# A Boosting Approach to Exploit Instance Correlations for Multi-Instance Classification

Yali Li, Shengjin Wang, *Member, IEEE*, Qi Tian, *Senior Member, IEEE*, and Xiaoqing Ding, *Fellow, IEEE*

***Abstract*— We propose a Boosting approach for multi-instance (MI) classification. $L_p$-norm is integrated to localize the witness instances and formulate the bag scores from classifier outputs. The contributions are twofold. First, a flexible and concise model for Boosting is proposed by the $L_p$-norm localization and exponential loss optimization. The scores for bag-level classification are directly fused from the instance feature space without probabilistic assumptions. Second, gradient and Newton descent optimizations are applied to derive the weak learners for Boosting. In particular, the instance correlations are exploited by fitting the weights and Newton updates for the weak learner construction. The final Boosted classifiers are the sums of iteratively chosen weak learners. Experiments demonstrate that the proposed $L_p$-norm-localized Boosting approach significantly improves the MI classification performance. Compared with the state of the art, the approach achieves the highest MI classification accuracy on 7/10 benchmark data sets.**

***Index Terms*— Boosting, instance correlations, $L_p$-norm-based localization, multi-instance (MI) classification.**

## I. Introduction

Multi-instance (MI) classification, which extends from standard supervised learning, is an important issue in machine learning. In standard supervised learning, each instance corresponds to an example, and the true label is assigned for each example. But the true labels for individual instances are not available in MI classification. Multiple instances are grouped into an example bag, and the labels are assigned for bags. If a bag is labeled as positive, it indicates that at least one positive instance is included. A bag is labeled as negative indicates that all included instances are negative. MI classification is mainly challenged by instance labeling ambiguity. It has been emerging as a useful tool in a number of application domains, such as drug activity prediction [1], visual recognition [2], [3], and content-based image retrieval [4].

Existing approaches for MI classification can be distinguished by the optimization criteria and procedures. In this brief, we focus on Boosting approaches for MI classification. The previous efforts include MI-LR [5], Noisy-or Boost [6], ISR Boost [6], ordered weighted averaging (OWA)-real miBoost [7], and other Boosting algorithms for specific tasks [2]–[4]. The outputs of Boosted classifiers for the individual instances are first fused into bag scores/probabilities. Loss function is then formulated to minimize the bag-level MI classification error. Weak learners are iteratively constructed and selected to minimize the loss and aggregated together for strong MI classification power. In general, there are two limitations of existing approaches. First, the mappings from the instance feature space to bag predictive scores are mostly probabilistic, with the assumption that the instances in the same bag are independent of each other or have an equal importance. These assumptions are hard to satisfy, which would reduce the flexibility in handling different instance distributions. Second, the instance correlations are not sufficiently exploited even ignored in deriving the weak learners. Since the instances in the same bags are highly correlated, the neglect of instance correlations would limit the MI classification performance.

In this brief, we address the issue of Boosting-based MI classification. A $L_p$-norm-localized MI Boosting approach is proposed. Our main contributions are as follows.

1) Our approach is the first Boosting approach that employs $L_p$-norm to localize the witness instances and fuse the instance scores for bag-level classification. We propose a flexible and concise mathematical optimization criterion by integrating the $L_p$-norm and exponential loss.
2) We propose Boosting algorithms aggregating iterative weak learners for strong MI classification power. Gradient descent optimization is used to derive the binary weak learners from instance weights. Newton descent optimization is used to derive the real-value weak learners by fitting Newton updates. In particular, we show that the instance correlations are directly represented and sufficiently exploited in Boosting iterations, which further provide a soft way to handle labeling ambiguity.
3) Empirical evaluation demonstrates the effectiveness of the proposed approach. Compared with the state of the art, our approach achieves the highest MI classification accuracy on 7/10 benchmark data sets and increases the accuracy by 4%. Compared with existing Boosting approaches, the average accuracy on ten data sets is increased by 5%.

The remainder of the brief is organized as follows. Section II presents the prior related work on MI classification. Section III describes the proposed Boosting approach for MI classification. The experimental results are given in Section IV. The conclusion is drawn in Section V.

## II. Related Work

MI classification is capable of handling the problems where the labels are associated with sets of samples instead of individual samples. Several instances are composed into a bag, and a label is associated with the bag. MI classification is to predict the labels of unknown bags. An example for an MI classification illustration is to determine whether a certain key chain is useful (positive) to enter a certain room [8]. If at least one key in the chain can enter the room, the key chain is positive; otherwise, it is negative. MI classification has displayed effectiveness in solving the problems where it is difficult to provide clear and unambiguous annotations. Due to the importance in various applications, MI classification has gained much attention.

Generative MI classification approaches choose witness instances in positive bags and model the region of interest (ROI) for the

Manuscript received January 12, 2015; revised June 1, 2015 and September 12, 2015; accepted September 17, 2015. This work was supported in part by the Initiative Science Research Program of Tsinghua University under Grant 20141081253 and in part by the National Science and Technology Support Program under Grant 2013BAK02B04.

Y. Li, S. Wang, and X. Ding are with the State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: wgsgj@tsinghua.edu.cn; liyali@ocrserv.ee.tsinghua.edu.cn).

Q. Tian is with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249-1604 USA (e-mail: qitian@cs.utsa.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org (File size: 1MB).

Digital Object Identifier 10.1109/TNNLS.2015.2497318

chosen positive instances. An axis-parallel rectangle [8] uses an axis-parallel hyper-rectangle to represent the ROI that covers at least one instance from each positive bag and no negative instances. Diverse density [9], [10] finds the concept points that are close to at least one instance from each positive bag while far from all instances in negative bags. Citation $k$NN [11], [12] uses both reference bags and citer bags to predict the labels of test bags. Discriminative MI classification approaches extend standard supervised learning to optimize MI classification loss. MI-SVM (support vector machine) [13] is based on maximizing the margin of the most positive instances and the least negative instances. There are several algorithms combining the instance selection for MI classification, such as MILES [14], SMILE [15], MILD [16], and MILIS [17]. Sparse-kernel classifiers [18] are learned for MI classification. Ensemble learning is employed to aggregate MI learners with Bagging/Boosting strategy [19]–[21]. Instance selection and instance labeling are two typical ways to handle labeling ambiguity in existing approaches. The former discards ambiguous instances, but neglecting the ambiguous instances might cause information loss. The latter labels instances in positive bags to convert MI classification into standard supervised learning issues. However, the prelabeling might add noise and reduce the classification accuracy.

There are Boosting approaches for MI classification. In MI-OptBall [22], hyper-balls/hyper-rectangles-based weak learners are combined with Adaboost. MI-LR [5] models the probability of a bag being positive as the average of the probabilities that included instances are positive. Noisy-or Boost [6] is based on the NOR probabilistic modeling and logistic loss optimization. ISR Boost [6] introduces a less probabilistic model to connect instance scores with bag probabilities. Both Noisy-or and ISR Boost are based on gradient descent optimization aggregating binary weak classifiers. OWA-real miBoost [7] uses OWA and NOR model to map from instance scores to bag probabilities, with weak learners the same as in the traditional real-Adaboost [23]. The adaptive resampling in Boosting provides a natural and soft way to associate instances with determinacy, which makes Boosting appropriate for MI classification. However, there are two limitations of existing approaches. First, the mapping from instance feature space to bag predictive scores is not flexible. The averaging and NOR probabilistic assumptions are difficult to satisfy in applications. Second, it lacks of instance correlation exploitation in the weak learners for Boosting.

To handle these issues, we propose a novel Boosting approach for MI classification in this brief. $L_p$-norm is integrated with exponential loss to formulate a flexible and concise model for optimization. Furthermore, gradient and Newton descent optimizations are adopted to derive the weak learners exploiting instance correlations for Boosting. The experimental results on benchmark data sets prove the effectiveness of the proposed $L_p$-norm-localized MI Boosting approach.

## III. $L_p$-Norm-Localized Multi-Instance Boosting

### A. Preliminaries

Let $x_{i,j}$ represent the instance with $i$ as the bag index and $j$ as the instance index. In MI classification, instances are grouped in bags. Hence, instances in the same bag share the same $i$. We further let $X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,N_i}\}$ denote the group of instances in bag $i$, where $N_i$ is the number of instances. Suppose $y_{i,j}$ is the instance score for $x_{i,j}$ (i.e., the classifier output for $x_{i,j}$), the score $y_i$ for classifying bag $X_i$ is formulated by $L_p$-norm as

$$a^{y_i} = \left(\sum_j a^{p y_{i,j}}\right)^{1/p}, \quad \text{i.e., } y_i = \frac{1}{p}\log_a\left(\sum_j a^{p y_{i,j}}\right) \tag{1}$$

where $a > 1$ is the base of exponential transformation. First, we use a monotone exponential transformation to introduce the positive rectified instance score $\tilde{y}_{i,j} = a^{y_{i,j}} > 0$. Then, $L_p$-norm is involved to fuse the rectified instance scores into the rectified bag score $\tilde{y}_i = a^{y_i}$. $L_p$-norm is flexible to localize the witness instances (i.e., instances with high scores). Its output is significantly determined by the maximal element. As $p$ increases, $\tilde{y}_i$ gets closer to $\max_{j\in i}\tilde{y}_{i,j}$, which is consistent with the MI classification assumption. Since $a = e^{\log a}$, we let $\lambda = p\log a$ and simplify $y_i$ as

$$y_i = \frac{1}{\lambda}\log\left(\sum_j e^{\lambda y_{i,j}}\right). \tag{2}$$

Based on this $L_p$-norm-localized instance-bag mapping, we define the exponential loss function as

$$C = \sum_i e^{-t_i y_i} = \sum_i \left(\sum_j e^{\lambda y_{i,j}}\right)^{-\frac{t_i}{\lambda}} \tag{3}$$

where $t_i \in \{-1, 1\}$ is the bag label. Let $N_B$ be the number of bags. Since $e^{-t_i y_i} \geq I(\text{sign}(y_i) \neq t_i)$ ($I(\text{sign}(y_i) \neq t_i) = 1$ only if $\text{sign}(y_i) \neq t_i$; otherwise, it is 0), $(C/N_B) \geq (1/N_B)\sum_i I(\text{sign}(y_i) \neq t_i)$. Therefore, $(C/N_B)$ is the upper bound of the bag classification error. We propose the optimization criterion for Boosting based on the $L_p$-norm instance-bag mapping and exponential loss. This optimization criterion directly maps from instance scores to bag scores without probabilistic modeling. Compared with the existing optimization criteria for MI classification, the formulation is more concise. We remove the probabilistic assumptions such as instances are independent of each other or contribute equally in the same bag. $L_p$-norm provides a flexible way to fuse the classifier outputs for bag-level classification. The combination of $L_p$-norm and exponential loss would also simplify the derivations for weak learners. In order to further develop the Boosting approach for MI classification, we present the following two theorems.

*Theorem 1:* The first derivative of $C$ against $y_{i,j}$ is

$$s_{i,j} = \frac{\partial C}{\partial y_{i,j}} = -t_i e^{-t_i y_i}(e^{y_{i,j}}/e^{y_i})^{\lambda}. \tag{4}$$

*Theorem 2:* The second derivative of $C$ with respect to $y_{i,j}$ consists in a block diagonal matrix $H$ as

$$H = \begin{bmatrix} H_1 & & & O \\ & H_2 & & \\ & & \ddots & \\ O & & & H_{N_B} \end{bmatrix}. \tag{5}$$

The elements in $H_i$ are computed as

$$\begin{aligned} h_{i,j,j} &= \frac{\partial^2 C}{\partial y_{i,j}^2} = -t_i\lambda e^{-t_i y_i}\left(\frac{e^{y_{i,j}}}{e^{y_i}}\right)^{\lambda} + t_i(t_i+\lambda)e^{-t_i y_i}\left(\frac{e^{y_{i,j}}}{e^{y_i}}\right)^{2\lambda} \\ h_{i,j,k} &= \frac{\partial^2 C}{\partial y_{i,j}\partial y_{i,k}} = t_i(t_i+\lambda)e^{-t_i y_i}\left(\frac{e^{y_{i,j}}}{e^{y_i}}\right)^{\lambda}\left(\frac{e^{y_{i,k}}}{e^{y_i}}\right)^{\lambda}, \quad j \neq k. \end{aligned} \tag{6}$$

The proof is simplified here. For $s_{i,j}$, $(e^{y_{i,j}}/e^{y_i})^{\lambda}$ is the proportion of the rectified instance score $\tilde{y}_{i,j}$ against the rectified bag score $\tilde{y}_i$, which represents the importance of the instance in a bag. We refer to it as the instance importance score. $e^{-t_i y_i}$ is the exponential bag classification error. The nonzero nondiagonal elements in $H_i$ indicate the correlations of instances in the same bag. $(e^{y_{i,j}}/e^{y_i})^{\lambda}(e^{y_{i,k}}/e^{y_i})^{\lambda}$ is the inner product of instance importance scores. We further derive the weak learners based on these theorems and involve the instance correlations for MI Boosting.

**Algorithm 1** Algorithm of ExpBin-miBoost

Given the training set $\{X_i, t_i\}, i = 1, \cdots, N_B$, in which $X_i = \{x_{i,j}\}, j = 1, \cdots, N_i$ denotes the bag with instances and $x_{i,j}$ denotes the instance.

1) Initialize the instance score as $y_{i,j}^1 = -\log N_i \,/\lambda$.
2) For $t = 1, \cdots, T$
   a) Compute $y_i^t, w_{i,j}^t$ with $y_{i,j}^t$.
   b) For each weak learner
      i) Partition the feature space $\chi$ into subspaces as $\chi_1, \chi_2, \cdots, \chi_l$.
      ii) Compute $\sum_{x_{i,j} \in \chi_m} w_{i,j}^t$ for each partitioned subspace and set $f_t(x) = sign\left(\sum_{x_{i,j} \in \chi_m} w_{i,j}^t\right)$, *if* $x_{i,j} \in \chi_m$.
   c) Find $\rho_t$ to make $y_{i,j}^t + \rho_t f_t\left(x_{i,j}\right)$ minimize $C$. Update the instance score $y_{i,j}^{t+1} = y_{i,j}^t + \rho_t f_t(x_{i,j})$.
3) The final classifier is the weighted sum of iteratively learned weak classifiers as $F(x) = \sum_t \rho_t f_t(x)$.

### *B. Boosting Approach for MI Classification*

Based on the above theorems, we propose the Boosting approach for MI classification. Two $L_p$-norm-localized MI Boosting algorithms are introduced. One is based on gradient descent optimization combining binary weak learners and linear step search. The other is based on Newton descent optimization aggregating real-value weak learners. To reduce the imbalance effect, an initialization step is first introduced.

*1) Initialization:* The imbalance between positive and negative bags exists in Boosting. That is, if the instance score $y_{i,j}$ is initialized as 0, the bag score $y_i > 0$. More instances lead to a larger initial bag score before learning, which is a numerical issue related to the instance-bag ratio (i.e., the instances number per bag). From the exponential loss in (3), the misclassification loss for negative bags is much higher than that for positive bags. Since Boosting focuses more on misclassified examples, the imbalance makes the weights of negative instances much higher than those of positive ones at the first Boosting iterations. Furthermore, weak learners at the first Boosting iterations are more determined by negative instances, while they ignore the information of positive ones. The weak learners are biased and less effective. In order to reduce this imbalance effect, we set the initial bag score to 0. Combined with (2), the instance score is initialized as $y_{i,j} = -(1/\lambda)\log N_i$.

*2) Gradient Descent-Based MI Boosting:* Gradient descent-based AnyBoost framework is based on fitting the first derivatives. It is quite useful in aggregating binary weak classifiers. Based on Theorem 1 and AnyBoost framework, we propose a novel Boosting algorithm for MI classification, referred to ExpBin-miBoost. Here, we let $w_{i,j} = -s_{i,j} = t_i e^{-t_i y_i}(e^{y_{i,j}}/e^{y_i})^{\lambda}$ denote the weight for the instance $x_{i,j}$ at each iteration of Boosting. $w_{i,j}$ can be divided into two parts. $(e^{y_{i,j}}/e^{y_i})^{\lambda}$ represents the instance importance in the bag. $t_i e^{-t_i y_i}$ is the bag weight for the MI misclassification penalty. Binary weak classifier $f(x_{i,j}) \in \{-1, 1\}$ is learned from the feature space $x_{i,j}$ to maximize $\sum_{i,j} w_{i,j} f(x_{i,j})$. Another 1-D linear search is imposed to find the step $\rho$ with $y_{i,j} + \rho f(x_{i,j})$ to minimize $C$. The algorithm framework is summarized in Algorithm 1.

Boosting aggregates the weak learners by adaptive resampling on instances. For ExpBin-miBoost, the instance weight $w_{i,j}$ is proportional to $e^{-t_i y_i}$ and $(e^{y_{i,j}}/e^{y_i})^{\lambda}$. The former is related to a bag classification error, which makes Boosting focus more on misclassified bags. The latter is the ratio of the rectified instance score against the rectified bag score for exploiting instance-bag correlations. It indicates that ExpBin-miBoost assigns higher weights for the instances with higher scores in the same bag, which provides a soft way of instance selection. The parameter $\lambda$ is related to the distribution of instance weights. As shown in Fig. 1, if $\lambda$ is large, the assigned weights for instances in the same bag are diverse, and the importance for the witness instances is higher. Hence, Boosting focuses more on the witness instances chosen by the Boosted weak learners. If $\lambda$ is chosen to be small (i.e., near 0), the weights for instances in the same bag are uniform, and Boosting assigns approximately equal importance for instances in the same bag. This means that the difference of instances in the same bag is disregarded. In particular, ExpBin-miBoost is flexible to handle different distributions of ambiguous instances. If the ambiguous instances are mostly positive, a small $\lambda$ is set to utilize all the included instances for learning. On the contrary, if a large proportion of negative instances are in positive bags, a large $\lambda$ is required to effectively choose the witness instances and exclude the negative instances.

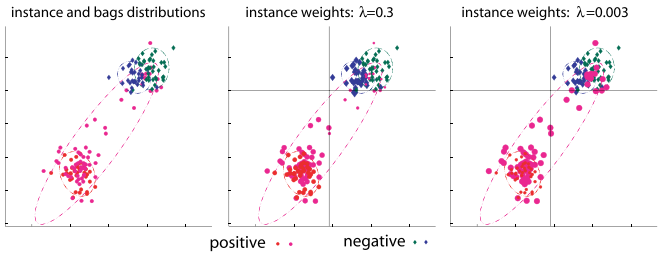


Fig. 1. Example to illustrate the effect of $\lambda$. Larger points indicate higher weights. For $\lambda = 0.3$, the weights for instances in the same bag significantly vary, and higher weights are assigned for chosen witness instances. For $\lambda = 0.003$, the weights for instances are nearly the same in the same bag.

*3) Newton Descent-Based MI Boosting:* We propose another novel Boosting algorithm by applying Newton descent optimization to the exponential loss minimization, referred to ExpReg-miBoost. First, the Newton updates for instances in bag $i$ can be concatenated into a vector $\mathbf{d}_i = [d_{i,1}, d_{i,2}, \ldots, d_{i,N_i}]^T$. Then, the Newton updates for all instances are composed into a long vector $\mathbf{d} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_{N_B}]$. They are computed as

$$\mathbf{d} = -H^{-1}\mathbf{s}, \quad \mathbf{d}_i = -H_i^{-1}\mathbf{s}_i \tag{7}$$

where $\mathbf{s}_i = [s_{i,1}, s_{i,2}, \ldots, s_{i,N_i}]^T$. At each iteration of Boosting, $d_{i,j}$ can be fitted with a stagewise function as

$$d(x_{i,j}) = \sum_m g_m I(x_{i,j} \in \chi_m) \tag{8}$$

where $I(x_{i,j} \in \chi_m)$ is the indication function. It is 1 only if $x_{i,j} \in \chi_m$; otherwise, it is 0. We need to find $\mathbf{g} = [g_m]$ that fits $-H^{-1}\mathbf{s}$. Suppose $\tilde{\mathbf{d}}$ is the approximation of $d_{i,j}$ with elements $d(x_{i,j})$, the problem is $\min_{\tilde{\mathbf{d}}} \varphi(\tilde{\mathbf{d}}) = (1/2)\tilde{\mathbf{d}}^T H \tilde{\mathbf{d}} + \mathbf{s}^T \tilde{\mathbf{d}}$. We further replace $d(x_{i,j})$ with $g_m$ in (8) and convert $\varphi(\tilde{\mathbf{d}})$ into $\psi(\mathbf{g})$ as

$$\psi(\mathbf{g}) = \frac{1}{2}\sum_{m,n} g_m g_n \left(\sum_i \sum_{j,k \in i} h_{i,j,k} I(x_{i,j} \in \chi_m, x_{i,k} \in \chi_n)\right) + \sum_m g_m \left(\sum_i \sum_{j \in i} s_{i,j} I(x_{i,j} \in \chi_m)\right). \tag{9}$$

**Algorithm 2** Algorithm of ExpReg-miBoost

Given the training set $\{X_i, t_i\}, i = 1, \cdots, N_B$, in which $X_i = \{x_{i,j}\}, j = 1, \cdots, N_i$ denotes the bag with instances and $x_{i,j}$ denotes the instance.

1) Initialize the instance score as $y_{i,j}^1 = -\log N_i \,/\lambda$.
2) For $t = 1, \cdots, T$
   a) Compute $y_i^t, s_{i,j}^t, h_{i,j,j}^t, h_{i,j,k}^t$ with $y_{i,j}^t$.
   b) For each weak learner
      i) Partition the feature space $\chi$ into subspaces as $\chi_1, \chi_2, \cdots, \chi_l$.
      ii) Compute $U_t, \mathbf{v}_t$ based on the instance distributions by $s_{i,j}^t, h_{i,j,j}^t, h_{i,j,k}^t$.
      iii) Solve the optimization problem as $\min \psi(\mathbf{g}_t) = \frac{1}{2}\mathbf{g}_t^T U_t \mathbf{g}_t + \mathbf{g}_t^T \mathbf{v}_t$.
   c) Find $\mathbf{g}_t$ which minimizes $C$, and update instance score $y_{i,j}^{t+1} = y_{i,j}^t + g_t(x_{i,j})$.
3) The final classifier is the sum of iteratively learned weak hypotheses as $G(x) = \sum_t g_t(x)$.

Suppose $U = [u_{m,n}]$, $\mathbf{v} = [v_m]$, and

$$u_{m,n} = \sum_i \sum_{j,k \in i} h_{i,j,k} I(x_{i,j} \in \chi_m, x_{i,k} \in \chi_n)$$
$$v_m = \sum_i \sum_{j \in i} s_{i,j} I(x_{i,j} \in \chi_m). \tag{10}$$

The solution of $\min_{\mathbf{g}} \psi(\mathbf{g})$ is $\mathbf{g}_{\min} = -U^{-1}\mathbf{v}$. Based on Newton descent optimization, the proposed ExpReg-miBoost fits the Newton updates to build an additive regression model to minimize $C$. Weak learners represented by $\mathbf{g}$ are learned iteratively, and the final classifier is the sum of iterative weak learners. In summary, the algorithm framework for ExpReg-miBoost is presented in Algorithm 2.

The ExpReg-miBoost is based on fitting the Newton updates that are computed from the second derivatives. From Theorem 2, we can find that the second derivative matrices $H_i$ are with nonzero nondiagonal elements indicating the instance correlations. Thus, the Newton updates of the instances in the same bag are correlated with each other, which further integrate the instance correlations into the weak learners. Similarly, $\lambda$ is related to the distribution of instance importance in ExpReg-miBoost. The larger $\lambda$ the higher importance is associated with the witness instances. In contrast, if $\lambda$ is set as near 0, the importance for the instances of the same bag is hardly affected by the instance importance scores. Since $\mathbf{g}$ is related to $1/\lambda$, $\lambda$ also affects the range of the weak learner values and the steps of Boosting iterations. There is no explicit representation of instance weights in ExpReg-miBoost, which is significantly different from ExpBin-miBoost and existing Boosting algorithms. The weights and fitting responses of the instances are implicitly and jointly represented by the Newton updates. Compared with ExpBin-miBoost, ExpReg-miBoost can achieve faster convergence speed, since it is based on Newton descent optimization with a quadratic convergence rate. Furthermore, the weak learners for ExpBin-miBoost are constrained to be binary, whereas for ExpReg-miBoost, they are real-value. The latter leads to more sufficient exploitation of instance distribution in the feature space.

### C. Stump-Based Weak Learners

Both ExpBin-miBoost and ExpReg-miBoost need to partition the feature space into disjoint subspaces. Here, we use thresholding to divide feature space and construct stump-based weak learners. For each feature, a single threshold is chosen to partition the feature space into two subspaces. The stagewise values are computed in partitioned subspaces for weak learner representation. For ExpBin-miBoost, the weak learners are decision stumps, and the values are the signs of summed instance weights. For ExpReg-miBoost, the weak learners are regression stumps that fit the Newton updates with minimal errors. In the training phase, a set of thresholds is chosen for each feature. At each iteration of Boosting, these thresholds are tested, and a 1-D search is imposed to find the best threshold with minimal loss $C$. The stump-based weak learner is constructed from the feature and the associated threshold. The weak learners, which achieve the minimal loss, are iteratively chosen. In the testing phase, the output of weak learners is computed by determining the falling subspaces with the chosen threshold. The final classifier output is the sum of the outputs of the Boosted weak learners. It is noteworthy that the stump-based weak learner is applied but is not the unique choice for the proposed MI Boosting approach.

### D. Computational Complexity

The computational complexity of the proposed algorithms is related to the type of weak learners. The training cost for the ExpBin-miBoost and ExpReg-miBoost algorithms can be divided into two parts. The first part is for computing the instance weights from the instance scores. The time cost is determined by the number of bags and the instance-bag ratio. Suppose the instance-bag ratio is $r$. For ExpBin-miBoost and ExpReg-miBoost, $N_B r$ times of exponential computation are required. For ExpBin-miBoost, the computational complexity for calculating instance weights is $O(N_B r)$. For ExpReg-miBoost, the complexity is increased by calculating the second derivative matrices, which is $O(N_B r^2)$. The second part of computational cost is for finding the best weak learners. It is related to the number of candidate weak learners, denoted by $N_w$. Since the stump-based weak learner is constructed from the feature with the associated threshold, $N_w$ is further determined by the number of features and the number of thresholds. For ExpBin-miBoost, the construction of each weak learner needs $N_B r$ times of comparison and addition. However, the 1-D step search adds an extra computational burden. Suppose the number of possible steps is $N_\rho$, then it requires $N_\rho N_B$ times of exponential computation and addition. For ExpReg-miBoost, $N_B r^2$ times of comparison and addition are required for each weak learner. If the number of candidate weak learners is not large, the computational cost of ExpBin-miBoost and ExpReg-miBoost is mainly determined by the time spent on exponential computation. Since ExpReg-miBoost does not require the step search, it is more time-saving. In contrast, if the training data are with a high instance-bag ratio and a large set of candidate weak learners, ExpBin-miBoost is preferred. In the testing phase, the computational complexity is the same for both ExpBin-miBoost and ExpReg-miBoost. Since the comparison and addition is needed only once for each weak learner, the computational complexity for online decision with the stump-based weak learners is quite low.

## IV. Experiments

We evaluate the proposed MI Boosting approach on ten benchmark data sets,[1,2] referred to *mutagenesis-atoms, mutagenesis-bonds, mutagenesis-chains, eastWest, westEast, musk1, musk2, elephant, tiger, fox,* respectively. The three *mutagenicity* data sets [1] are for drug activity prediction. The *eastWest/westEast* data sets [24] are to classify whether a train is eastbound or westbound. The two *musk* data sets [8] are to predict whether a given molecule emits a musky

[1] http://www.uco.es/grupos/kdis/mil/dataset.html

[2] Complementary experiments are provided in supplementary material.

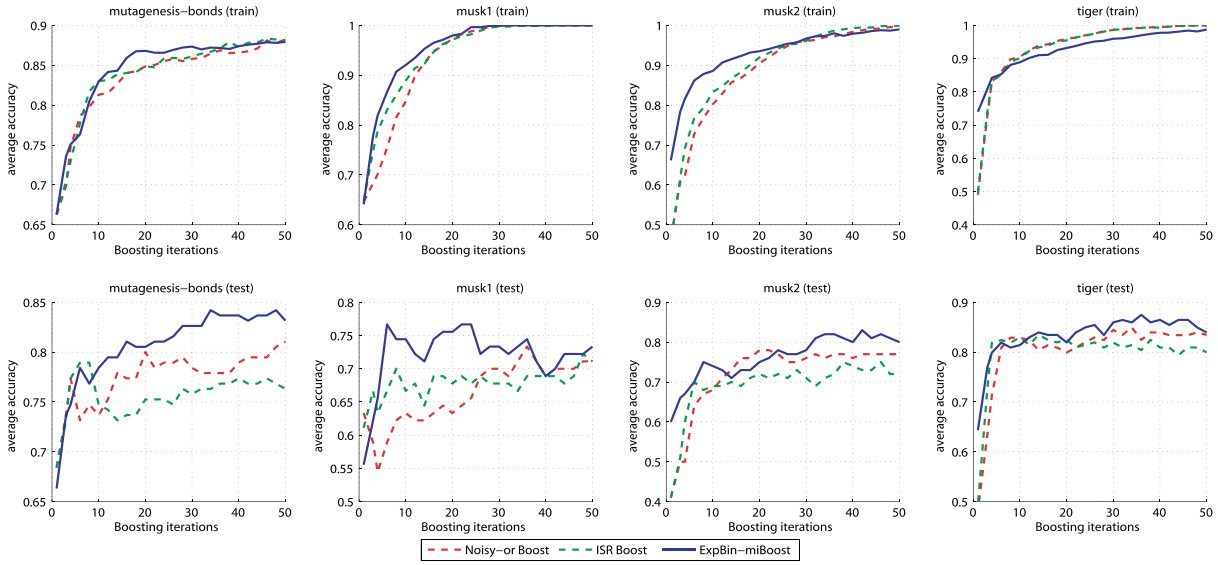


Fig. 2. Comparison of different Boosting algorithms, to illustrate the impact of optimization criteria.

odor. The three image retrieval data sets [24], such as *elephant, tiger* and *fox*, are to identify whether the target animals exist in images. Following the existing literature, tenfold cross-validation is applied for evaluation. The bags are equally divided into ten parts. Among them, nine parts are for training, and the rest is for testing. The average MI classification accuracy is used as the performance measurement.

### A. Comparison on Optimization Criteria

In order to illustrate the impact of optimization criteria, we compare ExpBin-miBoost with existing MI Boosting algorithms, such as Noisy-or Boost [6] and ISR Boost [6]. Following the notations in Section III, we further denote $p_i$ as the probability for bag $X_i$. The instance-bag mapping models for the Noisy-or Boost and the ISR Boost are mathematically formed as follows.

1) *NOR Mapping:* $p_i = 1 - \prod_{j\in i}\left(1-p_{i,j}\right), p_{i,j} = 1/\left(1+\exp(-y_{i,j})\right)$.
2) *ISR Mapping:* $p_i = \left(\sum_{j\in i}\exp(y_{i,j})\right) / \left(1+\sum_{j\in i}\exp(y_{i,j})\right)$.

Both Noisy-or Boost and ISR Boost are based on the logistic loss function as follows.

*Logistic Loss:* $C = -\sum_i[\tilde{t_i}\log(p_i) + (1-\tilde{t_i})\log(1-p_i)]$, where $\tilde{t_i} = (1+t_i)/2 \in \{0,1\}$. It can be seen that our proposed optimization criterion is significantly different from existing algorithms. The direct mapping from instance scores to bag scores reduces the complexity of subsequent optimization procedures. With the proposed optimization criterion, we introduce ExpBin-miBoost based on gradient descent optimization aggregating binary weak classifiers, just as Noisy-or Boost and ISR Boost. In Fig. 2, we present the average accuracy of tenfold cross-validation in both training and testing sets of *mutagenesis-bonds, musk1, musk2, tiger* data sets. The average accuracy increases with Boosting iterations. Compared with Noisy-or Boost and ISR Boost, ExpBin-miBoost can achieve faster convergence speed in training on *mutagenesis-bonds, musk1, musk2* data sets. Furthermore, the average accuracy in testing on *mutagenesis-bonds, musk1, musk2, tiger* data sets is also higher than that of Noisy-or Boost and ISR Boost. The experiment validates that the proposed optimization criterion is effective to construct the Boosting approach for MI classification.

### B. Performance of $L_p$-Norm-Localized MI Boosting

In Fig. 3, we present the average accuracy versus the number of Boosting iterations on four groups of data sets. The average accuracy curves on training sets are shown in Fig. 3(a), and those on test sets are shown in Fig. 3(b). Besides, the average accuracy curves of Noisy-or Boost [6], ISR Boost [6], and OWA-real miBoost [7] under the same experimental conditions are plotted for comparison. ExpBin-miBoost, Noisy-or Boost, and ISR Boost are all based on gradient descent optimization and binary weak classifiers. Compared ExpBin-miBoost with Noisy-or Boost and ISR Boost, it can be found that the MI classification accuracy is improved by the proposed $L_p$-norm and exponential loss-based optimization criterion. ExpBin-miBoost achieves higher training convergence rate in *eastWest/westEast* data sets and *musk* data sets. In addition, the MI classification accuracy in test with the four groups of data sets is increased by ExpBin-miBoost.

To compare ExpBin-miBoost with ExpReg-miBoost, the impact of optimization procedures is further illustrated. From Fig. 3(a), we can find that ExpReg-miBoost achieves higher average accuracy in training on all four groups of data sets. It validates that the Newton descent optimization can achieve higher convergence rates. The testing performance of ExpReg-miBoost and ExpBin-miBoost is comparable. The average accuracy on *mutagenesis* data sets and *musk* data sets is improved by ExpReg-miBoost. Comparing the $L_p$-norm-localized MI Boosting with existing algorithms, we can find that ExpReg-miBoost achieves the best training convergence performance. In testing, ExpBin-miBoost performs the best on *eastWest/westEast* data sets and image retrieval data sets. Meanwhile, ExpReg-miBoost significantly improves the classification accuracy on *mutagenesis* data sets and *musk* data sets.

The average accuracy of tenfold cross-validation on ten data sets is reported in Table I. Since the feature dimension on *eastWest/westEast* data sets (i.e., 11) is quite small and the convergence of Boosting is fast, we set the number of Boosting iterations on *eastWest/westEast* data sets to 50. The number of Boosting iterations on the other data sets is set to 100. $\lambda$ is fixed as 0.1. It can be seen from Table I that the average accuracy of ExpReg-miBoost is higher than that of ExpBin-miBoost on three *mutagenesis* data sets, two *musk* data sets, and *elephant* data set. By ExpReg-miBoost, the average accuracy achieved on *mutagenesis-atoms*, *mutagenesis-bonds*, and

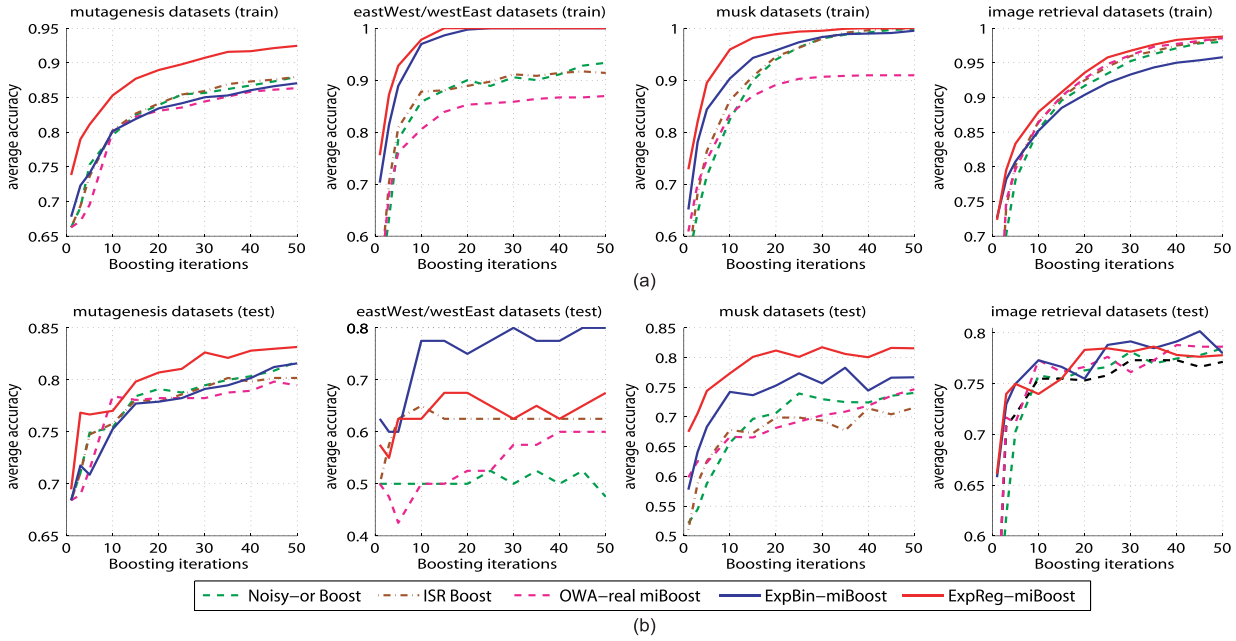


Fig. 3. Comparison with existing Boosting algorithms for MI classification on groups of data sets. (a) Average accuracy curves in training. (b) Average accuracy curves in testing.

TABLE I

COMPARISON OF AVERAGE ACCURACY (%) BY TENFOLD CROSS-VALIDATION ON MI CLASSIFICATION BENCHMARK DATA SETS

| algorithms | atom | bond | chain | east | west | musk1 | musk2 | eleph. | tiger | fox | avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ExpBin | 80.5 | 85.3 | 84.2 | **85.0** | **75.0** | 72.2 | 79.0 | 83.0 | **86.5** | **70.5** | **80.12** |
| ExpReg | **84.2** | **85.8** | **87.4** | 75.0 | 60.0 | 82.2 | 83.0 | 87.0 | 85.9 | 64.5 | **79.50** |
| NOR Boost[6] | 81.1 | 82.1 | 86.3 | 70.0 | 25.0 | 71.1 | 78.0 | 86.5 | 85.9 | 65.5 | 73.15 |
| ISR Boost[6] | 83.2 | 80.5 | 84.7 | 60.0 | 65.0 | 77.8 | 77.0 | 83.0 | 79.9 | 69.0 | 76.01 |
| OWA-real[7] | 75.8 | 82.6 | 84.7 | 80.0 | 40.0 | 83.3 | 66.0 | **88.0** | 81.9 | 61.0 | 74.33 |
| MI-OptBall[22] | 72.6 | 74.2 | 71.6 | 79.0 | 70.0 | 80.0 | 77.0 | 77.5 | 67.0 | 53.0 | 72.19 |
| MI-LR [5] | 70.0 | 71.1 | 76.8 | 60.0 | 60.0 | 87.8 | 85.0 | 78.0 | 77.0 | 56.0 | 72.17 |
| mi-DS [25] | 80.0 | 74.7 | 79.5 | 64.2 | 60.0 | 86.7 | 77.0 | 79.5 | 73.4 | 64.5 | 73.95 |
| MI-SVM[13] | 66.5 | 66.5 | 66.5 | 52.0 | 22.0 | 89.2 | 83.9 | 81.6 | 75.7 | 49.4 | 65.33 |
| MI-EMDD[10] | 73.2 | 72.1 | 73.7 | 45.0 | 30.0 | 88.9 | 90.0 | 73.0 | 74.5 | 59.0 | 67.94 |
| MI-NND[11] | 45.6 | 31.2 | 41.2 | 57.0 | 74.0 | 75.1 | 72.8 | 74.8 | 66.5 | 58.5 | 59.67 |
| MI-SMO[26] | 70.0 | 84.2 | 78.4 | 75.0 | 65.0 | 87.8 | 84.0 | 79.0 | 82.5 | 58.0 | 76.39 |
| BCKNN[12] | 75.4 | 72.8 | 74.5 | 65.0 | 65.0 | **92.2** | 83.2 | - | - | - | - |
| MI-Forest[21] | - | - | - | - | - | 85.0 | 82.0 | 84.0 | 82.0 | 64.0 | - |
| HSMILE[20] | - | - | - | - | - | 83.5 | 75.3 | 84.5 | 76.5 | 59.5 | - |
| MI-Graph[27] | - | - | - | - | - | 90.0 | 90.0 | 85.1 | 81.9 | 61.2 | - |
| mi-Graph[27] | - | - | - | - | - | 88.9 | **90.3** | 86.8 | 86.0 | 61.6 | - |
| AW-SVM[28] | - | - | - | - | - | 86.0 | 84.0 | 82.0 | 83.0 | 64.0 | - |
| AL-SVM[28] | - | - | - | - | - | 86.0 | 83.0 | 79.0 | 78.0 | 63.0 | - |
| ExpBin(optimal $\lambda$) | 81.6 | 85.8 | 84.7 | 85.0 | 75.0 | 74.4 | 82.0 | 86.5 | 89.5 | 72.0 | 81.65 |
| optimal $\lambda$ | 0.03 | 0.03 | 0.03 | 0.05 | 0.1 | 0.3 | 0.05 | 0.5 | 0.3 | 0.03 | - |
| ExpReg(optimal $\lambda$) | 84.2 | 87.9 | 88.4 | 75.0 | 65.0 | 82.2 | 84.0 | 87.5 | 88.0 | 68.0 | 81.02 |
| optimal $\lambda$ | 0.1 | 0.03 | 0.5 | 0.1 | 0.05 | 0.1 | 0.3 | 0.3 | 0.3 | 0.03 | - |

*mutagenesis-chains* is 84.2%, 85.8%, and 87.4%, respectively. ExpBin-miBoost achieves the average accuracy as 85% and 75% on *eastWest* and *westEast* data sets, respectively. The average accuracy by ExpBin-miBoost is 86.5% and 70.5% on *tiger* and *fox* data sets.

Next, we compare the MI Boosting algorithms with existing algorithms. The baselines include Boosting-based MI-OptBall [22], MI-LR [5], Noisy-or Boost [6], ISR Boost [6], OWA-real miBoost [7], SVM-based MI-SMO [26], mi-SVM [13], AW-SVM [28], AL-SVM [28], bagging-based MI-Forest [21], and HSMILE [20], as well as the other rule-based mi-DS [25], graph-theory-based MI-Graph [27], and mi-Graph [27]. We report the comparative results of Noisy-or Boost, ISR Boost, and OWA-real miBoost with re-implementation under the same conditions as ExpBin-miBoost and ExpReg-miBoost. The average accuracy of the other algorithms is either from the original literature or the existing results in [24] and [25]. It can be seen from Table I that the proposed $L_p$-norm-localized MI Boosting algorithms achieve the highest average accuracy on 7/10 data sets. ExpBin-miBoost achieves the highest average accuracy on four data sets as *eastWest, westEast, tiger* and *fox*. In addition, ExpReg-miBoost achieves the highest average accuracy on three *mutagenesis* data sets.

Compared with the state-of-the-art results on respective data sets, the average accuracy is increased by 1.2%, 1.9%, and 1.3% on three *mutagenesis* data sets with ExpReg-miBoost. By ExpBin-miBoost, the average accuracy is increased by 6.3% and 1.4% on *eastWest* and *westEast* data sets, respectively. In addition, the average accuracy rates on *tiger* and *fox* data sets are increased by 0.6% and 0.7%, respectively. In addition, the average of tenfold cross-validation accuracy over ten data sets is applied for statistical evaluation. The proposed ExpBin-miBoost achieves the highest average accuracy on ten benchmark data sets as 80.12%. The average accuracy achieved by ExpReg-miBoost is 79.50%. Compared with the existing MI Boosting algorithms, the average accuracy is increased by 5.4% with ExpBin-miBoost and 4.6% with ExpReg-miBoost. Compared with the state of the art, the average accuracy is increased by 4.9% and 4.1% with ExpBin-miBoost and ExpReg-miBoost, respectively.

### C. Effects of Parameter $\lambda$

We investigate the effects of $\lambda$ for the proposed $L_p$-norm-localized MI Boosting on ten data sets. $\lambda$ is jointly determined by the exponential base $a$ and $p$. A range of $\lambda$ values are chosen in approximately log-scale as $[0.03, 0.05, 0.1, 0.3, 0.5, 1]$. The average accuracy on ten benchmark data sets with fixed Boosting iterations is recorded. The optimal $\lambda$ varies by different data sets. On *mutagenesis-atoms* data set, the average accuracy remains stable with $\lambda$ ranging from 0.1 to 0.5. On *mutagenesis-bonds* and *mutagenesis-chains* data sets, both ExpBin-miBoost and ExpReg-miBoost obtain the comparable accuracy when $\lambda$ ranges from 0.03 to 0.5. The optimal $\lambda$ on *eastWest/westEast* data sets is smaller, and both algorithms

TABLE II
COMPARISON OF TRAINING TIME FOR ExpBin-miBoost AND ExpReg-miBoost

| datasets | bags number | feature length | instance-bag ratio | ExpBin train time(s) | ExpReg train time(s) |
|---|---|---|---|---|---|
| chains | 188 | 25 | 28.5 | 34 | 50 |
| eastWest | 20 | 24 | 8.9 | 3.0 | 0.6 |
| musk1 | 92 | 166 | 5.2 | 77 | 191 |
| elephant | 200 | 230 | 7 | 122 | 152 |

perform better when $\lambda$ is chosen from 0.05 to 0.1. On *musk1* data set, the average accuracy reaches the peak when $\lambda = 0.3$ with ExpBin-miBoost. On *musk2* data set, ExpBin-miBoost performs better when $\lambda$ is smaller than 0.1, and ExpReg-miBoost performs better when $\lambda$ ranges from 0.1 to 1. On *elephant* and *tiger* data sets, ExpBin-miBoost and ExpReg-miBoost perform better when $\lambda$ is bigger than 0.3. On *fox* data set, the accuracy slightly changes with $\lambda$.

As analyzed in Section III, the choice of $\lambda$ is related to the instance distributions. If the ambiguous instances in positive bags are close to positive, a small $\lambda$ is needed to sufficiently utilize the information of all instances. Reversely, if the ambiguous instances in positive bags are close to negative, a large $\lambda$ is required to effectively select the witness instances. In addition, we present the average accuracy for respective data sets with optimal $\lambda$ choosing in the last four rows of Table I. The average accuracy on ten data sets is further increased to 81.65% by ExpBin-miBoost and 81.02% by ExpReg-miBoost.

### D. Time Complexity

The computational complexity of the proposed ExpBin-miBoost and ExpReg-miBoost was discussed in Section III. The time cost is related to the type of weak learners. As discussed above, the time complexity in the testing phase is the same for ExpBin-miBoost and ExpReg-miBoost. Hence, we mainly compare the training time cost. The computer setting is Intel(R) Core(TM) i7-2600 CPU with 3.40-GHz processor and 2.99-GB RAM. We report the time cost for training 100 iterations of Boosting classifiers in Table II. The related parameters are listed for reference. It can be found that the training time of ExpReg-miBoost is generally longer than that of ExpBin-miBoost. For ExpBin-miBoost, a linear step search is required, which adds extra time cost. However, the computation of second derivative matrices costs more time, especially under the larger instance-bag ratio conditions. On *mutagenesis-chains, musk1* and *elephant* data sets, the training time of ExpReg-miBoost is longer than that of ExpBin-miBoost. On *eastWest* data set with few bags and instances, the training time cost of ExpReg-miBoost is shorter than that of ExpBin-miBoost.

## V. CONCLUSION

In this brief, we propose a novel Boosting approach for MI classification. The optimization criterion for Boosting is formulated from the $L_p$-norm and exponential loss. First, gradient descent optimization is applied to derive the binary weak learners for ExpBin-miBoost. Second, we introduce the Newton descent optimization-based ExpReg-miBoost. The real-value weak learners for ExpReg-miBoost are derived from the Newton updates fitting. For ExpBin-miBoost, the instance correlations are denoted by iteratively updated instance weights. For ExpReg-miBoost, the instance correlations are represented by the Newton updates that are computed from correlated second derivative matrices. By aggregating iteratively chosen binary and real-value weak learners, the $L_p$-norm-localized Boosting algorithms are developed for MI classification. Experiments on benchmark data sets for MI classification demonstrate the effectiveness of the proposed approach. Compared with the state of the art, the proposed $L_p$-norm-localized MI Boosting achieves the highest average accuracy on 7/10 data sets as well as the best average MI classification performance. In the future, we would like to introduce the termination criterion in order to choose the optimal number of Boosting iterations. The appropriate termination criterion is required to learn compact and effective Boosted classifier for MI classification. Besides, we would like to apply the proposed $L_p$-norm-localized MI Boosting algorithms into practical applications.

## REFERENCES

[1] A. Srinivasan, S. H. Muggleton, R. D. King, and M. J. E. Sternberg, "Mutagenesis: ILP experiments in a non-determinate biological domain," in *Proc. 4th Int. Workshop Inductive Logic Program. (ILP)*, Bonn, Germany, Sep. 1994, pp. 217–232.

[2] K. Ali and K. Saenko, "Confidence-rated multiple instance boosting for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 2433–2440.

[3] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?—Weakly-supervised learning with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 685–694.

[4] R. Hong, M. Wang, Y. Gao, D. Tao, X. Li, and X. Wu, "Image annotation by multiple-instance learning with discriminative feature mapping and selection," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 669–680, May 2014.

[5] X. Xu and E. Frank, "Logistic regression and boosting for labeled bags of instances," in *Proc. 8th Pacific-Asia Conf. Knowl. Discovery Data Mining (PAKDD)*, Sydney, NSW, Australia, May 2004, pp. 272–281.

[6] P. Viola, J. C. Platt, and C. Zhang, "Multiple instance boosting for object detection," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, Dec. 2006, pp. 1419–1426.

[7] H. Hajimirsadeghi and G. Mori, "Multiple instance real boosting with aggregation functions," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Tsukuba, Japan, Nov. 2012, pp. 2706–2710.

[8] 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.

[9] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proc. Conf. Adv. Neural Inf. Process. Syst. (NIPS)*, Denver, CO, USA, Dec. 1999, pp. 570–576.

[10] Q. Zhang and S. A. Goldman, "EM-DD: An improved multiple-instance learning technique," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, Dec. 2002, pp. 1073–1080.

[11] J. Wang and J.-D. Zucker, "Solving multiple-instance problem: A lazy learning approach," in *Proc. 17th Int. Conf. Mach. Learn. (ICML)*, Stanford, CA, USA, Jun. 2000, pp. 1119–1125.

[12] L. Jiang, Z. Cai, D. Wang, and H. Zhang, "Bayesian citation-KNN with distance weighting," *Int. J. Mach. Learn. Cybern.*, vol. 5, no. 2, pp. 193–199, 2014.

[13] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, Dec. 2003, pp. 577–584.

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

[15] Y. Xiao, B. Liu, L. Cao, J. Yin, and X. Wu, "SMILE: A similarity-based approach for multiple instance learning," in *Proc. 10th IEEE Int. Conf. Data Mining (ICDM)*, Sydney, NSW, Australia, Dec. 2010, pp. 589–598.

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

[17] Z. Fu, A. Robles-Kelly, and J. Zhou, "MILIS: Multiple instance learning with instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 958–977, May 2011.

[18] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "Learning sparse kernel classifiers for multi-instance classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 9, pp. 1377–1389, Sep. 2013.

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

[20] H. Yuan, M. Fang, and X. Zhu, "Hierarchical sampling for multi-instance ensemble learning," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 12, pp. 2900–2905, Dec. 2012.

[21] C. Leistner, A. Saffari, and H. Bischof, "MIForests: Multiple-instance learning with randomized trees," in *Proc. 11th Eur. Conf. Comput. Vis. (ECCV)*, Heraklion, Greece, Sep. 2010, pp. 29–42.

[22] P. Auer and R. Ortner, "A boosting approach to multiple instance learning," in *Proc. 15th Eur. Conf. Mach. Learn.*, Pisa, Italy, Sep. 2004, pp. 63–74.

[23] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *Ann. Statist.*, vol. 28, no. 2, pp. 337–407, 2001.

[24] L. Dong, "A comparison of multi-instance learning algorithms," M.S. thesis, School Comput. Math. Sci., Univ. Waikato, Hamilton, New Zealand, 2006.

[25] D. T. Nguyen, C. D. Nguyen, R. Hargraves, L. A. Kurgan, and K. J. Cios, "mi-DS: Multiple-instance learning algorithm," *IEEE Trans. Cybern.*, vol. 43, no. 1, pp. 143–154, Feb. 2013.

[26] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA, USA: MIT Press, 1998.

[27] 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. (ICML)*, Montreal, QC, Canada, Jun. 2009, pp. 1249–1256.

[28] P. V. Gehler and O. Chapelle, "Deterministic annealing for multiple-instance learning," in *Proc. 11th Int. Conf. Artif. Intell. Statist. (AISTATS)*, San Juan, PR, USA, Mar. 2007, pp. 123–130.