Label Stability in Multiple Instance Learning

Veronika Cheplygina[1,3], Lauge Sørensen[2], David M.J. Tax[1], Marleen de Bruijne[2,3], and Marco Loog[1,2]

[1] Pattern Recognition Laboratory, Delft University of Technology, The Netherlands
[2] The Image Section, University of Copenhagen, Copenhagen, Denmark
[3] Biomedical Imaging Group Rotterdam, Erasmus MC, Rotterdam, The Netherlands

Abstract. We address the problem of *instance label stability* in multiple instance learning (MIL) classifiers. These classifiers are trained only on globally annotated images (bags), but often can provide fine-grained annotations for image pixels or patches (instances). This is interesting for computer aided diagnosis (CAD) and other medical image analysis tasks for which only a coarse labeling is provided. Unfortunately, the instance labels may be unstable. This means that a slight change in training data could potentially lead to abnormalities being detected in different parts of the image, which is undesirable from a CAD point of view. Despite MIL gaining popularity in the CAD literature, this issue has not yet been addressed. We investigate the stability of instance labels provided by several MIL classifiers on 5 different datasets, of which 3 are medical image datasets (breast histopathology, diabetic retinopathy and computed tomography lung images). We propose an unsupervised measure to evaluate instance stability, and demonstrate that a performance-stability trade-off can be made when comparing MIL classifiers.

1 Introduction

Obtaining ground-truth annotations for patches, which can be used to train supervised classifiers for localization of abnormalities in medical images can be very costly and time-consuming. This hinders the use of supervised classifiers for this task. Fortunately, global labels for whole images, such as the overall condition of the patient, are available more readily. Multiple instance learning (MIL) is an extension of supervised learning which can train classifiers using such weakly labeled data. For example, a classifier trained on images (*bags*), where each bag is labeled as healthy or abnormal and consists of unlabeled image patches (*instances*), would be able to label patches of a novel image as healthy or abnormal.

MIL is becoming more and more popular in CAD [9,13,6,20,16,3,21,18,12]. In many of these applications, it is desirable to obtain instance labels, and to inspect the instances which are deemed positive. For example, in [13], weakly labeled x-ray images of healthy subjects and patients affected by tuberculosis are used to train a MIL classifier which can provide local abnormality scores, which can be visualized across the lungs. Furthermore, the MIL classifier *outperforms*

N. Navab et al. (Eds.): MICCAI 2015, Part I, LNCS 9349, pp. 539–546, 2015.
DOI: 10.1007/978-3-319-24553-9_66

its supervised counterpart which has access to fine-grained labels, showing the potential of MIL for CAD applications.

A pitfall in using MIL classifiers to obtain instance labels is that these labels might be unstable, for example, if a different subset of the data is used for training. This is clearly undesirable in a diagnostic setting, because abnormalities would be highlighted in different parts of the image. For example, in [12] a MIL classifier is used to identify which of the 8 regions (instances) of the tibial trabecular bone (bag) are most related to cartilage loss. The "most positive" region labeled positive by only 20% of the classifiers, trained on different subsets of the data. We have not been able to identify other research where this phenomenon is investigated, which emphasizes the importance of the present work.

In rare cases where instance-level annotations are available, such as in [9], instance labels can be evaluated using AUC. The results here show that the best bag classifier does not correspond to the best instance classifier, emphasizing that bag-level results are not reliable if instance labels are needed. Another approach is to evaluate the instances qualitatively. However, this is typically done for a single run of the classifier, which raises the question whether the same abnormalities would be found if the training set would change slightly.

We propose to evaluate the *stability* of instance-labeling MIL classifiers as an additional measure for classifier comparison. We evaluate two stability measures on three CAD datasets: computed tomography lung images with chronic obstructive pulmonary disease (COPD), histopathology images with breast cancer and diabetic retinopathy images. We demonstrate how stability varies in popular MIL classifiers, and show that choosing the classifier with the best bag-level performance may not lead to reliable instance labels.

2 Multiple Instance Learning

In multiple instance learning, a sample is a bag or set $B_i = \{\mathbf{x}_k^i | k = 1, ..., n_i\} \subset \mathbb{R}^d$ of n_i instances, each instance is thus a d-dimensional feature vector. We are given labeled training bags $\{(B_i, y_i) | i = 1, ...N_{tr}\}$ where $y_i \in \{0, 1\}$. The standard assumption is that there exist hidden instance labels $z_k^i \in \{0, 1\}$ which relate to the bag labels as follows: a bag is positive if and only if it contains at least one positive instance.

Originally, the goal in MIL is to train a bag classifier f_B to label previously unseen bags. Several MIL classifiers do this by inferring an instance classifier f_I, and combining the outputs of the bag's instances, for example by the noisy-or rule, $f_B(B_i) = \max_k\{f_I(\mathbf{x}_k^i)\}$. An example of such an *instance-level* classifier is SimpleMIL, which propagates the bag label to its instances and simply trains a supervised classifier on the, possibly noisy, instance labels. Classifiers which explicitly use the MIL assumption are miSVM [1] and milBoost [19], which are MIL adaptations of popular learning algorithms. For example, miSVM extends the SVM by not only searching for the optimal hyperplane $\mathbf{w}$ which defines f_I, but also for the instance labels $\{z_i^k\}$ which are consistent with the bag label assumptions:

$$\min_{\{z_i^k\}} \min_{\mathbf{w},\xi} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i,k} \xi_i^k \qquad \text{s.t.} \qquad (1)$$
$$\forall i,k : z_i^k(\langle \mathbf{w}, \mathbf{x}_i^k \rangle) \geq 1 - \xi_i^k, \xi_i^k \geq 0, z_i^k \in \{-1,1\}, \max\{z_i^k\} = y_i.$$

Another group, *bag-level* classifiers, typically represent each bag as a single feature vector and use supervised classifiers for training f_B directly [4,5]. Such classifiers are often robust, but usually can not provide instance labels. A notable exception is MILES [4], which represents each bag by its similarities to a set of prototype instances, $\mathbf{s}_i = [s(B_i, \mathbf{x}_1^1), \ldots, s(B_i, \mathbf{x}_{n_1}^1), \ldots s(B_i, \mathbf{x}_{n_{N_{tr}}}^{N_{tr}})]$ where $s(B_i, \mathbf{x}) = \exp(-\min_k ||\mathbf{x} - \mathbf{x}_k^i||)$ or any other kernel. A sparse classifier then selects the most discriminative features, which correspond to instance prototypes. It is assumed that discriminative prototypes from positive bags are positive, instances can therefore be classified based on their similarity to these prototypes.

The interest in **MIL for computer aided diagnosis** has grown over the past decade, as illustrated by Table 1. Supervised evaluation of instances is only performed in a few studies – where (a part) of the data has been annotated at the instance level. Otherwise, papers examine the instances qualitatively, such as displaying the most abnormal instances [13], or not at all, although instance labels would be interesting from a diagnostic point of view [6,9]. As our proposed evaluation is unsupervised, it can easily be adopted in all these studies.

Table 1. Evaluation of MIL in CAD tasks. Columns show bag (B) and instance (I) evaluation: supervised (+), qualitative (∘) or none (–).

Task	B	I	Task	B	I
Cancer histology [9]	+	+	Diabetic retinopathy [9]	+	–
COPD in CT [6]	+	–	Tuberculosis in XR [13]	+	∘
Cancer histopathology [22]	+	+	Osteoarthritis in MRI [12]	+	∘
Diabetic retinopathy [16]	+	+	Pulmonary embolism in CT [11]	+	–
Myocardial infarction in ECG [18]	+	–	COPD in CT [17]	+	–
Colorectal cancer in CT [8]	+	–			

3 Instance Stability

We are interested in evaluating the similarity of a labeling, or vector of outputs of two classifiers $\mathbf{z} = f_I(X)$ and $\mathbf{z}' = f_I'(X)$, trained on slightly different subsets of the training data, for the test set $X = [\mathbf{x}_1^1, \ldots, \mathbf{x}_{n_N}^N]^\intercal$. The stability measure should be **monotonically increasing** with the number of instances the classifiers agree on, have **limits**, and most importantly, be **unsupervised**, i.e. not dependent on the hidden instance labels z_i.

The general concept of stability is important in machine learning, and different aspects of it have been addressed in the literature. Leave-one-out stability [15] measures to what extent a decision boundary changes when a sample is removed from the training data, but is not appropriate because it is supervised. An unsupervised version where bags are left out, and true labels are substituted by classifier outputs, is related to the measures we propose. The kappa statistic is unsupervised, but does not follow the monotonicity property in class imbalance settings which could occur in MIL. Clustering stability [2] compares the outputs of two clustering procedures and is unsupervised. It is appropriate for our goal and is in fact related to the measures proposed in what follows.

Let $n_{00} = |\{i|z_i = 0 \wedge z'_i = 0\}|$, $n_{01} = |\{i|z_i = 0 \wedge z'_i = 1\}|$, $n_{10} = |\{i|z_i = 1 \wedge z'_i = 0\}|$ and $n_{11} = |\{i|z_i = 1 \wedge z'_i = 1\}|$. An intuitive measure that satisfies the properties above is the agreement fraction:

$$S(\mathbf{z}, \mathbf{z}') = (n_{00} + n_{11})/(n_{01} + n_{10} + n_{11} + n_{00}). \tag{2}$$

In a situation with many true negative instances, the value of S would be inflated due to the negative instances that the classifiers agree on. As a result, the classifier can still be unstable with respect to the positive instances. Due to the nature of CAD tasks, we might consider it more important for the classifiers to agree on the positive instances. Therefore we also consider the agreement on positive labels only, or Jaccard distance:

$$S_{+}(\mathbf{z}, \mathbf{z}') = n_{11}/(n_{01} + n_{10} + n_{11}). \tag{3}$$

We emphasize that the novelty does not lie in the measures themselves, well-known as they are. The novelty resides in what they measure in this context: we derive these measures as the appropriate ones for the stability that we want to quantify.

Classifier Selection. If instance classification stability is a crucial issue, one can study our measure in combination with bag-level AUC (or any other accuracy measure) and select a MIL classifier with a good trade-off of AUC and instance stability. We can see each classifier as a possible solution, parametrized by these two values. Intermediate solutions between classifiers f_I and f'_I can in theory be obtained by designing a randomized classifier, which trains classifier f_I with probability p and classifier f'_I with probability $1-p$. In the AUC-stability plane, the Pareto frontier is the set of classifiers which are Pareto efficient, i.e. no improvement can be made in AUC without decreasing instance stability and vice versa. Optimal classifiers can therefore be selected from this Pareto frontier. While the classifier with the highest AUC is in this set, it is not necessarily the only desirable solution, if the instance labels are of importance.

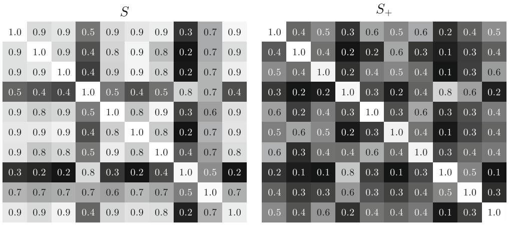
4 Experiments and Results

Datasets. The datasets are shown in Table 2. The Musk datasets are benchmark problems of molecule activity prediction. In Breast, an instance is a 7 × 7 patch from a 896×768 tissue microarray analysis image from a patient with a malignant (+) or benign (–) tumor. In Messidor, an instance is a 135 × 135 patch from a 700 × 700 fundus image of a diabetes (+) or healthy (–) subject. In COPD, a bag is a CT image of a lung of a subject with COPD (+) or a healthy subject (–). An instance is a region of interest (ROI) of 41 × 41 × 41 voxels, with the center inside the segmentation of the lung field.

Table 2. Datasets and their properties. Musk, Breast and Messidor can be downloaded from a MIL data repository [5] (`http://www.miproblems.org`).

Dataset	Bags	Instances	Inst per bag	Features
Musk 1	47+, 45–	476	2 to 40	166
Musk 2	39+, 63–	6598	1 to 1024	166
Breast [10]	26+, 32–	2002	21 to 40	657 (intensity, LBP, SIFT)
Messidor [9,7]	654+, 546–	12352	8 to 12	687 (intensity, LBP, SIFT)
COPD [17,14]	231+, 231–	26200	50	287 (Gaussian filter bank)

Illustrative Example. Fig. 1 shows the pairwise stability measures for the COPD validation data, for 10 MILES classifiers, each trained on random 80% of the training data. There is considerable disagreement for both measures, which is surprising because of the large overlap of the training sets. The measures are quite correlated ($\rho = 0.76$), but S has higher values because it is inflated by agreement on negative instances.



S

1.0	0.9	0.9	0.5	0.9	0.9	0.9	0.3	0.7	0.9
0.9	1.0	0.9	0.4	0.8	0.9	0.8	0.2	0.7	0.9
0.9	0.9	1.0	0.4	0.9	0.9	0.8	0.2	0.7	0.9
0.5	0.4	0.4	1.0	0.5	0.4	0.5	0.8	0.7	0.4
0.9	0.8	0.9	0.5	1.0	0.8	0.9	0.3	0.6	0.9
0.9	0.9	0.9	0.4	0.8	1.0	0.8	0.2	0.7	0.9
0.9	0.8	0.8	0.5	0.9	0.8	1.0	0.4	0.7	0.8
0.3	0.2	0.2	0.8	0.3	0.2	0.4	1.0	0.5	0.2
0.7	0.7	0.7	0.7	0.6	0.7	0.7	0.5	1.0	0.7
0.9	0.9	0.9	0.4	0.9	0.9	0.8	0.2	0.7	1.0

S_+

1.0	0.4	0.5	0.3	0.6	0.5	0.6	0.2	0.4	0.5
0.4	1.0	0.4	0.2	0.2	0.6	0.3	0.1	0.3	0.4
0.5	0.4	1.0	0.2	0.4	0.5	0.4	0.1	0.3	0.6
0.3	0.2	0.2	1.0	0.3	0.2	0.4	0.8	0.6	0.2
0.6	0.2	0.4	0.3	1.0	0.3	0.6	0.3	0.3	0.4
0.5	0.6	0.5	0.2	0.3	1.0	0.4	0.1	0.3	0.4
0.6	0.3	0.4	0.4	0.6	0.4	1.0	0.3	0.4	0.4
0.2	0.1	0.1	0.8	0.3	0.1	0.3	1.0	0.5	0.1
0.4	0.3	0.3	0.6	0.3	0.3	0.4	0.5	1.0	0.3
0.5	0.4	0.6	0.2	0.4	0.4	0.4	0.1	0.3	1.0

Fig. 1. Pairwise stability for 10 MILES classifiers for agreement (left) and positive agreement (right) for the COPD dataset

Fig. 2 shows how the instance classifications change in a true positive bag, i.e., CT image from a COPD patient. This bag is always classified as positive, but the instance labels are unstable. A perfectly stable classifier would have a bimodal distribution, classifying instances as positive either 0 or 10 times. We also show a number of ROIs with stable and unstable labels. Several ROIs containing emphysema have unstable classifications, and one emphysemous patch is even consistently classified as negative. This shows that while the bag is always classified correctly, the instance labels may not be very reliable.

Evaluation. We evaluate a number of classifiers (please see Sec. 2 for descriptions) from the MIL toolbox[1], which we modified to output instance labels:

- simpleMIL with SVM, nearest mean (NM) and 1-nearest neighbor (1NN)
- miSVM and its variants miNM and mi1NN (based on NM and 1NN)
- MILBoost

[1] `http://prlab.tudelft.nl/david-tax/mil.html`

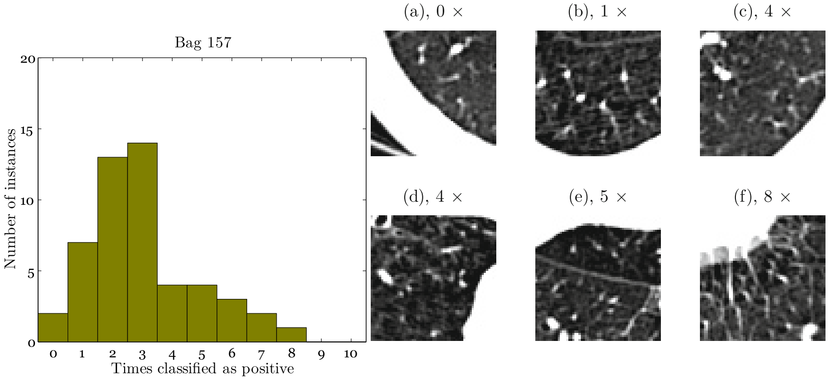


Fig. 2. "Positiveness" of 50 instances (ROIs) from a positive bag. **Left:** How often an ROI is classified positive, and for how many ROIs this holds. **Right**: Examples of 6 ROIs, for which the axial slice with most lung voxels below -910 hounsfield units is shown. ROIs (b,d,e,f) contain emphysema (low intensity areas within the lung tissue), but only (f) is often classified as positive. ROIs (a,c) are largely unaffected, but only (a) is consistently classified as negative.

- MILES

We use a linear kernel and regularization parameter $C = 1$ for SVM, miSVM and MILES. For each train/test split, we do the following 10 times: randomly sample 80% of the training bags (bag = subject), train the classifier, and evaluate on the test dataset. The splits are done randomly for Musk, Breast and Messidor and based on predefined sets for COPD.

The average bag AUCs, average pairwise instance stabilities, and the corresponding Pareto frontiers for $S+$ (S provided similar plots, but with inflated values) are shown in Fig. 3. Note that the (1, 0.5) point can be achieved by a classifier which labels all instances as positive. The main observation is that the most accurate classifier is often not the most stable one. This trade-off is especially well-illustrated in the COPD datasets. Here we see similar behavior between the two sets, which shows that if we were to use the validation set results for classifier selection, we would obtain a classifier with similar performance and stability on the test set.

With regard to the classifiers, miSVM and its variants seem to be relatively good choices. MILES, which is a popular classifier due to its good performance, can indeed be quite accurate, but at the same time unstable. The difference between the mi- classifiers and MILES is probably due to the fact that MILES trains a bag classifier f_B first, and infers f_I from f_B, while the mi- classifiers train f_I directly. MILBoost is both inaccurate and unstable, especially for COPD there is high disagreement on which instances to label as positive.

Note that the goal of these experiments is to demonstrate the trade-off between AUC and stability, not to maximize the AUC. Nevertheless, the best performances achieved by classifiers tested here are [0.91, 0.80, 0.72, 0.72] for

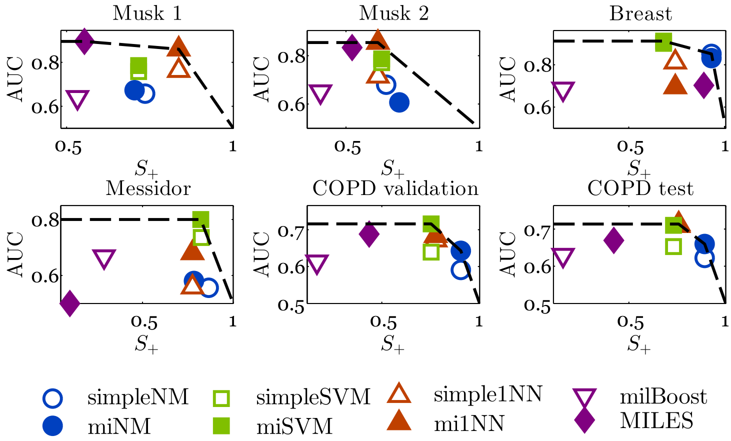


Fig. 3. Bag AUC vs positive instance stability and the corresponding Pareto frontiers

Breast, Messidor, and the COPD datasets. In previous works, the highest[2] performances for the same datasets were [0.90, 0.81, 0.74, 0.74]. This shows that our result are on par with state of the art, despite using less data and optimization.

5 Conclusions

We addressed the issue of stability of instance labels provided by MIL classifiers. We examined two unsupervised measures of agreement: S based on all labels, and S_+ based on positive (abnormal) labels, which might be more interesting from a CAD point of view. Our experiments demonstrate a trade-off between bag performance and instance label stability, and miSVM is a classifier which provides a good trade-off. In general, we propose to use instance label stability as an additional evaluation measure when applying MIL classifiers in CAD.

Acknowledgements. This research is partially financed by The Netherlands Organization for Scientific Research (NWO). We thank Dr. Melih Kandemir for kindly providing the Breast and Messidor datasets.

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: NIPS, pp. 561–568 (2002)

[2] Note that [6] reports several higher AUCs for COPD, but these correspond to a larger version of the dataset, which was not used in our study.

2. Ben-Hur, A., Elisseeff, A., Guyon, I.: A stability based method for discovering structure in clustered data. In: Pac. Symp. Biocomput., pp. 6–17 (2001)
3. Bi, J., Liang, J.: Multiple instance learning of pulmonary embolism detection with geodesic distance along vascular structure. In: CVPR, pp. 1–8 (2007)
4. Chen, Y., Bi, J., Wang, J.: MILES: Multiple-instance learning via embedded instance selection. IEEE T. Pattern. Anal. Mach. Intel. 28(12), 1931–1947 (2006)
5. Cheplygina, V., Tax, D.M.J., Loog, M.: Multiple instance learning with bag dissimilarities. Pattern Recognition 48(1), 264–275 (2015)
6. Cheplygina, V., et al.: Classification of COPD with multiple instance learning. In: ICPR, pp. 1508–1513 (2014)
7. Decencière, E., et al.: Feedback on a publicly distributed image database: the Messidor database. Image Anal. Stereol., 231–234 (2014)
8. Dundar, M.M., Fung, G., et al.: Multiple-instance learning algorithms for computer-aided detection. IEEE T. Biomed. Eng. 55(3), 1015–1021 (2008)
9. Kandemir, M., Hamprecht, F.A.: Computer-aided diagnosis from weak supervision: A benchmarking study. Comput. Med. Imag. Grap. (2014) (in press)
10. Kandemir, M., Zhang, C., Hamprecht, F.A.: Empowering multiple instance histopathology cancer diagnosis by cell graphs. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part II. LNCS, vol. 8674, pp. 228–235. Springer, Heidelberg (2014)
11. Liang, J., Bi, J.: Computer aided detection of pulmonary embolism with tobogganing and mutiple instance classification in CT pulmonary angiography. In: Karssemeijer, N., Lelieveldt, B. (eds.) IPMI 2007. LNCS, vol. 4584, pp. 630–641. Springer, Heidelberg (2007)
12. Marques, J.: Osteoarthritis imaging by quantification of tibial trabecular bone. Ph.D. thesis, Københavns Universitet (2013)
13. Melendez, J., et al.: A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays. TMI 31(1), 179–192 (2014)
14. Pedersen, J.H., et al.: The Danish randomized lung cancer CT screening trial-overall design and results of the prevalence round. J. Thorac. Oncol. 4(5), 608–614 (2009)
15. Poggio, T., Rifkin, R., Mukherjee, S., Niyogi, P.: General conditions for predictivity in learning theory. Nature 428(6981), 419–422 (2004)
16. Quellec, G., et al.: A multiple-instance learning framework for diabetic retinopathy screening. MedIA 16(6), 1228–1240 (2012)
17. Sørensen, L., Nielsen, M., Lo, P., Ashraf, H., Pedersen, J.H., de Bruijne, M.: Texture-based analysis of COPD: a data-driven approach. TMI 31(1), 70–78 (2012)
18. Sun, L., Lu, Y., Yang, K., Li, S.: ECG analysis using multiple instance learning for myocardial infarction detection. IEEE T. Biomed. Eng. 59(12), 3348–3356 (2012)
19. Viola, P., Platt, J., Zhang, C.: Multiple instance boosting for object detection. In: NIPS, pp. 1417–1424 (2005)
20. Wang, S., et al.: Seeing is believing: Video classification for computed tomographic colonography using multiple-instance learning. TMI 31(5), 1141–1153 (2012)
21. Wu, D., Bi, J., Boyer, K.: A min-max framework of cascaded classifier with multiple instance learning for computer aided diagnosis. In: CVPR, pp. 1359–1366 (2009)
22. Xu, Y., et al.: Weakly supervised histopathology cancer image segmentation and classification. MedIA 18(3), 591–604 (2014)