The Link between Multiple-Instance Learning and Learning from Only Positive and Unlabelled Examples

Yan Li, David M.J. Tax, Robert P.W. Duin, and Marco Loog

Pattern Recognition Laboratory, Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands
http://prlab.tudelft.nl

Abstract. This paper establishes a link between two supervised learning frameworks, namely multiple-instance learning (MIL) and learning from only positive and unlabelled examples (LOPU). MIL represents an object as a bag of instances. It is studied under the assumption that its instances are drawn from a mixture distribution of the concept and the non-concept. Based on this assumption, the classification of bags can be formulated as a classifier combining problem and the Bayes classifier for instances is shown to be closely related to the classification in LOPU. This relationship provides a possibility to adopt methods from LOPU to MIL or vice versa. In particular, we examine a parameter estimator in LOPU being applied to MIL. Experiments demonstrate the effectiveness of the instance classifier and the parameter estimator.

1 Introduction

Multiple-instance learning (MIL) [1] is a generalised supervised-learning framework that represents an object as a bag consisting of many feature vectors called instances. Only some of the instances in the bag are informative about the label of the object, while others share the same probability distribution for objects from different classes. In the training phase, only the labels of bags (not instances) are known, and a classifier is trained to separate bags into different classes. Many problems can be formulated as MIL problems, such as image annotation [2, 3], text categorization [2, 4], and visual tracking [5].

Learning with only positive and unlabelled examples (LOPU) deals with another limitation of supervised learning [6]. Here one is given a set of examples which are all positive, and another set of examples which include both positive and negative examples. It is the task of the classifier to learn from the unlabelled examples a model of the negative class. For example, to classify users' preference of web pages, an user's bookmarks can be seen as positive examples, while other web pages in the internet include both negative and potentially positive examples. LOPU has been applied to information retrieval [7], document or web page classification [8–10], and biomedical classification [11], among others.

Z.-H. Zhou, F. Roli, and J. Kittler (Eds.): MCS 2013, LNCS 7872, pp. 157–166, 2013.

In MIL, it is typically assumed that there is an underlying concept, whose instances distinguish between positive bags an negative bags [1, 12, 2]. The classical assumption of MIL is that a bag is classified to be positive if *at least one* of its instances is from the concept [1]. Many MIL algorithms (e.g. Diverse Density [12], MI-SVM [2], and the method in [13]), however, usually use the information of only one concept instance from each positive bag [14]. In order to exploit the information from concept instances more effectively, numerous new assumptions have been proposed for MIL [15].

We study MIL based on the assumption proposed by the authors in [16]. This assumption helps to effectively combine the information from all concept instances, which can be formulated as a classifier combining problem. The MIL model and method have been studied in [16]. The focus of this paper is on the link between this MIL model and LOPU. We examine how the instance classifier of this MIL model is related to the classification in LOPU, and how to adapt methods from LOPU to MIL or vice versa. In particular, a parameter estimator from LOPU is modified and applied to MIL.

The paper is organized as follows. Section 2 introduces MIL with our assumption and derives the instance classifier. The relationship between MIL and LOPU is elaborated in Section 3. Section 4 presents experiment results with the derived instance classifier and parameter estimator. Finally, Section 5 concludes the paper.

2 Instance Classifier for MIL

The MIL model with the mixture assumption is introduced, based on which a Bayes classifier is derived for instance classification.

2.1 The Mixture Assumption

Denote an object as a bag $B_i = \{\mathbf{B}_{i1}, \mathbf{B}_{i2}, \cdots, \mathbf{B}_{iJ_i}\}$, which contains J_i feature vectors $\mathbf{B}_{ij}$ of dimensionality D. Assume that instances in a bag are conditionally statistically independent [17]. That is, given the label of a bag, its instances are drawn independently: $p(\mathbf{B}_{i1}, \mathbf{B}_{i2}, \cdots, \mathbf{B}_{iJ_i} | \omega(B_i)) = \prod_{j=1}^{J_i} p(\mathbf{B}_{ij} | \omega(B_i))$, where $\omega(B_i)$ is the label of the bag B_i and can be either positive '+' or negative '−'.

Denote the probability density function (pdf) $p(\mathbf{x}|+)$, from which the instances $\mathbf{x}$ in positive bags are drawn, as $f^+(\mathbf{x})$ and the pdf for negative bags as $f^-(\mathbf{x})$. Assume further that a concept $\mathcal{C}$ exists, which defines the difference between positive and negative bags. The non-concept $\bar{\mathcal{C}}$ denotes the background region which is shared by both positive and negative bags. Formally, there are two distinct distributions to generate instances, one is $f^{\mathcal{C}}(\mathbf{x})$ for the concept $\mathcal{C}$ and the other is $f^{\bar{\mathcal{C}}}(\mathbf{x})$ for the non-concept $\bar{\mathcal{C}}$.

We follow the mixture assumption for MIL in [16]. It assumes that the instances in a negative bag are all drawn from $\bar{\mathcal{C}}$, while the instances in a positive bag are drawn from both $\mathcal{C}$ and $\bar{\mathcal{C}}$,

$$\begin{aligned} f^-(\mathbf{x}) &= f^{\bar{\mathcal{C}}}(\mathbf{x}), \\ f^+(\mathbf{x}) &= \alpha f^{\mathcal{C}}(\mathbf{x}) + (1-\alpha) f^{\bar{\mathcal{C}}}(\mathbf{x}), \quad 0 < \alpha < 1, \end{aligned} \tag{1}$$

where the mixing coefficient α represents the fraction of instances sampled from the concept $\mathcal{C}$ in a positive bag.

2.2 The Instance Classifier

Based on the mixture assumption, the Bayes classifier for instances has been briefly presented in [16]. For completeness, a more detailed derivation is provided, which is followed with an analysis.

Denote the prior probabilities for positive and negative *bags* as β and $1-\beta$ respectively,

$$\begin{aligned} P(+) &= \beta, \\ P(-) &= 1-\beta, 0 < \beta < 1. \end{aligned} \tag{2}$$

Then the prior probabilities for the concept $\mathcal{C}$ and the non-concept $\bar{\mathcal{C}}$ are

$$\begin{aligned} P(\mathcal{C}) &= \alpha\beta, \\ P(\bar{\mathcal{C}}) &= (1-\beta) + (1-\alpha)\beta = 1 - \alpha\beta. \end{aligned} \tag{3}$$

From Bayesian decision theory, we know that an instance should be classified to the class with the largest posterior. That is, for a new test instance $\mathbf{x}$,

$$\begin{aligned} &\text{assign } \mathbf{x} \text{ to the } \textit{concept } \mathcal{C}, \text{ if} \\ &P(\mathcal{C}|\mathbf{x}) \geq P(\bar{\mathcal{C}}|\mathbf{x}) \Longleftrightarrow \\ &p(\mathbf{x}|\mathcal{C})P(\mathcal{C}) \geq p(\mathbf{x}|\bar{\mathcal{C}})P(\bar{\mathcal{C}}) \Longleftrightarrow \\ &f^{\mathcal{C}}(\mathbf{x}) \cdot \alpha\beta \geq f^{\bar{\mathcal{C}}}(\mathbf{x}) \cdot (1-\alpha\beta) \end{aligned} \tag{4}$$

From Eq. (1), the density of the concept $f^{\mathcal{C}}(\mathbf{x})$ can be obtained as

$$f^{\mathcal{C}}(\mathbf{x}) = \frac{f^+(\mathbf{x}) - (1-\alpha) f^-(\mathbf{x})}{\alpha}. \tag{5}$$

Substituting it into (4), the decision rule becomes

$$\begin{aligned} \frac{f^+(\mathbf{x}) - (1-\alpha) f^-(\mathbf{x})}{\alpha} \cdot \alpha\beta \geq f^-(\mathbf{x}) \cdot (1-\alpha\beta) \Longleftrightarrow \\ f^+(\mathbf{x}) \geq \left(\frac{1}{\beta} + 1 - 2\alpha\right) f^-(\mathbf{x}) \end{aligned} \tag{6}$$

Since $f^+(\mathbf{x}) = p(\mathbf{x}|+) = \frac{p(\mathbf{x})P(+|\mathbf{x})}{P(+)} = \frac{P(+|\mathbf{x})}{\beta} p(\mathbf{x})$ and similarly $f^-(\mathbf{x}) = \frac{P(-|\mathbf{x})}{1-\beta} p(\mathbf{x})$, the decision rule (6) can be expressed using the posteriors $P(+|\mathbf{x})$ and $P(-|\mathbf{x})$:

$$\begin{aligned} \frac{P(+|\mathbf{x})}{\beta} p(\mathbf{x}) \geq \left(\frac{1}{\beta} + 1 - 2\alpha\right) \frac{P(-|\mathbf{x})}{1-\beta} p(\mathbf{x}) \Longleftrightarrow \\ P(+|\mathbf{x}) \geq \frac{1+\beta-2\alpha\beta}{1-\beta} P(-|\mathbf{x}) \end{aligned} \tag{7}$$

Note that as $(1+\beta-2\alpha\beta)-(1-\beta)=2\beta(1-\alpha)>0$, $\frac{1+\beta-2\alpha\beta}{1-\beta}>1$ always holds.

Equation (7) provides a way to adapt traditional classifiers for instance classification. Applying a traditional classifier to instances labelled according to their bag label ('+' or '−'), the posteriors $P(+|\mathbf{x})$ or $P(-|\mathbf{x})$ can be obtained. By weighting the obtained posteriors according to (7), the instances can then be classified to the concept $\mathcal{C}$ or the non-concept $\bar{\mathcal{C}}$.

Traditional supervised classifiers have been shown to work well on many MIL problems, despite the fact that they ignore the assumptions underlying MIL [14]. By taking such MIL assumptions into account, our approach can improve the performance of standard supervised classifiers when applied to MIL. In a comparable setting, [14] proposed to use higher cost for false positives in order to improve the classification performance. A higher cost for false positives is analogous to the term $\frac{1+\beta-2\alpha\beta}{1-\beta}$ in (7). The reason is that increasing the cost for false positives is equivalent to increasing the threshold of posteriors to classify an instance to the concept, which has the same meaning as (7).

The k-NN for MIL. To illustrate the decision rule (7), we present an instance classifier using k-NN (k-nearest neighbour) [18], which is also used in our experiments. If k nearest neighbours are found for an instance $\mathbf{x}$, k_+ of them are from positive bags and $k_-(=k-k_+)$ from negative bags, then the estimates of posteriors are

$$\hat{P}(+|\mathbf{x})=\frac{k_+}{k},\ \text{and}\ \hat{P}(-|\mathbf{x})=\frac{k_-}{k}. \tag{8}$$

From (7) the decision rule for the concept becomes

$$\begin{gathered}\text{assign } \mathbf{x} \text{ to the } \textit{concept } \mathcal{C}, \text{ if} \\ k_+ \geq \frac{1+\beta-2\alpha\beta}{1-\beta}\cdot k_-.\end{gathered} \tag{9}$$

3 Relation between MIL and LOPU

The instance classifier (7) is trained by assigning bag labels to their instances. The instances from negative bags are known to be from the non-concept, while those from positive bags are partly from the concept and partly from the non-concept. Consequently, this problem turns out to be very similar to LOPU. In LOPU, all labelled examples are positive, while unlabelled examples may be positive or negative. In terms of LOPU, the classification of instances in MIL can be considered as a learning problem with *only unlabelled and negative* examples and we can use this relation to our benefit.

One of the widely used methods in LOPU is to identify unlabelled examples that are most likely to to be negative with some heuristics, and then train a classifier based on the identified negative examples and the given positive examples [9, 11]. Similar heuristics have been used in Diverse Density to explicitly search for the concept area [12], in MI-SVM to identify the instance in each positive

bag which is mostly likely to be from the concept [2], and in the disambiguation method to identify concept instances in each positive bag [13].

The LOPU method proposed in [11] shares additional similarities with our instance classifier. It is based on the so-called "selected completely at random" assumption [6], which means that any positive example has a constant probability to be chosen by the user and then labelled. Consequently, the unlabelled examples include other positive examples which are not chosen by the user and all the negative examples. The pdf of the unlabelled examples can thus be expressed as a mixture of distributions, which is similar to the pdf $f^+(\mathbf{x})$ in (1). In addition, a central result in that paper is a lemma which relates the classifier trained by *treating all unlabelled examples as negative* and the "ideal" classifier if the labels of all unlabelled examples are provided. This lemma can be derived from the decision rule (7). Besides, they proposed a method to assign weights to unlabelled examples in LOPU, which can also be used to weight instances from positive bags in MIL.

Based on the relationship between LOPU and MIL, a parameter estimator of α has been proposed in [16]. This estimator is based on another estimator proposed in [11], but modified to make it robust in the MIL setting. The basic idea is as follows. In MIL terms, the estimator in [11] is for the parameter

$$\theta = \frac{1-\beta}{(1-\beta)+\beta(1-\alpha)}, \tag{10}$$

which is the probability that a non-concept instance is from a negative bag. It is estimated as the average posteriors of instances in negative bags B^- from a validation set

$$\hat{\theta} = \operatorname*{mean}_{\mathbf{x}\in B^-} P(-|\mathbf{x}). \tag{11}$$

This results in an estimator

$$\hat{\alpha} = \frac{\hat{\theta}+\hat{\beta}-1}{\hat{\theta}\cdot\hat{\beta}}, \tag{12}$$

which is, however, is very sensitive to estimation errors, as θ and β appear in the denominator. A more robust estimator is proposed in [16]. It is based on the assumption that in positive bags, 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 $1-\theta$. Therefore, by definition, α can be estimated as the fraction of instances in positive bags whose posteriors are larger than a threshold τ,

$$\hat{\alpha} = \frac{\#\{\mathbf{x}\,|\,\mathbf{x}\in B^+, P(+|\mathbf{x})>\tau\}}{\#\{\mathbf{x}\,|\,\mathbf{x}\in B^+\}}, \tag{13}$$

where # returns the number of elements in a set and τ takes value as the mean of $1-\hat{\theta}$ and the maximum posterior of instances in positive bags $\max_{\mathbf{x}\in B^+} P(+|\mathbf{x})$.

MIL was linked to semi-supervised learning (SSL) in [19], by viewing MIL as a problem with unlabelled data but positive constraints. The relation with LOPU

makes it clearer that in MIL, there is no labelled positive instances. Besides, the link with SSL is based on the classical MIL assumption, while the mixture assumption is used to relate it to LOPU.

The crucial difference between MIL and LOPU is that the instances in MIL are from bags and are not individual objects. Therefore, MIL has to consider the problem of bag classification in addition to instance classification.

4 Experiments

The derived instance classifier (7) and parameter estimator (13) are tested with a synthetic data and various benchmark MIL datasets. To obtain the label of a bag from its instances, the method derived in [16] is used, which is based on classifier combining. Its basic idea is to compute the fraction of a bag's instances which are classified to the concept and label the bag as positive if the fraction is large than a threshold. The threshold is derived to be $\frac{(1-\alpha\beta)(1-2\beta)}{2(1-\beta)J_i} + \alpha\beta$, which can be approximated by $\alpha\beta$ if the number of instances in a bag J_i is sufficiently large or is $\alpha\beta$ if $\beta = 0.5$. We use k-NN as the instance classifier, with k set to half of the average number of instances in a bag. The corresponding MIL algorithm is as follows:

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

Instance classification. For a test bag $B_i = \{\mathbf{B}_{i1}, \mathbf{B}_{i2}, \cdots, \mathbf{B}_{iJ_i}\}$, classify each instance $\mathbf{B}_{ij}$ according to (9).

Bag classification by Combining. Calculate the fraction of concept instances in the bag B_i, and classify it as positive if the fraction is larger than the threshold $\frac{(1-\alpha\beta)(1-2\beta)}{2(1-\beta)J_i} + \alpha\beta$, and negative otherwise.

4.1 Synthetic Data

We use the synthetic data in [20, 21]. Figure 1(a) shows the ground-truth of the problem, where the black area is the non-concept and the white area is the concept. With $\alpha = 0.3$ and $\beta = 0.5$, one realisation is shown in Figure 1(b).

Trained on the data in Fig. 1(b), the decision boundaries of the instance classifier (9) is shown in Fig. 1(c), where the cyan area is for the concept and the purple area is for the non-concept. We can see that the instance classifier separates the concept and the non-concept very well. In comparison, the decision boundaries for the classifier without the weighting factor $\frac{1+\beta-2\alpha\beta}{1-\beta}$ in (9) is shown in Fig. 1(d), where much non-conept area is misclassified to the concept.

The parameter estimator (13) and the instance classifier (9) are tested with α changing from 0.05 to 0.9. The prior β is fixed to 0.5, and there are 30 positive and 30 negative bags, with 40 instances in each bag. The average results of ten times 10-fold cross-validation are reported. Figure 2(a) shows the estimated α. Overall, the estimator works very well for all different αs, though it seems that

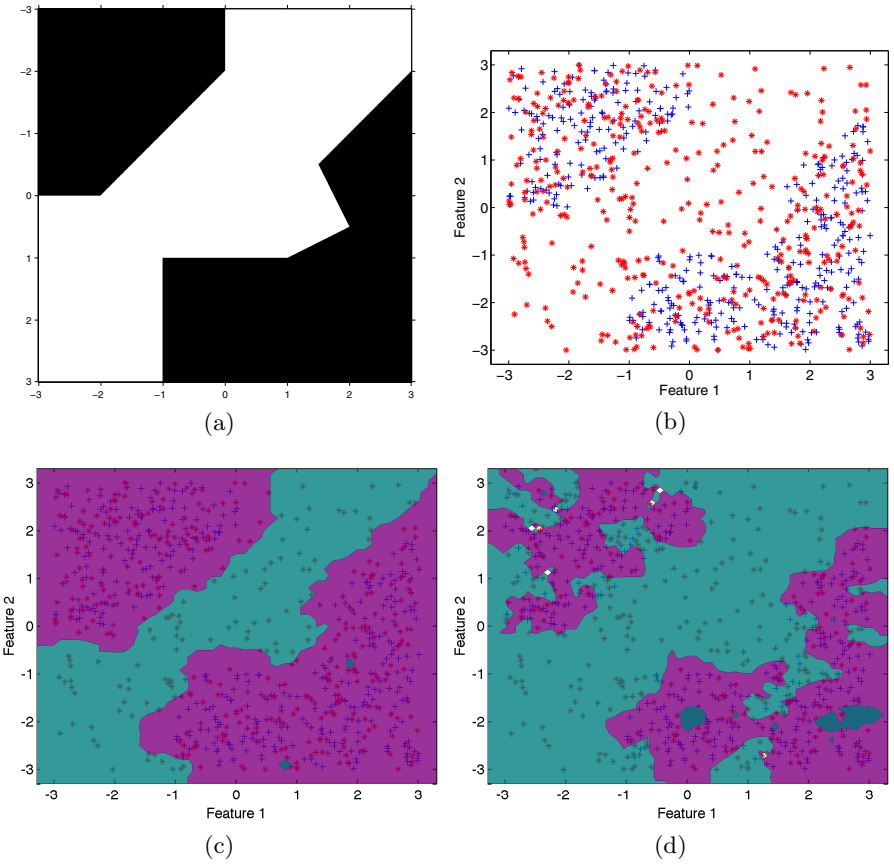

Fig. 1. A synthetic MIL dataset. (a) The ground-truth decision boundary between the concept (white) and the non-concept (black). (b) Scatter plot of all instances, where red instances (*) are from positive bags and blue ones (+) are from negative bags. (c) Decision boundaries of the instance classifier (9). (d) Decision boundary of (9), but without the weighting factor $\frac{1+\beta-2\alpha\beta}{1-\beta}$. The cyan area is for the concept and the purple area (the dark area, if viewed in black and white) is for the non-concept.

there is a small tendency of underestimation (around 0.03). Figure 2(b) shows the classification error of instances with our approach (9) and the traditional k-NN, i.e., (9) without the weighting factor $\frac{1+\beta-2\alpha\beta}{1-\beta}$. The results clearly demonstrate the necessity of this weighting factor, especially when α is small. When α is very large (e.g. 0.8 or 0.9), this weighting factor goes close to one and the difference between the two methods becomes small. Based on our instance classification, the results of bag classification are shown in Figure 2(c). It shows that when α is large than 0.2, perfect classification are obtained. When α is very small (e.g. 0.05), the error is relatively high, as there are only one or two concept instances

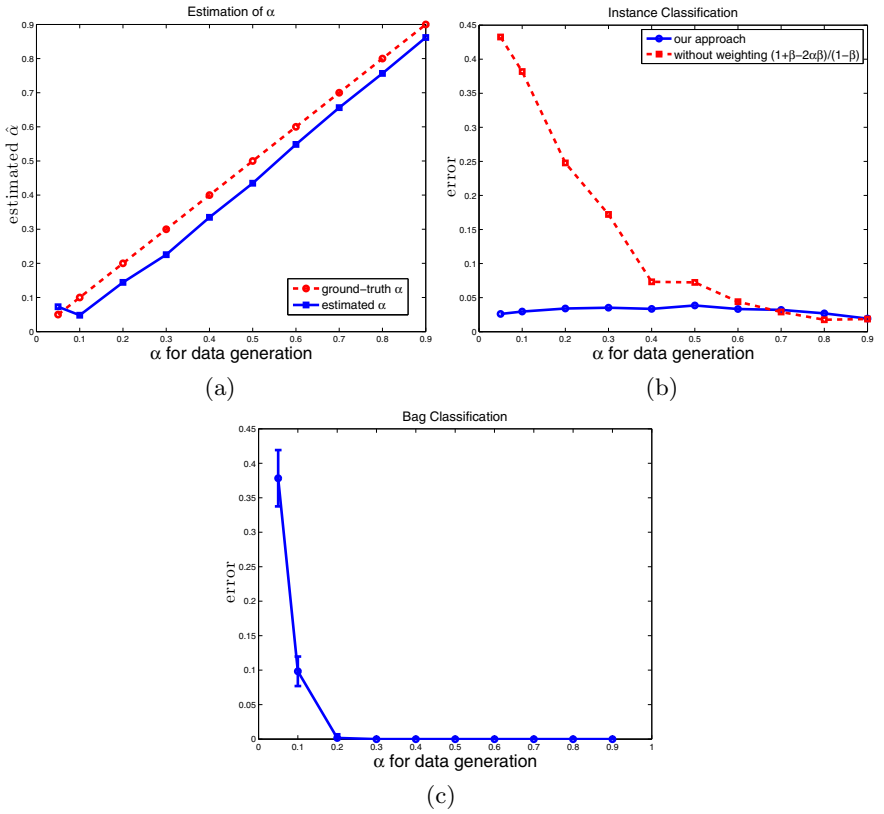

Fig. 2. Results with different αs. (a) The estimated α. (b) The classification of instances. (c) The classification of bags.

in a positive bag. As a result, there are very few concept instances in the training set. Using k-NN with a smaller k, or increasing the number of bags for training, may improve the performance.

4.2 Benchmark MIL Datasets

Our method has been applied to a few mostly tested MIL datasets in [16]. The data sets are the MUSK1 and MUSK2 collected in [1], and the Elephant, Tiger, and Fox from the COREL dataset [2]. A support vector machine with RBF kernel is used as the instance classifier. It has been shown that our method, though relatively simple, achieves results comparable to other state-of-the-art methods.

Table 1 shows the classification results of our method, and another two methods which output estimations of a parameter very similar to α. This parameter is the so-called witness rate, which is the fraction of "true positive" instances in all the positive bags. The witness rate in [21] was automatically estimated, while

Table 1. Accuracy on the five benchmark MIL datasets. The results of our method are comparable to other state-of-the-art methods [16].

	MUSK1	*MUSK2*	*Elephant*	*Tiger*	*Fox*
ALP-SVM [20]	86.3	86.2	83.5	86.0	66.0
witness rate	1.00	0.28	0.58	0.6	0.71
SVR-SVM [21]	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	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

that in [20] was tuned manually. We can see that the estimated witness rates and αs are quite close to each other (except for Elephant). In addition, 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 large values of α may also explain why using traditional supervised classifiers without taking MIL assumptions can work well for some data sets [14]. When α is very large, the weighting factor in the instance classifier (7) becomes close to one, and thus the traditional classifier without this weighting factor can already work relatively well (c.f. Figure 2(b)).

5 Conclusion

Based on the assumption that instances from positive bags follow a mixture distribution, the Bayes classifier is derived for instance classification. A relationship is then established between the classification of instances in MIL and another learning framework called LOPU. It is shown how numerous results in both fields can be linked together. This link also makes it possible to apply methods from MIL to LOPU, or the other way around. In particular, it is studied how to adopt a parameter estimator proposed in LOPU for estimating the parameter α in MIL. The derived instance classifier and the parameter estimator are shown to perform well in the experiments.

References

1. Dietterich, T., Lathrop, R., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 89(1-2), 31–71 (1997)
2. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Adv. Neu. Inf. Proc. Sys., pp. 577–584 (2003)
3. Chen, Y., Bi, J., Wang, J.: MILES: Multiple-instance learning via embedded instance selection. IEEE Trans. PAMI 28(12), 1931–1947 (2006)
4. Zhou, Z., Sun, Y., Li, Y.: Multi-instance learning by treating instances as non-IID samples. In: Proc. 26th ICML, pp. 1249–1256 (2009)
5. Babenko, B., Yang, M., Belongie, S.: Visual tracking with online multiple instance learning. In: IEEE CVPR, pp. 983–990 (2009)

6. Denis, F.: PAC learning from positive statistical queries. In: Richter, M.M., Smith, C.H., Wiehagen, R., Zeugmann, T. (eds.) ALT 1998. LNCS (LNAI), vol. 1501, pp. 112–126. Springer, Heidelberg (1998)
7. Lee, W., Liu, B.: Learning with positive and unlabeled examples using weighted logistic regression. In: Proc. 20th ICML, pp. 448–455 (2003)
8. Liu, B., Dai, Y., Li, X., Lee, W., Yu, P.: Building text classifiers using positive and unlabeled examples. In: Proc. Int'l Conf. Data Mining, pp. 179–188 (2003)
9. Yu, H., Han, J., Chang, K.: PEBL: Web page classification without negative examples. IEEE Trans. Know. and Data Eng. 16(1), 70–81 (2004)
10. Zhou, K., Xue, G., Yang, Q., Yu, Y.: Learning with Positive and Unlabeled Examples Using Topic-Sensitive PLSA. IEEE Trans. on Knowledge and Data Engineering 22(1), 46–58 (2010)
11. Elkan, C., Noto, K.: Learning classifiers from only positive and unlabeled data. In: Proc. 14th ACM Conf. Knowledge Discovery and Data Mining, pp. 213–220 (2008)
12. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. In: Adv. Neu. Inf. Proc. Sys., pp. 570–576 (1998)
13. Li, W.J., Yeung, D.Y.: MILD: Multiple-instance learning via disambiguation. IEEE Transactions on Knowledge and Data Engineering 22(1), 76–89 (2010)
14. Ray, S., Craven, M.: Supervised versus multiple instance learning: An empirical comparison. In: Proc. 22nd Int'l Conf. Mach. Learn., pp. 697–704 (2005)
15. Foulds, J., Frank, E.: A review of multi-instance learning assumptions. The Knowledge Engineering Review 25(01), 1–25 (2010)
16. Li, Y., Tax, D., Duin, R., Loog, M.: Multiple-instance learning as a classifier combining problem. Pattern Recognition 46(3), 865–874 (2013)
17. Blum, A., Kalai, A.: A note on learning from multiple-instance examples. Machine Learning 30(1), 23–29 (1998)
18. Bishop, C.: Pattern Recognition and Machine Learning. Springer, New York (2006)
19. Zhou, Z., Xu, J.: On the relation between multi-instance learning and semi-supervised learning. In: Proc. 24th ICML, pp. 1167–1174 (2007)
20. Gehler, P., Chapelle, O.: Deterministic annealing for multiple-instance learning. In: Proc. 11th Int'l Conf. AISTAT, pp. 123–130 (2007)
21. Li, F., Sminchisescu, C.: Convex Multiple-Instance Learning by Estimating Likelihood Ratio. In: Adv. Neu. Inf. Proc. Sys., pp. 1–8 (2010)