DOI: 10.1109/ICPR.2014.268

Classification of COPD with Multiple Instance Learning

Veronika Cheplygina*, Lauge Sørensen† , David M. J. Tax*,
Jesper Holst Pedersen‡, Marco Loog*† and Marleen de Bruijne†§
*Pattern Recognition Laboratory, Delft University of Technology, Delft, The Netherlands
†The Image Group, Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
‡Department of Thoracic Surgery, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark
§Biomedical Imaging Group Rotterdam, Erasmus MC, Rotterdam, The Netherlands
Email: {v.cheplygina, d.m.j.tax, m.loog}@tudelft.nl, {lauges,marleen}@diku.dk

***Abstract*—Chronic obstructive pulmonary disease (COPD) is a lung disease where early detection benefits the survival rate. COPD can be quantified by classifying patches of computed tomography images, and combining patch labels into an overall diagnosis for the image. As labeled patches are often not available, image labels are propagated to the patches, incorrectly labeling healthy patches in COPD patients as being affected by the disease. We approach quantification of COPD from lung images as a multiple instance learning (MIL) problem, which is more suitable for such weakly labeled data. We investigate various MIL assumptions in the context of COPD and show that although a concept region with COPD-related disease patterns is present, considering the whole distribution of lung tissue patches improves the performance. The best method is based on averaging instances and obtains an AUC of 0.742, which is higher than the previously reported best of 0.713 on the same dataset. Using the full training set further increases performance to 0.776, which is significantly higher (DeLong test) than previous results.**

***Keywords*—Computer aided diagnosis, chronic obstructive pulmonary disease, supervised learning, multiple instance learning**

I. Introduction

Chronic obstructive pulmonary disease (COPD) is a disease of the lungs that is caused, among others, by smoking and air pollution. COPD is characterized by chronic inflammation of the lung airways, and degradation of lung tissue, called emphysema, both of which result in airflow limitation [1], [2]. The disease progresses in several stages and can eventually lead to death, however, detecting the disease at an early stage can increase the survival rate [3].

Due to limitations of traditional spirometry and visual assessment of computed tomography (CT) scans, texture classification was proposed to quantify COPD [4], [5], [6], [7], [8]. One approach is to classify patches of lung tissue, or regions of interest (ROIs) in the image, and combine the classifications into an overall probability for COPD [4], [6]. However, these supervised approaches require manually annotated ROIs, which are difficult and costly to obtain.

An alternative is to use weakly labeled medical images, i.e., where only a global image label is provided, for training an image classifier. In the absence of labeled ROIs, the image label can be propagated to its ROIs, and an ROI classifier can be trained as usual [5]. We call this straightforward approach SimpleMIL. However, this disregards the fact that in scans of patients with COPD, only a subset of the ROIs may be affected, while signs of COPD may be already apparent in some regions for subjects not yet diagnosed with the disease. This increases the label noise for the ROI classifier.

A technique which can handle learning with such weakly labeled data is called multiple instance learning (MIL) [9], [10]. The goal is to build a classifier for a collection, or bag, of feature vectors, also referred to as instances. Often it is assumed that a bag is positive if and only if at least one of its instances is positive. A further assumption is that positive instances are found in a region of the feature space called the concept. For COPD, the concept could be a part of the feature space, containing ROIs that are typical for, for example, emphysema. In this scenario, as soon as a CT scan contains such an ROI, the whole image is diagnosed as COPD.

MIL methods can be broadly divided into two categories: instance-based and bag-based. Instance-based methods use the constraints posed by the bag labels and the MIL assumptions to build an instance classifier, and combine instance classifications to classify bags [10], [11], [12], [13]. On the other hand, bag-based methods aim to classify bags directly, often by defining kernels [14] or dissimilarities [15], [16] between bags.

Every MIL classifier makes explicit or implicit assumptions about the data. Instance-based classifiers typically rely on the assumption that there is a concept, and that positive bags contain instances from this concept. Therefore, only concept instances are important for determining the bag label. Bag-based classifiers assume that bags from the same class are similar, and the similarity definition further specifies this assumption. In most definitions, all of the bag's instances are involved in defining the bag similarity, therefore the whole distribution of instances is important for the bag label. In [17] we have shown that many well-known MIL problems fall into these two categories (concept and distribution) and that this property determines how many MIL methods perform on the data.

Detection of COPD using lung texture has been tackled by classifying patches and combining their outputs, an approach we call SimpleMIL, in [5]. A more specialized MIL method, applied to this problem, is a dissimilarity-based approach in [18], and it shows promising results. Other dissimilarity or kernel-based approaches, which focus on the airways rather

than lung texture, have also been successful for COPD classification [19], [20]. In this work we investigate a broader range of MIL methods for classification of the lung texture in COPD. We examine which assumptions, commonly used for the instance-based and bag-based methods, are more suitable for this problem, and demonstrate state-of-the-art results on a COPD dataset from the Danish Lung Cancer Screening Trial [21].

II. Multiple Instance Learning

In multiple instance learning (MIL), an object is represented by a bag $B_i = \{\mathbf{x}_{ik} | k = 1, ..., n_i\} \subset \mathbb{R}^d$ of n_i instances, where the k-th instance is described by a d-dimensional feature vector $\mathbf{x}_{ik}$. The training set $\mathcal{X}_{tr} = \{(B_i, y_i) | i = 1, ...N\}$ consists of positive ($y_i = +1$) and negative ($y_i = -1$) bags. One way to deal with this type of input is to propagate the bag labels to the instances, and building an instance classifier. A bag label is obtained by classifying that bag's instances, and combining the instance classifications, for example by fusing the posterior probabilities [22]. The noisy-or rule,

$$\frac{p(y=1|B_i)}{p(y=-1|B_i)} = \frac{1 - \prod_{k=1}^{n_i}(1 - p(z_{ik}=1|\mathbf{x}_{ik}))}{\prod_{k=1}^{n_i} p(z_{ik}=-1|\mathbf{x}_{ik})} \quad (1)$$

reflects the standard assumption that a bag is positive if and only if at least one of the instances is positive. On the other hand, the average rule,

$$\frac{p(y=1|B_i)}{p(y=-1|B_i)} = \frac{1}{n_i} \sum_{k=1}^{n_i} \frac{p(z_{ik}=1|\mathbf{x}_{ik})}{p(z_{ik}=-1|\mathbf{x}_{ik})} \quad (2)$$

assumes that all instances contribute to the bag label. This fusion rule has been used in [5], by classifying ROIs with a nearest neighbor classifier, and combining the outputs to classify the entire image. We refer to this strategy as SimpleMIL in the experiments.

The standard assumption for MIL is that there are hidden instance labels z_{ik} which relate to the bag labels as follows: a bag is positive if and only if it contains at least one positive, or *concept* instance [9]. The strategy of earlier MIL approaches was to model the concept: a region in feature space which contains at least one instance from each positive bag, but no instances from negative bags. Diverse density [10] (DD) has been proposed to measure this property. For a given point t in the feature space, $DD(t)$ measures the ratio between the number of positive bags which have instances near t, and the sum of distances of the negative instances to t. The point where DD is maximized, t^* therefore corresponds to the target concept. Instances can be classified using their distance to t^*. However, the optimization problem suffers from local optima and, for the original DD algorithm, several restarts of the algorithm are needed. Therefore, an expectation-maximization version of this algorithm EM-DD [11] has been proposed. EM-DD has shown to perform well on a range of MIL problems, but is also very computationally intensive.

Several regular supervised classifiers have been extended to work in the MIL setting. One example is mi-SVM [12], an extension of support vector machines which attempts to find hidden labels of the instances under constraints, such as (1) or (2), posed by the bag labels. Another example is MILBoost [13], where the instances are reweighed in each of the boosting rounds. The instance weights indicate how informative the instances are in predicting the bag labels.

It has been recognized that the standard assumption might be too strict for certain types of MIL problems. Therefore, relaxed assumptions have emerged [23], where a fraction or a particular number of positive instances are needed to satisfy a concept, and where multiple concept regions are considered. In the case of COPD, this would correspond to the presence of a certain fraction of ROIs containing affected tissue, and/or different types of disease patterns. However, if the number of concepts and the fraction of positives per concept, are not given in advance, these extra parameters also need to be set using the training data, further increasing the risk of overtraining.

Therefore, methods which compare bags without explicitly relying on the standard, or relaxed assumptions, have been proposed. Such methods include Citation-kNN [24], and bag kernels [14]. Citation-kNN uses the Hausdorff distance between bags. For classification, both the k_R "referencing" nearest neighbors of a bag B, and the k_C "citing" neighbors (bags for which B is nearest neighbor) are taken into account. In [14], a bag kernel is defined either as a sum of the instance kernels, or as a standard (linear or radial basis) kernel on a summarized representation of the bag. This summary is created by, for each feature, averaging the bag's instances (which we refer to by mean-inst), or using both the minimum and the maximum instance values (which we refer to by extremes). In all cases, the way a kernel is defined affects which (implicit) assumptions are made about the problem. A drawback for real-world applications is that kernels must be positive semi-definitive, therefore excluding some domain-specific similarity functions.

Other bag-based methods have addressed MIL by representing each bag by (dis)similarities to a set of prototypes $\mathcal{R} = \{R_1, \ldots, R_M\}$ in a so-called dissimilarity space [25]. Therefore, each bag is represented by a single feature vector $\mathbf{d}(B_i, \mathcal{R}) = [d(B_i, R_1), \ldots, d(B_i, R_M)]$, where d is a (dis)similarity measure. In this space, any supervised classifier can be used. In MILES [26], all training instances are used as prototypes, creating a very high-dimensional representation. A sparse classifier is used to select the most discriminative similarities, and therefore, instances. In a bag-of-words approach, prototypes are "words", or clusters of instances, and the dissimilarity measure between a bag and a word is the number of instances, belonging to that cluster.

Using bags as prototypes [15], [17] reduces the dimensionality, and therefore, the possibility of overtraining. In this paper, we use all training bags as prototypes, i.e., $\mathcal{R} = \mathcal{X}_{tr}$, but prototype selection can be further used to reduce $|\mathcal{R}|$. The advantage of such methods is that more can be gained from the training data than in nearest neighbor approaches, and that there are no restrictions on the bag similarity function [27]. In this paper, for example, we use two definitions of d that are not necessarily metric: the average minimum instance distance (3), and the earth mover's distance (EMD) [28], defined in (4). Herein, the instance dissimilarity d is the squared Euclidean distance.

$$d_{meanmin}(B_i, B_j) = \frac{1}{n_i} \sum_{k=1}^{n_i} \min_l d(\mathbf{x}_{ik}, \mathbf{x}_{jl}) \quad (3)$$

$$d_{\text{EMD}}(B_i, B_j) = \sum_{\mathbf{x}_k \in B_i, \mathbf{x}_l \in B_j} f(\mathbf{x}_k, \mathbf{x}_l) d(\mathbf{x}_k, \mathbf{x}_l) \quad (4)$$

where $f(\mathbf{x}_k, \mathbf{x}_l)$ is the flow that minimizes the overall distance, and that is subject to constraints that ensure that the only available amounts of "earth" (instances of B_i) are transported into available "holes" (instances of B_j), and that all of the instances are indeed transported: $f(\mathbf{x}_k, \mathbf{x}_l) \geq 0$, $\sum_{\mathbf{x}_k \in B_i} f(\mathbf{x}_k, \mathbf{x}_l) \leq 1/n_j$, $\sum_{\mathbf{x}_l \in B_j} f(\mathbf{x}_k, \mathbf{x}_l) \leq 1/n_i$ and $\sum_{\mathbf{x}_k \in B_i, \mathbf{x}_l \in B_j} f(\mathbf{x}_k, \mathbf{x}_l) = 1$.

III. Experiments

We use the dataset from [5], which describes how CT lung images from the Danish Lung Cancer Screening Trial [21] have been processed. Parts of such images highlighting healthy and emphysemous lung tissue, are shown in Figure 1.

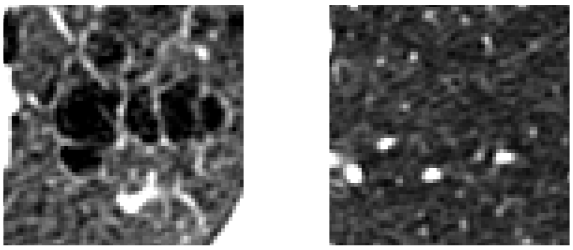

Fig. 1. Examples of patches containing centrilobular emphysema (left), characterized by black holes within the lung tissue, and healthy tissue (right). Both images are approximately 1.5 times the size of the ROIs used for classification and the intensity values have been rescaled to facilitate viewing.

The dataset consists of three parts: training set $\mathcal{X}_{tr}$, validation set $\mathcal{X}_{val}$ and test set $\mathcal{X}_{te}$. Originally, each of these parts consists of 100 COPD (positive) and 100 healthy (negative) images. In previous work [5], a subset of the training data with 31 COPD and 31 healthy images was selected to improve the class separability in the training set. We therefore refer to the full training data as $\mathcal{X}_{tr}$ and to the subsampled training data as $\mathcal{X}_{sub}$.

Each image is represented by 50 ROIs, sampled at random locations within the lungs. Each ROI is described by histograms of responses of 8 filters at 7 scales, which aim to capture the texture of the image. The filters used the following: Gaussian, gradient magnitude, Laplacian of Gaussian, first, second and third eigenvalue of the Hessian, Gaussian curvature and eigen magnitude. The scales range from 0.6 to 4.8 mm. The responses of each filter at each scale are stored in a histogram with 41 bins. This approach creates a 2296-dimensional feature vector for each ROI. In [5], the validation set was used to select the most appropriate filters and scales. Because these features are selected for a particular classifier only (SimpleMIL with nearest neighbor classifier), we use the full feature set here for all the classifiers.

The evaluated classifiers are available from the MIL toolbox [29] and PRTools [30]. We evaluate the following selection:

- SimpleMIL with a logistic (regularization parameter $C \in \{0.01, 0.1, 1, 10\}$) and nearest neighbor ($k \in \{25, 35, 45\}$) classifiers. We consider both noisy-or and average fusion rules.
- EM-DD with 10% of instances used for initialization
- miSVM with a polynomial kernel, where $p \in \{1, 2\}$ is the degree of the polynomial and $C \in \{0.01, 0.1, 1, 10\}$ is a regularization parameter. We consider both noisy-or and average fusion rules.
- MILBoost with 100 reweighting rounds
- Citation k-NN, $k_R \in \{1, 5, 10\}$, $k_C \in \{1, 5, 10\}$
- Averaging the instances (mean-inst), and minimum and maximum feature values for each bag (extremes) with an SVM, $p \in \{1, 2\}$, $C = \{0.01, 0.1, 1, 10\}$.
- Bag-of-words (BoW) with {50, 100, 200} words and an SVM, $p \in \{1, 2\}$, $C \in \{0.01, 0.1, 1, 10\}$
- MILES with a polynomial kernel, $p \in \{1, 2\}$, $C \in \{0.01, 0.1, 1, 10\}$
- Bag dissimilarities (meanmin and emd) with k-NN, $k \in \{1, 5, 10\}$, and in the dissimilarity space with an SVM, $p \in 1, 2$, $C = \{0.01, 0.1, 1, 10\}$

We perform evaluation in three ways. First, each classifier (with different parameter settings) is trained on $\mathcal{X}_{sub}$ and $\mathcal{X}_{tr}$, depending on the experiment. Each classifier is then evaluated on $\mathcal{X}_{val}$. The evaluation metric is the area under the receiver-operating characteristic curve, or AUC. We report the best of these performances on $\mathcal{X}_{val}$, and select the corresponding parameters. We then report the performance of this classifier with the best parameters on an independent test set $\mathcal{X}_{te}$. The difference in AUC on $\mathcal{X}_{val}$ and $\mathcal{X}_{te}$ is an indicator of overtraining, i.e., fitting the parameters too well to the validation set. Lastly, we randomly select half of the bags in $\mathcal{X}_{sub}$ or $\mathcal{X}_{tr}$, 10 times. For each subsample, we train a classifier, select parameters using $\mathcal{X}_{val}$, and evaluate on $\mathcal{X}_{te}$. The average and the standard deviations of the 10 performances are reported. This result gives an indication of a situation where less training data is available, and of the variance in performance due to a different sampling of the data.

The performances are shown in Table 1. For each training dataset, we compare the performances per column. The best performance and performances not significantly worse than best, are shown in bold. We test for significant differences using the DeLong test for ROC curves [31] for the performances on $\mathcal{X}_{val}$ and $\mathcal{X}_{te}$, and using a dependent t-test for the 10 cross-validation performances, both at a significance level of 0.05. A few results are not reported. For EM-DD, time requirements were too high for both datasets. For miSVM, the instance kernel matrix for $\mathcal{X}_{tr}$ was too large to fit in memory.

IV. Discussion

A. Classifier Performance

Across the different training datasets, we can see similar trends in the classifier performances. It is clear that some

TABLE I. AUC PERFORMANCES ($\times 100$) OF MIL CLASSIFIERS, TRAINED ON $\mathcal{X}_{sub}$ (TOP) AND $\mathcal{X}_{tr}$ (BOTTOM). FROM LEFT TO RIGHT: BEST PARAMETERS ON $\mathcal{X}_{val}$, SAME PARAMETERS ON $\mathcal{X}_{te}$, MEAN(STD) WHEN SUBSAMPLING $\mathcal{X}_{sub}$ OR $\mathcal{X}_{tr}$ TO 50% 10 TIMES

	Trained on $\mathcal{X}_{sub}$		
Classifier	AUC $\mathcal{X}_{val}$	AUC $\mathcal{X}_{te}$	10x AUC $\mathcal{X}_{te}$
Simple logistic noisy	50.0	50.0	50.2 (0.7)
Simple logistic avg	**71.9**	**70.5**	67.9 (1.3)
Simple k-NN noisy	61.0	65.9	63.7 (2.3)
Simple k-NN avg	67.0	67.8	66.0 (1.5)
miSVM noisy	**69.7**	65.4	62.0 (3.1)
miSVM avg	**74.5**	**71.7**	69.4 (1.5)
MILBoost	55.8	61.4	59.3 (10.2)
Citation k-NN	65.2	61.5	63.5 (1.5)
mean-inst SVM	**74.0**	**74.2**	**72.3 (2.7)**
extremes SVM	**70.8**	68.6	68.3 (2.7)
BoW SVM	50.0	50.0	50.0 (0.0)
MILES	65.8	68.2	64.3 (4.2)
meanmin SVM	70.8	**71.3**	69.6 (2.1)
meanmin k-NN	65.0	**69.1**	65.7 (1.6)
emd SVM	**73.7**	**74.6**	69.3 (3.3)
emd k-NN	65.1	67.1	64.6 (1.8)

	Trained on $\mathcal{X}_{tr}$		
Classifier	AUC $\mathcal{X}_{val}$	AUC $\mathcal{X}_{te}$	10x AUC $\mathcal{X}_{te}$
Simple logistic noisy	60.9	60.7	50.0 (0.0)
Simple logistic avg	**73.5**	**75.8**	72.3 (3.1)
Simple k-NN noisy	64.3	68.2	66.9 (2.1)
Simple k-NN avg	66.8	69.7	68.5 (0.8)
MILBoost	54.6	54.3	62.3 (7.8)
Citation k-NN	65.9	56.9	60.4 (2.4)
mean-inst SVM	**77.2**	**77.6**	**76.5 (3.8)**
extremes SVM	**73.1**	65.2	67.2 (1.2)
BoW SVM	50.0	50.0	50.0 (0.0)
MILES	50.0	50.0	67.6 (2.5)
meanmin SVM	**74.0**	**75.4**	73.8 (2.6)
meanmin k-NN	59.0	53.5	53.5 (4.6)
emd SVM	**74.2**	72.9	**75.1 (2.7)**
emd k-NN	63.9	54.4	51.2 (4.3)

classifiers suffer from the high dimensionality. For example, the bag of words approach, which uses a mixture of Gaussians to estimate words in feature space, is not able to do so in 2296 dimensions. For BoW, MILBoost, and EM-DD which could not handle the dimensionality computationally, we performed additional experiments with the 287-dimensional feature set that resulted from a feature selection procedure used in [5]. The results on $\mathcal{X}_{te}$ are 0.657 (BoW), 0.641 (EM-DD) and 0.551 (MILBoost), suggesting that these classifiers benefit from feature selection. It may of course generally be interesting to study feature selection for the other classifiers as well.

The full training dataset $\mathcal{X}_{tr}$ has two main differences with respect to $\mathcal{X}_{sub}$: higher class overlap, and more bags, and therefore instances in total. Several classifiers, such as SimpleMIL logistic, mean-inst, extremes, and SVM in the dissimilarity space, show increases in performances due to the higher sample size. On the other hand, MILES suffers from the increased sample size, because the dimensionality of the dissimilarity representation is equal to the number of instances. This also explains why MILES performs better when the training set is subsampled to 50%.

The performances of many methods do not degrade very much when only 50% of the bags are used for training. This suggests that the subsampled dataset is still representative for the whole data distribution, and each class can be described well with only a few samples. Furthermore, most classifiers do not suffer a lot from overtraining, as the difference in performance on $\mathcal{X}_{val}$ and on $\mathcal{X}_{te}$ is quite small. Notable exceptions are the k-NN classifiers trained on the full training set $\mathcal{X}_{tr}$, where the parameter k is overfit to the validation set, causing lower performances on $\mathcal{X}_{te}$.

SimpleMIL performs quite well, especially when posterior probabilities of all instances are taken into account, as in the averaging fusion rule. Methods which assume a concept, such as EM-DD and miSVM, also perform reasonably, which suggests that there is a region in feature space with a high density of disease patches and low density of normal patches. However, the performances are lower than those of bag-based methods, suggesting that detecting the concept is not sufficient for the diagnosis of COPD. This is also supported by the fact that miSVM with the averaging rule outperforms miSVM with the noisy or rule, which shows that it is beneficial to take all instance classifications into account.

Methods with assumptions on bag level have the best performances, in particular, averaging all the instances in a bag is already able to separate the bag classes quite well. This suggests that negative instances in positive bags, and the negative instances in negative bags, do not originate from the same distribution. In other words, scans affected with COPD do not contain the same types of healthy patches, as healthy scans. The disease appears to be more diffuse, affecting a large part of the lung rather than small isolated regions.

For the bag-based methods, the mean-inst bag representation and the dissimilarity-based SVM perform particularly well. MILES suffers from the high dimensionality, but we expect that the performance would improve if instance selection techniques would be used. Another interesting observation is that the dissimilarity-based SVM significantly outperforms k-NN on the same dissimilarities. SVM is able to use the dissimilarities of the training set to create a more robust classifier, which is consistent with results in [18], although slightly different dissimilarities are used there. We expect that further investigation into different bag dissimilarity measures could further improve these results.

Unfortunately, our results are not directly comparable to the dissimilarity-based approach of [18], because an earlier version of the dataset was used. The results in [5], however, are obtained by training on $\mathcal{X}_{sub}$ and the performances can be compared. There, the best approach obtains an AUC of 0.713. Our results show superior performances when training on $\mathcal{X}_{sub}$, with an AUC 0.742 for mean-inst and 0.746 for d_{EMD} in the dissimilarity space. However, these performances are not significantly better than the result in [5]. Using $\mathcal{X}_{tr}$ further improves the results, for an AUC of 0.758 for SimpleMIL with a logistic classifier, 0.776 for mean-inst, and 0.754 for $d_{meanmin}$ in the dissimilarity space. The best approach using $\mathcal{X}_{tr}$ in [5] obtains an AUC of 0.690, and our performances are significantly better according to the DeLong test.

Furthermore, we examined the output of the best-performing classifiers to see which images still get misclassified. Rather than only looking at the positive label for COPD, we now use the COPD stages [2], from mild (I) to moderate (III). The results show that most of the confusion is between the healthy scans, and stage I scans, which supports our intuition about where the class overlap is largest. Because the classifiers differ in some of their errors, it may be of interest to combine their decisions.

B. Concept Region

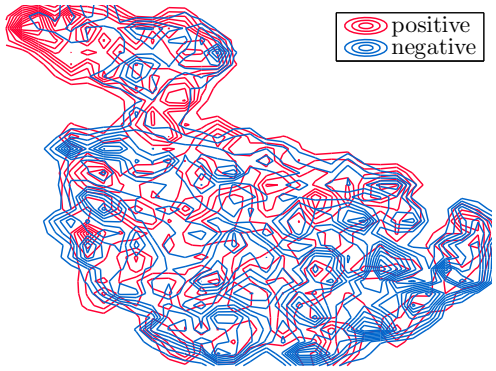

Fig. 2. Density contours of t-SNE projection of instances of $\mathcal{X}_{sub}$

In order to better understand the classifier performances, we examine a 2D projection of the instances, obtained by t-distributed stochastic neighbor embedding (t-SNE) [32] (Figure 2). We see two clusters of instances, a smaller cluster in the top left and a larger cluster. In the small cluster the density of instances from positive bags is clearly higher, which suggests that part of it could be a concept region. To investigate whether these patches display emphysema, we examined the intensity histograms of the Gaussian filters at the smallest scale. As emphysema results in darker patches, we would expect patches with emphysema to have intensity histograms skewed to the left. This is exactly what we find when averaging all the instances per cluster and plotting the two corresponding histograms in Figure 3.

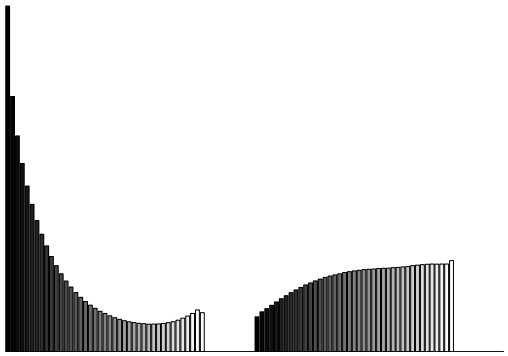

Fig. 3. Histograms of Gaussian filter responses at the scale 0.6, for the averaged instances in the two clusters found in the t-SNE plot.

Visual inspection of patches from both the small cluster and from the lower right part of the big cluster in Figure 2 confirmed the tendency we saw in the average Gaussian filter response histograms in Figure 3. The patches in the small cluster were generally affected by emphysema whereas the patches in the lower part of the big cluster showed no or only faint signs of emphysema.

It is important to note that the dataset mainly contains mild to moderate COPD patients, and no patients with very severe emphysema. We expect that if this was the case, the concept or concepts would be more pronounced.

C. Interpretability

Next to the classifier performances, it is important to consider how these classifiers would be used in a medical setting. Despite slightly lower performances, instance-based methods are of interest because of their ability to provide instance labels for the ROIs. An expert could then inspect the instance labels in different regions of the lungs, allowing for better diagnosis or treatment planning. The instance labels, however, should be used with caution. Specialized MIL (i.e., except SimpleMIL) methods are trained to classify bags correctly, not instances, and the best bag classifier is not necessarily the best instance classifier [33]. Therefore, correct instance labels would be sacrificed for the greater good of correct bag labels.

Although bag-based methods perform better, their interpretability may be more difficult. For example, the average histograms (as in mean-inst) separate the classes very well, but this method can not provide information on how the affected tissue is distributed within the lungs, which could be important for determining the best treatment as well as for monitoring disease progression and therapy effect.

Dissimilarity-based methods provide more opportunities in terms of interpretability compared to mean-inst or extremes. For these methods, we can investigate which prototypes, i.e. CT images, patch clusters or individual patches, correspond to typical healthy or COPD cases. By using linear classifiers in the dissimilarity space, the diagnosis would be explained in terms of a linear combination of dissimilarities to such prototypes.

V. Conclusions

We have studied the possibility of classifying COPD by means of various classical and more recent MIL approaches. The study revealed that MIL offers classification methods for this problem that are potentially better than the techniques previously proposed. The diversity of methods also enabled us to reason about the nature of COPD as a MIL problem. Although we found a concept region with patches showing typical disease patterns, considering the whole distribution of instances for bag classification improved the results. The best performing method is an SVM with a kernel based on the average instance per bag. This method obtains an AUC of 0.742 which is higher (but not significantly) than the previous best performance of 0.713 on the same dataset. By using the full training data we achieve a significantly higher AUC of 0.776.

References

[1] P. M. Calverley, "COPD: Early detection and intervention," *CHEST Journal*, vol. 117, no. 5_suppl_2, pp. 365S–371S, 2000.

[2] K. F. Rabe, S. Hurd, A. Anzueto, P. J. Barnes, S. A. Buist, P. Calverley, Y. Fukuchi, C. Jenkins, R. Rodriguez-Roisin, C. van Weel *et al.*, "Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: Gold executive summary," *American Journal of Respiratory and Critical Care Medicine*, vol. 176, no. 6, pp. 532–555, 2007.

[3] R. A. Pauwels, A. S. Buist, P. M. A. Calverley, C. R. Jenkins, and S. S. Hurd, "Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease," *Am. J. of Respir. and Crit. Care Med.*, vol. 163, no. 5, 2012.

[4] Y. S. Park, J. B. Seo, N. Kim, E. J. Chae, Y. M. Oh, S. Do Lee, Y. Lee, and S.-H. Kang, "Texture-based quantification of pulmonary emphysema on high-resolution computed tomography: comparison with density-based quantification and correlation with pulmonary function test," *Investigative Radiology*, vol. 43, no. 6, pp. 395–402, 2008.

[5] L. Sørensen, M. Nielsen, P. Lo, H. Ashraf, J. H. Pedersen, and M. de Bruijne, "Texture-based analysis of COPD: a data-driven approach," *IEEE Transactions on Medical Imaging*, vol. 31, no. 1, pp. 70–78, 2012.

[6] L. Sørensen, S. B. Shaker, and M. de Bruijne, "Quantitative analysis of pulmonary emphysema using local binary patterns," *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 559–569, 2010.

[7] R. Uppaluri, T. Mitsa, M. Sonka, E. A. Hoffman, and G. McLennan, "Quantification of pulmonary emphysema from lung computed tomography images," *American Journal of Respiratory and Critical Care Medicine*, vol. 156, no. 1, pp. 248–254, 1997.

[8] C. S. Mendoza, G. R. Washko, J. C. Ross, A. A. Diaz, D. A. Lynch, J. D. Crapo, E. K. Silverman, B. Acha, C. Serrano, and R. S. J. Estepar, "Emphysema quantification in a multi-scanner HRCT cohort using local intensity distributions," in *International Symposium on Biomedical Imaging*, 2012, pp. 474–477.

[9] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[10] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in Neural Information Processing Systems*, 1998, pp. 570–576.

[11] Q. Zhang, S. A. Goldman *et al.*, "EM-DD: An improved multiple-instance learning technique," in *Advances in Neural Information Processing Systems*, 2001, pp. 1073–1080.

[12] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems*, 2002, pp. 561–568.

[13] P. Viola, J. Platt, and C. Zhang, "Multiple instance boosting for object detection," in *Advances in Neural Information Processing Systems*, 2006, pp. 1417–1424.

[14] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, "Multi-instance kernels," in *International Conference on Machine Learning*, 2002, pp. 179–186.

[15] D. M. J. Tax, M. Loog, R. P. W. Duin, V. Cheplygina, and W. J. Lee, "Bag dissimilarities for multiple instance learning," in *Similarity-Based Pattern Recognition*, 2011, pp. 222–234.

[16] M. L. Zhang and Z. H. Zhou, "Multi-instance clustering with applications to multi-instance prediction," *Applied Intelligence*, vol. 31, no. 1, pp. 47–68, 2009.

[17] V. Cheplygina, D. M. J. Tax, and M. Loog, "Does one rotten apple spoil the whole barrel?" in *International Conference on Pattern Recognition*, 2012, pp. 1156–1159.

[18] L. Sørensen, M. Loog, P. Lo, H. Ashraf, A. Dirksen, R. P. W. Duin, and M. de Bruijne, "Image dissimilarity-based quantification of lung disease from CT," in *Medical Image Computing and Computer-Assisted Intervention*, 2010, pp. 37–44.

[19] L. Sørensen, P. Lo, A. Dirksen, J. Petersen, and M. De Bruijne, "Dissimilarity-based classification of anatomical tree structures," in *Information Processing in Medical Imaging*, 2011, pp. 475–485.

[20] A. Feragen, J. Petersen, D. Grimm, A. Dirksen, J. H. Pedersen, K. Borgwardt, and M. de Bruijne, "Geometric tree kernels: Classification of COPD from airway tree geometry," in *Information Processing in Medical Imaging*, 2013, pp. 171–183.

[21] J. H. Pedersen, H. Ashraf, A. Dirksen, K. Bach, H. Hansen, P. Toennesen, H. Thorsen, J. Brodersen, B. G. Skov, M. Døssing *et al.*, "The Danish randomized lung cancer CT screening trial-overall design and results of the prevalence round," *Journal of Thoracic Oncology*, vol. 4, no. 5, pp. 608–614, 2009.

[22] M. Loog and B. van Ginneken, "Static posterior probability fusion for signal detection: applications in the detection of interstitial diseases in chest radiographs," in *International Conference on Pattern Recognition*, vol. 1, 2004, pp. 644–647.

[23] N. Weidmann, E. Frank, and B. Pfahringer, "A two-level learning method for generalized multi-instance problems," in *Europearn Conference on Machine Learning*, 2003, pp. 468–479.

[24] J. Wang and J. D. Zucker, "Solving multiple-instance problem: A lazy learning approach," in *International Conference on Machine Learning*, 2000, pp. 1119–1125.

[25] E. Pękalska and R. P. W. Duin, *The dissimilarity representation for pattern recognition: foundations and applications*. World Scientific Pub Co Inc, 2005, vol. 64.

[26] Y. Chen, J. Bi, and J. Z. Wang, "MILES: multiple-instance learning via embedded instance selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1931–1947, 2006.

[27] E. Pękalska, A. Harol, R. P. W. Duin, B. Spillmann, and H. Bunke, "Non-euclidean or non-metric measures can be informative," in *Structural, Syntactic, and Statistical Pattern Recognition*, 2006, pp. 871–880.

[28] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[29] D. M. J. Tax, "MIL, a MATLAB toolbox for Multiple Instance Learning," May 2011, version 0.7.9. [Online]. Available: http://prlab.tudelft.nl/david-tax/mil.html

[30] R. P. W. Duin, P. Juszczak, P. Paclik, E. Pękalska, D. de Ridder, D. M. J. Tax, and S. Verzakov, "PRTools, a MATLAB toolbox for pattern recognition," http://www.prtools.org, 2010.

[31] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, pp. 837–845, 1988.

[32] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.

[33] V. Tragante do O, D. Fierens, and H. Blockeel, "Instance-level accuracy versus bag-level accuracy in multi-instance learning," in *Proceedings of the 23rd Benelux Conference on Artificial Intelligence (BNAIC)*, 2011, p. 8.