b_{p-1} \geq 1 - \zeta_i^p$ and $\mathbf{w}^T B_{ij}^p - b_p \leq -1 + \zeta_i^{'p}$, as seen from the first and the second sets of constraints. This is because according to Theorem 1, if bag B_i^p has at least one instance located between hyperplanes $(\mathbf{w}, b_{p-1})$ and $(\mathbf{w}, b_p)$, the instance B_{ij}^p which is nearest to the middle hyperplane $(\mathbf{w}, ((b_{p-1}+b_p)/2))$ must be inside this range.

However, problem (3) is still difficult to resolve, since the "min" function is included in the subscripts of instances. To deal with this problem, we transform problem (3) into an equivalent form, as shown in Theorem 2.

Theorem 2: Problem (3) is equivalent to

$$\begin{aligned} \min_{\zeta_i^p, \eta \geq 0} \ & \frac{1}{2}\|\mathbf{w}\|^2 + C_0\eta + C_1 \sum_p \sum_i \zeta_i^p \\ \text{s.t.} \ & \min_{j\in B_i^p} \left| \mathbf{w}^T B_{ij}^p - \frac{\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p}{2} \right| \\ & \leq \frac{\mathbb{b}^T\mathbf{e}_p - \mathbb{b}^T\mathbf{e}_{p-1}}{2} - 1 + \zeta_i^p \\ & i = 1, \ldots, n_p, \quad p = 1, \ldots, K \\ & \mathbb{b}^T\mathbf{e}_K - \mathbb{b}^T\mathbf{e}_0 \leq \eta \\ & \mathbb{b}^T\mathbf{e}_{p-1} \leq \mathbb{b}^T\mathbf{e}_p, \quad p = 1, \ldots, K \end{aligned} \tag{4}$$

where it has $\mathbb{b} = \{b_0, b_1, \ldots, b_K\}^T \in R^{K+1}$, and $\mathbf{e}_p \in R^{K+1}$ is a $(K+1)$-dimensional column vector with the pth element being 1 and the others being 0. Hence, $\mathbb{b}^T\mathbf{e}_p$ is equal to b_p.

Proof: Let $j = \arg\min_{j\in B_i^p} |\mathbf{w}^T B_{ij}^p - ((\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p)/2)|$, and the first set of constraints in problem (4) is changed into $|\mathbf{w}^T B_{ij}^p - ((\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p)/2)| \leq ((\mathbb{b}^T\mathbf{e}_p - \mathbb{b}^T\mathbf{e}_{p-1})/2) - 1 + \zeta_i^p$. When $\mathbf{w}^T B_{ij}^p - ((\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p)/2) \geq 0$ holds, it can be reduced to $\mathbf{w}^T B_{ij}^p - \mathbb{b}^T\mathbf{e}_p \leq -1 + \zeta_i^p$. When $\mathbf{w}^T B_{ij}^p - ((\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p)/2) \leq 0$ comes true, it is rewritten as $\mathbf{w}^T B_{ij}^p - \mathbb{b}^T\mathbf{e}_{p-1} \geq 1 - \zeta_i^p$.

In problem (3), an instance should satisfy both of the constraints $\mathbf{w}^T B_{ij}^p - \mathbb{b}^T\mathbf{e}_{p-1} \geq 1 - \zeta_i^p$ and $\mathbf{w}^T B_{ij}^p - \mathbb{b}^T\mathbf{e}_p \leq -1 + \zeta_i^{'p}$, as seen from the first and the second sets of constraints. However, in problem (4), when $\mathbf{w}^T B_{ij}^p - ((\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p)/2) \geq 0$ holds, it only needs to meet the constraint $\mathbf{w}^T B_{ij}^p - \mathbb{b}^T\mathbf{e}_p \leq -1 + \zeta_i^p$. This is because as shown in Fig. 5, when it has $\mathbf{w}^T B_{ij}^p - ((\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p)/2) \geq 0$, the instance (e.g., B_{1j}^p) lies on the positive (right) side of hyperplane $(\mathbf{w}, ((\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p)/2))$. Considering that hyperplane $(\mathbf{w}, ((\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p)/2))$ is on the positive side of hyperplane $(\mathbf{w}, \mathbb{b}^T\mathbf{e}_{p-1})$, the instance (e.g., B_{1j}^p) is naturally on the positive side of $(\mathbf{w}, \mathbb{b}^T\mathbf{e}_{p-1})$ and satisfies the constraint $\mathbf{w}^T B_{ij}^p - \mathbb{b}^T\mathbf{e}_{p-1} \geq 1 - \zeta_i^p$. Hence, the constraint $\mathbf{w}^T B_{ij}^p - \mathbb{b}^T\mathbf{e}_{p-1} \geq 1 - \zeta_i^p$ is implicitly met and thus can be neglected from the learning problem.

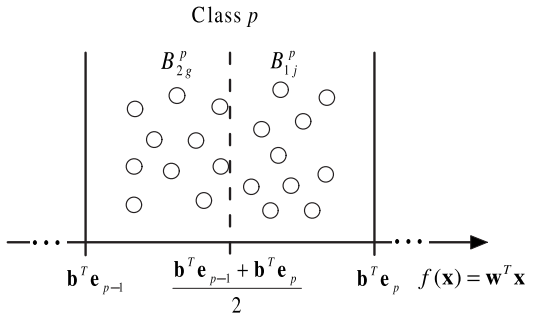

Fig. 5. Illustration of the first set of constraints in problem (4). To be clear, for each bag, only the instance B_{ij}^p ($j = \arg\min_{j\in B_i^p} |\mathbf{w}^T B_{ij}^p - ((\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p)/2)|$) which has the minimum distance to the middle hyperplane $(\mathbf{w}, ((\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p)/2))$ is drawn out, and the other instances in the bag are omitted. Moreover, only class p is given, and the other classes are neglected.

Likewise, when $\mathbf{w}^T B_{ij}^p - ((\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p)/2) \leq 0$ comes true, it only needs to satisfy the constraint $\mathbf{w}^T B_{ij}^p - \mathbb{b}^T\mathbf{e}_{p-1} \geq 1 - \zeta_i^p$. When it has $\mathbf{w}^T B_{ij}^p - ((\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p)/2) \leq 0$, the instance (e.g., B_{2g}^p) lies on the negative (left) side of hyperplane $(\mathbf{w}, ((\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p)/2))$. Since $(\mathbf{w}, ((\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p)/2))$ is on the negative side of $(\mathbf{w}, \mathbb{b}^T\mathbf{e}_p)$, the instance (e.g., B_{2g}^p) is naturally on the negative side of $(\mathbf{w}, \mathbb{b}^T\mathbf{e}_p)$ and meets $\mathbf{w}^T B_{ij}^p - \mathbb{b}^T\mathbf{e}_p \leq -1 + \zeta_i^{'p}$. Hence, this constraint can be ignored.

C. CCCP Decomposition

In the learning problem (4), the first set of constraints, i.e., $\min_{j\in B_i^p} |\mathbf{w}^T B_{ij}^p - ((\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p)/2)| \leq ((\mathbb{b}^T\mathbf{e}_p - \mathbb{b}^T\mathbf{e}_{p-1})/2) - 1 + \zeta_i^p$, involves the "min" function and the "absolute" function, which make the learning problem nonconvex. To solve this nonconvex problem, we employ the CCCP technique [26]–[29]. It is proposed to solve the nonconvex problem whose objective function can be expressed as a difference of convex functions. CCCP has been extensively applied to solve the SVM-based nonconvex problems [25], [50]–[56].

Define $g(\mathbf{w}, \mathbb{b}, i, j, p) = |\mathbf{w}^T B_{ij}^p - ((\mathbb{b}^T\mathbf{e}_{p-1} + \mathbb{b}^T\mathbf{e}_p)/2)|$, and $h(\mathbf{w}, \mathbb{b}, i, p) = \min_{j\in B_i^p} g(\mathbf{w}, \mathbb{b}, i, j, p)$. Hence, the first set of constraints in problem (4) turns to be

$h(\mathbf{w}, \mathbb{b}, i, p) \leq ((\mathbb{b}^T\mathbf{e}_p - \mathbb{b}^T\mathbf{e}_{p-1})/2) - 1 + \zeta_i^p$. Based on this, the CCCP technique can be used to decompose the nonconvex learning problem (4) into a series of convex quadratic programming (QP) problems. Given an initial point $(\mathbf{w}^{(0)}, \mathbb{b}^{(0)})$, CCCP calculates $(\mathbf{w}^{(t+1)}, \mathbb{b}^{(t+1)})$ from $(\mathbf{w}^{(t)}, \mathbb{b}^{(t)})$ iteratively by replacing $h(\mathbf{w}, \mathbb{b}, i, p)$ with its corresponding first-order Taylor expansion at $(\mathbf{w}^{(t)}, \mathbb{b}^{(t)})$. The resulting QP problem is then solved until the stopping criterion is met.

To solve problem (4) using CCCP, we first compute the first-order Taylor expansion of $h(\mathbf{w}, \mathbb{b}, i, p)$ at $(\mathbf{w}^{(t)}, \mathbb{b}^{(t)})$. However, it is a nonsmooth function. As done in [25] and [50]–[53], we replace its gradient with the corresponding subgradient

$$\begin{aligned}
&\left.\frac{\partial h(\mathbf{w}, \mathbb{b}, i, p)}{\partial \mathbf{w}}\right|_{\substack{\mathbf{w}=\mathbf{w}^{(t)}\\ \mathbb{b}=\mathbb{b}^{(t)}}} \\
&\quad = \frac{\partial h(\mathbf{w}, \mathbb{b}, i, p)}{\partial g(\mathbf{w}, \mathbb{b}, i, j, p)} \times \left.\frac{\partial g(\mathbf{w}, \mathbb{b}, i, j, p)}{\partial \mathbf{w}}\right|_{\substack{\mathbf{w}=\mathbf{w}^{(t)}\\ \mathbb{b}=\mathbb{b}^{(t)}}} \\
&\quad = \sum_{j\in B_i^p} \theta_{ijp}^{(t)}\lambda_{ijp}^{(t)} B_{ij}^p \qquad (5) \\
&\left.\frac{\partial h(\mathbf{w}, \mathbb{b}, i, p)}{\partial \mathbb{b}}\right|_{\substack{\mathbf{w}=\mathbf{w}^{(t)}\\ \mathbb{b}=\mathbb{b}^{(t)}}} \\
&\quad = \frac{\partial h(\mathbf{w}, \mathbb{b}, i, p)}{\partial g(\mathbf{w}, \mathbb{b}, i, j, p)} \times \left.\frac{\partial g(\mathbf{w}, \mathbb{b}, i, j, p)}{\partial \mathbb{b}}\right|_{\substack{\mathbf{w}=\mathbf{w}^{(t)}\\ \mathbb{b}=\mathbb{b}^{(t)}}} \\
&\quad = -\frac{1}{2}\sum_{j\in B_i^p} \theta_{ijp}^{(t)}\lambda_{ijp}^{(t)}(\mathbf{e}_{p-1} + \mathbf{e}_p) \qquad (6)
\end{aligned}$$

where it has

$$\theta_{ijp}^{(t)} = \begin{cases} 1, & \text{if } j = \arg\min\limits_{j\in B_i^p} g(\mathbf{w}^{(t)}, \mathbb{b}^{(t)}, i, j, p) \\ 0, & \text{otherwise.} \end{cases} \qquad (7)$$

$$\lambda_{ijp}^{(t)} = \begin{cases} 1, & \text{if } (\mathbf{w}^{(t)})^T B_{ij}^p - \dfrac{(\mathbb{b}^{(t)})^T\mathbf{e}_{p-1} + (\mathbb{b}^{(t)})^T\mathbf{e}_p}{2} \geq 0 \\ -1, & \text{otherwise.} \end{cases} \qquad (8)$$

We substitute (5) and (6) into the first-order Taylor expression of $h(\mathbf{w}, \mathbb{b}, i, p)$ at $(\mathbf{w}^{(t)}, \mathbb{b}^{(t)})$, and the following equation can be obtained:

$$\begin{aligned}
&h(\mathbf{w}, \mathbb{b}, i, p) \\
&\approx h(\mathbf{w}^{(t)}, \mathbb{b}^{(t)}, i, p) + (\mathbf{w} - \mathbf{w}^{(t)})^T \left.\frac{\partial h(\mathbf{w}, \mathbb{b}, i, p)}{\partial \mathbf{w}}\right|_{\substack{\mathbf{w}=\mathbf{w}^{(t)}\\ \mathbb{b}=\mathbb{b}^{(t)}}} \\
&\quad + (\mathbb{b} - \mathbb{b}^{(t)}) \left.\frac{\partial h(\mathbf{w}, \mathbb{b}, i, p)}{\partial \mathbb{b}}\right|_{\substack{\mathbf{w}=\mathbf{w}^{(t)}\\ \mathbb{b}=\mathbb{b}^{(t)}}} \\
&= \mathbf{w}^T \sum_{j\in B_i^p} \theta_{ijp}^{(t)}\lambda_{ijp}^{(t)} B_{ij}^p - \frac{1}{2}\mathbb{b}^T \sum_{j\in B_i^p} \theta_{ijp}^{(t)}\lambda_{ijp}^{(t)}(\mathbf{e}_{p-1} + \mathbf{e}_p).
\end{aligned} \qquad (9)$$

By replacing $h(\mathbf{w}, \mathbb{b}, i, p)$ with its first-order Taylor expression (9) at $(\mathbf{w}^{(t)}, \mathbb{b}^{(t)})$, the learning problem for the tth CCCP iteration can be given by

$$\begin{aligned}
&\min_{\zeta_i^p, \eta\geq 0} \ \frac{1}{2}\|\mathbf{w}\|^2 + C_0\eta + C_1\sum_p\sum_i \zeta_i^p \\
&\text{s.t. } \mathbf{w}^T \sum_{j\in B_i^p} \theta_{ijp}^{(t)}\lambda_{ijp}^{(t)} B_{ij}^p - \frac{1}{2}\mathbb{b}^T \sum_{j\in B_i^p} \theta_{ijp}^{(t)}\lambda_{ijp}^{(t)}(\mathbf{e}_{p-1} + \mathbf{e}_p) \\
&\quad \leq \frac{\mathbb{b}^T\mathbf{e}_p - \mathbb{b}^T\mathbf{e}_{p-1}}{2} - 1 + \zeta_i^p \\
&\quad i = 1, \ldots, n_p, \quad p = 1, \ldots, K \\
&\quad \mathbb{b}^T\mathbf{e}_K - \mathbb{b}^T\mathbf{e}_0 \leq \eta \\
&\quad \mathbb{b}^T\mathbf{e}_{p-1} \leq \mathbb{b}^T\mathbf{e}_p, \quad p = 1, \ldots, K. \qquad (10)
\end{aligned}$$

D. Dual Form

Problem (10) is a QP problem. The QP problem can be solved via the the Lagrange function. In order to derive the Lagrange function, we introduce the Lagrange multipliers $\alpha_i^p \geq 0$, $\beta_i^p \geq 0$, $\rho \geq 0$, $\gamma \geq 0$, and $\mu_p \geq 0$. By introducing these Lagrange multipliers, the Lagrange function of problem (10) can be acquired. We deviate the Lagrange function with $\mathbf{w}$, $\mathbb{b}$, η, and ζ_i^p, and obtain

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} + \sum_{p=1}^{K}\sum_{i=1}^{n_p}\sum_{j\in B_i^p} \alpha_i^p\theta_{ijp}^{(t)}\lambda_{ijp}^{(t)} B_{ij}^p = 0 \qquad (11)$$

$$\begin{aligned}
\frac{\partial L}{\partial \mathbb{b}} &= \rho(\mathbf{e}_K - \mathbf{e}_0) + \sum_{p=1}^{K}\mu_p(\mathbf{e}_{p-1} - \mathbf{e}_p) - \frac{1}{2} \\
&\quad \times \sum_{p=1}^{K}\sum_{i=1}^{n_p}\sum_{j\in B_i^p} \alpha_i^p\left[\mathbf{e}_p\left(\theta_{ijp}^{(t)}\lambda_{ijp}^{(t)} + 1\right)\right. \\
&\quad \left. + \mathbf{e}_{p-1}\left(\theta_{ijp}^{(t)}\lambda_{ijp}^{(t)} - 1\right)\right] = 0 \qquad (12)
\end{aligned}$$

$$\frac{\partial L}{\partial \eta} = C_0 - \rho - \gamma = 0 \qquad (13)$$

$$\frac{\partial L}{\partial \zeta_i^p} = C_1 - \alpha_i^p - \beta_i^p = 0. \qquad (14)$$

From (11), (13), and (14), it is easy to deduce

$$\mathbf{w} = -\sum_{p=1}^{K}\sum_{i=1}^{n_p}\sum_{j\in B_i^p} \alpha_i^p\theta_{ijp}^{(t)}\lambda_{ijp}^{(t)} B_{ij}^p \qquad (15)$$

$$0 \leq \rho \leq C_0 \qquad (16)$$

$$0 \leq \alpha_i^p \leq C_1. \qquad (17)$$

Furthermore, it can be seen that (12) is a linear combination of $\mathbf{e}_p$ ($p = 0, \ldots, K$), and $\mathbf{e}_p$ is a $(K+1)$-dimensional column vector with the pth dimension being 1 and the other dimensions being 0. Hence, (12) can be separated into $K+1$ equations according to different $\mathbf{e}_p$ values. The equations obtained from $\mathbf{e}_0$, $\mathbf{e}_1$, $\ldots$, $\mathbf{e}_K$ are as follows.

From $\mathbf{e}_0$, (18) can be acquired

$$-\frac{1}{2}\sum_{i=1}^{n_1}\sum_{j\in B_i^1} \alpha_i^1\left(\theta_{ij1}^{(t)}\lambda_{ij1}^{(t)} - 1\right) - \rho + \mu_1 = 0. \qquad (18)$$

From $\mathbf{e}_i$ $(i = 1, \ldots, K-1)$, we can obtain

$$-\frac{1}{2}\sum_{i=1}^{n_p}\sum_{j\in B_i^p}\alpha_i^p\left(\theta_{ijp}^{(t)}\lambda_{ijp}^{(t)}+1\right)-\mu_p$$
$$-\frac{1}{2}\sum_{i=1}^{n_{p+1}}\sum_{j\in B_i^{p+1}}\alpha_i^{p+1}\left(\theta_{ij,p+1}^{(t)}\lambda_{ij,p+1}^{(t)}-1\right)+\mu_{p+1}=0. \quad (19)$$

From $\mathbf{e}_K$, the following equation can be derived:

$$-\frac{1}{2}\sum_{i=1}^{n_K}\sum_{j\in B_i^K}\alpha_i^K\left(\theta_{ijK}^{(t)}\lambda_{ijK}^{(t)}+1\right)+\rho-\mu_K=0. \quad (20)$$

By substituting formulas (12)–(20) into the Lagrange function, the dual from of problem (10) is given by

$$\begin{aligned}
\max\ & \sum_{p=1}^{K}\sum_{i=1}^{n_p}\alpha_i^p-\frac{1}{2}\\
&\times\sum_{p,q=1}^{K}\sum_{i=1}^{n_p}\sum_{u=1}^{n_q}\sum_{\substack{j\in B_i^p\\ v\in B_u^q}}\alpha_i^p\delta_{ijp}^{(t)}\left(B_{ij}^p\right)^T B_{uv}^q\delta_{uvq}^{(t)}\alpha_u^q\\
\text{s.t. }& \mu_1+\frac{1}{2}\sum_{i=1}^{n_1}\sum_{j\in B_i^1}\alpha_i^1\left(1-\delta_{ij1}^{(t)}\right)-\rho=0\\
&\mu_{p+1}+\frac{1}{2}\sum_{i=1}^{n_{p+1}}\sum_{j\in B_i^{p+1}}\alpha_i^{p+1}\left(1-\delta_{ij,p+1}^{(t)}\right)\\
&=\mu_p+\frac{1}{2}\sum_{i=1}^{n_p}\sum_{j\in B_i^p}\alpha_i^p\left(1+\delta_{ijp}^{(t)}\right),\quad p=1,\ldots,K-1\\
&\mu_K+\frac{1}{2}\sum_{i=1}^{n_K}\sum_{j\in B_i^K}\alpha_i^K\left(1+\delta_{ijK}^{(t)}\right)-\rho=0\\
&0\le\rho\le C_0\\
&0\le\alpha_i^p\le C_1,\quad i=1,\ldots,n_p,\ \ p=1,\ldots,K\\
&\mu_p\ge 0,\quad p=1,\ldots,K
\end{aligned} \quad (21)$$

where it has $\delta_{ijp}^{(t)}=\theta_{ijp}^{(t)}\lambda_{ijp}^{(t)}$. $\delta_{ijp}^{(t)}$ can be computed from (7) and (8), and are known in the dual form.

The dual form (21) is a convex QP problem with linear constraints, which can be solved by a variety of optimization methods, such as conjugate gradient, active set, interior point methods, and so on. After solving the dual form (21), we can obtain $\mathbf{w}$ and $\mathbb{b}$. For a test bag B_i, its label Y_i is predicted as

$$Y_i=\arg\min_{p=1,\ldots,K}\min_{j\in B_i}\left|\mathbf{w}^T B_{ij}-\frac{\mathbb{b}^T\mathbf{e}_{p-1}+\mathbb{b}^T\mathbf{e}_p}{2}\right|. \quad (22)$$

To predict the label Y_i of bag B_i, we first evaluate the minimum distance from bag B_i to the middle hyperplane $(\mathbf{w}, ((\mathbb{b}^T\mathbf{e}_{p-1}+\mathbb{b}^T\mathbf{e}_p)/2))$ of each class, which is determined by the value of $\min_{j\in B_i}|\mathbf{w}^T B_{ij}-((\mathbb{b}^T\mathbf{e}_{p-1}+\mathbb{b}^T\mathbf{e}_p)/2)|$. Then, B_i is assigned to the class which is with the minimum distance. Furthermore, when the training data are nonlinearly separated, it can be mapped into a higher dimensional feature space via a mapping function $\phi(\cdot)$. The inner product of two instances in the feature space can be calculated as $K(B_{ij}, B_{uv})=\phi(B_{ij})\cdot\phi(B_{uv})$. To extend our method to the nonlinear cases, we need to replace $(B_{ij}^p)^T B_{uv}^q$ with $K(B_{ij}, B_{uv})$ in problem (21), and $\mathbf{w}^T B_{ij}^p$ with $-\sum_{q=1}^{K}\sum_{u=1}^{n_q}\sum_{v\in B_u^q}\alpha_u^q\theta_{uvq}^{(t)}\lambda_{uvq}^{(t)}K(B_{uv}^q, B_{ij}^p)$ in (7), (8), and (22).

TABLE I
PROPOSED ALGORITHM

Algorithm: The overview of the MIOR algorithm.

Input:
1. Dataset: K classes of bags: $\{B_1^p,\ldots,B_{n_p}^p\}_{p=1}^K$;
2. Parameters: regularization parameters C_0 and C_1, CCCP precision ς;

Output: The predicted label Y;

CCCP Iterations:

1: Set $\Delta J=10^{-3}$, $J^{(-1)}=10^{-3}$;
2: Initialize the values of $\mathbf{w}^{(0)}$ and $\mathbb{b}^{(0)}$;
3: Set $t=0$;
4: **While** $|\Delta J/J^{(t-1)}|>\varsigma$ **do**
5: Compute $\theta_{ijp}^{(t)}$ and $\lambda_{ijp}^{(t)}$ by substituting $\mathbf{w}^{(t)}$ and $\mathbb{b}^{(t)}$ according to Equations (7) and (8);
6: Decompose $h(\mathbf{w},\mathbb{b},i,p)$ at $(\mathbf{w}^{(t)},\mathbb{b}^{(t)})$ as $\mathbf{w}^T\sum_{j\in B_i^p}\theta_{ijp}^{(t)}\lambda_{ijp}^{(t)}B_{ij}^p-\frac{1}{2}\mathbb{b}^T\sum_{j\in B_i^p}\theta_{ijp}^{(t)}\lambda_{ijp}^{(t)}(\mathbf{e}_{p-1}+\mathbf{e}_p)$;
7: Obtain problem (10) by replacing $h(\mathbf{w},\mathbb{b},i,p)$ with that in line 5;
8: Derive the dual form of problem (10) as (21);
9: Resolve the dual form (21) to obtain $\mathbf{w}$ and $\mathbb{b}$;
10: t=t+1;
11: Let $\mathbf{w}^{(t)}$=$\mathbf{w}$ and $\mathbb{b}^{(t)}=\mathbb{b}$;
12: Let $\Delta J=J^{(t-2)}-J^{(t-1)}$;
13: **End while**
14: **Label Prediction:**
For a test bag B_i, its predicted label is
$Y_i=\arg\min_{p=1,\ldots,K}\min_{j\in B_i}\left|\mathbf{w}^T B_{ij}-\frac{\mathbb{b}^T\mathbf{e}_p+\mathbb{b}^T\mathbf{e}_{p-1}}{2}\right|$;

E. Algorithm Overview

Table I presents the overview of MIOR, which consists of a series of CCCP iterations. Here, $J^{(t)}=(1/2)\|\mathbf{w}^{(t)}\|^2+C_0\eta^{(t)}+C_1\sum_p\sum_i\zeta_i^{p(t)}$ is the value of the objective function (10) at the tth CCCP iteration, and $\Delta J=J^{(t-2)}-J^{(t-1)}$ is the difference between the objective function values in the $(t-1)$th and $(t-2)$th CCCP iterations. When the proportion of ΔJ and $J^{(t-1)}$ is smaller than a threshold ς, the algorithm terminates. As in [25] and [51], ΔJ is set to be 0.1 in the experiments.

IV. Experiments

In this section, substantial experiments on real-world text and image data sets are conducted to investigate the performance of MIOR. The objectives of experiments are: 1) to evaluate the effectiveness of our method in solving the multi-instance OR problem when only a small amount of labeled data is available to train the classifier and 2) to evaluate the performance variation of our method when different numbers of CCCP iterations are adapted.

A. Evaluation Metrics

To evaluate the performance of our MIOR model, the mean zero-one error and mean absolute error are used as the evaluation metrics. They are popularly used in the OR work [16], [18], [19]. The ways to compute the mean zero-one error and mean absolute error are as follows.

1) *Mean zero-one error:* The error rate of the classifier, that is

$$\frac{1}{n}\sum_{i=1}^{n} I(Y_i^* \neq Y_i) = 1 - \text{Acc} \tag{23}$$

where n is the number of bags in the training set; Y_i^* is the true label of the bag and Y_i is the predicted label; and Acc is the classifier accuracy. It has $I(Y_i^* \neq Y_i) = 1$ when the true label is not equal to the predicted label, i.e., $Y_i^* \neq Y_i$. Otherwise, it has $I(Y_i^* \neq Y_i) = 0$. Hence, the mean zero-one error implies the fraction of incorrect prediction on individual bags, namely the error rate of the classifier. The smaller the mean zero-one error is, the better performance the algorithm obtains.

2) *Mean Absolute Error:* The average deviation of the predicted label from the true one which is treated as a consecutive integer, that is

$$\frac{1}{n}\sum_{i=1}^{n} \left|Y_i^* - Y_i\right| \tag{24}$$

where $|Y_i^* - Y_i|$ is the distance between the true label Y_i^* and the predicted label Y_i. Instead of using 0 or 1 as the mean zero-one error does, the mean absolute error uses the difference of Y_i^* and Y_i to represent how far the predicted label is from the true one. In OR, there is some kind of ordering information among the labels, and using the mean zero-one error cannot exactly evaluate whether the classifier tends to predict the data close to the true class. The smaller the mean absolute error is, the better performance the algorithm acquires.

B. Baselines and Experimental Settings

Experiments are conducted to compare MIOR with the single-instance OR methods (i.e., SVOR-EXC [15], SVOR-IMC [15], and RED-SVM [40], [41]), the traditional multiple-instance classification methods (i.e., mi-SVM [31] and MI-SVM [31]), as well as the multiple-instance regression method (i.e., MI-reg [36]).

The first is SVOR-EXC [15], which introduces the label ordering information into the learning problem via explicit constraints on label orders. The second is SVOR-IMC [15], which implicitly incorporates the label ordering information by considering all the training instances to determine each hyperplane. The third is RED-SVM [40], [41], which is similar to SVOR-IMC in spirit. The main difference is that RED-SVM minimizes the threshold term θ_i^2 in the objective function. By considering $\langle \mathbf{w}, \theta_i \rangle$ as a new vector, RED-SVM can be resolved by the standard SVM solvers.

These three baselines are utilized to compare the performance of the multi-instance OR method (i.e., MIOR) with the single-instance OR methods (i.e., SVOR-EXC, SVOR-IMC, and RED-SVM). SVOR-EXC, SVOR-IMC, and RED-SVM are proposed for single-instance OR problems where each instance is explicitly associated with a label. Nevertheless, in multi-instance OR, the label is associated with a bag of instances and the instance label is unknown. Hence, as done in [24], [25], and [37], they learn the classifiers straightforwardly on the single-instance data where a text document or an image is considered as a global feature vector.

The fourth is mi-SVM [31], which initializes all instances in the positive bag as positive and trains an SVM classifier iteratively until each positive bag has at least one instance classified to be positive. The fifth is MI-SVM [31], which selects one instance from each positive bag and learns the classifier iteratively using the selected instances and those in negative bags.

These two baselines are used to compare the performance of the multi-instance OR method (i.e., MIOR) with the traditional multiple-instance classification methods (i.e., mi-SVM and MI-SVM) on dealing with the OR data where there exists an ordering between different labels. As stated in Section I, the OR problem can be considered as a multiclass classification paradigm with ordered labels. mi-SVM and MI-SVM which are proposed for binary class classification can be extended to multiclass classification by adapting the one-against-all decomposition strategy [57]–[59]. However, mi-SVM and MI-SVM are not designed for OR problems. They consider the OR problem as a traditional multiclass classification problem and neglect the ordering information of labels, which may limit their performance on the OR data when only a small amount of labeled data is available.

The sixth is MI-reg [36], which learns a regression function on multiple-instance data. As done in [15], MI-reg is extended to multiple-instance classification problems by treating the class labels as continuous target values.

For SVOR-EXC, SVOR-IMC, RED-SVM, mi-SVM, MI-SVM, MI-reg, and MIOR, the radial basis function (RBF) kernel is used. The kernel parameter is picked up from $10^{[-5:1:5]}$. For the regularization parameters, we let $C_0 = C_1$ and select the value from $2^{[-5:1:5]}$. The threshold ς is set to be 0.1. In MI-reg, the tolerance ϵ is set to be 0.1, as the same setting in [15]. Experiments run on a laptop with 2.8-GHz processor and 3-GB DRAM. SVOR-EXC, SVOR-IMC, and MIOR are implemented based on the sequential minimal optimization (SMO) algorithm in [15]. All algorithms are implemented in the MATLAB environment.

TABLE II
MEAN ZERO-ONE ERRORS ON THE EXPERIMENTAL DATA SETS

Dataset	MIOR	MI-reg (p-value)	mi-SVM (p-value)	MI-SVM (p-value)	SVOR-EXC (p-value)	SVOR-IMC (p-value)	RED-SVM (p-value)
Cameras	**0.574**	0.678 (0.013)	0.659 (0.024)	0.642 (0.035)	0.606 (0.018)	0.612 (0.018)	0.611 (0.015)
Laptops	**0.583**	0.702 (0.008)	0.661 (0.033)	0.665 (0.036)	0.626 (0.015)	0.618 (0.027)	0.616 (0.021)
Tablets	**0.577**	0.689 (0.027)	0.674 (0.018)	0.659 (0.029)	0.614 (0.038)	0.619 (0.025)	0.623 (0.026)
Mobile phone	**0.594**	0.688 (0.018)	0.666 (0.024)	0.661 (0.032)	0.628 (0.023)	0.635 (0.022)	0.633 (0.017)
TVs	**0.585**	0.723 (0.009)	0.690 (0.018)	0.693 (0.015)	0.614 (0.034)	0.611 (0.019)	0.608 (0.023)
Video surveillance	**0.591**	0.689 (0.014)	0.664 (0.022)	0.642 (0.026)	0.622 (0.035)	0.625 (0.015)	0.626 (0.012)
MSRA-MM	**0.304**	0.426 (0.021)	0.392 (0.019)	0.373 (0.035)	0.340 (0.027)	0.345 (0.028)	0.347 (0.026)

TABLE III
MEAN ABSOLUTE ERRORS ON THE EXPERIMENTAL DATA SETS

Dataset	MIOR	MI-reg (p-value)	mi-SVM (p-value)	MI-SVM (p-value)	SVOR-EXC (p-value)	SVOR-IMC (p-value)	RED-SVM (p-value)
Cameras	**0.736**	1.186 (0.008)	1.109 (0.014)	1.032 (0.019)	0.839 (0.031)	0.978 (0.023)	0.976 (0.028)
Laptops	**0.777**	1.359 (0.017)	1.118 (0.025)	1.133 (0.028)	0.934 (0.012)	0.899 (0.036)	0.895 (0.032)
Tablets	**0.841**	1.280 (0.024)	1.247 (0.033)	1.138 (0.038)	1.028 (0.042)	1.057 (0.011)	1.063 (0.017)
Mobile phone	**0.895**	1.532 (0.005)	1.094 (0.019)	1.030 (0.026)	0.998 (0.021)	0.983 (0.032)	0.979 (0.036)
TVs	**0.867**	1.144 (0.014)	1.027 (0.032)	1.091 (0.025)	1.026 (0.039)	1.065 (0.028)	1.060 (0.023)
Video surveillance	**0.925**	1.468 ($<$0.001)	1.221 (0.023)	1.217 (0.018)	1.032 (0.027)	1.111 (0.016)	1.112 (0.012)
MSRA-MM	**0.402**	0.737 (0.022)	0.561 (0.031)	0.548 (0.029)	0.446 (0.019)	0.455 (0.031)	0.458 (0.033)

C. Results on Amazon Sentiment Data Sets

The Amazon sentiment data sets[1] are generated from *Amazon.com* and contain six categories of product reviews: "Cameras," "Laptops," "Mobile phones," "Tablets," "TVs," and "Video surveillance." Each review is associated with one of the five ordinal rating labels $\{1, 2, 3, 4, 5\}$. A higher rating indicates a better review feedback. The task of sentiment classification is to predict the rating for a new review. The numbers of reviews in the Cameras, Laptops, Mobile phones, Tablets, TVs, and Video surveillance data sets are 7673, 2473, 4471, 1049, 2365, and 2790, respectively. The review documents are preprocessed as follows: 1) convert uppercase text to lowercase; 2) replace common nonword pattern with a unique identifier (e.g., mapping smileys ":-)" to "⟨happy⟩"); 3) remove HTML tags and any character that is neither alphanumeric nor a punctuation; and 4) normalize contractions (e.g., transforming "don't" to "do not"). Then, "function words" on the SMART stoplist [60] are removed from the vocabulary, and the remaining words are stemmed. Following the same operations in [33], each review is considered as a bag, and each sentence in the review is treated as an instance. Each instance is represented as a TF-IDF feature vector, and each feature is normalized. Finally, Principal component analysis (PCA) is performed to reduce the data to 200 features.

[1]Available at http://sifaka.cs.uiuc.edu/∼wang296/Data/index.html.

To investigate the classification performance of MIOR and the baselines when only a small amount of labeled data is available to train the classifier, similar to [10] and [11], we randomly select 100 bags to form the training set, and the other bags are used as the testing set. These operations are repeated for 20 times, and the results are summarized. Table II presents the mean zero-one errors and p-values on the Amazon sentiment data sets. The best performance is in bold. The p-values are computed by performing the paired t-test comparing all other classifiers to MIOR under the null hypothesis that there is no difference between the testing zero-one error distributions. When the p-value is smaller than the confidence level 0.05, there is a significant difference between MIOR and the method compared. Likewise, Table III shows the mean absolute errors and p-values.

The mean zero-one errors and mean absolute errors are reported in Tables II and III, respectively. We can observe from Table II that MI-reg delivers higher mean zero-one errors than the other methods. This may be due to the fact that MI-reg is designed for regression problems where a hyperplane is learned to fit the data, rather than classification problems. mi-SVM and MI-SVM show improved classification performance than MI-reg. SVOR-IMC and RED-SVM exhibit similar results on the experimental data sets, which is in line with that obtained in [40] and [41]. Notably, MIOR illustrates the best classification performance across

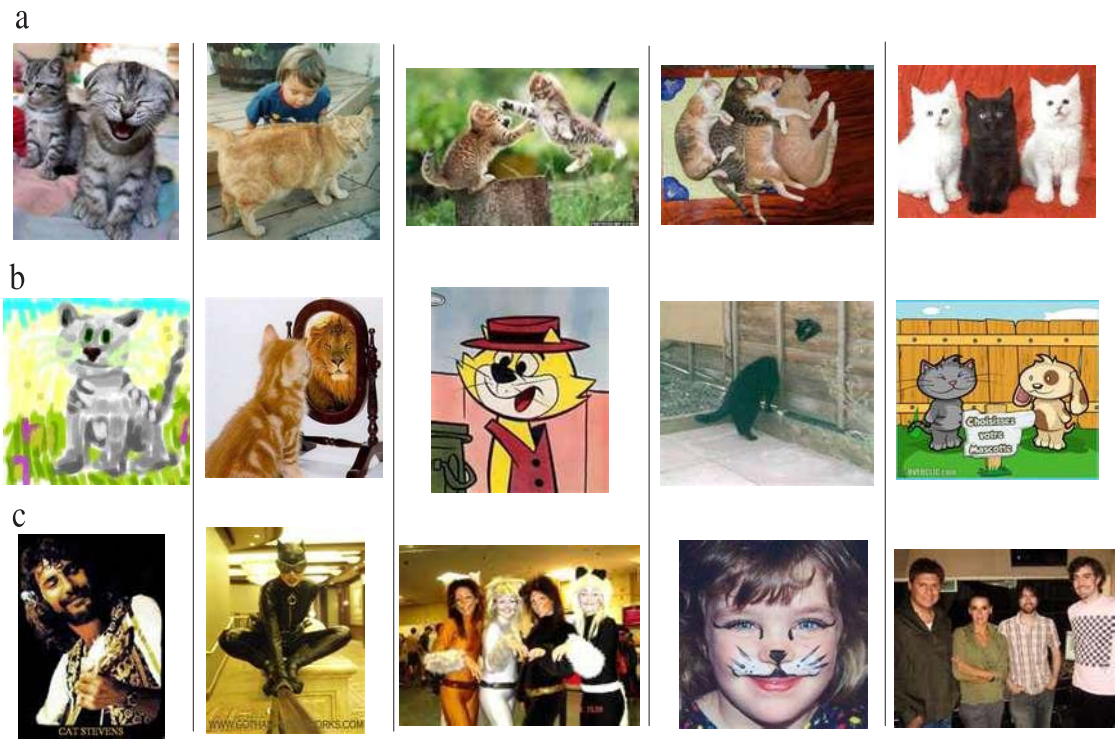

Fig. 6. Samples of the ground truth for the query "Cat" from the MSRA-MM data set. (a) Relevant. (b) Partially relevant. (c) Irrelevant.

all the Amazon sentiment data sets. For example, on the Cameras data set, MIOR reports a minimum of 0.032 and up to 0.104 improvements, relative to the baselines in the mean zero-one errors.

It is noted that MIOR obtains significantly better classification performance than the traditional multiple-instance learning methods (i.e., mi-SVM and MI-SVM). Taking the Cameras data set in Table II as an example, the mean zero-one error of MIOR is 0.574, which is lower than mi-SVM (0.659) and MI-SVM (0.642) at 0.085 and 0.068, respectively. Meanwhile, it can be seen in Table III that the mean absolute error of MIOR is significantly lower than mi-SVM and MI-SVM, since the p-values are smaller than 0.05. In mi-SVM and MI-SVM, the label ordering information is not considered, and the hyperplanes are usually disordered and intersected. Distinguished from mi-SVM and MI-SVM, MIOR incorporates the label ordering information into the training phase and learns a set of parallel hyperplanes. The better performance of MIOR over the traditional multiple-instance classification methods (i.e., mi-SVM and MI-SVM) indicates that for the data with ordered labels, when only a small number of training bags are available, learning a set of ordered parallel hyperplanes can enable better classification performance than constructing a group of disordered intersected hyperplanes.

D. Results on MSRA-MM Data Set

The MSRA-MM data set[2] is collected by Microsoft Research. It is an image retrieval data set with 68 queries and 19 436 images. The different queries include "Cat," "Angel," "Dogs," "Earth," "Snakes," and so on. For each image, its relevance to the corresponding query is labeled with three levels: "relevant," "partially relevant," and "irrelevant." Some samples of the ground truth are shown in Fig. 6.

We follow the same routine in [31] and [61]–[65] to create bags from the MSRA-MM image data set. Specifically, each image is segmented into several regions (blobs) by employing the Blobworld system [66], which can automatically segment the images and require no parameter tuning of regions. Then, the segmented results are converted to multiple-instance data by conducting the codes[3] provided by Tsochantaridis *et al.* [31]. In the obtained multiple-instance data set, each bag contains a number of instances, and each instance has 230 dimensions, representing the color, texture, and shape information. More detailed descriptions of the data features can refer to the documentation within Stuart Andrews's codes.

For each query, 60 bags are selected as the training set, and the other bags are as the testing set. These operations are repeated for 20 times. The results of all the queries are averaged and reported. Tables II and III show the mean zero-one errors and mean absolute errors on the MSRA-MM data set. After investigating the details in Tables II and III, we can find that MIOR obtains significantly lower mean zero-one errors and mean absolute errors than the single-instance OR methods (i.e., SVOR-EXC, SVOR-IMC, and RED-SVM). On the one hand, as presented in Table II, the mean zero-one errors of MIOR, SVOR-EXC, SVOR-IMC, and RED-SVM are 0.304, 0.340, 0.345, and 0.347, respectively. MIOR reports a minimum of 0.036 and up to 0.043 improvements, relative to the single-instance OR methods (i.e., SVOR-EXC, SVOR-IMC,

[2]Available at http://research.microsoft.com/en-us/projects/msrammdata/.

[3]The codes are available at http://www.cs.columbia.edu/~andrews/mil/data/MIL-Data-2002-Musk-Corel-Trec9.tgz.

and RED-SVM). On the other hand, as shown in Table III, the mean absolute errors of MIOR, SVOR-EXC, SVOR-IMC, and RED-SVM are 0.402, 0.446, 0.455, and 0.458, respectively. MIOR displays superior performance over the single-instance OR methods (i.e., SVOR-EXC, SVOR-IMC, and RED-SVM), with at least 0.044 and up to 0.056 improvements in the mean absolute errors. The improvement of MIOR over the single-instance OR methods (i.e., SVOR-EXC, SVOR-IMC, and RED-SVM) implies that considering the image as multiple-instance data can lead to a better classification result than treating it as single-instance data, which is consistent with that observed in [42], [43], [47], and [67]. In the single-instance OR methods (i.e., SVOR-EXC, SVOR-IMC, and RED-SVM), the whole image is considered as a single instance. Nevertheless, the user is usually interested in a portion of the image. If the whole image is transformed into a single instance, the useful information that the user is of interest may be overridden by noisy information from irrelevant regions. Different from the single-instance OR methods (i.e., SVOR-EXC, SVOR-IMC, and RED-SVM), MIOR is designed for multi-instance OR problems, where an image is segmented into a number of regions, and it is classified as of the user interest if it contains at least one targeted region. By considering image retrieval as a multi-instance OR problem, MIOR is able to obtain better classification performance than the single-instance OR methods (i.e., SVOR-EXC, SVOR-IMC, and RED-SVM).

E. Results on IMDB Movie Review Data Set

The IMDB movie review data set[4] [68] contains 50 000 movie reviews downloaded from the IMDB movie Web site *imdb.com*. Each review document is associated with one of the ten star ratings $\{1, 2, \ldots, 10\}$. Similar to the Amazon sentiment data sets, each review is considered as a bag, and each sentence in the review is treated as an instance. The text data are preprocessed by removing all stop words, normalizing each feature and performing stemming. PCA is used and each instance is represented as a vector with 200 features.

To investigate the performance variation of MIOR with different numbers of training bags, we follow the same routine in [10] and [11] to conduct the experiments. Specially, we randomly select N_{train} bags for training, and make sure that each class has at least one bag. Except for the selected N_{train} bags, the other bags in the data set are used for testing. The above-mentioned operations are repeated for 20 times, and the results are averaged. Similar to [10] and [11], we let N_{train} vary from 30 to 300. Figs. 7 and 8 present the mean zero-one errors when the number of training bags (N_{train}) varies from 30 to 300. Figs. 9 and 10 show the mean absolute errors.

Figs. 7 and 9 present the results of the OR methods (i.e., MIOR, SVOR-EXC, SVOR-IMC, and RED-SVM). Among all the OR methods, the multi-instance OR method (i.e., MIOR) obtains lower classification errors than the single-instance OR methods (i.e., SVOR-EXC, SVOR-IMC, and RED-SVM). The IMDB movie review data set is about the categorization of text review data.

[4]Available at http://ai.stanford.edu/∼amaas/data/sentiment/.

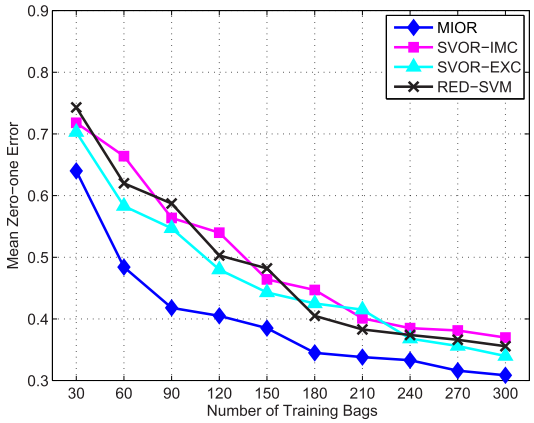

Fig. 7. Mean zero-one errors on the IMDB data set.

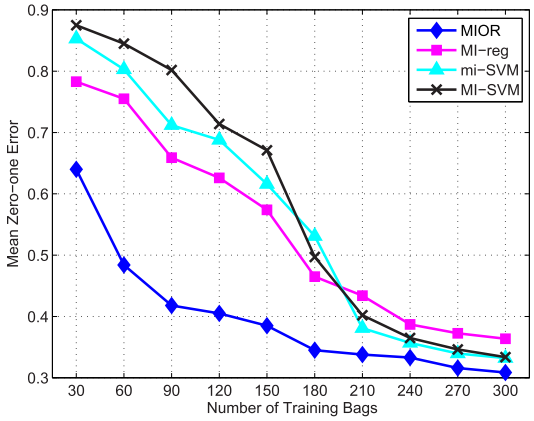

Fig. 8. Mean zero-one errors on the IMDB data set (cont.).

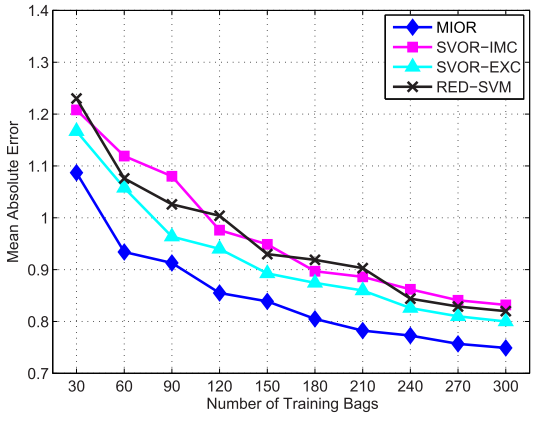

Fig. 9. Mean absolute errors on the IMDB data set.

The superior performance of the multi-instance OR method (i.e., MIOR) over the single-instance OR methods (i.e., SVOR-EXC, SVOR-IMC, and RED-SVM) indicates the effectiveness of transforming the text review data into multiple-instance data. Rather than treating the text data as single instances, considering it as multiple instances can obtain improved classification performance, which is in line with the observations in [23], [25], [31], and [33].

Figs. 8 and 10 show the results of the multi-instance OR method (i.e., MIOR) and the traditional multiple-instance

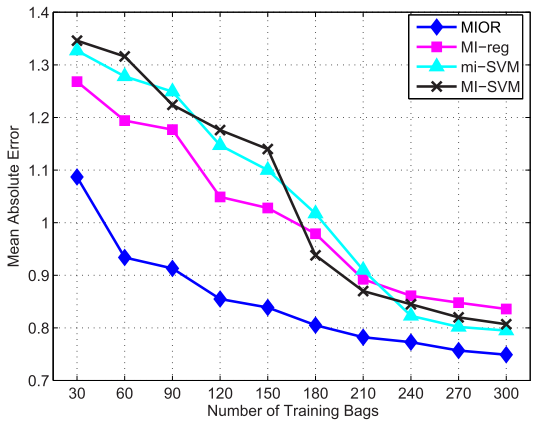

Fig. 10. Mean absolute errors on the IMDB data set (cont.).

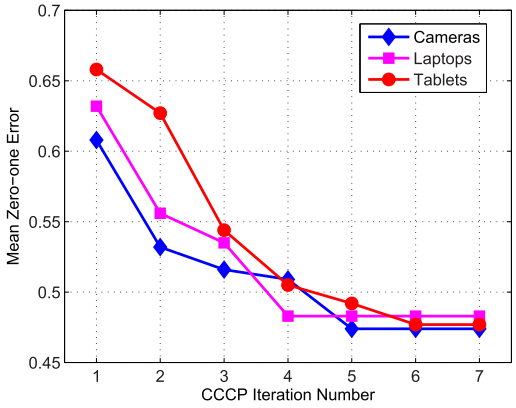

Fig. 11. Mean zero-one errors with different CCCP iteration numbers on the Amazon Sentiment data set.

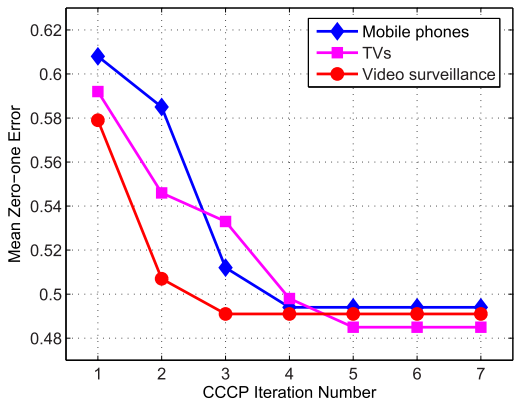

Fig. 12. Mean zero-one errors with different CCCP iteration numbers on the Amazon Sentiment data set (cont.).

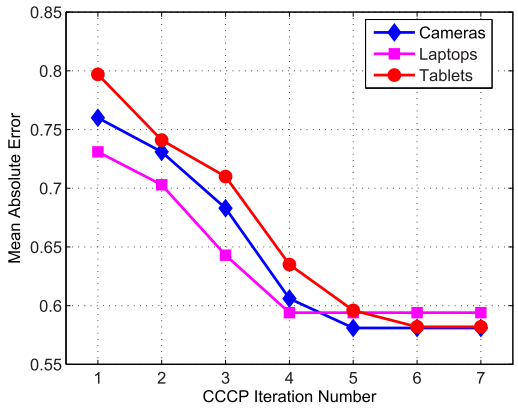

Fig. 13. Mean absolute errors with different CCCP iteration numbers on the Amazon Sentiment data set.

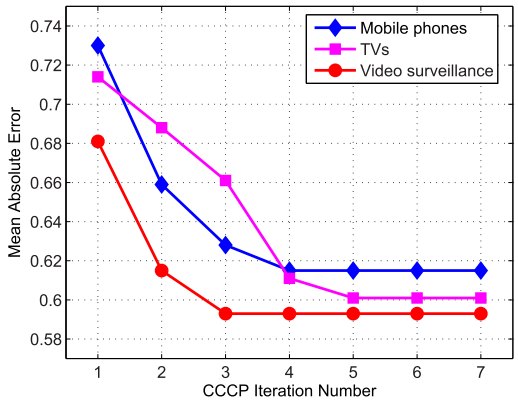

Fig. 14. Mean absolute errors with different CCCP iteration numbers on the Amazon Sentiment data set (cont.).

classification methods (i.e., mi-SVM and MI-SVM). It can be seen that the multi-instance OR method (i.e., MIOR) can obtain markedly lower classification errors than the traditional multiple-instance classification methods (i.e., mi-SVM and MI-SVM), when only a small number of training bags are available to learn the classifier. For example, in Fig. 8, when the number of training bags is 60, the mean zero-one error of the multi-instance OR method (i.e., MIOR) is explicitly lower than the traditional multiple-instance classification methods (i.e., mi-SVM and MI-SVM).

Furthermore, Fig. 7 shows the mean zero-one errors of the OR methods (i.e., MIOR, SVOR-EXC, SVOR-IMC, and RED-SVM). Fig. 8 presents the mean zero-one errors of the traditional multiple-instance classification methods (i.e., mi-SVM and MI-SVM). By comparing Figs. 7 and 8, it can be observed that when the number of training bags is relatively small (e.g., 60), the mean zero-one errors of the OR methods in Fig. 7 are generally lower than the traditional multiple-instance classification methods (i.e., mi-SVM and MI-SVM) in Fig. 8. This may be because mi-SVM and MI-SVM consider the OR problem as a traditional multiclass classification problem and utilize a group of intersecting hyperplanes to classify the data. Distinctively, the OR methods (i.e., MIOR, SVOR-EXC, SVOR-IMC, and RED-SVM) take the label ordering information into account and learn a set of parallel hyperplanes to separate the data. For the data where the classes are ranked in orders, it is more desirable to classify it using a set of ordered parallel hyperplanes, rather than a group of disordered intersecting hyperplanes, as the mi-SVM and MI-SVM do. This is especially the case when the number of training bags available is relatively small. The similar findingscan also be observed in [10] and [11].

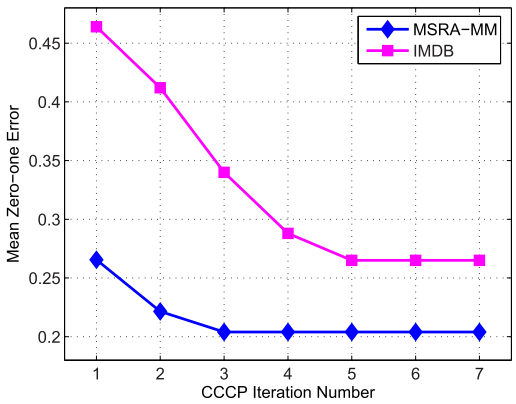

Fig. 15. Mean zero-one errors with different CCCP iteration numbers on the MSRA-MM and IMDB data sets.

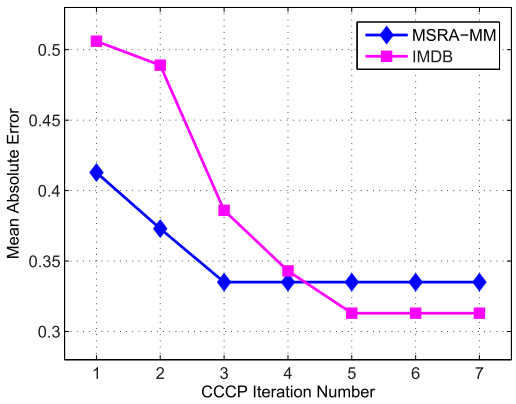

Fig. 16. Mean absolute errors with different CCCP iteration numbers on the MSRA-MM and IMDB data sets.

F. Performance With Different Numbers of CCCP Iterations

The performance of MIOR with varying numbers of CCCP iterations is investigated. To do this, we record the corresponding mean zero-one error and mean absolute error when the number of CCCP iterations goes up from 1 to 7, as shown from Figs. 11–16. It is seen that the mean zero-one error and mean absolute error decrease with the increasing number of CCCP iterations. This is because when more CCCP iterations are carried out, the solution of the optimization function is closer to an optimal one, and thus, the mean zero-one error and mean absolute error go down. Furthermore, the mean zero-one error and mean absolute error remain relatively stable after a few number of CCCP iterations, which indicates the convergence of our proposed MIOR method. By employing the termination criterion in Table I, MIOR stops after 4.02 iterations on the Amazon sentiment data sets, 2.69 iterations on the MSRA-MM data set, and 4.34 iterations on the IMDB movie review data set.

V. Conclusion and Future Work

A. Contributions of This Paper

In this paper, we address the multi-instance OR problem and propose a novel MIOR method, which can be used to solve applications, such as sentiment classification and image retrieval. In contrast to the existing single-instance OR methods, MIOR is able to learn the OR classifier on multiple-instance data where only the bag label is available and the specific instance label is not given. Since the derived MIOR problem is nonconvex, we relax the original problem by using CCCP. Moreover, experimental evaluations on real-world text and image data sets show that MIOR outperforms the existing single-instance OR, multiple-instance classification, and multiple-instance regression methods in terms of the mean zero-one error and mean absolute error.

B. Limitations and Future Work

The learning problem (4) of MIOR is nonconvex, and the CCCP technique is used to decompose it into several QP problems. Assuming that the computational cost of solving a QP problem is $t(n)$, the time complexity of MIOR is roughly $O(T*t(n))$, where T is the number of CCCP iterations. On the one hand, the number of CCCP iterations (T) is a relatively small value compared with the sample set size. As shown in Section IV-F, the average numbers of CCCP iterations (T) on the Amazon sentiment, MSRA-MM, and IMDB movie review data sets are 4.02, 2.69, and 4.34, respectively. It can be seen that the number of CCCP iterations (T) is a relatively small value. The similar observations can also be found in [23], [25], [31], and [33]. On the other hand, $t(n)$ is the computational cost of solving a QP problem. In our experiments, we employ the SMO algorithm in [15] to solve the QP problem, and the computational cost of solving a QP problem via SMO is $t(n) = O(n^2)$. Hence, the time complexity of our method is $O(Tn^2)$, where n is the number of training bags.

First, MIOR is much more efficient than mi-SVM and MI-SVM. Similar to MIOR, mi-SVM and MI-SVM obtain the classifier by resolving a series of QP problems. However, the QP problem size of MIOR is smaller than mi-SVM and MI-SVM. For a training set which consists of K classes, with each class having (n/K) bags and each bag having m instances, MIOR solves a QP problem of size n in each iteration, while mi-SVM and MI-SVM need to resolve QP problems of sizes nm and $(1/K)n + ((K-1)/K)nm$ [31], respectively. As a result, MIOR delivers lower computational cost than mi-SVM and MI-SVM. Let us take the Tablets data set as an example. The running time of MIOR is 3.95 s, which is lower than mi-SVM (192.43 s) and MI-SVM (165.72 s).

Second, the training cost of MIOR is generally lower than SVOR-IMC and RED-SVM, especially when the class number K is relatively large. Given a training set with K classes, the time complexity of SVOR-IMC and RED-SVM is $O((K-1)^2\, n^2)$ [15], [40], and that of MIOR is $O(Tn^2)$. On the one hand, in MIOR, T is a relatively small number. As shown in Section IV-F, T is averagely 4.02 on the Amazon sentiment data sets, 2.69 on the MSRA-MM data set, and 4.34 on the IMDB movie review data set. On the other hand, in SVOR-IMC and RED-SVM, the training efficiency is closely related to the class number K. When the class number K is large, the running time of SVOR-IMC and RED-SVM will become high. Considering an example in the Tablets data

set which contains five classes, the running time of MIOR is 3.95 s, which is markedly lower than SVOR-IMC (16.95 s) and RED-SVM (12.68 s).

Finally, the time complexity of SVOR-EXC is about $O(4n^2)$, and the training efficiency of MIOR is comparable to SVOR-EXC. For example, on the Tablets data set, the CCCP number T is around 3.86, and the running time of MIOR is 3.95 s, which is slightly lower than SVOR-EXC (4.17 s). Although the training efficiency of MIOR is comparable to SVOR-EXC, MIOR delivers significantly lower classification errors than SVOR-EXC. In the Tablets data set, the mean zero-one error and mean absolute error of MIOR are 0.577 and 0.841, respectively, which are significantly lower than SVOR-EXC (0.614 and 1.028).

The time complexity of MIOR is $O(T * t(n))$. We employ the SMO algorithm in [15] to solve the QP problem, and it has $t(n) = O(n^2)$. Hence, the time complexity of MIOR is $O(Tn^2)$. Besides SMO, there exist a number of optimization methods [38], [69]–[71], which can solve the QP problem efficiently. For example, we can extend Pegasos [70] or NORMA [71] to solve the QP problem. By applying these methods, the running time of solving a QP problem (i.e., $t(n)$) can increase linearly with the number of nonzero features in each instance and does not depend directly on the data set size. These optimization techniques can be utilized to speed up the resolving of QP problems (i.e., $t(n)$), and thus the time complexity of MIOR (i.e., $O(T * t(n))$) can be largely reduced. The application of these efficient QP optimization methods will be a valuable consideration in our future work.

REFERENCES

[1] B.-Y. Sun, J. Li, D. D. Wu, X.-M. Zhang, and W.-B. Li, "Kernel discriminant learning for ordinal regression," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 6, pp. 906–910, Jun. 2010.

[2] S. Yu, K. Yu, V. Tresp, and H.-P. Kriegel, "Collaborative ordinal regression," in *Proc. 23rd Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, Jun. 2006, pp. 1089–1096.

[3] S. Fouad and P. Tino, "Adaptive metric learning vector quantization for ordinal classification," *Neural Comput.*, vol. 24, no. 11, pp. 2825–2851, Nov. 2012.

[4] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, and S. Li, "Incremental support vector learning for ordinal regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1403–1416, Jul. 2014.

[5] J. S. Cardoso and J. F. P. da Costa, "Learning to classify ordinal data: The data replication method," *J. Mach. Learn. Res.*, vol. 8, no. 12, pp. 1393–1429, 2007.

[6] J. S. Cardoso, J. F. P. da Costa, and M. J. Cardoso, "Modelling ordinal relations with SVMs: An application to objective aesthetic evaluation of breast cancer conservative treatment," *Neural Netw.*, vol. 18, nos. 5–6, pp. 808–817, 2005.

[7] O. Pavlovic, V. Rudovic, and M. Pantic, "Context-sensitive dynamic ordinal regression for intensity estimation of facial action units," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 944–958, Sep. 2015.

[8] S. Baccianella, A. Esuli, and F. Sebastiani, "Feature selection for ordinal text classification," *Neural Comput.*, vol. 26, no. 3, pp. 557–591, Mar. 2014.

[9] C. Li, Q. Liu, J. Liu, and H. Lu, "Ordinal distance metric learning for image ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1551–1559, Jul. 2015.

[10] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," in *Advances in Large Margin Classifiers*, A. J. Smola and P. J. Bartlett, Eds. Cambridge, MA, USA: MIT Press, 2000.

[11] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," in *Proc. 9th Int. Conf. Artif. Neural Netw.*, Edinburgh, U.K., Sep. 1999, pp. 97–102.

[12] S. Corrente, S. Greco, M. Kadziński, and R. Slowiński, "Robust ordinal regression in preference learning and ranking," *Mach. Learn.*, vol. 93, nos. 2–3, pp. 381–422, 2013.

[13] W. Kotlowski and R. Slowiński, "On nonparametric ordinal classification with monotonicity constraints," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 11, pp. 2576–2589, Oct. 2013.

[14] Y. Liu, Y. Liu, and K. C. C. Chan, "Ordinal regression via manifold learning," in *Proc. 25th AAAI Conf. Artif. Intell. Learn.*, San Francisco, CA, USA, Aug. 2011, pp. 398–403.

[15] W. Chu and S. S. Keerthi, "Support vector ordinal regression," *Neural Comput.*, vol. 19, no. 3, pp. 792–815, 2007.

[16] F. Fernández-Navarro, A. Riccardi, and S. Carloni, "Ordinal neural networks without iterative tuning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 11, pp. 2075–2085, Feb. 2014.

[17] W. Chu and Z. Ghahramani, "Gaussian processes for ordinal regression," *J. Mach. Learn. Res.*, vol. 6, pp. 1019–1041, Dec. 2005.

[18] F. Fernández-Navarro, P. A. Gutiérrez, C. Hervás-Martínez, and X. Yao, "Negative correlation ensemble learning for ordinal regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 11, pp. 1836–1849, Jun. 2013.

[19] A. Riccardi, F. Fernández-Navarro, and S. Carloni, "Cost-sensitive adaboost algorithm for ordinal regression based on extreme learning machine," *IEEE Trans. Cybern.*, vol. 44, no. 10, pp. 1898–1909, Jan. 2014.

[20] M. Pérez-Ortiz, P. A. Gutiérrez, and C. Hervás-Martínez, "Projection-based ensemble learning for ordinal regression," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 681–694, Jun. 2014.

[21] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez, "Ordinal regression methods: Survey and experimental study," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 127–146, Jan. 2016.

[22] R. Rahmani, S. A. Goldman, H. Zhang, S. R. Cholleti, and J. E. Fritts, "Localized content-based image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1902–1912, Nov. 2008.

[23] L. Si, J. Zhang, D. Zhang, Y. Liu, and R. D. Lawrence, "Multiple instance learning on structured data," in *Proc. 25th Conf. Neural Inf. Process. Syst.*, Granada, Spain, Dec. 2011, pp. 145–153.

[24] Z.-H. Zhou and M.-L. Zhang, "Multi-instance multi-label learning with application to scene classification," in *Proc. 20th Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2006, pp. 1609–1616.

[25] D. Zhang, F. Wang, L. Si, and T. Li, "Maximum margin multiple instance clustering with applications to image and text clustering," *IEEE Trans. Neural Netw.*, vol. 22, no. 5, pp. 739–751, May 2011.

[26] A. J. Smola, S. V. N. Vishwanathan, and T. Hofmann, "Kernel methods for missing variables," in *Proc. 10th Int. Workshop Artif. Intell. Stat.*, 2005, pp. 325–332.

[27] B. K. Sriperumbudur and G. R. G. Lanckriet, "A proof of convergence of the concave-convex procedure using Zangwill's theory," *Neural Comput.*, vol. 24, no. 6, pp. 1391–1407, 2012.

[28] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Comput.*, vol. 15, no. 4, pp. 915–936, 2003.

[29] B. K. Sriperumbudur and G. R. G. Lanckriet, "On the convergence of the concave-convex procedure," in *Proc. 22nd NIPS Workshop Optim. Mach. Learn.*, Vancouver, BC, Canada, Dec. 2009, pp. 1759–1767.

[30] A. Shashua and A. Levin, "Ranking with large margin principle: Two approaches," in *Proc. 15th Conf. Neural Inf. Process. Syst.*, Cambridge, U.K., Dec. 2002, pp. 937–944.

[31] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. 25th Annu. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2002, pp. 561–568.

[32] X. Xu and B. Li, "Evaluating multi-class multiple-instance learning for image categorization," in *Proc. 8th Asian Conf. Comput. Vis.*, Tokyo, Japan, Nov. 2007, pp. 155–165.

[33] J. C. Platt, Y. Zhang, A. C. Surendran, and M. Narasimhan, "Learning from multi-topic Web documents for contextual advertisement," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Las Vegas, NV, USA, Aug. 2008, pp. 1051–1059.

[34] Y.-P. Hung, Y.-T. Chen, C.-S. Chen, and K.-Y. Chang, "Multi-class multi-instance boosting for part-based human detection," in *Proc. 12th IEEE Int. Conf. Comput. Vis. Workshops*, Kyoto, Japan, Sep. 2009, pp. 1177–1184.

[35] J. C. Platt, P. Viola, and C. Zhang, "Multiple instance boosting for object detection," in *Proc. 19th Conf. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, Dec. 2005, pp. 1417–1424.

[36] S. Ray and D. Page, "Multiple instance regression," in *Proc. 18th Int. Conf. Mach. Learn.*, Williamstown, MA, USA, Jun. 2001, pp. 425–432.

[37] T. Lane, K. L. Wagstaff, and A. Roper, "Multiple-instance regression with structured data," in *Proc. 8th IEEE Int. Conf. Data Mining Workshops*, Pisa, Italy, Dec. 2008, pp. 291–300.

[38] B. Zhao, F. Wang, and C. Zhang, "Block-quantized support vector ordinal regression," *IEEE Trans. Neural Netw.*, vol. 20, no. 5, pp. 882–890, May 2009.

[39] F. Fernández-Navarro, P. Campoy-Muñoz, M. de la Paz-Marín, C. Hervás-Martínez, and X. Yao, "Addressing the eu sovereign ratings using an ordinal regression approach," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 2228–2240, Jun. 2013.

[40] L. Li and H.-T. Lin, "Ordinal regression by extended binary classification," in *Proc. 20th Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2006, pp. 865–872.

[41] H.-T. Lin and L. Li, "Reduction from cost-sensitive ordinal ranking to weighted binary classification," *Neural Comput.*, vol. 24, no. 5, pp. 1329–1367, 2012.

[42] Y. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," *J. Mach. Learn. Res.*, vol. 5, pp. 913–939, Aug. 2004.

[43] Y. Chen, J. Bi, and J. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, Dec. 2006.

[44] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proc. 11th Conf. Neural Inf. Process. Syst.*, Denver, CO, USA, Dec. 1997, pp. 570–576.

[45] Z. Zhou and J. Xu, "On the relation between multi-instance learning and semi-supervised learning," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvalis, OR, USA, Jun. 2007, pp. 1167–1174.

[46] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, nos. 1–2, pp. 31–71, 1997.

[47] Q. Zhang and S. Goldman, "EM-DD: An improved multiple-instance learning technique," in *Proc. 15th Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2001, pp. 1073–1080.

[48] O. Maron and A. L. Ratan, "Multiple-instance learning for natural scene classification," in *Proc. 15th Int. Conf. Mach. Learn.*, Madison, WI, USA, Jun. 1998, pp. 341–349.

[49] P. Yan, Q. Wang, Y. Yuan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 660–672, Mar. 2013.

[50] H. Zeng and Y.-M. Cheung, "Semi-supervised maximum margin clustering with pairwise constraints," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 926–939, May 2012.

[51] B. Zhao, F. Wang, and C. Zhang, "Linear time maximum margin clustering," *IEEE Trans. Neural Netw.*, vol. 21, no. 2, pp. 319–332, Feb. 2010.

[52] G. Fung and O. L. Mangasarian, "Semi-supervised support vector machines for unlabeled data classiffication," *Optim. Methods Softw.*, vol. 15, no. 1, pp. 29–44, 2001.

[53] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-I.I.D. Samples," in *Proc. 26th Int. Conf. Mach. Learn.*, Montreal, CA, USA, Jun. 2009, pp. 1249–1256.

[54] Y. Hu, J. Wang, N. Yu, and X.-S. Hua, "Maximum margin clustering with pairwise constraints," in *Proc. 8th Int. Conf. Data Mining*, Pisa, Italy, Dec. 2008, pp. 253–262.

[55] B. Zhao, F. Wang, and C. Zhang, "Efficient multiclass maximum margin clustering," in *Proc. 25th Int. Conf. Mach. Learn.*, Helsinki, Finland, Jun. 2008, pp. 1248–1255.

[56] C.-N. J. Yu and T. Joachims, "Learning structural SVMs with latent variables," in *Proc. 26th Int. Conf. Mach. Learn.*, Montreal, CA, USA, Jun. 2009, pp. 1169–1176.

[57] C.-W. Hse and C.-J. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[58] B. Fei and J. Liu, "Binary tree of SVM: A new fast multiclass training and classification algorithm," *IEEE Trans. Neural Netw.*, vol. 17, no. 3, pp. 696–707, May 2006.

[59] M. Gğnen, A. G. Tanuğur, and E. Alpaydin, "Multiclass posterior probability support vector machines," *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 130–139, Jan. 2008.

[60] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Boston, MA, USA: Addison-Wesley, 1989.

[61] A. Zafra and S. Ventura, "G3P-MI: A genetic programming algorithm for multiple instance learning," *Inf. Sci.*, vol. 180, no. 23, pp. 4496–4513, 2010.

[62] R. C. Bunescu and R. J. Mooney, "Multiple instance learning for sparse positive bags," in *Proc. 24th IEEE Int. Conf. Mach. Learn.*, Corvallis, OR, USA, Jun. 2007, pp. 105–112.

[63] J. Li, D. Wang, and B. Zhang, "Multiple-instance learning via random walk," in *Proc. 17th Eur. Conf. Mach. Learn.*, Berlin, Germany, Sep. 2006, pp. 473–484.

[64] P. M. Cheung and J. T. Kwok, "A regularization framework for multiple-instance learning," in *Proc. 23rd Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, Jun. 2006, pp. 193–200.

[65] M.-A. Carbonneau, E. Granger, A. J. Raymond, and G. Gagnon, "Robust multiple-instance learning ensembles using random subspace instance selection," *Pattern Recognit.*, vol. 58, pp. 83–99, Oct. 2017.

[66] S. Belongie, J. M. Hellerstein, C. Carson, M. Thomas, and J. Malik, "Blobworld: A system for region-based image indexing and retrieval," in *Proc. 3rd Int. Conf. Vis. Inf. Inf. Syst.*, Amsterdam, The Netherlands, Jun. 1999, pp. 509–516.

[67] W. Li and D. Yeung, "MILD: Multiple-instance learning via disambiguation," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 1, pp. 76–89, Feb. 2010.

[68] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. 49th Annu. Meet. Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Portland, OR, USA, Jun. 2011, pp. 142–150.

[69] R. Babaria, J. S. Nath, S. Krishnan, K. R. Sivaramakrishnan, C. Bhattacharyya, and M. N. Murty, "Focused crawling with scalable ordinal regression solvers," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, USA, Jun. 2007, pp. 57–64.

[70] S. Shalev-Shwartz, Y. Singer, and N. Srebro, "Pegasos: Primal estimated sub-gradient solver for SVM," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, USA, Jun. 2007, pp. 807–814.

[71] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2165–2176, Aug. 2004.

Yanshan Xiao received the Ph.D. degree in computer science from the University of Technology Sydney, Ultimo, NSW, Australia, in 2011.

She is currently a Professor with the School of Computers, Guangdong University of Technology, Guangzhou, China. She has published papers in the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and so on. Her current research interests include multiple-instance learning, support vector machine, and data mining.

Bo Liu is currently a Professor with the School of Automation, Guangdong University of Technology, Guangzhou, China. He has published papers in the IEEE TRANSACTIONS ON NEURAL NETWORKS, the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, and so on. His current research interests include machine learning and data mining.

Zhifeng Hao received the B.Sc. degree in mathematics from Sun Yat-sen University, Guangzhou, China, in 1990, and the Ph.D. degree in mathematics from Nanjing University, Nanjing, China, in 1995.

He is currently a Professor with the School of Mathematics and Big Data, Foshan University, Foshan, China, and the School of Computers, Guangdong University of Technology, Guangzhou. His current research interests include various aspects of algebra, machine learning, data mining, and evolutionary algorithms.