Accepted in Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), 2016

# Thresholding for Accurate Instance Classification in Multiple Instance Learning

Marc-André Carbonneau, Eric Granger and Ghyslain Gagnon
École de technologie supérieure, Université du Québec, Montréal, Canada
e-mail: marcandre.carbonneau@gmail.com, eric.granger@etsmtl.ca,ghyslain.gagnon@etsmtl.ca

*Abstract*— **Multiple instance learning (MIL) is a form of weakly supervised learning for problems in which training instances are arranged into bags, and a label is provided for whole bags but not for individual instances. Most proposed MIL algorithms focused on bag classification, but more recently, the classification of individual instance has attracted the attention of the pattern recognition community. While these two tasks are similar, there are important differences in the consequences of instance misclassification. In this paper, the scoring function learned by MIL classifier for the bag classification task is exploited for instance classification by adjusting the decision threshold. A new criterion for the threshold adjustment is proposed and validated using 7 reference MIL algorithms on 3 real-world data sets from different application domains. Experiments show a considerable improvement in accuracy of these algorithms for instance classification. In some application, the unweighted average recall increases by as much as 18%, while the $\mathbf{F_1}$-score increases by 12%.**

*Keywords*— **Multiple instance learning, Weakly-supervised learning, Instance classification, set classification.**

## I. Introduction

In multiple instance learning problems, instances are grouped into sets called bags. A label is provided for bags, but not for individual instances. The so-called standard MIL assumption [1] states that if a bag contains at least one positive instance, it is labeled as positive. Therefore, positive bags can contain a mixture of negative and positive instances, while negative bags contain only negative instances. Problems from many application domains can be formulated as MIL. In the past, it has been used for molecule activity prediction [2], image classification [3], computer aided diagnosis [4], visual object tacking [5] and document classification [6]. MIL research traditionally focuses on bag classification, however, more recently, several authors considered problems in which instance must be classified individually [4], [7], [8].

Typically, when MIL is applied to computer vision problems, images (or video) are divided in segments or patches. These segments correspond to instances, which are grouped in a bag representing the whole image. In this regards, MIL encompasses bag-of-words methods [1]. For content-based image retrieval (CBIR) tasks, a label has to be attributed to bags and the exact label of the instances is not important. However, for image annotation tasks, such as object localization and tracking [5], the instances must be classified individually [9]. This task is of significant importance, especially for computer-aided diagnosis, where region of images are annotated as healthy or not. In this context, when using traditional supervised algorithm, the training data requires fine grained expert annotation which is costly [4]. With MIL, entire image can be used for learning and the patient diagnosis serves as weak supervision. This enables the use of a enormous quantity of training data otherwise unexploited.

It has been shown that the performance of MIL algorithms for bag classification are not representative of the performance for instance classification [10]. This is due to a combination of factors such as working assumptions on instance labels, the use of bag classification accuracy as optimization objective, and the data properties such as the witness rate (WR). Also, it can be shown experimentally that some algorithms perform well in terms of area under the ROC curve (AUC) but provide low accuracy [11]. This suggests that some algorithms learn to score the instances correctly, but learn a sub-optimal decision thresholds to predict to instance or bag labels.

In this paper, the optimal decision threshold for bag classification is shown to be different from the optimal threshold for instance classification. Also, the threshold obtained by training the MIL algorithms is experimentally shown to be sub-optimal for the instance classification task. Finally, a criterion for the selection of the decision threshold is proposed to increase instance classification accuracy performance. The proposed criterion leverages the MIL assumption that labels of instances in negative bags are fully known. These instances are considered individually instead of in bags in the criterion, which modifies the cost of misclassification, and thus, allows for a higher accuracy at instance level. The proposed criterion is used to adjust the decision threshold of 7 well-known reference MIL algorithms. Experiments are conducted on real-world data from 3 application domains.

The remainder of this paper is organized as follows: the next section surveys MIL algorithms applicable to instance classification problems. Section III shows how optimal threshold for instance and bag classification are different, and introduce the proposed criterion for threshold adjustment. Finally, Section IV presents the experimental methodology and the results are analyzed in Section V.

## II. Instance Classification in MIL

Many MIL methods originally proposed for bag classification, can be used directly for instance classification. These methods typically classify instances individually and then, under the standard MIL assumption, check for the presence of positive instance in bags. If a bag contains positive instance, it is labeled as positive, otherwise, it is labeled as negative. This is the case for APR [2], MI-SVM and mi-SVM [12], RSIS [11] and many diverse density based methods [13], [14]. When classifying bags with these methods, some type of instance classification error have no impact. For instance, in a positive bags, as long as at least one positive instance has been identified, false negatives and false positives have no effect on the bag label. This means that all but one positive instance per positive bags can be mislabeled, and yet, perfect bag accuracy can still be achieved. This is exploited directly by MI-SVM which select only the most positive instance per positive bag to train the SVM. Other methods, like MILBoost [15], which was proposed for instance classification, and EM-DD [14] use bag classification accuracy during their optimization process. This is a reasonable strategy in bag classification tasks but can be sub-optimal for instance classification tasks.

Many MIL methods do not attempt at classifying all instances individually, but instead, consider entire bags as single objects. Some of these methods use kernels or set distance metrics to compare entire bags [6], [16]–[18], while other methods, embed bags in a single vector representation (e.g. using distances to prototypes [3]). Since these methods do not attempt to discover the label of individual instances, they generally cannot be applied to instance classification problems. There are however some bag-level methods that can be used for instance classification. For instance, MILES [3] represents bag as a set of distances from selected instances. The authors proposed to use the contribution of each instance to the bag label as a witness identification mechanism. Some methods are adaptation of bag-level methods for instance classification. For instance, Citation-kNN-ROI [7] classifies bags using the minimal Hausdorff distance and the reference and citations scheme of Citation-kNN [18]. Once a bag is deemed positive, each instance it contains is treated as a bag, and is classified individually. All of these methods were proposed to classify bags and thus, have a consequent optimization objective and working assumption, which limits their accuracy on the instance classification task.

## III. Threshold for Instance Classification

This section describes why decision thresholds learned by MIL algorithms are often sub-optimal for instance classification. Then, a new threshold selection criterion is proposed to increase the instance-level accuracy by making better use of the weak supervision available in MIL problems.

### A. Decision Thresholds: Bags vs. Instances

Following the standard MIL assumption, the label of instances from negative bags are known without ambiguity although the labels of the instances in positive bags are unknown. Instance-based MIL methods infer the label of instances in order to predict bag labels. To assign a hard label to an instance or a bag, a decision threshold is applied to a score. The optimal threshold for instance classification is often different than for bag classification. There are several reasons why this is the case.

Firstly, in many MIL problems, the proportion of positive instances in positive bags is low. In images, for instance, in images, most of the regions do not correspond to the object of interest and thus the positive bags exhibit low WR [19]. This affect many MIL algorithms, like SI-SVM, EMDD, APR and Citation-kNN, that assume all the instances in positive bags to be positive. Also many MIL algorithm implicitly assume that the instances are independent and identically distributed (i.i.d.) in bags. However, this is rarely the case in practice. In many application, there is some correlation between the positive and negative instances of the same bag [6]. For instance in image classification, a tiger is most likely to be found in the jungle than in a spaceship. While instance corresponding to the jungle are as negative as instance from spaceship, the jungle instances are correlated with tiger instances. Moreover the different segments of the same image share some similarities because of capture conditions. All the segments of an image with low illumination will be darker. In the drug activity prediction problem [2], each bag contains many conformations of the same molecule. Only some of these conformations produce an effect of interest, but since bags are conformations of the same molecule, it is likely that the instances are similar to some extent. Finally, as stated in Section II, many MIL algorithms, like MI-SVM and MIL-Boost, use the bag-level classification accuracy as an optimization criterion which is often sub-optimal for instance classification.

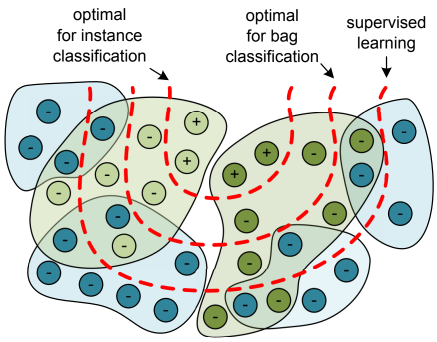
optimal for instance classification
optimal for bag classification
supervised learning
+
-

Fig. 1. Illustrative example of how different optimization objectives yield different threshold value in non i.i.d. instances and low witness rate MIL data sets.

Fig. 1 illustrates how low WR, correlation of instances in bags and optimization on bag-level accuracy can cause MIL to learn a sub-optimal threshold for instance-level classification. In this example, positive bags are represented by green regions and the negative bags by blue regions. The instances in each bag are grouped together (correlated), and there only a small number of positive instances in both positive bags. The dotted red lines are iso-contour of the score function learned by the classifier. In this illustrative example, there is a value for the decision threshold that can achieve a perfect classification of the instances, and thus, the bags. However, MIL algorithms optimizing bag-level accuracy will learn a

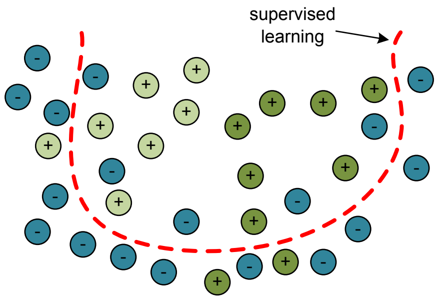

Fig. 2. The problem of Fig. 1 as seen by regular supervised algorithms: All instances inherited the label of their respective bag.

different decision threshold, which also achieves a perfect bag classification. It only requires the exclusion of all instances belonging to negative bags from the positive region. This produces false positives (FPs) in positive bags, which have no consequence when performing bag classification. However, in instance classification problems these FPs hinder performance. Finally, Fig. 1 also shows the decision threshold that would be learned by a supervised algorithm like SI-SVM. In that case, all instances in positive bags are considered positive and the problem reverts to a regular supervised problem as illustrated in Fig. 2. This shows why supervised algorithm are not suitable for instance classification in problems with low WR and non i.i.d. instances.

### *B. Proposed Strategy for Threshold Adjustment*

The proposed procedure can be used to increase the performance of existing MIL in the context of instance classification. It is applied after an algorithm has undergone its usual process. The decision threshold is then updated to maximize the proposed criterion. Following the standard MIL assumption, two sources of information are reliable: the bag labels and the labels of the instances in negative bags. Both these sources are considered in the criterion, instead of only bag labels like in most existing MIL methods.

Let $\mathcal{B}^+ = \{B_1, ..., B_{N^+}\}$ and $\mathcal{B}^- = \{B_1, ..., B_{N^-}\}$ be sets containing all positive and negative bags respectively. Each bag $B_i = \{\mathbf{x}_i^1, ..., \mathbf{x}_i^{M_i}\}$ is a set of instances. Finally, $\mathcal{I}^-$ is a set containing all instances of all negative bags $\mathcal{B}^-$. The threshold $\beta$ is obtained by maximizing:

$$\beta = \underset{\beta}{\text{argmax}} \, \{A_i(\beta) + A_b(\beta)\}. \tag{1}$$

$A_b(\beta)$ is the instance level accuracy on all instances contained in negative bags:

$$A_i = \frac{TN^{\mathcal{I}^-}(\beta)}{|\mathcal{I}^-|} \tag{2}$$

where $TN^{\mathcal{I}^-}$ is the number of correctly classified instances (true negatives). This diminishes the impact of the misclassification of a single instance in a negative bag, which leads to a higher level of accuracy at instance level. For example, if 1% of the instances in each negative bag are misclassified, then all these bags are misclassified, while 99% of the instance are correctly classified. The accuracy on the positive class must also be enforced. Since in MIL the instance labels in positive bags are unknown, the positive bag accuracy is used instead as the second term of the objective function:

$$A_b = \frac{TP^{\mathcal{B}^+}(\beta)}{|\mathcal{B}^+|} \tag{3}$$

where $TP^{\mathcal{B}^+}$ is the number of correctly classified positive bags. By considering instance from negative bags individually the criterion reduces the penalty for FPs, which allows the identification of more positive instances. This results in an improved recall and ultimately an increased accuracy. In many application increasing recall is important. For example, in computer aided diagnosis, a false negative could mean that a patient will not be diagnosed, and thus not treated.

## IV. Experimental Methodology

To measure the impact of the new threshold adjustment procedure on the performance of MIL algorithms, it has been applied to 7 reference algorithms, and on data sets from 3 application domains. In MIL instance classification tasks, the classes are often imbalanced. Classification performance will therefore be compared using two metrics that are appropriate for this context: the unweighted average recall (UAR), which is equivalent to averaging the accuracy for each class, and the $F_1$-score which is the harmonic mean of precision and recall. Precision, recall, the area under the precision-recall curve (AUPRC) and the false positive rate (FPR) will also be reported to better understand the impact of the proposed threshold adjustment procedure for each class.

A bag-level stratified 10-fold cross-validation process was used a to measure average performance. The hyper-parameters of all algorithms were optimized in each experiment via grid search in a nested cross-validation. The adjustment of the decision threshold is performed on the training data.

### *A. Data Sets*

In this subsection, the data sets used in the experiments are described in detail. They are some of the few MIL benchmarks providing instance labels. They have been chosen because they each pose different types of difficulties.

**Birds [8]:** In this data set, each bag corresponds to a 10 seconds recording of bird songs from one or more species. The recording is temporally segmented, and each part corresponds to a particular bird, or to background noises. These 10232 segments are the instances, each of represented by 38 features. Details on the extraction of the features are given in the original paper. There are 13 types of bird in the data set. If one specie at a time is considered as the positive class, 13 MIL problems can be generated from this data set. The difficulty for MIL is that the WR is low and is not constant across positive bags. Also there are sometimes a severe class imbalance at bag level. This data set show

**Newsgroups [20]:** This set was derived from the *20 Newsgroups* data set corpus. It contains posts from newsgroups on 20 subjects represented by 200 term frequency-inverse document frequency features. These features are generally sparse vectors, where each element represents a word frequency in a text. When one of the subject is selected as the positive class, all of the 19 other subjects are used as the negative class. The average WR of the data set is 3.7% which makes the problem difficult. Moreover the instance are not i.i.d. and the feature are sparse and their distribution is highly multi-modal.

**Spatially Independent, Variable Area, and Lighting (SIVAL) [21]:** This benchmark data set is often used to compare MIL algorithms on image retrieval tasks. It contains 1500 images each segmented and manually labeled by [20]. There are 25 classes of complex objects photographed from different view points in various environments. Each object is in turn considered as the positive class thus yielding 25 different learning problems. The bags correspond to images partitioned in approximately 30 segments, each corresponding to an instance. The segments are described by a 30-dimensional feature vector encoding color, texture and information about the neighboring segments. There are 60 images in each class, which makes 60 positive bags, and 5 images are randomly selected from each of the other 24 classes to create 120 negative bags. The WR of the data set is in 25.5% in average but ranges from 3.1 to 90.6% and the instances are non i.i.d. as in many image data sets.

### B. Reference Methods

This subsection describes the 7 reference methods used in the experiments. These methods were selected because they are well known and represent a wide spectrum of MIL algorithms suitable for instance classification.

**SI-SVM and SI-kNN:** A simple approach for instance classification is to transpose the MIL problem into a one-sided noise supervised classification problem. Each instance inherits the label of its bag and a classifier is trained on all instances. While not a MIL method *per se*, this method have been used as a reference point in many MIL papers [22], [23] because it indicates the pertinence of using MIL methods instead of regular supervised algorithm in such problems. In this paper an SVM (SI-SVM) and a nearest neighbor classifiers (SI-kNN) will be used in the experiments. These methods are interesting in the context of this paper because they discard bag information and treat instances individually.

**MI-SVM and mi-SVM [12]:** For mi-SVM, a label is assigned to each instances. An SVM is trained based on the instance label assignation. The instances are then reclassified using the newly obtain SVM. The resulting labels are then assigned to each instance and the SVM is retrained. This procedure is repeated until the labels are stable. The training procedure is similar for MI-SVM except that only the most positive instance of the positive bags is used for training. These two methods were selected because they are established MIL reference methods, they both use transductive learning and are different from each other in their optimization objective: mi-SVM focuses on instances while MI-SVM focuses on bags.

**EM-DD [14]:** Diverse Density (DD) [13] is a measure of the probability that a given point in the input feature space belongs to the positive class. It depends on the proportion of instances from positive and negative bags in the neighborhood. The highest point of the DD function corresponds to the positive concept from which are generated the witnesses Instances are classified based on their proximity to this point. In EM-DD [14], the Expectation-Maximization algorithm is used to locate the maximum of the DD function. This algorithm has been selected to represent DD-based methods because it is the most widely used as reference method. The implementation from [24] is used in the experiments.

**MIL-Boost [15]:** The MIL-Boost algorithm used in this paper [15] is essentially the same as gradient boosting [25] except that the loss function is computed on bag classification error. The instances are classified individually, and their labels are combined to obtained bag labels using a derivable approximation of the max function. This methods has been selected because it has been proposed to perform instance classification. The implementation from [24] is used in the experiments.

**Citation-kNN-ROI [7]:** Citation-kNN [18] is an adaption of kNN to MIL problems. The distance between the bags $X$ and $Y$ is measured using the minimal Hausdorff distance. Intuitively, it is the shortest distance between any of the instances contained in the two bags. In addition to using a distance measure for bags, the neighborhood is a combination of the $r$-nearest bags to the test bag, and the bags containing the test bag in their $c$-nearest bags. Each of the $r + c$ bags cast a vote on the label of the test bag, and the majority rule is applied. The algorithm was adapted in [7] to perform classification of instances. Basically it consists of classifying all bags using Citation-kNN, and then, in positive bags, classifying instances individually as if they were bags. Citation-kNN was selected because it is a well-known non parametric method, which have been adapted for instance classification. The implementation of [24] was used in the experiments.

## V. Results

### A. Decision Thresholds Instance and Bag Classification

The two top graphs in Fig. 3 show the accuracy performance at bag- and instances-level obtained with different threshold values with MI-SVM on the Brown Creeper data set from the Birds data set collection. There are two curves for each fold: a blue one obtained on the training data and a red curve obtained with test data. The similar shapes of the UAR curves obtained with the training and test data indicates that there is not a significant loss of generalization when using the training data to adjust the threshold instead of a held out validation fold.

When comparing these two graphs, it is clear that the optimal threshold for instance and bags classification are different. MI-SVM aims at classifying all instance from negative bags as negative and at least one instance per positive bag as

Table 1. Difference in performance of MIL methods following the application of the proposed threshold adjustment method.

| | | **Bag Level** | | | | | | **Instance Level** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Reference Methods** | **Dataset** | UAR (%) | Prec. (%) | Rec. (%) | FPR (%) | $F_1$ (%) | AUPRC (×100) | UAR (%) | Prec. (%) | Rec. (%) | FPR (%) | $F_1$ (%) |
| Citation-kNN-ROI | Birds | -4.2 | -20.2 | **16.4** | 5.6 | -4.5 | -5.5 | -6.7 | -16.4 | **9.3** | 1.8 | -5.5 |
| | SIVAL | **0.3** | **2.2** | **7.3** | **-2.6** | **2.7** | -1.0 | -1.4 | **5.4** | **5.0** | **-5.8** | -1.6 |
| EM-DD | Newsgroups | -2.6 | -8.2 | **58.8** | 2.2 | **25.7** | **13.9** | **1.9** | -19.3 | **10.9** | 13.3 | -4.2 |
| | Birds | **4.0** | -35.5 | **27.8** | 21.6 | -3.9 | **14.2** | **9.6** | -34.7 | **26.7** | 20.8 | **5.6** |
| | SIVAL | -2.4 | -24.7 | **42.5** | 23.1 | **5.1** | **28.5** | **4.9** | -24.1 | **13.9** | 22.5 | **11.2** |
| mi-SVM | Newsgroups | -5.2 | -15.2 | **16.4** | 11.7 | **2.8** | -4.3 | **14.5** | -21.9 | **9.4** | 18.4 | **0.2** |
| | Birds | -5.4 | -19.9 | **13.6** | 1.5 | -3.8 | -7.1 | -2.1 | -23.0 | **10.0** | 4.6 | -6.9 |
| | SIVAL | -2.3 | **0.6** | -21.1 | **-0.2** | -6.2 | -2.8 | -1.2 | -3.7 | **15.0** | 4.1 | -5.2 |
| MI-SVM | Newsgroups | -4.6 | -26.0 | **40.7** | 10.0 | **15.9** | 0.0 | **17.4** | -42.0 | **37.4** | 26.0 | **1.9** |
| | Birds | **1.8** | -39.7 | **26.0** | 29.0 | -9.5 | -6.2 | **5.0** | -34.5 | **18.5** | 23.8 | **2.8** |
| | SIVAL | -2.7 | -30.1 | **69.0** | **-3.5** | **27.3** | -5.0 | **4.8** | -32.1 | **64.7** | **-1.5** | **12.1** |
| MILBoost | Birds | -2.4 | -27.1 | **23.3** | 19.4 | -2.9 | -8.8 | **5.5** | -40.7 | **39.0** | 6.9 | **10.3** |
| | SIVAL | **0.4** | -20.8 | **22.9** | 20.0 | **1.8** | -1.0 | **7.1** | -18.3 | **15.4** | 15.9 | **16.6** |
| SI-kNN | Birds | **2.7** | -13.7 | **12.0** | 3.7 | -1.9 | -0.5 | **1.2** | -16.7 | **1.4** | 6.7 | -3.7 |
| | SIVAL | **6.3** | **4.5** | -0.7 | **-4.5** | **4.2** | -0.5 | -1.4 | **9.5** | -11.0 | **-9.5** | **3.1** |
| SI-SVM | Newsgroups | **1.0** | -13.4 | **20.0** | **-4.1** | **12.7** | -5.7 | **18.6** | -18.0 | **10.7** | 0.5 | **12.6** |
| | Birds | **9.7** | **7.2** | **6.1** | **-20.2** | **16.9** | 0.0 | -3.7 | -2.4 | -12.3 | **-10.7** | **5.1** |
| | SIVAL | **0.6** | **8.8** | -38.1 | **-6.8** | -8.8 | 0.0 | -2.8 | **4.5** | -11.7 | **-2.5** | -8.6 |

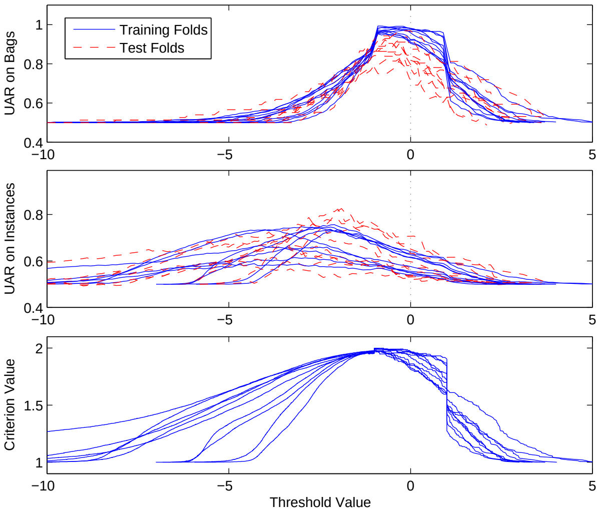

Fig. 3. Example of classification accuracy obtained using different decision threshold values. Each line represents the UAR obtained using MI-SVM on a different fold on the Brown Creeper (Birds) data set. The blue lines are obtained with training folds while the red lines are obtained with test folds.

positive. This is optimizes indirectly bag-level accuracy, and as a results, the optimal threshold for bag-level classification is near 0, which is the threshold value used by an SVM. The graph suggests that using a lower threshold would yield higher accuracy at instance-level and slightly higher accuracy at bag-level. As discussed in Section III the cost of misclassifying negative instances in negative bag, and positive instance in positive bags, are different in the two contexts which explain the different optimal threshold values.

The third curve, at the bottom, shows the value of the proposed criterion for the same threshold as the other two curves. While, the threshold is not optimal for instance classification, it represents an improvement on both performance measures in this case. The optimal threshold for instance classification cannot be found because of the uncertainty on the labels of instance belonging to positive bags.

### *B. Threshold Adjustment on Benchmark Data Sets*

Table 1 shows the difference on several performance metric on the 3 corpus of data sets after applying the proposed threshold adjustment procedure[1] (e.g.: $UAR_{after}$ - $UAR_{before}$). The numbers are in bold when an improvement is obtained. The results for Citation-kNN-ROI, SI-kNN and MILBoost are not reported for the Newsgroups data set because these algorithms failed to learn and consistently yielded an UAR of 50.0 % meaning that all bags were classified as the same class.

Results show that considerable improvement on instance classification performance can be obtained with the proposed criterion. For instance, on the Newsgroups data set, SI-SVM raises its UAR by 18.6% on average, or MILBoost increases its $F_1$-score by 16.6%. However, the table also indicates that the proposed method does not always leads to an improvement. Thus the method should not be applied blindly to all methods.

The adjustment strategy often lowers the decision threshold by the MIL algorithms. In other words, it makes the algorithm more sensitive to positive instances. As a results, recall is generally higher both for bag and instance classification, but precision is lower. Classes are highly imbalanced in MIL instance classification problems. For instance, in the Newsgroups data sets, the class imbalance ratio is 1:1 for bags but is 1:65 for instances. In that case, given a perfect recall, a precision loss of 50% means a decrease of less that 1% on accuracy. Thus diminishing precision can still result in an improved instance-level accuracy in this context. In many cases, the accuracy gain at instance-level does not reflect on bag-level accuracy. A more sensitive algorithm will be more susceptible to false positive and misclassification do not have the same impact when classifying instance or bag.

The proposed method is particularly successful with method using bag-level accuracy as an optimization criterion during learning. MI-SVM and MILBoost consistently improve their $F_1$-score and UAR for instance classification on all data sets.

[1] The results on all individual data sets can be found on the author website: https://sites.google.com/site/marcandrecarbonneau/

Similar results are observed for EM-DD, along with significant improvement on bag accuracy. The difference in maximizing the bag-level accuracy and the proposed criterion is that in the proposed criterion, bag accuracy is only measured on positive bags instead of on both classes. When computing bag accuracy on negative bags, a false positive has a great impact since it cause the entire bag to be misclassified. To correctly classify a positive bag, only one positive instance has to be identified. These two facts explain why algorithms maximizing bag accuracy are less sensitive. The proposed criterion lessens the penalty imposed to misclassified negative instance in negative bags by considering them individually instead of in groups.

Improvement were not consistently observed for all methods. The instance-level accuracy of the supervised methods, SI-SVM and SI-kNN, did not increase on the SIVAL data set. However UAR increased by 18.6% on the Newsgroups data set with SI-SVM, which suggests that the nature of the data distribution plays an important role in determining the success of applying the proposed method. In each experiment, the bag-level accuracy benefited from the threshold adjustment because these algorithms completely discard the structure of the MIL problem before learning. Therefore they already optimize instance-level accuracy during learning. The proposed criterion also enforces accuracy at bag-level which results in an accuracy improvement at this level. In essence, mi-SVM is similar to SI-SVM because the algorithm also classify each instance individually. As a matter a fact, SI-SVM is the first iteration of the mi-SVM algorithm. Bag structure is only used if a positive bag does not contain a positive instance. In that case, the most positive instance is labeled as positive. This explains why mi-SVM as a similar behavior to SI-SVM. Finally, the proposed adjustment strategy did not prove beneficial to the Citation-kNN-ROI algorithm on any data sets, perhaps because the algorithm makes predictions in two steps. It starts by classifying bags and then, classify instances. The proposed method is not equipped to deal with this kind of hierarchical decision process.

## VI. Conclusion

Instance and bag classification in MIL are different tasks that follow different objectives. It was shown that algorithms designed for bag classification can be used for instance classification, but higher accuracy is achievable by adjusting the decision threshold. A criterion for threshold adjustment has been proposed and experiments showed accuracy performance improvement for many methods for instance classification.

Different criteria considering the cluster arrangement of the instance in feature space could be proposed for threshold adjustment. Also, research should be devoted to new methods incorporating instance classification criteria for the learning phase of the MIL algorithms instead of adjusting the threshold as a post-processing step. Finally, experiment should be conducted using larger data sets for which the criterion could be computed on a held-out validation set.

## References

[1] J. Amores, “Multiple Instance Classification: Review, Taxonomy and Comparative Study,” *Artif. Intell.*, vol. 201, pp. 81–105, Aug. 2013.

[2] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, “Solving the Multiple Instance Problem with Axis-parallel Rectangles,” *Artif. Intell.*, vol. 89, no. 1-2, pp. 31–71, Jan. 1997.

[3] Y. Chen, J. Bi, and J. Z. Wang, “MILES: Multiple-Instance Learning via Embedded Instance Selection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, Dec. 2006.

[4] M. Kandemir and F. A. Hamprecht, “Computer-aided diagnosis from weak supervision: a benchmarking study.” *Computerized Medical Imaging and Graphics*, vol. 42, pp. 44–50, jun 2015.

[5] B. Babenko, M.-H. Yang, and S. Belongie, “Robust Object Tracking with Online Multiple Instance Learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011.

[6] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, “Multi-Instance Learning by Treating Instances As non-I.I.D. Samples,” in *Int. Conf. on Mach. Learn.* New York, USA: ACM, 2009, pp. 1249–1256.

[7] Z.-H. Zhou, X.-B. Xue, and Y. Jiang, “Locating Regions of Interest in CBIR with Multi-instance Learning Techniques,” in *Australian Joint Conf. on Artificial Intelligence*. Springer, 2005, pp. 92–101.

[8] F. Briggs, X. Z. Fern, and R. Raich, “Rank-loss support instance machines for MIML instance annotation,” in *Int. Conf. on Knowl. Discovery and Data Mining*. ACM, 2012, pp. 534–542.

[9] V. Cheplygina, D. M. Tax, and M. Loog, “On classification with bags, groups and sets,” *Pattern Recognition Lett.*, vol. 59, pp. 11–17, jul 2015.

[10] G. Vanwinckelen, V. do O, D. Fierens, and H. Blockeel, “Instance-level accuracy versus bag-level accuracy in multi-instance learning,” *Data Mining and Knowl. Discovery*, pp. 1–29, 2015.

[11] M.-A. Carbonneau, E. Granger, A. J. Raymond, and G. Gagnon, “Robust multiple-instance learning ensembles using random subspace instance selection,” *Pattern Recognition*, vol. 58, pp. 83–99, 2016.

[12] S. Andrews, I. Tsochantaridis, and T. Hofmann, “Support Vector Machines for Multiple-Instance Learning,” in *Advances in Neural Infor. Process. Syst. 15*. MIT Press, 2002, pp. 561–568.

[13] O. Maron and T. Lozano-Pérez, “A Framework for Multiple-Instance Learning,” in *Advances in Neural Infor. Process. Syst.* MIT Press, 1998, pp. 570–576.

[14] Q. Zhang and S. A. Goldman, “EM-DD : An Improved Multiple-Instance Learning Technique,” in *Advances in Neural Infor. Process. Syst.* MIT Press, 2001, pp. 1073–1080.

[15] B. Babenko, P. Dollár, Z. Tu, and S. Belongie, “Simultaneous Learning and Alignment: Multi-Instance and Multi-Pose Learning,” in *Workshop on Faces in ’Real-Life’ Images: Detection, Alignment, and Recognition*, Marseille, France, Oct. 2008.

[16] V. Cheplygina, D. M. J. Tax, and M. Loog, “Dissimilarity-Based Ensembles for Multiple Instance Learning,” *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–13, 2015.

[17] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, “Multi-Instance Kernels,” in *Int. Conf. on Mach. Learning*. ACM, 2002, pp. 179–186.

[18] J. Wang and J.-D. Zucker, “Solving the Multiple-Instance Problem: A Lazy Learning Approach,” in *Int. Conf. on Mach. Learning*. San Francisco, USA: ACM, 2000, pp. 1119–1126.

[19] Q. Zhang, S. A. Goldman, W. Yu, and J. Fritts, “Content-Based Image Retrieval Using Multiple-Instance Learning,” in *Int. Conf. on Mach. Learn.* San Francisco, USA: ACM, 2002, pp. 682–689.

[20] B. Settles, M. Craven, and S. Ray, “Multiple-Instance Active Learning,” in *Advances in Neural Infor. Process. Syst. 20*, 2008, pp. 1289–1296.

[21] R. Rahmani, S. A. Goldman, H. Zhang, J. Krettek, and J. E. Fritts, “Localized Content Based Image Retrieval,” in *Proc. of ACM SIGMM Int. Workshop on Multimedia Inform. Retrieval*, 2005, pp. 227–236.

[22] E. Alpaydin, V. Cheplygina, M. Loog, and D. M. Tax, “Single- vs. multiple-instance classification,” *Pattern Recognition*, vol. 48, no. 9, pp. 2831–2838, sep 2015.

[23] S. Ray and M. Craven, “Supervised Versus Multiple Instance Learning: An Empirical Comparison,” in *Int. Conf. on Mach. Learning*. New York, USA: ACM, 2005, pp. 697–704.

[24] D. Tax and V. Cheplygina, “MIL, a Matlab toolbox for multiple instance learning,” Jun 2015, version 1.1.0. [Online]. Available: http://prlab.tudelft.nl/david-tax/mil.html

[25] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.