Pattern Recognition 46 (2013) 865–874

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Multiple-instance learning as a classifier combining problem

Yan Li [a,*], David M.J. Tax [a], Robert P.W. Duin [a], Marco Loog [a,b]

[a] Pattern Recognition Laboratory, Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
[b] The Image Group, Department of Computer Science, University of Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 18 January 2012
Received in revised form
20 August 2012
Accepted 22 August 2012
Available online 6 September 2012

Keywords:
Multiple instance learning
Classifier combining

ABSTRACT

In multiple-instance learning (MIL), an object is represented as a bag consisting of a set of feature vectors called instances. In the training set, the labels of bags are given, while the uncertainty comes from the unknown labels of instances in the bags. In this paper, we study MIL with the assumption that instances are drawn from a mixture distribution of the concept and the non-concept, which leads to a convenient way to solve MIL as a classifier combining problem. It is shown that instances can be classified with any standard supervised classifier by re-weighting the classification posteriors. Given the instance labels, the label of a bag can be obtained as a classifier combining problem. An optimal decision rule is derived that determines the threshold on the fraction of instances in a bag that is assigned to the concept class. We provide estimators for the two parameters in the model. The method is tested on a toy data set and various benchmark data sets, and shown to provide results comparable to state-of-the-art MIL methods.

1. Introduction

In multiple-instance learning (MIL), an object is represented as a bag consisting of a set of feature vectors called instances, and the objective is to classify the object or bag into either the positive or the negative class. Typically, it is assumed that there is an underlying concept, and instances from the concept separate positive bags from negative bags [10,25,3]. The classical assumption of MIL, formulated in [10] for the musk odor prediction problem, is that for a positive bag there is at least one instance from the concept and all instances from a negative bag belong to the non-concept. In some papers an instance from the concept is called to be (truly) positive, but to avoid confusion in this paper we do not refer to an instance itself to be negative or positive. An instance can belong to the concept or the non-concept and come from a positive or negative bag.

The term MIL was first proposed in [10] to solve the problem of musk odor prediction. From that on, more and more problems are formulated as MIL ones, such as image categorization [3,8,14], object detection [5,30], text categorization [3,39], hard driver failure prediction [26], spam filtering [18], visual tracking [4,21,36,34], and human action recognition [2]. Take image annotation as an example, an image is represented as a set of patches extracted from it and is classified to a specific class (e.g., elephant) if at least one of the patches contains one object (e.g., elephant) of that class.

Numerous MIL methods have been proposed. Some of them focus only on the bag label without necessarily knowing the label of each instance, while others first work on the instance level and then combine the classifications of instances to obtain the bag label. Examples belonging to the former group include Gärtner et al. [15] and Wang et al. [31] that define kernels between bags, Wang and Zucker [32] that calculates nearest neighbours among bags, Chen et al. [8] and Sørensen et al. [28] that represent a bag with a derived feature vector by some dissimilarity measures, Zhou et al. [39] that defines a graph with instances from a bag, Deselaers and Ferrari [9] that treats bags as nodes and instances as the nodes' states in a conditional random field model, Babenko et al. [5] that models bags as manifolds in the instance space, and Zhang et al. [37] that incorporates structure information between bags or instances. For MIL methods belonging to the second group, some try to directly find the concept area in the space such as the axis-parallel rectangles method [10] and the diverse density method and its variation [25,38], and some introduce a latent variable to represent (or can be used to calculate) whether an instance is from the concept, such as the mi-SVM method [3,17] and the methods in [19,22].

Though based on the classical MIL assumption, many algorithms effectively use, as Ray and Craven [27] calls it, "nearly one" concept instance from each positive bag. The word "nearly one" means that an MIL algorithm uses mainly the information of only one concept instance from each positive bag. Examples are the Diverse Density based on the noise-or model [25] and the MI-SVM relying on the

* Corresponding author. Tel.: +31 15 2788433; fax: +31 15 2781843.
E-mail addresses: yan.li@tudelft.nl (Y. Li), d.m.j.tax@tudelft.nl (D.M.J. Tax), r.duin@ieee.org (R.P.W. Duin), m.loog@tudelft.nl (M. Loog).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.08.018

most positive instance in a bag [3]. For many applications, however, there are in fact more than one instances from the concept in any positive bag and much of the information contained in these instances is lost [27]. This problem is also discussed in [17] where they exploited the distribution of all instances inside the positive bags to improve the MI-SVM method [3].

Alternative modelling assumptions have been adopted for MIL algorithms and applications [13]. For example, a bag is defined to be positive if the number of instances from the concept is larger than a threshold or between an interval [33], or a bag's label is determined by a classifier taking into account all or part of its instances [8,39,28,9], or a bag's label depends on the average posterior probability of all instances in the bag [35]. Wang et al. [31] showed that determining a bag label by its instances can be different in different applications and does not necessarily obey the traditional assumption of MIL.

We propose an MIL solution based on the assumption that *positive bags can be modelled by a mixture of concept and non-concept distributions*. This assumption respects the fact that the number of instances from the concept in a positive bag can be different in different applications. A similar assumption was adopted in [16] (and also in [22]) by introducing in the criterion a new parameter denoting the expected fraction of concept instances in a positive bag. With this assumption the classification of a bag can then be considered as a classifier combining problem [20,23], which combines the classification results of all instances in the bag. A rule called the γ-rule is derived to decide the label of a bag, which compares the fraction of a bag's instances classified to the concept with a particular threshold. Besides, it is also shown that a standard supervised classifier can be adapted for instance classification by weighting the posteriors for positive and negative classes. The estimation of two parameters in the MIL model is analyzed, and estimators are proposed.

The paper is organized as follows. Section 2 formulates the MIL problem with our mixture assumption. The combining rules for the label of a bag are derived in Section 3, while the classification of instances is studied in Section 4. Section 5 introduces estimators for the two parameters used in our MIL model. Our MIL algorithm is summarized in Section 6 and tested in Section 7 with both artificial and real-world data sets. Section 8 discusses several related issues and finally, Section 9 concludes the paper.

2. MIL formulation and the mixture assumption

An object is represented as a bag $B_i=\{\mathbf{B}_{i1},\mathbf{B}_{i2},\ldots,\mathbf{B}_{ij_i}\}$, which contains a set of D-dimensional feature vectors called instances. As in [7], the instances in a bag are assumed to be conditionally statistically independent. That is, given the label of a bag $\omega(B_i)$ (either positive '+' or negative '−'), the instances are drawn independently: $p(\mathbf{B}_{i1},\mathbf{B}_{i2},\ldots,\mathbf{B}_{ij_i}|\omega(B_i))=\prod_{j=1}^{j_i}p(\mathbf{B}_{ij}|\omega(B_i))$. By denoting the probability density function (pdf) $p(\mathbf{x}|+)$ that an instance x is drawn from a positive bag as $f^+(\mathbf{x})$ and that from a negative bag as $f^-(\mathbf{x})$, the independence assumption can be rewritten as

$$p(\mathbf{B}_{i1},\mathbf{B}_{i2},\ldots,\mathbf{B}_{ij_i}|+)=\prod_{j=1}^{J_i}f^+(\mathbf{B}_{ij}),$$
$$p(\mathbf{B}_{i1},\mathbf{B}_{i2},\ldots,\mathbf{B}_{ij_i}|-)=\prod_{j=1}^{J_i}f^-(\mathbf{B}_{ij}). \quad (1)$$

Assume further that there exists a concept $\mathbf{C}$, which defines the difference between the positive and negative bags. The non-concept $\overline{\mathbf{C}}$ denotes the background region which is shared by both positive and negative bags. That is, there are two distinct distributions to generate instances, $f^{\mathbf{C}}(\mathbf{x})$ for the concept $\mathbf{C}$ and

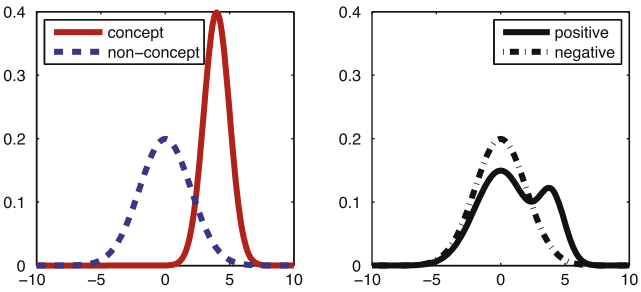

Fig. 1. An illustration of the pdfs in our MIL formulation. On the left are the pdfs for the concept $f^{\mathbf{C}}(\mathbf{x})$ and the non-concept $f^{\overline{\mathbf{C}}}(\mathbf{x})$, the right show the pdfs for the instances in positive bags $f^+(\mathbf{x})$ or in negative bags $f^-(\mathbf{x})$. The α used here is 0.25.

$f^{\overline{\mathbf{C}}}(\mathbf{x})$ for the non-concept $\overline{\mathbf{C}}$. The instances in a negative bag are all drawn from $\overline{\mathbf{C}}$, while the instances in a positive bag are from a mixture of both $\mathbf{C}$ and $\overline{\mathbf{C}}$, with a fraction α of instances sampled from the concept $\mathbf{C}$. That is,

$$f^-(\mathbf{x})=f^{\overline{\mathbf{C}}}(\mathbf{x}),$$
$$f^+(\mathbf{x})=\alpha f^{\mathbf{C}}(\mathbf{x})+(1-\alpha)f^{\overline{\mathbf{C}}}(\mathbf{x}),\quad 0<\alpha<1. \quad (2)$$

An illustration of the assumed distributions is shown in Fig. 1. From (2), the density of the concept $f^{\mathbf{C}}(\mathbf{x})$ can be written as

$$f^{\mathbf{C}}(\mathbf{x})=\frac{f^+(\mathbf{x})-(1-\alpha)f^-(\mathbf{x})}{\alpha}. \quad (3)$$

Denote the prior probabilities for positive and negative bags as $P(+)=\beta$ and $P(-)=1-\beta, 0<\beta<1$, respectively. Then the prior probabilities for the concept $\mathbf{C}$ and the non-concept $\overline{\mathbf{C}}$ are

$$P(\mathbf{C})=\alpha\beta,\quad P(\overline{\mathbf{C}})=1-\alpha\beta. \quad (4)$$

With the mixture assumption (2), there could exist a positive bag without any concept instance. This is especially the case when α is small and the number of instances in a bag is small. To avoid this, we could add a constraint that *a positive bag should have at least one concept instance*, as is the case for similar models such as ALP-SVM [16] and SVR-SVM [22]. We do not include this constraint explicitly in our model, however, which leads to an easy combining rule for bags and the flexibility to use any regular classifier for instances. Moreover, adding this constraint will not affect the derivation of the combining rule as it is automatically satisfied by our combining rule except under extreme situations, though it does provide additional information for instance classifier training. We will return to this issue in Section 8.

3. MIL by combining: the product, sum and γ rules

Following the independence assumption (1), the label of a bag should be combined with the product rule from the classification of instances. To increase the robustness, the product rule can be replaced with the sum rule. In Section 3.2, the combining rules are rewritten based on the mixture assumption (2), based on which we derive the so-called γ-rule in Section 3.3. The γ-rule deals with the situation when instances have been classified to the concept or non-concept. It determines a threshold, and a bag is labelled as positive if its fraction of concept instances is larger than the threshold.

The γ-rule can be used together with other MIL methods which output classification of instances [3,16,17,19,22,25,33,38]. Very different combining rules have been used in those methods, such as the classical presence rule [25,3,16], a learned SVM combiner [22], arithmetic or geometric mean of the posteriors [35], and the so-called threshold- or count-based rule [33].

As the concept is important to understand the MIL problem and it is useful to identify instances from the concept, classification

between the concept and non-concept is needed. This can be achieved in a separate procedure from the bag classification, as explained in Section 4. The estimation of parameters α and β will be investigated in Section 5.

3.1. Classification of a bag B_i: the product rule and the sum rule

With the assumption that the instances in a bag are conditionally statistically independent as defined in (1), the *product rule* should be used to combine the classifications of the instances for the label of a bag [20]. Specifically, the product of the posteriors of instances, weighted by the bag prior, is used to decide the bag label:

$$\text{classify } B_i=\{\mathbf{B}_{i1},\ldots,\mathbf{B}_{iJ_i}\} \text{ as } \textit{positive} \text{ if} P(+|B_i)\geq P(-|B_i)$$

$$\Leftrightarrow \beta^{-(J_i-1)}\prod_{j=1}^{J_i}P(+|\mathbf{B}_{ij})\geq(1-\beta)^{-(J_i-1)}\prod_{j=1}^{J_i}P(-|\mathbf{B}_{ij}), \tag{5}$$

where $P(+|B_i)$ or $P(-|B_i)$ denotes the posterior that the bag B_i is positive or negative, respectively. $P(+|\mathbf{B}_{ij})$ and $P(-|\mathbf{B}_{ij})$ reflect respectively the information that instance $\mathbf{B}_{ij}$ contains to predict its bag to be positive or negative. For example, if $\mathbf{B}_{ij}$ is from the concept, then $P(+|\mathbf{B}_{ij})$ is very likely to be larger than $P(-|\mathbf{B}_{ij})$. Therefore $\mathbf{B}_{ij}$ predicts that its bag is more likely to be positive.

The product rule is very sensitive to estimation errors of the posteriors: it amplifies the error from each instance and is heavily influenced by one individual instance. Experiments show that it performs well only when all the posteriors are well estimated [20,29,1,11]. In the case of large estimation errors, it is preferable to use the more robust sum rule, which can be derived as an approximation to the product rule [20,1]

$$(1-J_i)\beta+\sum_{j=1}^{J_i}P(+|\mathbf{B}_{ij})\geq(1-J_i)(1-\beta)+\sum_{j=1}^{J_i}P(-|\mathbf{B}_{ij}). \tag{6}$$

The posteriors $P(+|\mathbf{B}_{ij})$ or $P(-|\mathbf{B}_{ij})$ can be estimated by assigning the bag label to its instances and applying a traditional supervised classifier to the instances. If the posteriors are estimated well, then the product rule (5) can be used to estimate the label of a bag; otherwise the sum rule (6) should be used. Similar combining rules have been used in [35], where the arithmetic or geometric mean (i.e., sum or product) of the posteriors of all instances is used to determine a bag's label.

3.2. Bag classification with $P(\mathbf{C}|\mathbf{B}_{ij})$ and $P(\overline{\mathbf{C}}|\mathbf{B}_{ij})$

With the mixture assumption (2), the label of a bag can also be determined from the classification of instances into the concept or non-concept. This is based on the relationship between two sets of posteriors of an instance $\mathbf{x}$, $P(\mathbf{C}|\mathbf{x})$ and $P(\overline{\mathbf{C}}|\mathbf{x})$, and $P(+|\mathbf{x})$ and $P(-|\mathbf{x})$. $P(\mathbf{C}|\mathbf{x})$ and $P(\overline{\mathbf{C}}|\mathbf{x})$ denote respectively the probability that $\mathbf{x}$ is from the concept $\mathbf{C}$ or the non-concept $\overline{\mathbf{C}}$; they sum to one.

From Eq. (2), we have

$$P(+|\mathbf{x})=\frac{P(+)p(\mathbf{x}|+)}{p(\mathbf{x})}=P(+)f^{+}(\mathbf{x})\frac{1}{p(\mathbf{x})}$$

$$=(\alpha\beta f^{\mathbf{C}}(\mathbf{x})+(\beta-\alpha\beta)f^{\overline{\mathbf{C}}}(\mathbf{x}))\frac{1}{p(\mathbf{x})},$$

and similarly

$$P(-|\mathbf{x})=(1-\beta)f^{\overline{\mathbf{C}}}(\mathbf{x})\frac{1}{p(\mathbf{x})}.$$

By using the relation between the pdfs $f^{\mathbf{C}}(\mathbf{x})$ and $f^{\overline{\mathbf{C}}}(\mathbf{x})$ and their posteriors $P(\mathbf{C}|\mathbf{x})$ and $P(\overline{\mathbf{C}}|\mathbf{x})$, with straightforward deduction it follows

$$P(+|\mathbf{x})=P(\mathbf{C}|\mathbf{x})+\frac{\beta-\alpha\beta}{1-\alpha\beta}P(\overline{\mathbf{C}}|\mathbf{x}),$$

$$P(-|\mathbf{x})=\frac{1-\beta}{1-\alpha\beta}P(\overline{\mathbf{C}}|\mathbf{x}). \tag{7}$$

Substituting Eq. (7) into the decision rule (5), we have

$$\beta^{-(J_i-1)}\prod_{j=1}^{J_i}\left(P(\mathbf{C}|\mathbf{B}_{ij})+\frac{\beta-\alpha\beta}{1-\alpha\beta}P(\overline{\mathbf{C}}|\mathbf{B}_{ij})\right)$$

$$\geq(1-\beta)^{-(J_i-1)}\prod_{j=1}^{J_i}\frac{1-\beta}{1-\alpha\beta}P(\overline{\mathbf{C}}|\mathbf{B}_{ij}). \tag{8}$$

If the posteriors $P(\mathbf{C}|\mathbf{B}_{ij})$ and $P(\overline{\mathbf{C}}|\mathbf{B}_{ij})$ can be accurately estimated, then from (8) we can obtain an optimal solution (in the sense of Bayes error) for the label of a bag.

Consider the special situation when the concept $\mathbf{C}$ does not have any overlap with the non-concept $\overline{\mathbf{C}}$. For an instance sampled from the concept, we have $f^{\overline{\mathbf{C}}}(\mathbf{B}_{ij})=0$ and $f^{\mathbf{C}}(\mathbf{B}_{ij})>0$, and thereby $P(\overline{\mathbf{C}}|\mathbf{B}_{ij})=0$ and $P(\mathbf{C}|\mathbf{B}_{ij})=1$. Thus for a bag containing such an instance from the concept, the right part of (8) is zero and the bag is labelled as positive with posterior probability equaling to one. This actually follows the classical definition of MIL: if there exists one instance from the concept, then the bag is positive.

Similar to the approximation from (5) to (6), the product rule (8) can also be approximated by the more robust sum rule

$$(1-J_i)\beta+\sum_{j=1}^{J_i}\left(P(\mathbf{C}|\mathbf{B}_{ij})+\frac{\beta-\alpha\beta}{1-\alpha\beta}P(\overline{\mathbf{C}}|\mathbf{B}_{ij})\right)$$

$$\geq(1-J_i)(1-\beta)+\sum_{j=1}^{J_i}\frac{1-\beta}{1-\alpha\beta}P(\overline{\mathbf{C}}|\mathbf{B}_{ij}). \tag{9}$$

3.3. Bag classification with the γ-rule

If the posteriors $P(\mathbf{C}|\mathbf{B}_{ij})$ and $P(\overline{\mathbf{C}}|\mathbf{B}_{ij})$ cannot be estimated properly and we only know that an instance is more probable to be from the concept $\mathbf{C}$ or the non-concept $\overline{\mathbf{C}}$, then we have to decide the label of a bag by counting how many instances are estimated as belonging to the concept. A natural question is what would be the optimal threshold to label a bag as positive. We derive the threshold based on the sum rule (9), since there can be large estimation errors in the posteriors.

Assume that there are J_i instances in the bag, and a fraction γ of them, or γJ_i instances are estimated to be from the concept, then (9) can be rewritten as

$$(1-J_i)\beta+\sum_{\mathbf{B}_{ij}\in\mathbf{C}}\left(\hat{P}(\mathbf{C}|\mathbf{B}_{ij})+\frac{\beta-\alpha\beta}{1-\alpha\beta}\hat{P}(\overline{\mathbf{C}}|\mathbf{B}_{ij})\right)$$

$$+\sum_{\mathbf{B}_{ij}\in\overline{\mathbf{C}}}\left(\hat{P}(\mathbf{C}|\mathbf{B}_{ij})+\frac{\beta-\alpha\beta}{1-\alpha\beta}\hat{P}(\overline{\mathbf{C}}|\mathbf{B}_{ij})\right)$$

$$\geq(1-J_i)(1-\beta)+\sum_{\mathbf{B}_{ij}\in\mathbf{C}}\frac{1-\beta}{1-\alpha\beta}\hat{P}(\overline{\mathbf{C}}|\mathbf{B}_{ij})$$

$$+\sum_{\mathbf{B}_{ij}\in\overline{\mathbf{C}}}\frac{1-\beta}{1-\alpha\beta}\hat{P}(\overline{\mathbf{C}}|\mathbf{B}_{ij}),$$

where $\hat{P}(\mathbf{C}|\mathbf{B}_{ij})$ and $\hat{P}(\overline{\mathbf{C}}|\mathbf{B}_{ij})$ denote estimated posteriors, and there are γJ_i instances in the summation $\sum_{\mathbf{B}_{ij}\in\mathbf{C}}$, and $(1-\gamma)J_i$ instances in the summation $\sum_{\mathbf{B}_{ij}\in\overline{\mathbf{C}}}$.

The estimation of the posteriors is still problematic. For an instance $\mathbf{B}_{ij}$ classified to the concept, we only know that its estimated posterior $\hat{P}(\mathbf{C}|\mathbf{B}_{ij})$ is larger than 0.5. Without more

information, there is no particular reason to select one value over another in the interval (0.5,1]. To simplify, we assume here that the posterior $\hat{P}(\mathbf{C}|\mathbf{B}_{ij})$ equals to one if the instance $\mathbf{B}_{ij}$ is classified to the concept, or zero if to the non-concept. With this assumption the above equation becomes

$$\begin{aligned}&(1-J_i)\beta+\gamma J_i\left(\hat{P}(\mathbf{C}|\mathbf{B}_{ij})+\frac{\beta-\alpha\beta}{1-\alpha\beta}\hat{P}(\overline{\mathbf{C}}|\mathbf{B}_{ij})\right)\Big|_{\substack{\hat{P}(\mathbf{C}|\mathbf{B}_{ij})=1\\ \hat{P}(\overline{\mathbf{C}}|\mathbf{B}_{ij})=0}}\\ &+(1-\gamma)J_i\left(\hat{P}(\mathbf{C}|\mathbf{B}_{ij})+\frac{\beta-\alpha\beta}{1-\alpha\beta}\hat{P}(\overline{\mathbf{C}}|\mathbf{B}_{ij})\right)\Big|_{\substack{\hat{P}(\mathbf{C}|\mathbf{B}_{ij})=0\\ \hat{P}(\overline{\mathbf{C}}|\mathbf{B}_{ij})=1}}\\ &\geq(1-J_i)(1-\beta)+\gamma J_i\left(\frac{1-\beta}{1-\alpha\beta}\hat{P}(\overline{\mathbf{C}}|\mathbf{B}_{ij})\right)\Big|_{\substack{\hat{P}(\mathbf{C}|\mathbf{B}_{ij})=1\\ \hat{P}(\overline{\mathbf{C}}|\mathbf{B}_{ij})=0}}\\ &+(1-\gamma)J_i\left(\frac{1-\beta}{1-\alpha\beta}\hat{P}(\overline{\mathbf{C}}|\mathbf{B}_{ij})\right)\Big|_{\substack{\hat{P}(\mathbf{C}|\mathbf{B}_{ij})=0\\ \hat{P}(\overline{\mathbf{C}}|\mathbf{B}_{ij})=1}}.\end{aligned}$$

After some simple algebra it follows that

$$\gamma\geq\frac{(1-\alpha\beta)(1-2\beta)}{2(1-\beta)J_i}+\alpha\beta. \tag{10}$$

Thus a bag is classified to be positive if γ is larger than the threshold in (10). The decision criterion (10) is called the γ-rule in the paper.

3.3.1. Analysis of the γ-rule

The first term $(1-\alpha\beta)(1-2\beta)/2(1-\beta)J_i$ of the threshold in the γ-rule can have different signs and values. It equals to zero when $\beta=0.5$, and the γ-rule simplifies to

$$\gamma\geq\alpha\beta. \tag{11}$$

Since $0<(1-\alpha\beta)<1$, $-1<(1-2\beta)<1$ and thus $-1<(1-\alpha\beta)(1-2\beta)<1$. Therefore, the numerator is bounded between -1 and 1. So if the denominator $2(1-\beta)J_i$ is sufficiently large, then the first term in the threshold becomes very small and the γ-rule can be approximated also by (11). The term $2(1-\beta)J_i$ becomes large when the number of instances J_i in the bag is large and the prior β is not close to one. Note that the threshold in (11) equals to the prior of the concept: $\alpha\beta=P(\mathbf{C})$.

The threshold in (10) can be a negative number, under the condition

$$\beta\to 1$$

and

$$(1-\beta)J_i\leq\frac{(1-\alpha\beta)(2\beta-1)}{2\alpha\beta}\approx\frac{1-\alpha}{2\alpha}=\frac{1}{2}\left(\frac{1}{\alpha}-1\right). \tag{12}$$

A negative threshold means that the prior $(1-\beta)$ for a negative bag is very small, the mixing coefficient α is not close to one (thus $(1-\alpha)/2\alpha$ is not close to zero), and the number of instances J_i is not large enough to expect a single instance to be from the concept, even though the bag is positive. Consequently, the γ-rule is always satisfied and every bag is classified as positive.

It should be mentioned that condition (12) is rather extreme, and is rarely satisfied in real-world applications. In virtually all practical cases, the threshold in (10) is positive, and no bag without any concept instance will be classified to the positive class. In other words, the γ-rule makes sure that there is at least one concept instance in a positive bag.

3.3.2. Relation with the classical assumption

The γ-rule is equivalent to the criterion with the classical MIL assumption under the following condition:

$$\begin{aligned}&0<J_i\cdot\left(\frac{(1-\alpha\beta)(1-2\beta)}{2(1-\beta)J_i}+\alpha\beta\right)\leq 1\\ &\Leftrightarrow-\frac{(1-\alpha\beta)(1-2\beta)}{2\alpha\beta(1-\beta)}<J_i\leq\frac{1}{\alpha\beta}-\frac{(1-\alpha\beta)(1-2\beta)}{2\alpha\beta(1-\beta)}.\end{aligned}$$

If $\beta=0.5$, then it becomes

$$0<J_i\leq\frac{2}{\alpha}. \tag{13}$$

Equally, if we choose α to be $\alpha\leq 2/\max_i J_i$, then (13) is always satisfied and the γ-rule $\gamma\geq\alpha\beta=0.5\alpha$ is equivalent to detecting if there is at least one instance in the bag to be from the concept.

4. Discrimination between the concept and the non-concept

We study the classification of instances between the concept and the non-concept. For many problems it is useful to identify instances from the concept and to understand the concept $\mathbf{C}$ of the problem. For example, for musk odor prediction to know which specific molecule conformations smell musky [10] and for image annotation to locate the target objects inside an image [3,14]. In addition, the classification results can also be used in combination with the γ-rule to estimate the label of a bag.

From Bayesian decision theory, an instance is classified to the class with the highest posterior. Specifically, an instance $\mathbf{x}$ is classified to the concept if $P(\mathbf{C}|\mathbf{x})\geq P(\overline{\mathbf{C}}|\mathbf{x})$ or $p(\mathbf{x}|\mathbf{C})P(\mathbf{C})\geq p(\mathbf{x}|\overline{\mathbf{C}})P(\overline{\mathbf{C}})$, which is

$$f^{\mathbf{C}}(\mathbf{x})\cdot\alpha\beta\geq f^{\overline{\mathbf{C}}}(\mathbf{x})\cdot(1-\alpha\beta).$$

By substituting Eqs. (3) and (2), it becomes

$$f^{+}(\mathbf{x})\geq\left(\frac{1}{\beta}+1-2\alpha\right)f^{-}(\mathbf{x}). \tag{14}$$

Since $f^{+}(\mathbf{x})=p(\mathbf{x}|+)=p(\mathbf{x})P(+|\mathbf{x})/P(+)=p(\mathbf{x})P(+|\mathbf{x})/\beta$ and $f^{-}(\mathbf{x})=p(\mathbf{x})P(-|\mathbf{x})/(1-\beta)$, the rule (14) can be expressed with posteriors $P(+|\mathbf{x})$ and $P(-|\mathbf{x})$:

$$P(+|\mathbf{x})\geq\frac{1+\beta-2\alpha\beta}{1-\beta}P(-|\mathbf{x}) \tag{15}$$

Note that as $(1+\beta-2\alpha\beta)-(1-\beta)=2\beta(1-\alpha)>0$ and $(1-\beta)>0$, thus it always holds that $(1+\beta-2\alpha\beta)/(1-\beta)>1$.

Recall that the probability $P(+|\mathbf{x})$ or $P(-|\mathbf{x})$ can be estimated by assigning the bag label to its instances and applying a traditional supervised classifier to the instances. By weighting the obtained posteriors according to (15), we can classify an instance to the concept $\mathbf{C}$ or the non-concept $\overline{\mathbf{C}}$. Thus traditional classifiers can be easily adopted for discrimination between $\mathbf{C}$ and $\overline{\mathbf{C}}$.

In principle any classifier in traditional supervised learning can be used. But it should be noted that instances from positive bags are generated from two sources, the concept $\mathbf{C}$ and the non-concept $\overline{\mathbf{C}}$, as defined in Eq. (2). Therefore, if the assumption of a supervised classifier is that each class is sampled from *one* Gaussian, such as linear discriminant analysis, then the classifier can lead to inaccurate estimation and ultimately bad results. Keeping this property in mind, we should select the classifiers which can deal with two or even more sources in one class. Non-parametric techniques such as Parzen windows and k-nearest-neighbor, mixture (of Gaussian) models and kernel methods might be good candidates from this aspect. For example, the mixture model is used in [31] to represent the instances in a bag.

5. Estimation of parameters α and β

Parameter β can relatively easily be estimated by computing the fraction of positive bags among all bags. Estimating α is more complicated. One possible solution is to use the trained instance classifier to validate α. Specifically, an instance classifier is trained with a given α, which is then used to classify instances from positive bags. The fraction of instances classified to the concept can be obtained, which in principle should be equal to the given α. A similar problem of estimating α also occurs in [16].

The α can be estimated directly from the posteriors of instances. A related problem is studied in [12], which can be considered to estimate a parameter $\theta=(1-\beta)/((1-\beta)+\beta(1-\alpha))$ (and thus the inequality $\theta+\beta>1$ always holds). A non-concept instance can be from either a positive or negative bag, and θ is the expected probability that it is from a negative bag. Elkan and Noto [12] proposed to estimate θ as the average posteriors of instances in negative bags B^- from a validation set $\hat{\theta}=\text{mean}_{\mathbf{x}\in B^-}P(-|\mathbf{x})$. Accordingly, α can be estimated as $\hat{\alpha}=(\hat{\theta}+\hat{\beta}-1)/(\hat{\theta}\cdot\hat{\beta})$. This estimator, however, is very sensitive to the estimation errors, as θ and β appear in the denominator.

We propose to use the posteriors of instances in positive bags and the estimated $\hat{\theta}$ for a more robust estimator of α. In positive bags there are concept and non-concept instances, which have different posteriors $P(+|\mathbf{x},\mathbf{x}\in B^+)$. The posteriors of concept instances should be large (close to one) since these instances occur only in positive bags, and the posteriors of non-concept instances have an expectation of $1-\theta$. Thus by its definition, α can be estimated as *the fraction of instances in positive bags whose posteriors are larger than a threshold* τ. In practice, we set τ to the mean of $1-\hat{\theta}$ and the maximum posterior of instances in positive bags $\max_{\mathbf{x}\in B^+}P(+|\mathbf{x})$. This estimator is used in our experiments.

6. MIL algorithm

Our MIL algorithm is summarized in Fig. 2. It collects the training data and estimates necessary parameters and classification posteriors of instances. From the posteriors, bags can directly be classified with the product rule (5) or the sum rule (6), and instances be classified to the concept or non-concept with the decision rule (15). The classification of instances can also be combined to estimate the bag labels with the γ-rule.

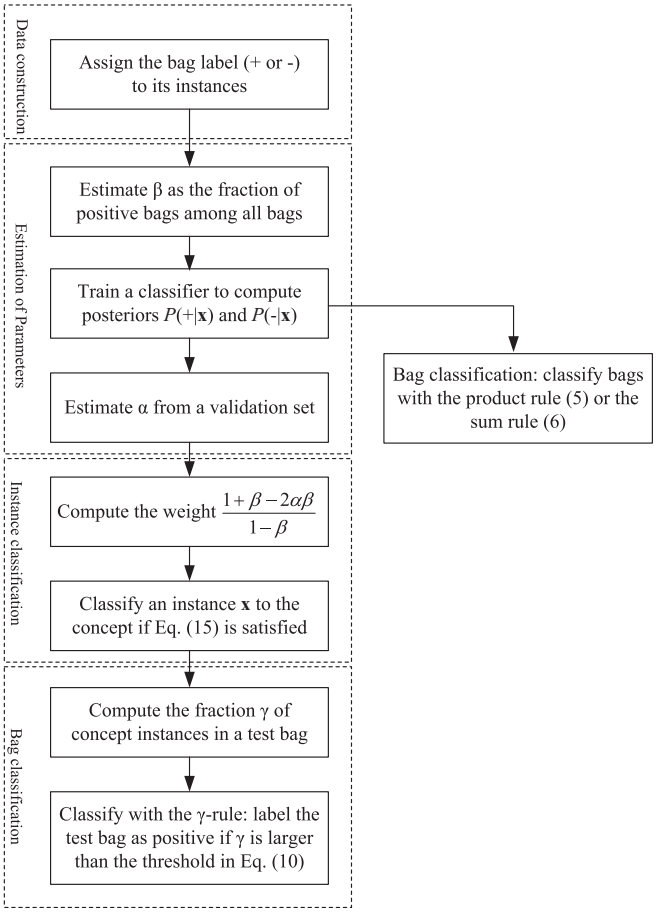

Fig. 2. Flowchart of our MIL algorithm.

The main steps of the algorithm in Fig. 2 are summarized in the following:

Training data construction: For all training bags, assign the bag label to its instances and use all instances for training.

Estimations of parameters: Estimate the parameters α and β.

Instance classification: Classify each instance $\mathbf{B}_{ij}$ according to (15).

Bag classification: Classify the bag B_i according to the γ-rule (10).

7. Experiments

7.1. Toy data: two Gaussians

The data is randomly sampled from two Gaussians with unit covariance matrices, one located at the origin (0,0) and the other at (4,1). For a negative bag, all of its instances are from the Gaussian at the origin, while for a positive bag, each instance is from the Gaussian at (4,1) with probability α and from the other Gaussian with probability $1-\alpha$. When α is very small (e.g., 0.05), a positive bag can have no concept instance as bags are labelled according to their generation schemes. The Gaussian at (4,1) may be viewed as the pdf for the concept. There are 30 positive bags and $30(1-\beta)/\beta$ negative bags, and the number of instances in a bag is a random number between 101 and 110. With $\alpha=0.05$ and $\beta=0.5$, one realization of the data is shown in Fig. 3(a).

The γ-rule is compared with four standard combining rules, the presence rule, vote rule, mean rule, and product rule. The presence rule is based on the classical MIL assumption, which classifies a bag as positive if there is at least one instance from the concept. The vote rule classifies a bag as positive if more than 50% of its instances are classified to the concept. The mean or product rule computes the mean or product of the posteriors of all instances in a bag and assigns the bag to the class with a larger mean or product. The data is generated with β fixed to 0.5, and α changing from 0.05 to 0.95 with a step size of 0.05. For each α, the experiment is repeated for 10 times and their average accuracy and standard deviation of the classification errors of bags are obtained. The k-NN is used for the instance classifier with k set to one-fourth of the number of instances in positive bags. The results are shown in Fig. 3(b). The presence rule works well when α is small. With increasing α, many of the negative bags can be easily misclassified as positive and the accuracy becomes lower. On the other hand, the vote rule, the mean rule and the product rule perform badly when α is small and become good only with larger α. In comparison, the γ-rule can adapt with respect to α and achieves good accuracy for different α's used to generate the data. Note that the fraction of concept instances in a positive bag only has an expectation of value α. When $\alpha=0.05$, the accuracy is low since with such small α, a positive bag can be generated to have very few or even no instance from the concept.

We use the estimator introduced in Section 5 to estimate α. To make the problem less trivial, the 2-D feature vector is increased to 10-D, where each of the added 8-D features has the identical Gaussian distribution. Quadratic discriminant classifier [6] is used, whose regularization parameters are optimized by cross-validation using only the training data set. The estimator is applied to data sets with different α's and β's to generate the data, and its means and standard deviations across 10 times 10-fold cross-validation are shown in Fig. 3(c). In general the estimator performs well, especially when α is large. A small α makes it more difficult to identify concept instances in positive bags, and thus more difficult to estimate α itself. The performance of the estimator also depends on the prior β, since it influences

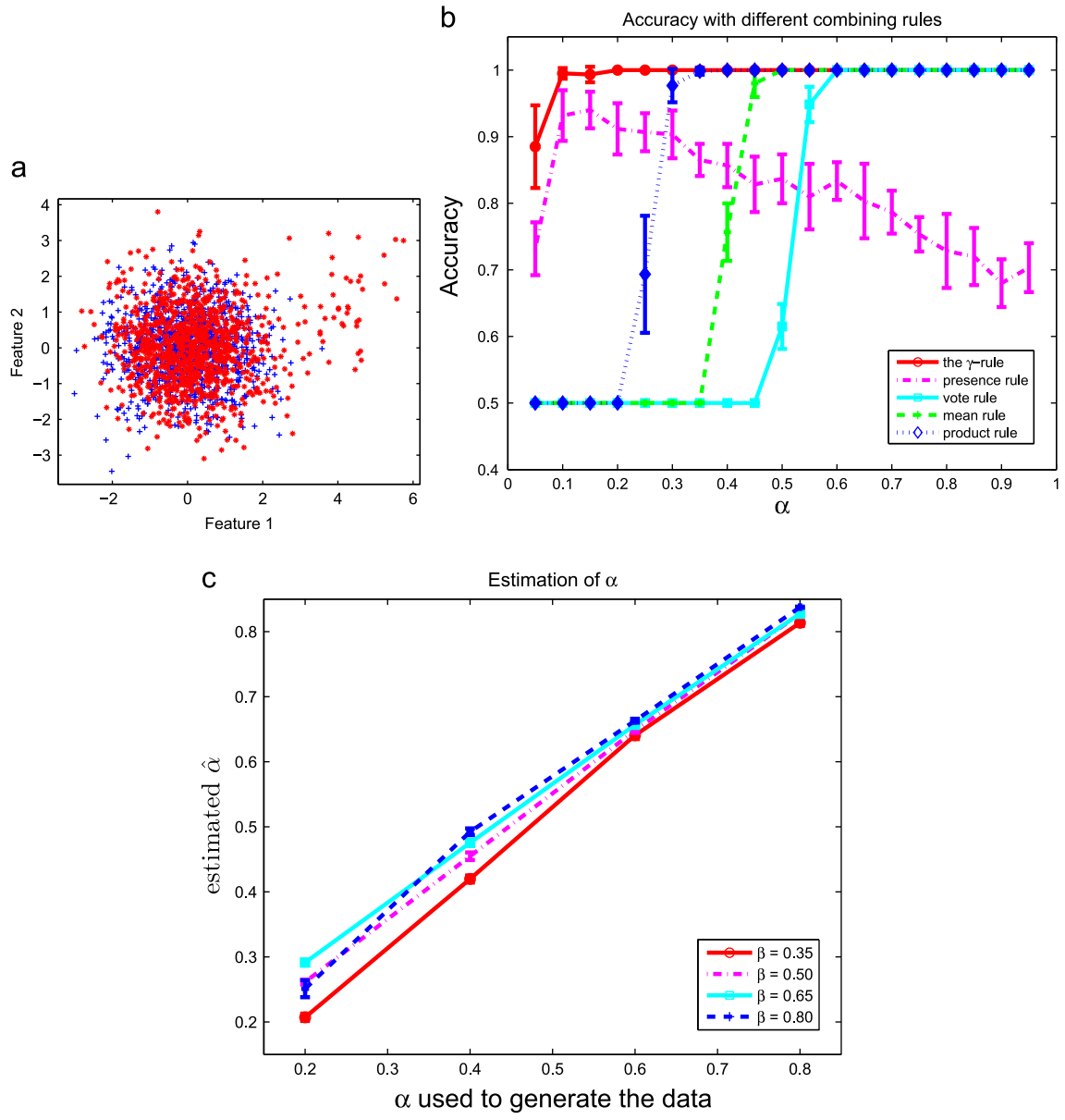

Fig. 3. Toy MIL data set with two Gaussians. (a) Scatter plot of instances in one realization. (b) Comparison of the γ-rule and other combining rules. (c) Estimation of α. The data sets are generated with $\alpha \in \{0.2,0.4,0.6,0.8\}$ and $\beta \in \{0.35,0.50,0.65,0.80\}$.

the estimation accuracy of the parameter θ, which is used in the threshold of the estimator of α.

Trained on data in Fig. 3(a), the decision boundary of the instance classifier defined in (15) is shown in Fig. 4(a), where the cyan area is for the concept and the purple area for the non-concept. Though 95% of the instances from positive bags are actually from the non-concept, the instance classifier successfully separates the concept and the non-concept. In comparison, the decision boundaries for the classifier without the weight in (15) is shown in Fig. 4(b), where much of the non-concept area is misclassified as the concept.

7.2. Five benchmark MIL data sets

We applied our method to five most tested MIL data sets, namely the *MUSK*1 and *MUSK*2 [10], and the *Elephant*, *Tiger* and *Fox* from the COREL data set [3]. *MUSK*1 and *MUSK*2 classify whether a molecule smell "musky". Each molecule is viewed as a bag, whose instances are the different low-energy conformations of the molecule. Surface properties of the conformation were extracted as feature vectors, which have 166 dimensions. *MUSK*1 has 47 positive and 45 negative bags, and 476 instances in total. *MUSK*2 has 39 positive and 63 negative bags, and 6598 instances in total. There are three bags which have around 1000 instances (1010, 1044, and 911, respectively), and the remaining bags have instances ranging from 1 to 383. *MUSK*2 shares 72 molecules with *MUSK*1, but includes more conformations (and thus more instances) for those shared molecules.

Elephant, *Tiger* and *Fox* classify whether an image contains such an animal. For each data set, 100 images containing the target animal were used as positive bags, and 100 images randomly drawn from a set of images of other animals as negative bags. Each image was represented by a set of segments, and each segment was described with a 230-D feature vector characterizing color, texture and shape. The number of instances in a bag ranges from 1 to 13.

The support vector machine (SVM) [6] with a RBF kernel was used as the instance classifier, with the scaling parameter in the

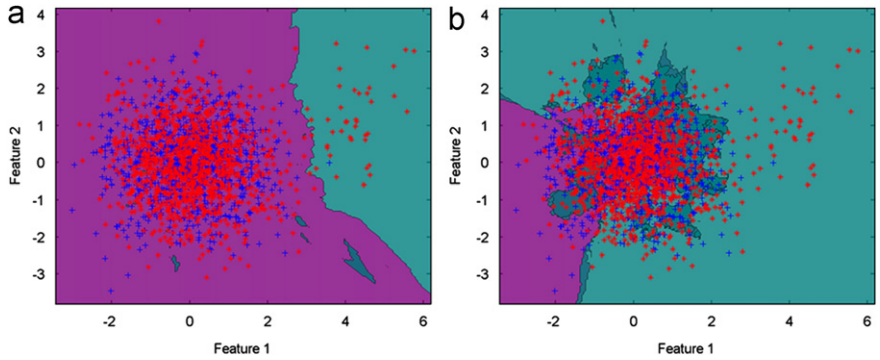

Fig. 4. Decision boundaries of instances between the concept and the non-concept. (a) Classifier defined as (15), and (b) classifier without the weight $(1+\beta-2\alpha\beta)/(1-\beta)$ in (15). The cyan area (the right part of the subimage) is for the concept and the purple area (the left part of the subimage) is for the non-concept. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Table 1
Accuracy on the five benchmark MIL data sets. Previous work is divided into two groups, depending whether or not the instance labels are outputted. For previous methods, the best results in each group are shown in bold. For our method, the result is shown in bold if it is comparable to other state-of-the-art methods.

	*MUSK*1	*MUSK*2	*Elephant*	*Tiger*	*Fox*
Previous methods without an instance classifier					
MI-kernel [15,39]	88.0 (3.1)	89.3 (1.5)	84.3 (1.6)	84.2 (1.0)	60.3 (1.9)
PPPM-kernel [31]	**95.6**	81.2	82.4	80.2	60.3
MIGraph [39]	90.0 (3.8)	90.0 (2.7)	85.1 (2.8)	81.9 (1.5)	61.2 (1.7)
miGraph [39]	88.9 (3.3)	**90.3** (2.6)	**86.8** (0.7)	**86.0** (2.8)	**61.6** (1.6)
Previous methods with instance classifiers					
EM-DD [38,3]	84.8	84.9	78.3	72.1	56.1
mi-SVM [3]	87.4	83.6	82.2	78.4	58.2
MI-SVM [3]	77.9	84.3	81.4	84.0	57.8
MICA [24,16]	84.4	**90.5**	82.5	82.0	62.0
AW-SVM [16]	85.7	83.8	82.0	83.0	63.5
ALP-SVM [16]	86.3	86.2	83.5	**86.0**	**66.0**
Witness rate	1.00	0.28	0.58	0.6	0.71
SVR-SVM [22]	**87.9** (1.7)	85.4 (1.8)	**85.3** (2.8)	79.8 (3.4)	63.0 (3.5)
Witness rate	1.00	0.895	0.378	0.427	1.00
Our method: SVM *with a* RBF *kernel*+					
Product rule	**88.42** (1.83)	83.67 (1.24)	**84.75** (1.11)	78.15 (1.18)	**62.55** (1.32)
Sum rule	**89.22** (0.46)	**85.04** (1.96)	**84.25** (1.06)	79.30 (0.98)	**63.05** (1.19)
The γ-rule	**88.38** (1.08)	**84.92** (2.18)	**84.35** (0.88)	80.75 (1.16)	**62.80** (0.86)
Estimated α	0.82	0.77	0.80	0.51	0.88

kernel as the medium of all pairwise distances between instances. The penalty parameter for *MUSK*1 and *Elephant* was 32, for *MUSK*2 was 64, and for *Tiger* and Fox was 2, which were chosen among a few numbers with a small pilot experiment. For all data sets, the estimated β was 0.5 for all but 0.4 for the MUSK2. The α was estimated using the quadratic discriminant analysis [6] as the classifier.

With the posteriors $P(+|\mathbf{B}_{ij})$ and $P(-|\mathbf{B}_{ij})$ estimated from the instance classifier, bags can directly be classified with the product rule (5) or the sum rule (6). An alternative is to classify instances into the concept and non-concept with (15) and then label a bag with the γ-rule. This is not the most natural way as it needs to estimate another parameter α, but its results can be used to examine how good the parameter estimators and the γ-rule are.

Table 1 shows the means of accuracies of these three combining rules across 10 times 10-fold cross-validation, with the standard deviations in the brackets. The classification results of three combining rules are quite similar, though the sum rule performs slightly better overall. This may due to the robustness of the sum rule to posterior estimation errors, as analysed before. As the method with the γ-rule needs to estimate α and β, its good performance demonstrates the effectiveness of the parameter estimators and the γ-rule itself.

The results of various other methods are also shown in Table 1 for comparison, which were taken from the corresponding papers. For MI-kernel [15] and MICA [24], results recomputed in papers [39,16] respectively were used here as they have better performances. For EM-DD, [38] used the test set for parameter selection, so the results of the correct version [3] were used here. All methods reported the results across 10 times 10-fold cross-validation, except [31,16], of which the number of trials were not explained in the paper. All the methods were divided into two groups, depending on whether they outputted the label of an instance.

Our method with the γ-rule achieves results comparable to other state-of-the-art methods which also provide instance labels. The standard deviation of the mean is used to determine whether or not two results significantly differ. If mean $\pm$ std$/\sqrt{10}$ (10 is the number of trials) overlaps with each other, then one result is not significantly (at significance level 5%) better than the other, and the results are comparable. If only the accuracies are compared, our method with the γ-rule has the highest accuracy for *MUSK*1, outperforms four out of seven for *MUSK*2 and *Fox*, six out of seven for *Elephant*, and three out of seven for *Tiger*, among the methods with instance classifiers. For *Tiger*, our performance is below state-of-the-art; a different choice of classifier could potentially fix this problem.

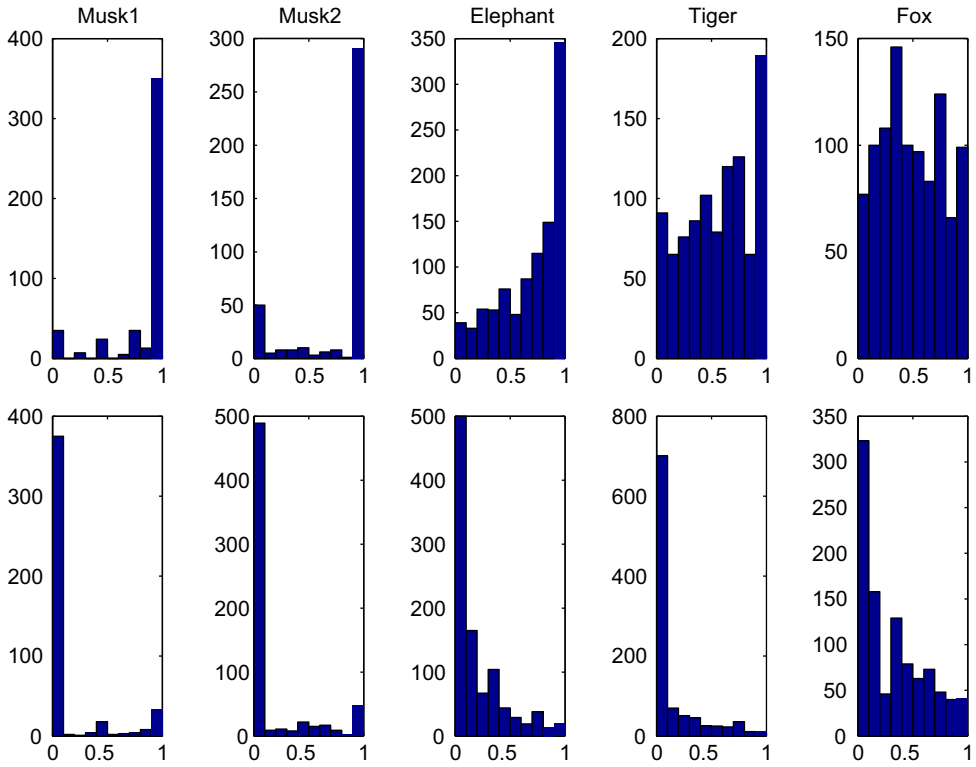

Fig. 5. Histograms of the fraction of concept instances in a bag for the five benchmark MIL data sets. The horizontal axis is the fraction of concept instances in a bag, and the vertical axis is the number of bags. The first row shows histograms for positive bags, and the second row for negative bags.

The results show that our method with the γ-rule performs well when α is relatively large, which is the case for all data sets in Table 1 except *Tiger*. A larger α means that there are more concept instances in a positive bag, which make a positive bag more different from a negative one. As a result, more concept instances can be used to train an instance classifier, and the label of a bag accumulates evidence from more (concept) instances and can be more reliable.

ALP-SVM [16], SVR-SVM [22] and our method have similar MIL assumptions. ALP-SVM has slightly better performance, but its results are based on optimally tuned parameters (witness rates), which are unknown in practical settings. SVR-SVM and our method have similar performances. Our algorithm with the γ-rule, however, is much simpler and can in principle be used to adopt any traditional supervised classifiers for MIL. SVR-SVM needs to solve a constrained support vector regression for instance classification, which optimizes two sets of parameters with alternating minimization, and to train another SVM for bag classification.

For *MUSK*1, *Elephant*, and *Fox*, ALP-SVM, SVR-SVM, and our method outperform EM-DD, mi-SVM, MI-SVM, MICA, and AW-SVM most of the time. The former group of methods use an MIL model similar to our mixture assumption, while the latter group effectively use one concept instance from each positive bag. As also observed in [22], this indicates that effectively use more concept instances from a positive bag is beneficial.

For *MUSK*1 and *MUSK*2, the methods without an instance classifier have better performance than those with an instance classifier in general. This indicates that classifying the bag as a whole is beneficial for these two data sets. The reason may come from the fact that a bag contains various conformations of the same molecule, and all the conformations can contribute to the label of the bag (or molecule). For *Fox*, the methods with instance classifiers can perform better than those without instance classifiers. One possible reason is that for this data set, instances in a bag are very heterogeneous (e.g., a segment that is not related to fox), and classifying between the concept and the non-concept (instead of considering a bag as a whole) is beneficial for the bag classification. For *Elephant* and *Tiger*, there is no significant difference between methods with or without instance classifiers.

To better understand different data sets, histograms of the fraction of concept instances in bags are shown in Fig. 5. In the test phase of classification, the instances were classified between the concept and the non-concept, and the fraction of concept instances in each test bag was computed. If the true label of the test bag was positive, then its concept fraction was added to the first pool, otherwise the fraction was added to the second pool. The test bags were from a 10 times 10-fold cross-validation in our experiments.

The histograms of *MUSK*1, *MUSK*2 differ significantly from those of the other three data sets: high peaks at 1 for positive bags and at 0 for negative bags, and very small values in between. For a positive bag, that the concept fraction equals to one means all the instances are classified to the concept. The high peak at 1 in the histogram means that for a majority of the positive bags, all the instances in a bag are from the concept. Similarly, a majority of the negative bags have all the instances from the non-concept. This may explain why the methods classifying a bag as a whole (without instance classification) can perform very well, and why the estimated witness rates (except *MUSK*2 for ALP-SVM) and α are high. For *Elephant*, we can still see the high peaks at 1 for positive bags and at 0 for negative bags, and the vast majority of the bags are close to those peaks. With the estimates $\alpha = 0.80$ and $\beta = 0.5$, the γ-rule (10) assigns a positive label to a bag if its concept fraction is larger than $0.4\ (= 0 + \alpha\beta)$. Correct labels are assigned for the vast majority of the bags which are close to the two peaks. For *Tiger*, though there are still peaks for the two histograms, the concept fractions for positive bags are much more uniformly distributed. It is possible that for this data set, the concept fractions in the positive bags are indeed very diverse, which leads to worse results of our method. It is also possible that the concept class is not identified very well and

many concept instances are misclassified to the non-concept. In this case, a different choice of classifier may improve the results. For *Fox*, the peak at 1 for positive bags disappears and there are many negative bags with high concept fractions. This demonstrate the difficulty in identifying concept instances or positive bags, which leads to the low accuracy for this data set.

Table 1 also reports our estimated α for each data set, which was used in our algorithm and led to good classification results. The method SVR-SVM [22] outputs the so-called witness rate, which is the fraction of "true positive" instances in all the positive bags. The estimated witness rates and our estimated α are relatively close to each other (except for *Elephant*). Besides, their values are quite large, which indicates that there are usually more than one concept instance in a positive bag and thus justifies our assumption to some extent. The witness rates used in [16] is also included in the table, which are also close to our estimated α except for *MUSK2*.

8. Discussions

In Section 2, we have mentioned that the constraint that *a positive bag should have at least one concept instance* could be added explicitly to our MIL model. The current model, however, leads to an easy combining rule for bags and the flexibility to use any regular classifier for instances. In the following, we discuss, from three aspects, the influence of the constraint when one would include it in the model. Firstly, this constraint provides additional information about positive bags, which could potentially improve instance classification if the assumption of such constraint is truly valid. The amount of additional information depends on the number of positive bags for which no instances are classified to the concept. Secondly, the derivation of the combining rule for a bag's label remains unchanged, but the constraint should be considered in addition to the γ-rule. That is, a bag is labelled positive if it has at least one concept instance and the γ-rule is satisfied. However, if the threshold in the γ-rule (10) is larger than zero, the constraint of having at least one concept instance is automatically satisfied. The threshold of the γ-rule is negative only under extreme situations, specified in Eq. (12) ($\beta \to 1$, small J_i, and $\alpha < 1$). If $\beta \to 1$, classifying every bag as positive is understandable. In real-world applications, the extreme condition (12) is rarely satisfied. For example, for the five data sets in Table 1, their thresholds in the γ-rule are much larger than zero (the minimum is $0.255 = \alpha\beta = 0.51 \times 0.5$ for *Tiger*), and thus no bag without any concept instance can be classified as positive. Thirdly, including that constraint can avoid generating a positive bag which does not have any concept instance. For real-world applications, however, we usually do not label a bag based on the setting it is generated, which is unknown. Instead, quite often we are given a data set labelled by an expert. If we need to label additional bags for training, then the γ-rule will be used and that constraint is satisfied except in extreme situations (12). To summarize, adding that constraint will not affect the derivation of the combining rule and it is automatically satisfied by our γ-rule except under extreme situations, though it does provide additional information to train an instance classifier.

The γ-rule is derived based on a simplified assumption, that is, the estimated posterior $\hat{P}(\mathbf{C}|\mathbf{B}_{ij})$ equals to one if the instance is classified to the concept and zero otherwise. It is worthwhile to investigate the combining rule when the posterior is approximated by another number in the interval (0.5,1]. For example, we may study how does the threshold in (10) change when the estimated posterior varies from 0.5 to 1. Being able to deal with such a more accurate assumption potentially leads to even more powerful combining rules for standard classifiers.

9. Conclusion

We have studied MIL by assuming that positive bags can be modelled as a mixture of concept and non-concept distributions. With this assumption, instances can be classified with any standard supervised classifier, and a bag is labelled by a simple γ-rule. The γ-rule classifies a bag as positive if the fraction of its concept instances is larger than a particular threshold, which is a parameter of the problem. We have proposed estimators for parameters used in instance classification and the γ-rule. The instance and bag classifiers, as well as the parameter estimators, have been tested with artificial and real-world data sets and showed performance which is comparable with state-of-the-art methods.

Acknowledgments

The authors thank all the reviewers for their insightful and constructive comments.

References

[1] L.A. Alexandre, A.C. Campilho, M. Kamel, On combining classifiers using sum and product rules, Pattern Recognition Letters 22 (12) (2001) 1283–1289.
[2] S. Ali, M. Shah, Human action recognition in videos using kinematic features and multiple instance learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2) (2010) 288–303.
[3] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, in: Advances in Neural Information Processing Systems, 2003, pp. 577–584.
[4] B. Babenko, M.H. Yang, S. Belongie, Visual tracking with online multiple instance learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 983–990.
[5] B. Babenko, N. Verma, P. Dollar, S, Belongie, Multiple instance learning with manifold bags, in: Proceedings of the 28nd International Conference on Machine Learning, 2011, pp. 81–88.
[6] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.
[7] A. Blum, A. Kalai, A note on learning from multiple-instance examples, Machine Learning 30 (1) (1998) 23–29.
[8] Y. Chen, J. Bi, J.Z. Wang, MILES: multiple-instance learning via embedded instance selection, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (12) (2006) 1931–1947.
[9] T. Deselaers, V. Ferrari, A Conditional Random Field for Multiple-Instance Learning, in: Proceedings of the 27th International Conference on Machine Learning, 2010, pp. 287–294.
[10] T.G. Dietterich, R.H. Lathrop, T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, Artificial Intelligence 89 (1–2) (1997) 31–71.
[11] R. Duin, D. Tax, Experiments with classifier combining rules, in: Multiple Classifier Systems, 2000, pp. 16–29.
[12] C. Elkan, K. Noto, Learning classifiers from only positive and unlabeled data, in: Proceedings of the 14th ACM Conference on Knowledge Discovery and Data Mining, 2008, pp. 213–220.
[13] J. Foulds, E. Frank, A review of multi-instance learning assumptions, Knowledge Engineering Review 25 (1) (2010) 1–25.
[14] Z. Fu, A. Robles-Kelly, J. Zhou, MILIS: multiple instance learning with instance selection, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (5) (2011) 958–977.
[15] T. Gärtner, P.A. Flach, A. Kowalczyk, A.J. Smola, Multi-instance kernels, in: Proceedings of the 19th International Conference on Machine Learning, 2002, pp. 179–186.
[16] P.V. Gehler, O. Chapelle, Deterministic annealing for multiple-instance learning, in: Proceedings of the 11th International Conference on AISTAT, 2007, pp. 123–130.
[17] Y. Han, Q. Tao, J. Wang, Avoiding false positive in multi-instance learning, in: Advances in Neural Information Processing Systems, 2010, pp. 1–8.
[18] Z. Jorgensen, Y. Zhou, M. Inge, A multiple instance learning strategy for combating good word attacks on spam filters, Journal of Machine Learning Research 9 (2008) 1115–1146.
[19] M. Kim, F. De la Torre, Gaussian processes multiple-instance learning, in: Proceedings of the 27th International Conference on Machine Learning, 2010, pp. 535–542.
[20] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (3) (1998) 226–239.

[21] C. Leistner, A. Saffari, H. Bischof, Miforests: multiple-instance learning with randomized trees, in: European Conference on Computer Vision, 2010, pp. 29–42.
[22] F. Li, C. Sminchisescu, Convex Multiple-Instance Learning by Estimating Likelihood Ratio, in: Advances in Neural Information Processing Systems, 2010, pp. 1–8.
[23] M. Loog, B. Van Ginneken, Static posterior probability fusion for signal detection: applications in the detection of interstitial diseases in chest radiographs, in: 17th International Conference on Pattern Recognition, vol. 1, 2004, pp. 644–647.
[24] O.L. Mangasarian, E.W. Wild, Multiple instance classification via successive linear programming, Journal of Optimization Theory and Applications 137 (3) (2008) 555–568.
[25] O. Maron, T. Lozano-Pérez, A framework for multiple-instance learning, in: Advances in Neural Information Processing Systems, 1998, pp. 570–576.
[26] J.F. Murray, G.F. Hughes, K. Kreutz-Delgado, Machine learning methods for predicting failures in hard drives: a multiple-instance application, Journal of Machine Learning Research 6 (1) (2006) 783.
[27] S. Ray, M. Craven, Supervised versus multiple instance learning: an empirical comparison, in: Proceedings of the 22nd International Conference on Machine Learning, 2005, pp. 697–704.
[28] L. Sørensen, M. Loog, D. Tax, W.J. Lee, M. de Bruijne, R. Duin, Dissimilarity-Based Multiple Instance Learning, in: S+SSPR, Lecture Notes in Computer Science, vol. 6218, 2010, pp. 129–138.
[29] D.M.J. Tax, M. Van Breukelen, R.P.W. Duin, J. Kittler, Combining multiple classifiers by averaging or by multiplying? Pattern recognition 33 (9) (2000) 1475–1485.
[30] P. Viola, J. Platt, C. Zhang, Multiple instance boosting for object detection, in: Advances in Neural Information Processing Systems, vol. 18, 2006, pp. 1417–1426.
[31] H.Y. Wang, Q. Yang, H. Zha, Adaptive p-posterior mixture-model kernels for multiple instance learning, in: Proceedings of the 25th International Conference on Machine Learning, 2008, pp. 1136–1143.
[32] J. Wang, J.D. Zucker, Solving the multiple-instance problem: a lazy learning approach, in: Proceedings of the 17th International Conference on Machine Learning, 2000, pp. 1119–1126.
[33] N. Weidmann, E. Frank, B. Pfahringer, A two-level learning method for generalized multi-instance problems, in: Proceedings of the 14th European Conference on Machine Learning, 2003, pp. 468-479.
[34] Y. Xie, Y. Qu, C. Li, W. Zhang, Online multiple instance gradient feature selection for robust visual tracking, Pattern Recognition Letters 33 (9) (2012) 1075–1082.
[35] X. Xu, E. Frank, Logistic regression and boosting for labeled bags of instances, in: PAKDD, Lecture Notes in Artificial Intelligence, vol. 3056, 2004.
[36] B. Zeisl, C. Leistner, A. Saffari, H. Bischof, On-line semi-supervised multiple-instance boosting, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 1879.
[37] D. Zhang, Y. Liu, L. Si, J. Zhang, R.D. Lawrence, Multiple Instance Learning on Structured Data, in: Advances in Neural Information Processing Systems, vol. 24, 2011, pp. 145–153.
[38] Q. Zhang, S.A. Goldman, EM-DD: an improved multiple-instance learning technique, in: Advances in Neural Information Processing Systems, vol. 2, 2002, pp. 1073–1080.
[39] Z.H. Zhou, Y.Y. Sun, Y.F. Li, Multi-instance learning by treating instances as non-IID samples, in: Proceedings of the 26th International Conference on Machine Learning, 2009, pp. 1249–1256.

Yan Li received his M.Sc. degree in 2007 from Peking University, Beijing, PR China, with the thesis ''Color Filter Arrays: Representation, Analysis and Design''. He is currently a Ph.D. student in the Pattern Recognition Laboratory at Delft University of Technology, Delft, the Netherlands. His current interests include multiple-instance learning, classifier combining, the dissimilarity approach, and multi-scale image analysis.

David M.J. Tax studied physics at the University of Nijmegen, The Netherlands, in 1996, and received Master degree with the thesis ''Learning of structure by Many-take-all Neural Networks''. After that he had his Ph.D. at the Delft University of Technology in the Pattern Recognition group, under the supervision of R.P.W. Duin. In 2001 he promoted with the thesis ''One-class classification''. After working for two years as a Marie Curie Fellow in the Intelligent Data Analysis group in Berlin, at present he is assistant professor in the Pattern Recognition Laboratory at the Delft University of Technology. His main research interest is in the learning and development of detection algorithms and (one-class) classifiers that optimize alternative performance criteria like ordering criteria using the Area under the ROC curve or a Precision–Recall graph. Furthermore, the problems concerning the representation of data, multiple instance learning, simple and elegant classifiers and the fair evaluation of methods have focus.

Robert P.W. Duin received in 1978 the Ph.D. degree in applied physics from Delft University of Technology, Delft, The Netherlands, for a thesis on statistical pattern recognition. He is currently an Associate Professor in the Faculty of Electrical Engineering, Mathematics and Computer Science of the same university.

During 1980–1990, he studied and developed hardware architectures and software configurations for interactive image analysis. After that he became involved with pattern recognition by neural networks. His current research interests are in the design, evaluation, and application of algorithms that learn from examples, which includes neural network classifiers, support vector machines, classifier combining strategies, and one-class classifiers. Especially complexity issues and the learning behavior of trainable systems receive much interest. From 2000 he started to investigate alternative object representations for classification and he thereby became interested in dissimilarity-based pattern recognition, trainable similarities, and the handling of non-Euclidean data.

Dr. Duin is an associated editor of Pattern Recognition Letters and a past-associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence. He is a Fellow of the International Association for Pattern Recognition (IAPR). In August 2006 he was the recipient of the Pierre Devijver Award for his contributions to statistical pattern recognition.

Marco Loog received an M.Sc. in mathematics from Utrecht University, the Netherlands, and in 2004 a Ph.D. degree from the Image Sciences Institute for the development and improvement of contextual statistical pattern recognition methods and their use in the processing and analysis of images. After this joyful event, he moved to Copenhagen, Denmark, where he acted as assistant and, eventually, associate professor next to which he worked as a research scientist at Nordic Bioscience. In 2008, after several splendid years in Denmark, Marco moved to Delft University of Technology, the Netherlands, where he now works as an assistant professor in the Pattern Recognition Laboratory. He currently is vice-chair of Technical Committee 1 of the IAPR, holds three associate editorships, and has been an Invited CNRS Research Scientist in 2008 and 2009. His current interests include multiscale image analysis, semi-supervised and multiple instance learning, saliency, computational perception, the dissimilarity approach, and black math.