Pattern Recognition 47 (2014) 3106–3119

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Active learning with multi-criteria decision making systems

Ran Wang [a,b], Sam Kwong [a,b,*]

[a] Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong
[b] City University of Hong Kong Shenzhen Research Institute, Shenzhen 518057, China

ARTICLE INFO

Article history:
Received 19 October 2012
Received in revised form
26 September 2013
Accepted 18 March 2014
Available online 27 March 2014

Keywords:
Active learning
Multiple-instance learning
Multi-criteria decision making
Support vector machine

ABSTRACT

In active learning, the learner is required to measure the importance of unlabeled samples in a large dataset and select the best one iteratively. This sample selection process could be treated as a decision making problem, which evaluates, ranks, and makes choices from a finite set of alternatives. In many decision making problems, it usually applied multiple criteria since the performance is better than using a single criterion. Motivated by these facts, an active learning model based on multi-criteria decision making (MCMD) is proposed in this paper. After the investigation between any two unlabeled samples, a preference preorder is determined for each criterion. The dominated index and the dominating index are then defined and calculated to evaluate the informativeness of unlabeled samples, which provide an effective metric measure for sample selection. On the other hand, under multiple-instance learning (MIL) environment, the instances/samples are grouped into bags, a bag is negative only if all of its instances are negative, and is positive otherwise. Multiple-instance active learning (MIAL) aims to select and label the most informative bags from numerous unlabeled ones, and learn a MIL classifier for accurately predicting unseen bags by requesting as few labels as possible. It adopts a MIL algorithm as the base classifier, and follows an active learning procedure. In order to achieve a balance between learning efficiency and generalization capability, the proposed active learning model is restricted to a specific algorithm under MIL environment. Experimental results demonstrate the effectiveness of the proposed method.

1. Introduction

In classification, the use of active learning technique [11,2] is an effective method of selecting a small number of training samples from a large unlabeled dataset. It iteratively measures the informativeness of unlabeled samples, selects the one with the highest preference level, makes query on its label, and updates the training set. The goal is to learn an accurate model by making as few queries as possible. Applying active learning to classification can efficiently reduce the effort for manual labeling and data redundancy.

Currently, most existing active learning strategies are based on single criterion. However, it is commented that the integration of multiple criteria is likely to give better performance than each single one. In [44], a multi-criteria based learning strategy is proposed for named entity recognition, which measures the informativeness of unlabeled samples by the weighted-sum of normalized uncertainty and diversity. This measurement tries to maximize the contribution of the selected samples. However, since the ranges of the criteria are quite different, the weighted-sum of their normalized values tends to weaken some important information when the weights are not well set, and may fail to reflect the sample's priority with respect to each criterion.

Multi-criteria decision making (MCDM) [56] is a well known branch of decision making, which evaluates a finite set of alternatives on the basis of two or more criteria. MCDM tasks include choosing the best alternative from the set of candidates, or sorting the alternatives into a preference preorder. Under some assumptions, the decision maker is required to maximize an utility function [50], or indicate the priority between any two alternatives with respect to each criterion. Due to the generality and universality, MCDM has been applied to various areas such as operational research [40,53,27], evolutionary multi-objective optimization [19,5,45,14], knowledge discovery [8], and preference modeling [18]. It is noteworthy that the main task in active learning is to measure the preference levels of unlabeled samples, which might be well handled by a MCDM system. In this paper, an active learning model with MCDM system is proposed. By investigating the priority between any two unlabeled samples, a preference preorder is determined for each criterion. The dominated indices and dominating indices of a sample in the preference preorders are then defined and computed to reflect its informativeness. Since the preference preorders are evaluated

* Corresponding author.
E-mail address: CSSAMK@cityu.edu.hk (S. Kwong).

http://dx.doi.org/10.1016/j.patcog.2014.03.011
0031-3203/

uniformly and independently, this strategy is likely to provide accurate priority information.

On the other hand, multiple instance learning (MIL) is a task that aims to train classification models from structured data. In MIL, the samples are grouped into bags, and each sample is taken as an instance of its bag. A bag is positive if at least one of its instances is positive, and is negative only if all of its instances are negative [36,37]. MIL was first introduced to drug activity prediction [15], and then applied to real-world machine learning scenarios such as content-based image retrieval (CBIR) [10] and document classification [43]. In CBIR and document classification, the images or documents are taken as the bags, and the segmented regions or short passages are their instances. The key step in MIL is to identify the instances that are responsible for the final decisions in each bag [9,42,29].

Many successful algorithms for training MIL classifiers have been developed [4,16,54], however, the problem of how to collect sufficient training bags has not been properly addressed. One solution to this problem is applying multiple-instance active learning (MIAL), which is able to select and label the most informative bags from numerous unlabeled ones. Currently, MIAL has been realized with three types of strategies: (1) instance-level learning; (2) bag-level learning; and (3) mixed mode. The distinction among them is lying in the query objects, which could be instances, bags, and their combinations. By analyzing the heuristic optimization of MIL algorithms, it is observed that applying MCDM system to MIAL is likely to achieve a good trade-off between the learning efficiency and the generalization capability. Thus, the proposed active learning model is further restricted to a specific algorithm under MIL environment.

The rest of this paper is organized as follows: in Section 2, background knowledge and motivation of this work are given. In Section 3, the active learning model with MCDM system is proposed. In Section 4, the proposed learning model is refined to a specific algorithm under MIL environment. In Section 5, experimental comparisons are conducted to show the feasibility and effectiveness of the proposed method. Finally, conclusions and future work directions are given in Section 6.

2. Backgrounds and motivation

2.1. Basic definitions in MCDM

Generally, a MCDM system consists of two phases [53]: (1) information input and construction; (2) aggregation and exploitation. In the first phase, a typical MCDM problem [24,6] is mathematically modeled as a decision matrix defined in Definition 1.

Definition 1 (*decision matrix*). Given a MCDM problem with n distinct alternatives $\mathbf{A}_1, \ldots, \mathbf{A}_n$ and m criteria $Cr_1, \ldots, Cr_m$, where the level of achievement of $\mathbf{A}_i$ $(i=1,\ldots,n)$ with regard to Cr_k $(k=1,\ldots,m)$ is denoted by $Cr_k(\mathbf{A}_i)$, then

$$\begin{bmatrix} Cr_1(\mathbf{A}_1) & Cr_2(\mathbf{A}_1) & \ldots & Cr_m(\mathbf{A}_1) \\ Cr_1(\mathbf{A}_2) & Cr_2(\mathbf{A}_2) & \ldots & Cr_m(\mathbf{A}_2) \\ \vdots & \vdots & \ddots & \vdots \\ Cr_1(\mathbf{A}_n) & Cr_2(\mathbf{A}_n) & \ldots & Cr_m(\mathbf{A}_n) \end{bmatrix}$$

is called the decision matric of this problem.

In the second phase, the aggregation and the exploitation are carried out on the basis of the decision matrix constructed in the first phase, as well as the decision maker's willingness. Generally, $Cr_k(\mathbf{A}_i)$ could be either cardinal [38] or ordinal [13,12,17]. If the decision matrix is composed of cardinal numbers, the system is required to maximize an utility function $u(\mathbf{A}_i)=\sum_{k=1}^{m} w_k Cr_k(\mathbf{A}_i)$, where w_k is the weight of Cr_k. While for preference modeling problems, the decision matrix is usually composed of ordinal numbers. In this case, the system is required to indicate that, regarding each criterion Cr_k, which of the following four relations holds for any two distinct alternatives $\mathbf{A}_i, \mathbf{A}_j$:

1. $\mathbf{A}_i \succ \mathbf{A}_j$: $\mathbf{A}_i$ is preferred to $\mathbf{A}_j$;
2. $\mathbf{A}_i \prec \mathbf{A}_j$: $\mathbf{A}_j$ is preferred to $\mathbf{A}_i$;
3. $\mathbf{A}_i ? \mathbf{A}_j$: $\mathbf{A}_i$ is incomparable to $\mathbf{A}_j$;
4. $\mathbf{A}_i \approx \mathbf{A}_j$: $\mathbf{A}_i$ is indifferent to $\mathbf{A}_j$.

Holding these relations, all the alternatives could be arranged into some preference preorders as defined in Definition 2.

Definition 2 (*preference preorder*). Given a MCDM problem with n distinct alternatives $\mathbf{A}_1, \ldots, \mathbf{A}_n$ and m criteria $Cr_1, \ldots, Cr_m$, where the level of achievement of $\mathbf{A}_i$ $(i=1,\ldots,n)$ with regard to Cr_k $(k=1,\ldots,m)$ is denoted by $Cr_k(\mathbf{A}_i)$. If for Cr_k there exists $Cr_k(\mathbf{A}_1^*) \succ \approx Cr_k(\mathbf{A}_2^*) \succ \approx \cdots \succ \approx Cr_k(\mathbf{A}_n^*)$, where $\mathbf{A}_1^* \neq \mathbf{A}_2^* \neq \cdots \neq \mathbf{A}_n^* \in \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n\}$, then the order $\mathbf{A}_1^*, \mathbf{A}_2^*, \ldots, \mathbf{A}_n^*$ is called a preference preorder of $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n$ with regard to Cr_k, denoted by $\mathcal{P}_k$.

It is obvious that the anterior alternative in $\mathcal{P}_k$ gets more priority and importance with respect to Cr_k. Based on the preference preorders, the following definitions could be further introduced.

Definition 3 (*k-th dominated index and k-th dominating index*). Given that $\mathcal{P}_k$ is a preference preorder of alternatives $\mathbf{A}_1, \ldots, \mathbf{A}_n$ with regard to Cr_k $(k=1,\ldots,m)$, then the k-th dominated index and the k-th dominating index of $\mathbf{A}_i$ $(i=1,\ldots,n)$, denoted by $\psi_k^{\succ}(\mathbf{A}_i)$ and $\psi_k^{\prec}(\mathbf{A}_i)$, are respectively defined as

$$\psi_k^{\succ}(\mathbf{A}_i) = \sum_{j=1,\ldots,n, j \neq i} d(\succ, R_{ij}^{(k)}), \tag{1}$$

and

$$\psi_k^{\prec}(\mathbf{A}_i) = \sum_{j=1,\ldots,n, j \neq i} d(\prec, R_{ij}^{(k)}). \tag{2}$$

where $d(\cdot,\cdot)$ is a distance metric, $R_{ij}^{(k)} \in \{\succ, \prec, \approx, ?\}$ is the relation of $\mathbf{A}_i$ and $\mathbf{A}_j$ with regard to Cr_k.

Definition 4 (*dominated index and dominating index*). Given that $\mathcal{P}_1, \ldots, \mathcal{P}_m$ are the preference preorders of alternatives $\mathbf{A}_1, \ldots, \mathbf{A}_n$ with regard to $Cr_1, \ldots, Cr_m$, then the dominated index and the dominating index of $\mathbf{A}_i$ $(i=1,\ldots,n)$, denoted by $\psi^{\succ}(\mathbf{A}_i)$ and $\psi^{\prec}(\mathbf{A}_i)$, are respectively defined as

$$\phi^{\succ}(\mathbf{A}_i) = \sum_{k=1}^{m} w_k \psi_k^{\succ}(\mathbf{A}_i), \tag{3}$$

and

$$\phi^{\prec}(\mathbf{A}_i) = \sum_{k=1}^{m} w_k \psi_k^{\prec}(\mathbf{A}_i), \tag{4}$$

where w_k is the weight of Cr_k.

The dominated index and the dominating index of $\mathbf{A}_i$ respectively reflect the degrees of $\mathbf{A}_i$ being dominated by others and dominating others in $\mathcal{P}_1, \ldots, \mathcal{P}_m$. Obviously, in a MCDM problem, an alternative with lower dominated index and higher dominating index is preferred.

2.2. Active learning

In active learning, the learner is required to measure the preference levels of unlabeled samples and select the best one from a large dataset iteratively, which is supposed to be well handled by a MCDM system. Generally, there are stream-based framework [20] and pool-based framework [48] for active learning. In stream-based learning, the alternatives are observed one by one, which is impossible to determine a preference preorder. While in pool-based learning, the learner is allowed to access all the candidates at one time, and select the most informative one during each iteration. Thus, we will mainly focus on pool-based learning as described in Algorithm 1.

Algorithm 1. Basic framework of pool-based active learning.

Input:
  Labeled set: $\mathbb{L} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ with l initially labeled instances;
  Unlabeled pool: $\mathbb{U} = \{\mathbf{x}_i\}_{i=l+1}^{l+u}$ with u unlabeled instances;
  Parameters for training the base-classifier;
  Criterion Cr for informativeness measurement.
while $\mathbb{U}$ is not empty **do**
  if stop criterion is met **then**
    stop;
  else
    For each $\mathbf{x}_i \in \mathbb{U}$, calculate its informativeness $Cr(\mathbf{x}_i)$.
    Select the optimal instance $\mathbf{x}^*$, i.e., $\mathbf{x}^* = \arg\max_{\mathbf{x}_i \in \mathbb{U}} Cr(\mathbf{x}_i)$;
    Query the label of $\mathbf{x}^*$, denoted by y^*;
    Let $\mathbb{U} = \mathbb{U} - \mathbf{x}^*$, and $\mathbb{L} = \mathbb{L} \cup (\mathbf{x}^*, y^*)$;
    Update h based on $\mathbb{L}$;
  end if
end while
Output:
  Classifier h trained on the final training set.

The key issue in Algorithm 1 is to design a proper sample selection criterion. Under traditional framework, various criteria have been proposed such as uncertainty [52,33], expected error [7,34,41], size of version space [48], diversity [48], density [39], and relevance [47]. Although these criteria have been proved to be effective in different scenarios, none of them can outperform others on all cases. In [44], an integration of uncertainty and diversity has demonstrated a possibility of multi-criteria based active learning, however, no framework is proposed in this work.

Let us first consider the simplest case with only two criteria, which could be either dependent or independent. Suppose that for these two criteria, a larger value represents a higher informativeness, thus three relations may happen as shown in Fig. 1. If the two criteria are independent, the unlabeled alternatives will be randomly located at the measure space, which will lead to an independent distribution as shown in Fig. 1(a). Otherwise, one criterion value will either decrease or increase with the increase of the other criterion value, which will lead to a conflict distribution as shown in Fig. 1(b) or a consistent distribution as shown in Fig. 1(c). It is noteworthy that under active learning environment, we are talking about some approximate relations.

It is trivial to discuss MCDM for the consistent case, since the two criteria are roughly accordant. While for the independent and conflict cases, further investigation seems to be necessary. In addition, it should be valid for a general case when more than two criteria are taken into consideration.

2.3. Multiple instance learning

Different from traditional learning problems, in MIL, the instances are grouped into bags, a bag is positive if at least one of its instances is positive, and negative only if all its instances are negative. Given a MIL problem on the training set $\{(\mathbf{B}_i, y_i)\}_{i=1}^{n}$ where $\mathbf{B}_i = \{\mathbf{B}_{ij}\}_{j=1}^{n_i}$ is the i-th bag with n_i instances, the instance-level labels could be either known or unknown. If only bag-level labels are known, $y_i \in \{+1, -1\}$ is the label of $\mathbf{B}_i$, otherwise, $y_i = \{y_{i1}, \dots, y_{in_i}\}$ contains the label information of all the instances in $\mathbf{B}_i$ where $y_{ij} \in \{+1, -1\}, j = 1, \dots, n_i$. The goal of MIL is to learn a classification model that can accurately predict new unlabeled bags.

MIL problems have been resolved with axis-parallel rectangles [15], modified random walk process [51], combinatorial margin maximization formulation [29], and other methods [9,4,16,22,55,32]. In this paper, we only focus on the learning methods with support vector machine (SVM). SVM is a well known classification technique based on statistical learning theory [49]. The main idea is to generate an optimal separating hyperplane that can maximize the margin between the two referred classes. Given a training set with single instance $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n} \in \mathbf{R}^d \times \{+1, -1\}$, by considering the soft-margin case, SVM is formulated as

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \ge 1 - \xi_i \\ & \xi_i \ge 0, \quad i = 1, \dots, n \end{aligned} \tag{5}$$

where C is a trade-off constant, and ξ_i is the slack variable introduced to $\mathbf{x}_i$.

Currently, two SVM-based MIL models have been proposed under the condition that only bag-level labels are known [1], i.e., mi-SVM and MI-SVM. Specifically, mi-SVM is a maximum instance margin formulation:

$$\begin{aligned} \min_{\{y_{ij}\}} \min_{\mathbf{x}, b, \xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i\sum_j \xi_{ij} \\ \text{s.t.} \quad & \forall i, j : y_{ij}(\mathbf{w} \cdot \mathbf{B}_{ij} + b) \ge 1 - \xi_{ij} \\ & \xi_{ij} \ge 0, \ y_{ij} \in \{-1, 1\} \\ & \sum_{j \in \mathbf{B}_i} \frac{y_{ij} + 1}{2} \ge 1, \quad \forall i \text{ s.t. } y_i = 1 \\ & y_{ij} = -1, \quad \forall i \text{ s.t. } y_i = -1 \end{aligned} \tag{6}$$

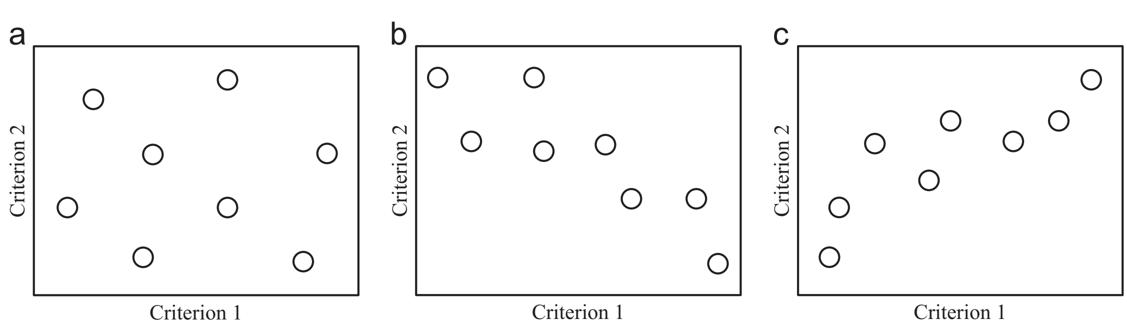

Fig. 1. Three different relations between two selection criteria: (a) relation1-independent, (b) relation2-conflict, and (c) relation3-consistent.

where ξ_{ij} is the slack variable introduced to $\mathbf{B}_{ij}$. The main idea of mi-SVM is that when training SVM, all the instances of the negative bags are treated as negative, and at least one instance of each positive bag is treated as positive. The optimization heuristic of mi-SVM is described in Appendix A.

Similarly, MI-SVM is a maximum bag margin formulation:

$$\begin{aligned} \min_{\mathbf{x},b,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \\ \text{s.t.} \quad & \forall i: y_i \max_{j \in \mathbf{B}_i} (\mathbf{w} \cdot \mathbf{B}_{ij} + b) \ge 1 - \xi_i \\ & \xi_i \ge 0 \end{aligned} \tag{7}$$

where ξ_i is the slack variable introduced to $\mathbf{B}_i$. This method applies the maximum margin idea to bags instead of instances, which directly puts the efforts on predicting the bag-level labels.

3. Active learning with multi-criteria decision making systems

3.1. Distance metric

Follow Algorithm 1, we further let $\mathbb{C} = \{Cr_1, Cr_2, \ldots, Cr_m\}$ be a set of m criteria on the unlabeled set $\mathbb{U}$. Iteratively, the task is to select the optimal alternative $\mathbf{A}^*$ from $\mathbb{U}$ that has the highest preference level regarding all the criteria. In this case, an effective distance metric $d(\cdot,\cdot)$ is required for applying a MCDM system to active learning.

Let $R, R' \in \{\succ, \prec, ?, \approx\}$, it is mentioned in [40] that the distance between R and R', i.e., $d(R, R')$, is a real number. There are six conditions for $d(\cdot,\cdot)$ to be a metric measure:

1. $d(\succ, ?) = d(\prec, ?)$ and $d(\succ, \approx) = d(\prec, \approx)$.
2. $d(\succ, \approx) + d(\approx, \prec) = d(\succ, \prec)$.
3. $d(\succ, ?) \ge d(\approx, ?)$.
4. $d(\approx, ?) \ge d(\approx, \succ)$.
5. $d(\succ, \prec) = \max\{d(R, R') \mid R, R' \in \{\succ, \prec, \approx, ?\}\}$.
6. $d(R, R') > 0$, if $R \ne R'$ and $d(R, R') = 0$ if $R = R'$.

In order to get a concrete form of $d(R, R')$, one further adopts the assumption $d(\succ, \prec) - d(\succ, ?) = d(\succ, ?) - d(\approx, ?) = d(\approx, ?) - d(\approx, \succ)$ [40,27]. Finally, the distances are derived as in Table 1, where a is a positive real number.

By applying the distance metrics in Table 1 to Definitions 3 and 4, the dominated index and the dominating index could be computed to measure the informativeness of unlabeled samples.

3.2. Active ranking procedures

Generally, in a preference preorder, the alternatives with smaller dominated indices and larger dominating indices should be ranked ahead. Thus, the ranking procedures based on dominated indices and dominating indices are respectively described in Algorithms 2 and 3. We denote the preference preorders led by Algorithms 2 and 3 as $\mathcal{P}^{\succ}$ and $\mathcal{P}^{\prec}$. Thus, the learner tends to select the anterior ones from $\mathcal{P}^{\succ}$ and $\mathcal{P}^{\prec}$.

Table 1
Distances between relations.

$d(\cdot,\cdot)$	$\mathbf{A}_i \succ \mathbf{A}_p$	$\mathbf{A}_i \prec \mathbf{A}_p$	$\mathbf{A}_i ? \mathbf{A}_p$	$\mathbf{A}_i \approx \mathbf{A}_p$
$\mathbf{A}_i \succ \mathbf{A}_p$	0	$2a$	$(5/3)a$	a
$\mathbf{A}_i \prec \mathbf{A}_p$	$2a$	0	$(5/3)a$	a
$\mathbf{A}_i ? \mathbf{A}_p$	$(5/3)a$	$(5/3)a$	0	$(4/3)a$
$\mathbf{A}_i \approx \mathbf{A}_p$	a	a	$(4/3)a$	0

Algorithm 2. Ranking procedure based on dominated indices.

Let $X_1 = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n\}$, and $I_1 = \{1, 2, \ldots, n\}$;
for $j = 1$:number of iteration **do**
  Compute $\phi_j^{\succ}(\mathbf{A}_i) = \sum_{k=1}^{m} w_k \psi_k^{\succ}(\mathbf{A}_i)$,
where $\mathbf{A}_i \in X_j$, $i \in I_j$, and $\psi_k^{\succ}(\mathbf{A}_i) = \sum_{k \in I_j} d(\succ, R_{ip}^{(k)})$;
  Let $X^* = \{\mathbf{A}^* | \phi_j^{\succ}(\mathbf{A}^*) = \min\{\phi_j^{\succ}(\mathbf{A}_i) | i \in I_j\}\}$.
The alternatives in X^* are ranked in the j-th place;
  Let $X_{j+1} = X_j - X^*$ and $I_{j+1} = \{i \in I_j | \mathbf{A}_i \in X_{j+1}\}$;
end for

Algorithm 3. Ranking procedure based on dominating indices.

Let $X_1 = \{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n\}$, and $I_1 = \{1, 2, \ldots, n\}$;
for $j = 1$:number of iteration **do**
  Compute $\phi_j^{\prec}(\mathbf{A}_i) = \sum_{k=1}^{m} w_k \psi_k^{\prec}(\mathbf{A}_i)$,
where $\mathbf{A}_i \in X_j$, $i \in I_j$, and $\psi_k^{\prec}(\mathbf{A}_i) = \sum_{k \in I_j} d(\prec, R_{ip}^{(k)})$;
  Let $X^* = \{\mathbf{A}^* | \phi_j^{\prec}(\mathbf{A}^*) = \max\{\phi_j^{\prec}(\mathbf{A}_i) | i \in I_j\}\}$.
The alternatives in X^* are ranked in the j-th place;
  Let $X_{j+1} = X_j - X^*$ and $I_{j+1} = \{i \in I_j | \mathbf{A}_i \in X_{j+1}\}$;
end for

The procedures in Algorithms 2 and 3 are accordant with active learning iterations. The alternatives ranked in the first place will be selected in the first iteration of active learning, the one ranked in the second place will be selected in the second iteration, etc. However, in a ranking problem, $\mathcal{P}^{\succ}$ and $\mathcal{P}^{\prec}$ may not be consistent, i.e., the least dominated alternative may not be the most dominating one. In this case, we simply use

$$info(\mathbf{A}_i) = \psi^{\prec}(\mathbf{A}_i) - \psi^{\succ}(\mathbf{A}_i), \tag{8}$$

as the informativeness measurement, which roughly reflects a higher dominating index and a lower dominated index. Specifically, we have

$$info(\mathbf{A}_i) = \sum_{k=1}^{m} w_k \sum_{j \ne i} [d(\prec, R_{ij}^{(k)}) - d(\succ, R_{ij}^{(k)})],$$

where $d(\prec, R_{ij}^{(k)})$ and $d(\succ, R_{ij}^{(k)})$ are in $[0, 2a]$. Furthermore, since $0 < w_k < 1$, the range of $info(\mathbf{A}_i)$ could be determined as

$$(-2am(u-1), 2am(u-1)),$$

where u represents the number of unlabeled alternatives to be ranked, m is the number of criteria, and a is a positive real number. Thus, during each active learning iteration, the sample $\mathbf{A}^* = \arg\max_{\mathbf{A}_i \in \mathbb{U}} info(\mathbf{A}_i)$ is selected. It is clear that this measurement is just determined by the parameter set (u, m, a), and is dependent on the range of any single criterion value.

3.3. Application considerations

In Sections 3.1 and 3.2, we have discussed how to realize active learning with MCDM systems. However, in real implementation, the following two issues should be further taken into consideration.

1. Usually, different criteria are computed with different learning structures. Thus, the calculation on the criteria values may be complicated.
2. Due to the first consideration, the time complexity will be much higher than single criterion based methods.

Simply speaking, in order to improve the generalization capability, the learning efficiency may be sacrificed to a great extent. In order to realize a good trade-off between generalization and efficiency, our focus further shifts to another application domain, i.e., active learning under MIL environment.

4. Applications to multiple-instance active learning

In this section, we first give the informativeness measurements for unlabeled bags, then propose the MIAL algorithm with MCDM system.

4.1. Settings for MIAL

Basically, there are two reasons for applying MCDM system to MIAL:

1. For a given unlabeled bag, different measurements are computed based on the same instances. Thus, the calculations could be finished with only one learning structure.
2. As demonstrated in Appendix A, the training procedures of MIL classifiers are composed of some iterations. We generally hope that a more informative bag will lead to a lower number of training iterations, which may achieve a trade-off between the generalization and the efficiency.

It is noteworthy that in MIAL, the query objects could be either bags or instances. If the bag-level labels are queried, the base classifier will be mi-SVM or MI-SVM, while if the instance-level labels are queried, the base classifier will be traditional SVM. Thus, the settings for MIAL could be concluded as in Table 2. In this paper, we only focus on the first two settings which covered most real cases.

4.2. Measurements for instance

Given a SVM classifier $f(\mathbf{x}) = \mathbf{w} \cdot \phi(\mathbf{x}) + b$ where $\mathbf{w}$ is a real-valued vector and ϕ denotes the kernel mapping, the decision value of the j-th instance in the i-th unlabeled bag is calculated as $f(\mathbf{B}_{ij}) = \mathbf{w} \cdot \phi(\mathbf{B}_{ij}) + b$. Since the decision value is proportional to its distance to the current decision boundary, with the logistic regression model [26], the conditional probabilities of the unlabeled instance $\mathbf{B}_{ij}$ are computed as

$$\begin{cases} P(y_{ij} = +1|\mathbf{B}_{ij}) = \frac{1}{1+e^{-(\mathbf{w}\cdot\phi(\mathbf{B}_{ij})+b)}} \\ P(y_{ij} = -1|\mathbf{B}_{ij}) = 1 - P(y_{ij} = +1|\mathbf{B}_{ij}) \end{cases}. \tag{9}$$

When we denote $P(y_{ij} = +1|\mathbf{B}_{ij})$ and $P(y_{ij} = -1|\mathbf{B}_{ij})$ by P_{ij}^{+} and P_{ij}^{-}, the uncertainty of $\mathbf{B}_{ij}$, i.e., $U(\mathbf{B}_{ij})$, is easily got by its label entropy:

$$U(\mathbf{B}_{ij}) = -P_{ij}^{+} \log P_{ij}^{+} - P_{ij}^{-} \log P_{ij}^{-}. \tag{10}$$

4.3. Measurements for bag

4.3.1. Bag margin

The margin of the unlabeled bag $\mathbf{B}_i$, i.e., $M(\mathbf{B}_i)$, could be measured either by its minimum instance margin or its average instance margin [57]. Then the informativeness of $\mathbf{B}_i$ is defined as $Cr(\mathbf{B}_i) = 1/M(\mathbf{B}_i)$.

4.3.2. Softmax model

Softmax model is a smooth approximation of maximum. For a given set of values $x_1, \ldots, x_n$, the softmax approximation is

$$softmax_\alpha(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i \cdot e^{\alpha \cdot x_i} \Bigg/ \sum_{i=1}^{n} e^{\alpha \cdot x_i}, \tag{11}$$

where α is a parameterized constant. When $\alpha \to \infty$, $softmax_\alpha \to \infty$; when $\alpha = 0$, $softmax_\alpha$ is a simple average; and when $\alpha \to -\infty$, $softmax_\alpha \to -\infty$.

In this case, the conditional probabilities of $\mathbf{B}_i$ are computed as

$$\begin{cases} P(y_i = +1|\mathbf{B}_i) = softmax_\alpha(P_{i1}^{+}, \ldots, P_{in_i}^{+}) \\ P(y_i = -1|\mathbf{B}_i) = 1 - softmax_\alpha(P_{i1}^{+}, \ldots, P_{in_i}^{+}) \end{cases}. \tag{12}$$

Similarly, when we denote $P(y_i = +1|\mathbf{B}_i)$ and $P(y_i = -1|\mathbf{B}_i)$ by P_i^{+} and P_i^{-}, the uncertainty of bag $\mathbf{B}_i$, i.e., $U(\mathbf{B}_i)$, is computed as

$$U(\mathbf{B}_i) = -P_i^{+} \log P_i^{+} - P_i^{-} \log P_i^{-}. \tag{13}$$

Another softmax-based model, i.e., combinU model [21], is proposed by making the softmax approximation of the instance uncertainties, thus the uncertainty of $\mathbf{B}_i$ could also be calculated as

$$U(\mathbf{B}_i) = softmax_\alpha(U(\mathbf{B}_{i1}), \ldots, U(\mathbf{B}_{in_i})). \tag{14}$$

And the informativeness of $\mathbf{B}_i$ is then defined as $Cr(\mathbf{B}_i) = U(\mathbf{B}_i)$.

4.3.3. Noisy-Or model

The generalized Noisy-Or structure [46] is a convenient tool for measuring uncertainty relationships. With this model, the relationship between $P(y_i|\mathbf{B}_i)$ and $P(y_{ij}|\mathbf{B}_{ij})$ could be characterized as

$$\begin{cases} P(y_i = +1|\mathbf{B}_i) = 1 - \prod_{j=1}^{n_i} (1 - P(y_{ij} = +1|\mathbf{B}_{ij})) \\ P(y_i = -1|\mathbf{B}_{ij}) = \prod_{j=1}^{n_i} (1 - P(y_{ij} = +1|\mathbf{B}_{ij})) \end{cases}. \tag{15}$$

The uncertainty of $\mathbf{B}_i$ is then computed by (13).

4.3.4. Fisher information

Fisher information [58] has been proved to be a good sample selection criterion, and successfully applied to traditional batch-mode active learning [25]. Given a data distribution $q(\mathbf{x})$ with n i.i.d samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and a classification model $p(y|\mathbf{x}, \theta)$, the Fisher Information Matrix is defined as

$$I_{q(\mathbf{x})}(\theta) = -\int q(\mathbf{x})\, d\mathbf{x} \int p(y|\mathbf{x}, \theta) \frac{\partial^2}{\partial \theta^2} \log p(y|\mathbf{x}, \theta)\, d\mathbf{x}, \tag{16}$$

where θ includes all the parameters of the classification model.

Based on the standard Cramér-Rao lower-bound, the co-variance matrix of $\mathbf{x}_1, \ldots, \mathbf{x}_n$, denoted by $cov(t_n)$, satisfies that $cov(t_n) \geq (1/n) I(\theta)^{-1}$, where $I(\theta)$ is the Fisher Information Matrix. This inequality gives the lower-bound of the co-variance matrix regarding a set of samples, which provides a reliable measure for selecting multiple samples in batch-mode active learning. Given that $p(\mathbf{x})$ is the distribution of all the unlabeled samples, and $q(\mathbf{x})$ is the distribution of a subset of the unlabeled samples for manual labeling, the Cramér-Rao lower-bound implies that the distribution q^* can most efficiently reduce the model uncertainty, which

Table 2
Different settings for MIAL with SVM classifier.

Setting	Bag label	Instance label	Query object	Base learner
1	Unknown	Unknown	Bag	mi-SVM and MI-SVM
2	Unknown	Unknown	Instance	SVM
3	Known	Unknown	Instance in positive bag	SVM
4	Some are known and some are unknown	unknown	Bag, instance in positive bag, and combinations	mi-SVM and MI-SVM

could be found by minimizing the trace of $I_{q(\mathbf{x})}(\theta)^{-1}I_{p(\mathbf{x})}(\theta)$. In MIAL, this measure is feasible for selecting valuable bag $\mathbf{B}^*$ with multiple instances, i.e.,

$$\mathbf{B}^* = \arg\min_{\mathbf{B}} \operatorname{tr}(I_{q(\mathbf{B})}(\theta)^{-1}I_{p(\mathbf{B})}(\theta)), \tag{17}$$

where $q(\mathbf{B})$ is the data distribution of bag $\mathbf{B}$, and $p(\mathbf{B})$ is the distribution of all the unlabeled bags.

As mentioned in [57], under MIAL environment, $I_{p(\mathbf{B})}(\theta)$ does not have any concrete form and cannot be regarded as a function of $q(\mathbf{B})$, thus is neglected. Then, for unlabeled bag $\mathbf{B}_i$, its Fisher information is defined as

$$fis(\mathbf{B}_i) = -\operatorname{tr}\left(\sum_{y_i=\pm 1} p(y_i|\mathbf{B}_i,\theta)\frac{\partial^2}{\partial\theta^2}\log\, p(y_i|\mathbf{B}_i,\theta)\right), \tag{18}$$

where a larger $fis(\mathbf{B}_i)$ represents a higher informativeness.

With different probability models, $fis(\mathbf{B}_i)$ will be different. In most cases, the concrete expression of $fis(\mathbf{B}_i)$ is impractical due to the second-order derivative [57]. Here, we propose an effective expression form:

$$\begin{aligned} fis(\mathbf{B}_i) &= -\operatorname{tr}\left(\sum_{y_i=\pm 1} p(y_i|\mathbf{B}_i,\theta)\frac{\partial^2}{\partial\theta^2}\log\, p(y_i|\mathbf{B}_i,\theta)\right) \\ &= (P_i^-/P_i^+)\times\sum_{j\in\mathbf{B}_i}(P_{ij}^+\times\phi(\mathbf{B}_{ij}))^T\sum_{j\in\mathbf{B}_i}(P_{ij}^+\times\phi(\mathbf{B}_{ij})) \\ &= (P_i^-/P_i^+)\times\sum_{j\in\mathbf{B}_i}\sum_{q\in\mathbf{B}_i}(P_{ij}^+\times\phi(\mathbf{B}_{ij}))^T(P_{iq}^+\times\phi(\mathbf{B}_{iq})) \\ &= (P_i^-/P_i^+)\times\sum_{j\in\mathbf{B}_i}\sum_{q\in\mathbf{B}_i}P_{ij}^+\times P_{iq}^+\times K(\mathbf{B}_{ij},\mathbf{B}_{iq}), \end{aligned} \tag{19}$$

where ϕ denotes a kernel mapping, and $K(\cdot,\cdot)$ is the corresponding kernel function. Usually, one can use the logistic regression model (9) to determine the conditional probabilities of instances, and use the softmax model (12) or the Noisy-Or model (15) to determine the conditional probabilities of bags. Then, the informativeness of $\mathbf{B}_i$ is defined as $Cr(\mathbf{B}_i) = fis(\mathbf{B}_i)$.

4.4. MIAL with MCDM

It is mentioned in [40] that when all the alternatives are arranged into a preference preorder, the preorder is called a complete preorder if the set of incomparable relation is empty, i.e., $? = \varnothing$, which is a particular case of partial preorder. Note that all the above introduced criteria are real-valued criteria, which make the bags comparable. Thus, the preference preorders determined by these criteria are complete preorders. In this case, two issues need to be resolved for each criterion:

1. How to define the relation of $\approx$.
2. How to design the weight w.

As for the first issue, we design a threshold between the relations $\approx$ and $\prec$ or $\succ$. The threshold for the k-th criterion is calculated as

$$T_k = \frac{\max_{\mathbf{B}_i\in\mathbb{U}}\{Cr_k(\mathbf{B}_i)\} - \min_{\mathbf{B}_i\in\mathbb{U}}\{Cr_k(\mathbf{B}_i)\}}{|\mathbb{U}|}, \tag{20}$$

where $k=1,\dots,m$, and $|\mathbb{U}|$ represents the number of unlabeled bags in $\mathbb{U}$. If the difference of Cr_k between two bags is smaller than T_k, they are taken as indifferent regarding the k-th criterion.

As for the second issue, it is investigated in [28] that the quality of decisions improves quickly with the improving on weights information. In [3], this problem is solved by a linear utility function in a L_p metric form, but this metric form is infeasible with the dominated and dominating indices. Similar to [23] and [30], we assign the importance of the criteria from a set of feasible weights W such that $w_k \in W$ and $\sum_{k=1}^m w_k = 1$.

By integrating the active ranking procedures discussed in Section 3.2, the criteria discussed in Section 4.3, and the above designed parameters, the MIAL scheme with MCDM system is described in Algorithm 4.

Algorithm 4. Multiple-instance active learning with multi-criteria decision making system

Input:
- Labeled set: $\mathbb{L} = \{(\mathbf{B}_i, y_i)\}_{i=1}^l$ with l initially labeled bags;
- Unlabeled pool: $\mathbb{U} = \{(\mathbf{B}_i\}_{i=l+1}^{l+u}$ with u unlabeled bags;
- Parameters for training the base-classifier;
- A set of criterion measurements $\mathbb{C} = \{Cr_1, Cr_2, \dots, Cr_m\}$.

while $\mathbb{U}$ is not empty **do**
  if stop criterion is met **then**
    stop;
  else
    For each unlabeled bag $\mathbf{B}_i \in \mathbb{U}$, calculate a set of criterion values $Cr_1(\mathbf{B}_i), Cr_2(\mathbf{B}_i), \dots, Cr_m(\mathbf{B}_i)$ by $\mathbb{C}$.
    For each criterion $Cr_k \in \mathbb{C}$, $k=1,\dots,m$, calculate the threshold T_k by Eq. (20).
    Based on T_k, get the relations between any two unlabeled bags regarding Cr_k from $\{\succ, \prec, \approx\}$.
    Generate the weights of the criteria, i.e., w_k, $k=1,\dots,m$.
    Compute the dominated index and the dominating index of each $\mathbf{B}_i$, i.e., $\phi^{\succ}(\mathbf{B}_i) = \sum_{k=1}^m w_k\psi_k^{\succ}(\mathbf{B}_i)$ and $\phi^{\prec}(\mathbf{B}_i) = \sum_{k=1}^m w_k\psi_k^{\prec}(\mathbf{B}_i)$;
    Calculate the informativeness of each $\mathbf{B}_i$, i.e., $info(\mathbf{B}_i) = \phi^{\prec}(\mathbf{B}_i) - \phi^{\succ}(\mathbf{B}_i)$;
    Select the bag $\mathbf{B}^*$ with the maximum value of $info(\mathbf{B}_i)$, i.e. $\mathbf{B}^* = \arg\max_{\mathbf{B}_i\in\mathbb{U}} info(\mathbf{B}_i)$;
    For bag-query-mode, query the label of $\mathbf{B}^*$; for instance-query-mode, query all the instances of $\mathbf{B}^*$;
    Let $\mathbb{U} = \mathbb{U} - \mathbf{B}^*$, and $\mathbb{L} = \mathbb{L} \cup (\mathbf{B}^*, y^*)$;
    Update h based on $\mathbb{L}$;
  end if
end while

Output:
Classifier h trained on the final training set.

5. Experimental comparisons

In this section, we conduct some experimental comparisons to show the feasibility and the effectiveness of the proposed algorithm.

5.1. Learning strategies for performance comparison

Nine learning strategies are listed in this section for performance comparison:

1. *Random sampling* (*Random*): During each iteration, the learner randomly selects a bag from the unlabeled pool and add it to the training set.
2. *Active learning SVM* (*SVMactive*): The learner selects the bag that contains the instance with the minimum margin.
3. *Bag margin based learning* (*BagMargin*): The learner selects the bag with the minimum average instance margin.
4. *Softmax based learning* (*SoftMax*): The learner selects the bag with the highest uncertainty based on Eq. (13).

5. *CombinU based learning* (*CombinU*): The learner selects the bag with the highest CombinU value based on Eq. (14).
6. *Noisy-Or based learning* (*NoisyOr*): This method selects the bag with the highest uncertainty based on Eq. (15).
7. *Fisher information based learning* (*Fisher*): The bag with the highest fisher information, i.e., Eq. (19), is selected during each iteration.
8. *Weighted-sum based learning* (*WeightedSum*): This is a learning strategy with multiple criteria. The values for each criterion are normalized into $[0,1]$ where 1 represents the highest preference. Then, the weighted-sum of these criteria is used as the informativeness measurement. In our experiment, we adopt three criteria as bag margin, combinU, and Fisher information, which have independent evaluating mechanisms.
9. *Learning with MCDM system* (*Proposed*): Algorithm 4 is realized, which also adopts bag margin, combinU, and Fisher information.

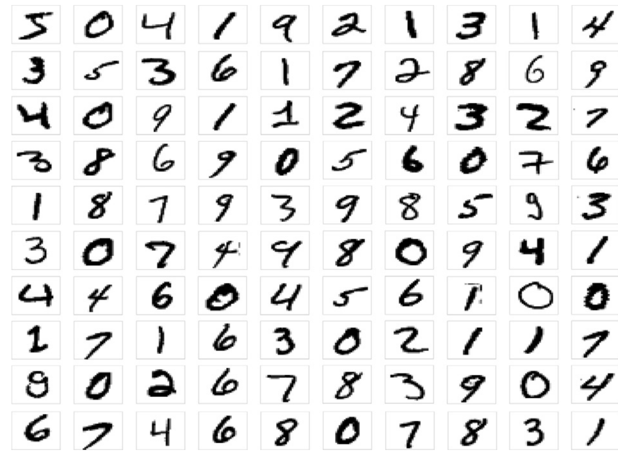

Fig. 2. The first 100 training examples of MNIST dataset.

5.2. Datasets

5.2.1. Handwritten digit image datasets

We first test the learning strategies on the MNIST handwritten digit image recognition problem.[1] This task is to distinguish 0–9 handwritten digits from approximately 250 writers that contains 60,000 training samples and 10,000 testing samples. Fig. 2 shows the first 100 training images of this dataset. The original black and white handwritten digit images from NIST were normalized to fit in a 20×20 pixel box while preserving their aspect ratio. Then, with an anti-aliasing technique and a normalization algorithm, the images were further centered in 28×28 gray level images. Thus the raw information of each sample in MNIST is composed of 784 gray level pixels with each pixel value $\in \{0,\dots,255\}$. We use the gradient-based method[2] presented in [31,35] to extract the gradient histogram features, and finally construct a 2172-dimensional feature vector. Furthermore, in order to generate a compact dataset, we perform a feature selection process and retain 65 features.

For each class, a MIL dataset is artificially generated. Each MIL dataset is composed of 100 positive bags and 100 negative bags with each bag containing 20 instances. Take digit 0 as an example, the instances in the negative bags are randomly selected from the samples of digits 1–9 without overlap; while for each positive bag, an integer n_{pos} $(1\le n_{pos}\le 20)$ that represents the number of positive instances in it is first generated, then the positive instances and the negative instances are respectively selected from the samples of digit 0 and digits 1–9 randomly. Finally, the detailed information of the 10 MIL datasets is listed in Table 3. The task for these datasets is to identify that whether a specific digit exists in a set of images.

5.2.2. Content-based image retrieval datasets

Three Corel MIL image datasets,[3] i.e., elephant, fox, and tiger, are used to simulate some CBIR tasks. The goal is to distinguish these three kinds of animals from other background pictures. In other words, given a set of images, identify that whether any picture of a specific animal exists. Detailed descriptions of these datasets are listed in Table 4.

[1] http://yann.lecun.com/exdb/mnist
[2] http://www.cs.berkeley.edu/~smaji/projects/digits
[3] http://www.cs.columbia.edu/~andrews/mil/datasets.html

Table 3
Detailed description of the generated datasets.

Datasets	#Bags		#Instances	
	Positive	Negative	Positive	Negative
Digit "0"	100	100	1013	2987
Digit "1"	100	100	1156	2844
Digit "2"	100	100	1104	2896
Digit "3"	100	100	1040	2960
Digit "4"	100	100	998	3002
Digit "5"	100	100	913	3087
Digit "6"	100	100	1025	2975
Digit "7"	100	100	1021	2979
Digit "8"	100	100	1096	2904
Digit "9"	100	100	994	3006

Table 4
Detailed description of the corel MIL image datasets.

Datasets	#Features (nonzero)	#Bags		#Instances		#Instances per bag	
		Positive	Negative	Positive	Negative	Min	Max
Elephant	230 (143)	100	100	762	629	1	13
Fox	230 (143)	100	100	647	673	2	13
Tiger	230 (143)	100	100	544	676	2	13

5.3. Experimental settings

For each handwritten digit image dataset, 50% data are randomly selected as the training set, and the other 50% are taken as the testing set. Since the instance-level labels are known, we implement MIAL with both instance-query-mode and bag-query-mode. With instance-query-mode, the learning starts with one positive bag and one negative bag, the learner selects one new bag during each iteration, and queries the labels of all its instances. The base classifier for instance-query-mode is traditional SVM. With bag-query-mode, the learning starts with two positive bags and two negative bags, the learner selects one new bag during each iteration, and queries its label directly. The base learner for bag-query-mode is mi-SVM.

Regarding the image retrieval datasets, 70% data are randomly selected as the training set, and the other 30% are taken as the testing set. Since the instance-level labels are unknown, we only implement bag-query-mode. The learning starts with 10 positive bags and 10 negative bags, and the learner selects one new bag during each iteration. The base learner is mi-SVM.

For fair comparison, Gaussian kernel is used consistently with parameters $C=100$ and $\sigma=1$ for all the learning strategies. The parameter α in softmax and combinU models is set as 1. For each task, the learning repeats 50 times, and the average value is observed. The experiments are performed under MATLAB 7.9.0 with the "svmtrain" and "svmpredict" functions of libsvm, which are executed on a computer with a 3.16-GHz Intel Core 2 Duo CPU, a maximum 4.00-GB memory, and 64-bit windows 7 operating system.

5.4. Conflict analysis

When applying multiple criteria to decision making problems, conflict should exist between at least two criteria. However, theoretical proof on this kind of conflict is very difficult under multiple-instance environment. Thus, we empirically investigate it by recording the rankings of different criteria on the unlabeled bags.

Take digit 0 of MNIST as an example, there are 100 training bags in total. Since one positive bag and one negative bag are included in the initial training set, there are 98 unlabeled bags at the beginning. We plot the ranking relations among the three criteria, i.e., Fisher information, bag margin, and combinU value, on the 98 unlabeled bags. Note that a higher ranking represents a lower preference, e.g., the bag ranked as 1 has the highest preference to be selected, while the bag ranked as 98 has the lowest preference. It is observed from Fig. 3 that obvious conflict exists between Fisher information and combinU value, as well as bag margin and combinU value. Similar cases also happen for other datasets. Thus, the integration of multiple criteria is a good choice to facilitate the learning.

5.5. Result and discussion

The performances of different learning strategies on the MNIST MIL datasets with instance-query-mode and bag-query-mode are respectively shown in Figs. 4 and 5. It can be seen that the results of these two query modes are quite different. The range of instance-query-mode is much smaller than that of bag-query-mode. The initial accuracy is around 90% for instance-query-mode and is between 60% and 70% for bag-query-mode. We assume that mi-SVM assigns labels to the instances by the optimization heuristic, but these labels may still differ a lot from the real ones. In this case, when instance-level labels are known, the learning is easier and more accurate at the beginning. However, although the bag-query-mode has a lower accuracy at the beginning, it gradually outperforms the instance-query-mode in the latter part of the learning. Besides, in most real cases, it is time-consuming and even impractical to query all the instance labels, thus we pay more attention to the bag-query-mode. It is observed from Fig. 5 that in the anterior part of the learning, the proposed algorithm outperforms the other strategies on all the datasets, which demonstrates a faster convergence rate. Consider the latter part of the learning, it also gives the best performance on most datasets. Besides, it can be seen that in some cases, the weighted-sum of multiple criteria performs worse than its single criterion. Since the ranges of the criteria are quite different, the weighted-sum of their normalized values may weaken some important information when the weights are not well set. While in the proposed model, the dominated and dominating indices of an unlabeled sample with regard to all the criteria have the same range, which is independent of any specific criterion value. Thus, it is able to perform better than its single criterion. In addition, applying multiple criteria can effectively improve the stability of the performance.

The average performances of the two query modes on the 10 datasets are shown in Fig. 6. Fig. 6(a) and (b) demonstrates the average accuracy and deviation of instance-query-mode respectively. The proposed method obviously outperforms others during the entire learning process. It exhibits a stable performance with a relatively low deviation, which overcomes the problems of both single criterion strategies and the weighted-sum multi-criteria strategy. Besides, it can be seen from Fig. 6(a) that SVMactive, which selects the bag that contains the minimum margin instance, is the second best strategy. This is mainly benefitted from the accurate information on instance-level labels. Fig. 6(c) and (d) demonstrates the average accuracy and deviation of bag-query-mode respectively. The advantage of the proposed method is more obvious under this mode. It has a very fast convergence rate in the anterior part of the learning, and tends to be stable after 10 bags are labeled. In addition, the weighted-sum method is the second best one, which further validates the effectiveness of applying multiple criteria in bag-query-mode active learning.

Tables 5 and 6 respectively report the average time for each learning iteration in instance-query-mode and bag-query-mode on the MNIST MIL problems. The time cost during each learning iteration mainly consists of two parts: training base classifiers, and selecting new bags. It is observed from Table 5 that for instance-query-mode, the weighted-sum method and the proposed method have the highest time complexities. This is easy to explain, since all the strategies use almost the same amount of time to train the base classifiers, while when measuring the unlabeled bags, these two methods spend longer time to calculate the values of multiple criteria. However, it is observed from Table 6 that the time evaluation on bag-query-mode is much different from that on instance-query-mode. In fact, the two multi-criteria based methods are not the most time-consuming ones. Their time complexities are similar to those of the single criterion based methods, and even much lower than that of the softmax model. The reason

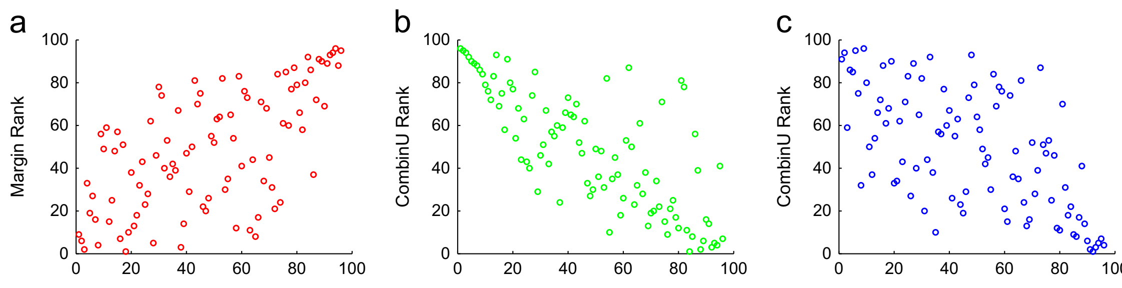

Fig. 3. Overall relations among the three selected criteria. (a) Fisher information vs. bag margin. (b) Fisher information vs. combinU value. (c) Bag margin vs. combinU value.

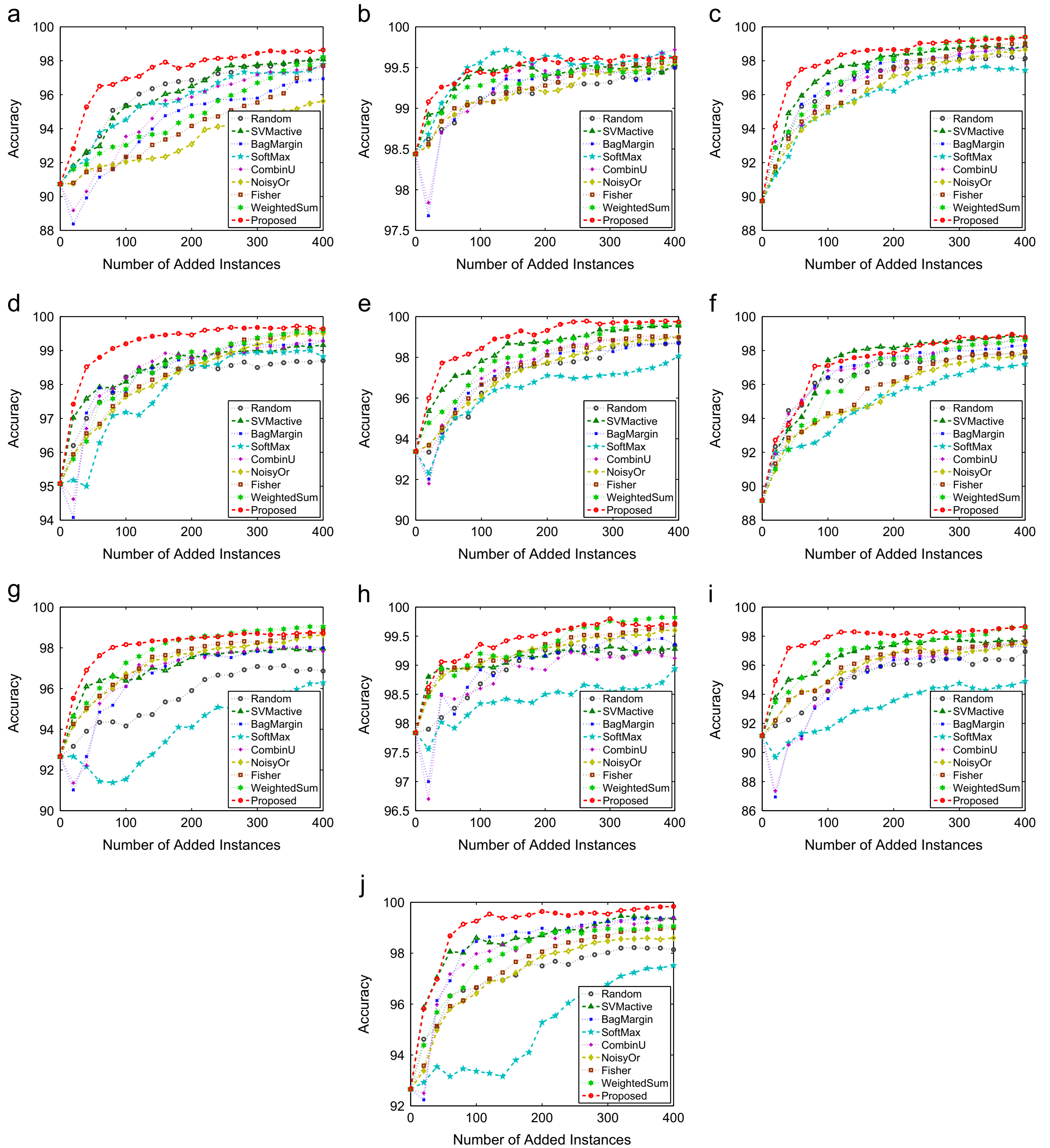

Fig. 4. Instance-query-mode: performance comparison of different learning strategies on recognizing handwritten digit images. (Base-learner: SVM). (a) Digit "0" (50 runs). (b) Digit "1" (50 runs). (c) Digit "2" (50 runs). (d) Digit "3" (50 runs). (e) Digit "4" (50 runs). (f) Digit "5" (50 runs). (g) Digit "6" (50 runs). (h) Digit "7" (50 runs). (i) Digit "8" (50 runs). (j) Digit "9" (50 runs).

could be explained by investigating the average number of iterations for training mi-SVM, which is also reported in Table 6. Although these two methods spend longer time in the bag selection process, this extra cost is compensated by the time used in the base classifier training process, since a more informative bag can lead to a smaller number of training iterations. It also explains why MCDM system is more suitable for MIAL than traditional active learning.

Fig. 7 reports the performances of the learning strategies on the three CBIR tasks. First, the average accuracy and deviation on the elephant dataset are given in Fig. 7(a) and (b) respectively. It can be seen that all the active learning strategies outperform the

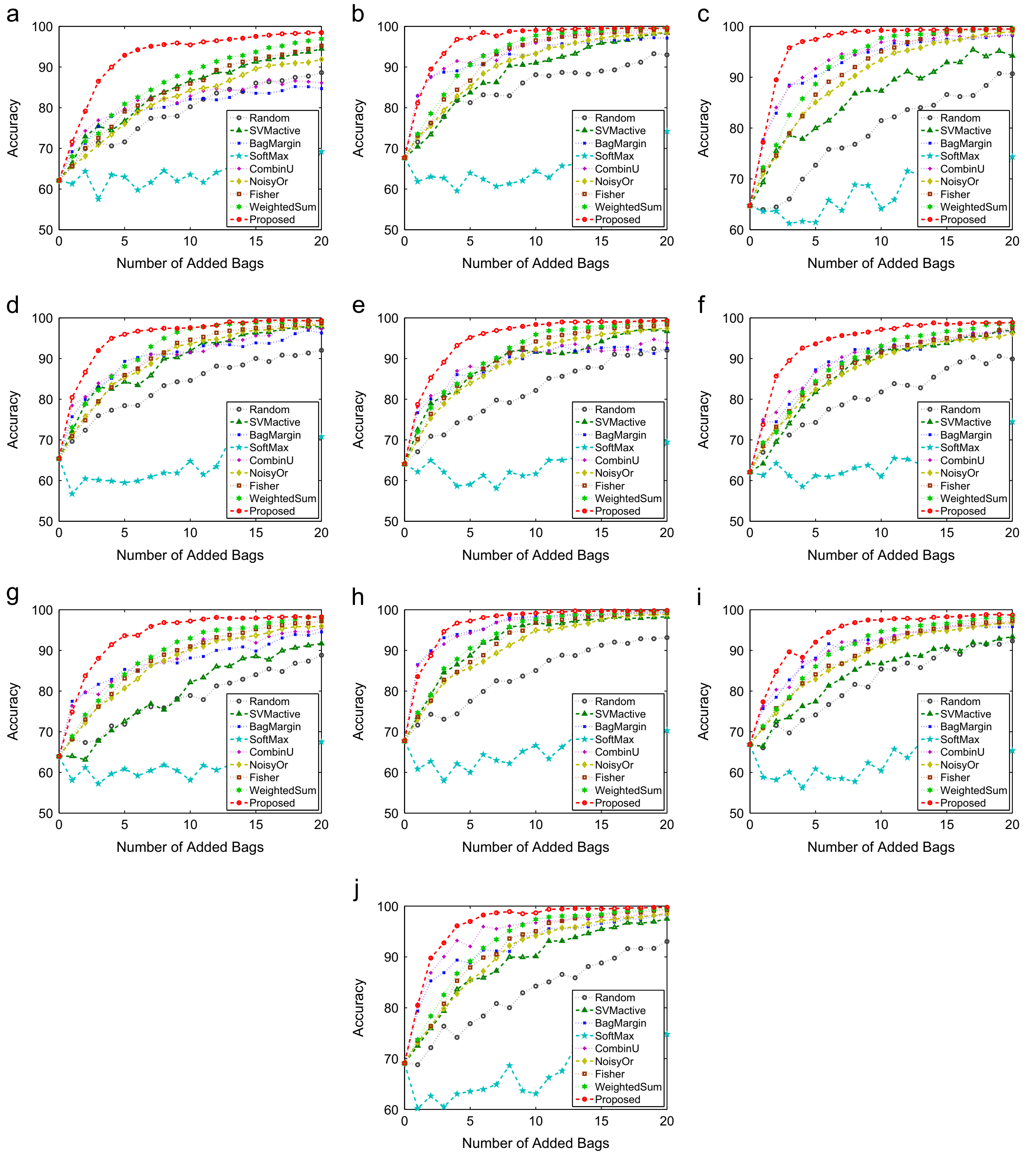

Fig. 5. Bag-query-mode: performance comparison of different learning strategies on recognizing handwritten digit images. (Base-learner: mi-SVM). (a) Digit "0" (50 runs). (b) Digit "1" (50 runs). (c) Digit "2" (50 runs). (d) Digit "3" (50 runs). (e) Digit "4" (50 runs). (f) Digit "5" (50 runs). (g) Digit "6" (50 runs). (h) Digit "7" (50 runs). (i) Digit "8" (50 runs). (j) Digit "9" (50 runs).

baseline random method. This validates the effectiveness of the selective sampling strategies. The proposed method gives the highest accuracy during the entire learning process with a relatively low deviation. Then, the average accuracy and deviation on the fox dataset are given in Fig. 7(c) and (d) respectively. It is found that all the methods exhibit an unstable performance. Generally, the NoisyOr model is the best one and the proposed method is the second best one. Finally, the average accuracy and deviation on the tiger dataset are given in Fig. 7(e) and (f) respectively. Our proposed method gives the highest accuracy. Although the NoisyOr model performs good on fox, its performance on tiger is the worst, which demonstrates a low robustness and stability. In summary, for these three CBIR tasks, the proposed method performs best on the two datasets and second best on the other.

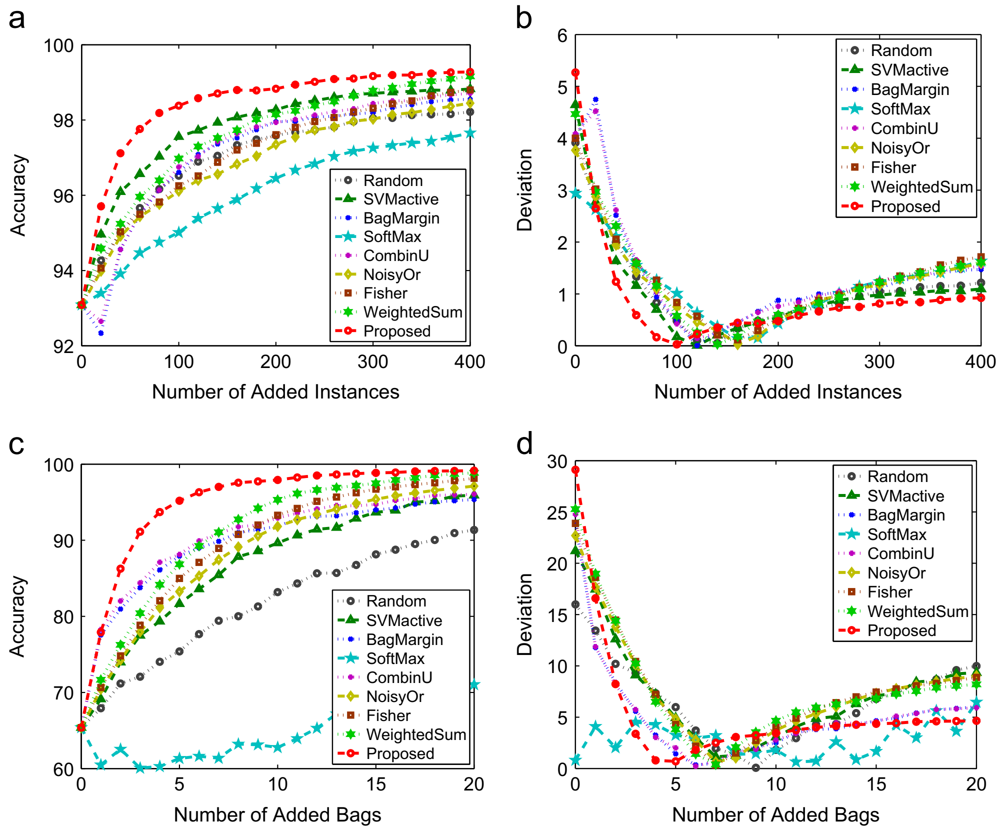

Fig. 6. Averaged performance for recognizing handwritten digit images. (a) Instance-query-mode: accuracy (50 runs). (b) Instance-query-mode: deviation (50 runs). (c) Bag-query-mode: accuracy (50 runs). (d) Bag-query-mode: deviation (50 runs).

Table 5
Instance-query-mode for handwritten digit image recognition: average time cost (seconds) for each active learning iteration.

Datasets	Random	SVMact	BagMargin	SoftMax	CombinU	NoisyOr	Fisher	WeightSum	Proposed
Digit “0”	0.0074	0.0222	0.0249	0.0243	0.0262	0.0186	0.0638	0.0722	0.0833
Digit “1”	0.0046	0.0175	0.0209	0.0197	0.0217	0.0138	0.0487	0.0552	0.0695
Digit “2”	0.0070	0.0237	0.0296	0.0262	0.0314	0.0190	0.0560	0.0608	0.0740
Digit “3”	0.0084	0.0225	0.0268	0.0242	0.0285	0.0192	0.0537	0.0603	0.0714
Digit “4”	0.0077	0.0235	0.0272	0.0250	0.0279	0.0189	0.0538	0.0597	0.0734
Digit “5”	0.0083	0.0231	0.0290	0.0269	0.0296	0.0194	0.0550	0.0632	0.0742
Digit “6”	0.0073	0.0234	0.0301	0.0276	0.0315	0.0179	0.0528	0.0589	0.0743
Digit “7”	0.0066	0.0193	0.0236	0.0214	0.0247	0.0167	0.0518	0.0577	0.0695
Digit “8”	0.0081	0.0262	0.0313	0.0275	0.0319	0.0208	0.0557	0.0617	0.0758
Digit “9”	0.0075	0.0234	0.0281	0.0251	0.0282	0.0191	0.0574	0.0617	0.0732
Avg.	0.0073	0.0225	0.0271	0.0248	0.0282	0.0183	0.0549	0.0611	0.0739

Table 6
Bag-query-mode for handwritten digit image recognition: average time cost (seconds) for each active learning iteration and average number of iterations for training mi-SVM.

Datasets	Random	SVMact	BagMargin	SoftMax	CombinU	NoisyOr	Fisher	WeightSum	Proposed
Digit “0”	0.2286 (5.1)	0.2289 (4.4)	0.3205 (5.6)	0.6243 (8.0)	0.3036 (5.4)	0.1404 (3.1)	0.2390 (3.2)	0.2479 (3.3)	0.2738 (3.5)
Digit “1”	0.1563 (4.0)	0.1398 (2.9)	0.2002 (4.3)	0.5051 (7.7)	0.1895 (4.0)	0.1000 (2.5)	0.1920 (2.5)	0.1992 (2.5)	0.2107 (2.5)
Digit “2”	0.2543 (5.3)	0.2254 (4.2)	0.3380 (5.6)	0.6036 (7.9)	0.3397 (5.7)	0.1534 (3.3)	0.2538 (3.3)	0.2634 (3.4)	0.2938 (3.7)
Digit “3”	0.2503 (5.4)	0.2508 (4.8)	0.3700 (6.2)	0.6215 (8.0)	0.3419 (5.8)	0.1470 (3.2)	0.2482 (3.3)	0.2648 (3.4)	0.3030 (3.9)
Digit “4”	0.2608 (5.5)	0.2536 (4.7)	0.3413 (5.9)	0.6464 (8.1)	0.3470 (6.0)	0.1497 (3.2)	0.2467 (3.3)	0.2600 (3.3)	0.2982 (3.7)
Digit “5”	0.2627 (5.6)	0.2632 (5.1)	0.3472 (6.1)	0.6592 (8.3)	0.3499 (6.1)	0.1455 (3.2)	0.2493 (3.3)	0.2638 (3.4)	0.3022 (3.8)
Digit “6”	0.2291 (5.3)	0.2367 (4.7)	0.2625 (5.2)	0.6063 (8.2)	0.2577 (5.1)	0.1228 (2.9)	0.2197 (3.0)	0.2294 (3.0)	0.2590 (3.3)
Digit “7”	0.2396 (5.3)	0.2218 (4.4)	0.2620 (4.9)	0.6309 (8.3)	0.2749 (5.0)	0.1277 (2.9)	0.2265 (3.0)	0.2398 (3.1)	0.2713 (3.3)
Digit “8”	0.2812 (5.5)	0.2726 (4.9)	0.4014 (5.9)	0.6561 (8.1)	0.4129 (6.0)	0.1818 (3.8)	0.2893 (3.8)	0.2963 (3.9)	0.3495 (4.4)
Digit “9”	0.2690 (5.6)	0.3031 (5.3)	0.3712 (6.1)	0.6399 (8.1)	0.3657 (5.9)	0.1621 (3.5)	0.2560 (3.5)	0.2841 (3.7)	0.3497 (4.5)
Avg.	0.2432 (5.3)	0.2396 (4.5)	0.3214 (5.6)	0.6193 (8.1)	0.3183 (5.5)	0.1431 (3.2)	0.2421 (3.2)	0.2549 (3.3)	0.2911 (3.7)

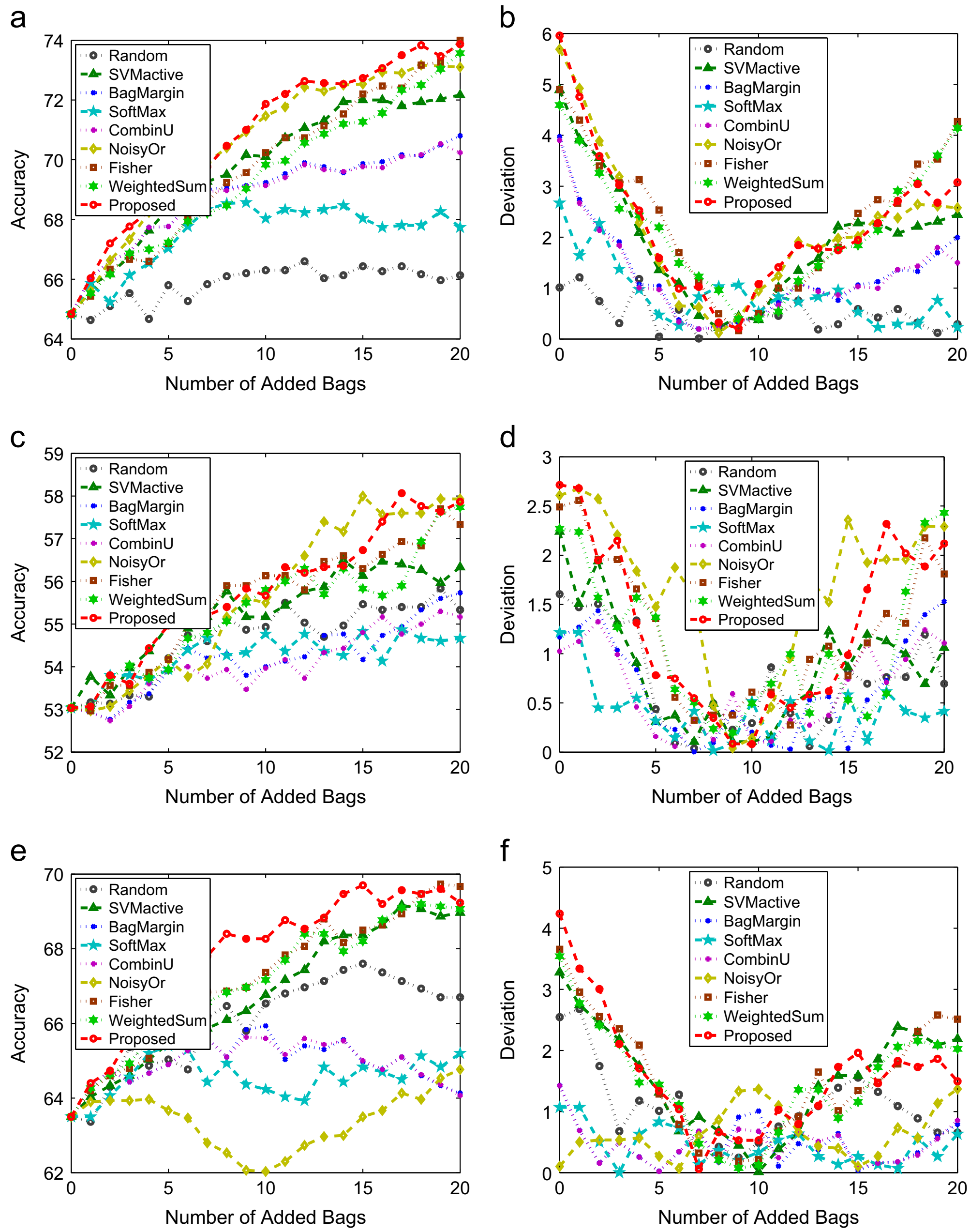

Fig. 7. Bag-query-mode: performance comparison of different learning strategies on content-based image retrieval tasks. (Base-learner: mi-SVM). (a) Dataset-Elephant: accuracy (50 runs). (b) Dataset-Elephant: deviation (50 runs). (c) Dataset-Fox: accuracy (50 runs). (d) Dataset-Fox: deviation (50 runs). (e) Dataset-Tiger: accuracy (50 runs). (f) Dataset-Tiger: deviation (50 runs).

6. Conclusion

In this paper, an active learning model with MCDM system is proposed, which measures the informativeness of unlabeled samples by their dominated and dominating indices. In order to achieve a trade-off between the generalization capability and the learning efficiency, the proposed model is applied to MIL environment. Experimental results on MNIST handwritten digit image MIL datasets and Corel CBIR MIL datasets demonstrate the feasibility and the effectiveness of the proposed method. Compared with many single criterion based strategies and the weighted-sum strategy, the proposed method not only gives higher generalization capability, but it is also highly stable. Meanwhile, the time complexity is in an acceptable range. Our future work regarding this topic may focus on applying it to more real-world problems.

Conflict of interest

None declared.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under the grant 61272289.

Appendix A

The heuristic optimization model of mi-SVM is presented as follows.

Algorithm 5. Pseudo-code for mi-SVM optimization heuristics.

initialize: $y_{ij} = y_i$ for $j \in \mathbf{B}_i$
repeat
compute SVM solution $\mathbf{w}, b$ with imputed labels;
compute $f_{ij} = \mathbf{w} \cdot \mathbf{B}_{ij} + b$ for all $\mathbf{B}_{ij}$ in positive bags;
set $y_{ij} = \text{sign}(f_{ij})$ for every $j \in \mathbf{B}_i$, $y_i = 1$;
for (every positive bag $\mathbf{B}_i$)
if $(\sum_{j \in \mathbf{B}_i} (1 + y_{ij})/2 == 0)$
compute $j^* = \arg\max_{j \in \mathbf{B}_i} f_{ij}$;
set $y_{j^*} = 1$;
end
end
while (imputed labels have changed)
output: $(\mathbf{w}, b)$

References

[1] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, in: Advances in Neural Information Processing Systems, vol. 15, 2002, pp. 561–568.
[2] D. Angluin, Queries and concept learning, Mach. Learn. 2 (4) (1988) 319–342.
[3] R. Battiti, A. Passerini, Brain–computer evolutionary multiobjective optimization: a genetic algorithm adapting to the decision maker, IEEE Trans. Evol. Comput. 14 (5) (2010) 671–687.
[4] J. Bolton, P. Gader, H. Frigui, P. Torrione, Random set framework for multiple instance learning, Inf. Sci. 181 (11) (2011) 2061–2070.
[5] J. Branke, T. Kaußler, H. Schmeck, Guidance in evolutionary multi-objective optimization, Adv. Eng. Softw. 32 (6) (2001) 499–507.
[6] G. Campanella, R.A. Ribeiro, A framework for dynamic multiple-criteria decision making, Decis. Support Syst. 52 (1) (2011) 52–60.
[7] C. Campbell, N. Cristianini, A. Smola, Query learning with large margin classifiers, in: Proceedings of 17th International Conference on Machine Learning, 2000, pp. 111–118.
[8] T.H. Chang, T.C. Wang, Using the fuzzy multi-criteria decision making approach for measuring the possibility of successful knowledge management, Inf. Sci. 179 (4) (2009) 355–370.
[9] Y. Chen, J. Bi, J.Z. Wang, Miles: Multiple-instance learning via embedded instance selection, IEEE Trans. Pattern Anal. Mach. Intell. 28 (12) (2006) 1931–1947.
[10] J.Y. Chiang, S.R. Cheng, Multiple-instance content-based image retrieval employing isometric embedded similarity measure, Pattern Recognit. 42 (1) (2009) 158–166.
[11] D. Cohn, L. Atlas, R. Ladner, Improving generalization with active learning, Mach. Learn. 15 (2) (1994) 201–221.
[12] W.D. Cook, J. Doyle, R. Green, M. Kress, Multiple criteria modelling and ordinal data: evaluation in terms of subsets of criteria, Eur. J. Oper. Res. 98 (3) (1997) 602–609.
[13] W.D. Cook, M. Kress, A multiple criteria decision model with ordinal preference data, Eur. J. Oper. Res. 54 (2) (1991) 191–198.
[14] D. Corne, J. Knowles, M. Oates, The pareto envelope-based selection algorithm for multiobjective optimization, in: Proceedings of 6th International Conference on Parallel Problem Solving from Nature, 2000, pp. 839–848.
[15] T.G. Dietterich, R.H. Lathrop, T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, Artif. Intell. 89 (1–2) (1997) 31–71.
[16] D. Dooly, S. Goldman, S. Kwek, Real-valued multiple-instance learning with queries, J. Comput. Syst. Sci. 72 (1) (2006) 1–15.
[17] Z.P. Fan, Q. Yue, B. Feng, Y. Liu, An approach to group decision-making with uncertain preference ordinals, Comput. Ind. Eng. 58 (1) (2010) 51–57.
[18] J. Figueira, S. Greco, M. Ehrgott, Multiple Criteria Decision Analysis: State of the Art Surveys, Springer Verlag, Berlin Heidelberg New York, 2005.
[19] C.M. Fonseca, P.J. Fleming, et al. Genetic algorithms for multiobjective optimization: formulation, discussion and generalization, in: Proceedings of 5th International Conference on Genetic Algorithms, vol. 1, San Mateo, California, 1993, p. 416.
[20] Y. Freund, H.S. Seung, E. Shamir, N. Tishby, Selective sampling using the query by committee algorithm, Mach. Learn. 28 (2) (1997) 133–168.
[21] J. Fu, J. Yin, Bag-level active multi-instance learning, in: 2011 International Conference on Fuzzy Systems and Knowledge Discovery, vol. 2, IEEE, 2011, pp. 1307–1311.
[22] S. Gao, Q. Sun, Exploiting generalized discriminative multiple instance learning for multimedia semantic concept detection, Pattern Recognit. 41 (10) (2008) 3214–3223.
[23] S. Greco, V. Mousseau, R. Slowinski, Ordinal regression revisited: multiple criteria ranking using a set of additive value functions, Eur. J. Oper. Res. 191 (2) (2008) 416–436.
[24] K.W. Hipel, K.J. Radford, L. Fang, Multiple participant-multiple criteria decision making, IEEE Trans. Syst. Man Cybern. 23 (4) (1993) 1184–1189.
[25] S.C.H. Hoi, R. Jin, M.R. Lyu, Batch mode active learning with applications to text categorization and image retrieval, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1233–1248.
[26] S.C.H. Hoi, R. Jin, J. Zhu, M.R. Lyu, Batch mode active learning and its application to medical image classification, in: Proceedings of 23rd International Conference on Machine Learning ACM, 2006, pp. 417–424.
[27] K. Jabeur, J.M. Martel, S.B. Khélifa, A distance-based collective preorder integrating the relative importance of the group's members, Group Decis. Negot. 13 (4) (2004) 327–349.
[28] A. Kangas, The risk of decision making with incomplete criteria weight information, Can. J. For. Res. 36 (1) (2006) 195–205.
[29] O.E. Kundakcioglu, O. Seref, P.M. Pardalos, Multiple instance learning via margin maximization, Appl. Numer. Math. 60 (4) (2010) 358–369.
[30] R. Lahdelma, K. Miettinen, P. Salminen, Ordinal criteria in stochastic multicriteria acceptability analysis (SMAA), Eur. J. Oper. Res. 147 (1) (2003) 117–127.
[31] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[32] D. Li, J. Peng, Z. Li, Q. Bu, LSA based multi-instance learning algorithm for image retrieval, Signal Process. 91 (8) (2011) 1993–2000.
[33] M. Li, I.K. Sethi, Confidence-based active learning, IEEE Trans. Pattern Anal. Mach. Intell. 28 (8) (2006) 1251–1261.
[34] M. Lindenbaum, S. Markovitch, D. Rusakov, Selective sampling for nearest neighbor classifiers, Mach. Learn. 54 (2) (2004) 125–152.
[35] S. Maji, J. Malik, Fast and Accurate Digit Classification, EECS Department, University of California, Berkeley, Technical Report UCB/EECS-2009-159, 2009.
[36] O. Maron, T. Lozano-Pérez, A framework for multiple-instance learning, in: Advances in Neural Information Processing Systems, 1998, pp. 570–576.
[37] O. Maron, A.L. Ratan, Multiple-instance learning for natural scene classification, in: Proceedings of 15th International Conference on Machine Learning, vol. 15, 1998, pp. 341–349.
[38] K. Meier, Methods for decision making with cardinal numbers and additive aggregation, Fuzzy Sets Syst. 88 (2) (1997) 135–159.
[39] H.T. Nguyen, A. Smeulders, Active learning using pre-clustering, in: Proceedings of 21st International Conference on Machine Learning, 2004.
[40] B. Roy, R. Slowinski, Criterion of distance between technical programming and socio-economic priority, RAIRO. Rech. Opér. 27 (1) (1993) 45–60.
[41] N. Roy, A. McCallum, Toward optimal active learning through sampling estimation of error reduction, in: Proceedings of 18th International Conference on Machine Learning, 2001, pp. 441–448.
[42] K. Sankaranarayanan, J. Davis, Attention-based target localization using multiple instance learning, in: Advances in Visual Computing, 2010, pp. 381–392.
[43] B. Settles, M. Craven, S. Ray, Multiple-instance active learning, in: Advances in Neural Information Processing Systems, Citeseer, 2008.
[44] D. Shen, J. Zhang, J. Su, G. Zhou, C.L. Tan, Multi-criteria-based active learning for named entity recognition, in: Proceedings of 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2004, p. 589.
[45] N. Srinivas, K. Deb, Muiltiobjective optimization using nondominated sorting in genetic algorithms, Evol. Comput. 2 (3) (1994) 221–248.
[46] S. Srinivas, A generalization of the noisy-or model, in: Proceedings of 9th Conference on Uncertainty in Artificial Intelligence, 1993, pp. 208–215.
[47] A. Tavares da Silva, A. Xavier Falcão, L. Pini Magalhães, Active learning paradigms for CBIR systems based on optimum-path forest classification, Pattern Recognit. 44 (12) (2011) 2971–2978.
[48] S. Tong, D. Koller, Support vector machine active learning with applications to text classification, J. Mach. Learn. Res. 2 (2002) 45–66.
[49] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer Verlag, New York, 2000.
[50] J. Wallenius, J.S. Dyer, P.C. Fishburn, R.E. Steuer, S. Zionts, K. Deb, Multiple criteria decision making, multiattribute utility theory: recent accomplishments and what lies ahead, Manag. Sci. 54 (7) (2008) 1336–1349.
[51] D. Wang, J. Li, B. Zhang, Multiple-instance learning via random walk, in: Machine Learning: ECML 2006, 2006, pp. 473–484.
[52] R. Wang, S. Kwong, D. Chen, Inconsistency-based active learning for support vector machines, Pattern Recognit. 45 (10) (2012) 3751–3767.
[53] X. Xu, J.M. Martel, B.F. Lamond, A multiple criteria ranking procedure based on distance between partial preorders, Eur. J. Oper. Res. 133 (1) (2001) 69–80.
[54] A. Zafra, C. Romero, S. Ventura, Multiple instance learning for classifying students in learning management systems, Expert Syst. Appl. 38 (12) (2011) 15020–15031.
[55] A. Zafra, S. Ventura, G3p-mi: a genetic programming algorithm for multiple instance learning, Inf. Sci. 180 (23) (2010) 4496–4513.
[56] M. Zeleny, Multiple Criteria Decision Making, vol. 25, McGraw-Hill, New York, 1982.
[57] D. Zhang, F. Wang, Z. Shi, C. Zhang, Interactive localized content based image retrieval with multiple-instance active learning, Pattern Recognit. 43 (2) (2010) 478–484.
[58] T. Zhang, F.J. Oles, A probability analysis on the value of unlabeled data for classification problems, in: Proceedings of 17th International Conference on Machine Learning, 2000, pp. 1191–1198.

Ran Wang (S'09) received the B.Sc. degree from the College of Information Science and Technology, Beijing Forestry University, China, in 2009, and the Ph.D degree from the Department of Computer Science, City University of Hong Kong, in 2014. She is currently a senior research associate in the Department of Computer Science, City University of Hong Kong. Her current research interests include support vector machines, extreme learning machines, decision tree induction, active learning, multiclass classification, and the related applications of machine learning.

Sam Kwong received his B.Sc. degree and M.Sc. degree in electrical engineering from the State University of New York at Buffalo, USA and University of Waterloo, Canada, in 1983 and 1985 respectively. In 1996, he later obtained his Ph.D. from the University of Hagen, Germany. From 1985 to 1987, he was a diagnostic engineer With the Control Data Canada where he designed the diagnostic software to detect the manufacture faults of the VLSI chips in the Cyber 430 machine. He later joined the Bell Northern Research Canada as a Member of Scientific staff. In 1990, he joined the City University of Hong Kong as a lecturer in the Department of Electronic Engineering. He is currently a Professor in the department of computer Science. His research interests are in Pattern Recognition, Evolutionary Algorithms and Video Coding.