ASYMMETRIC SIMILARITY-WEIGHTED ENSEMBLES FOR IMAGE SEGMENTATION

V. Cheplygina[*†] *A. van Opbroek*[*] *M. A. Ikram*[‡] *M. W. Vernooij*[‡] *M. de Bruijne*[*§]

[*] Biomedical Imaging Group Rotterdam, Dept. Medical Informatics and Radiology, Erasmus Medical Center, The Netherlands
[†] Pattern Recognition Laboratory, Delft University of Technology, The Netherlands
[‡] Dept. Epidemiology and Radiology, Erasmus Medical Center, The Netherlands
[§] The Image Section, Dept. Computer Science, University of Copenhagen, Denmark

ABSTRACT

Supervised classification is widely used for image segmentation. To work effectively, these techniques need large amounts of labeled training data, that is representative of the test data. Different patient groups, different scanners or different scanning protocols can lead to differences between the images, thus representative data might not be available. Transfer learning techniques can be used to account for these differences, thus taking advantage of all the available data acquired with different protocols.

We investigate the use of classifier ensembles, where each classifier is weighted according to the similarity between the data it is trained on, and the data it needs to segment. We examine 3 asymmetric similarity measures that can be used in scenarios where no labeled data from a newly introduced scanner or scanning protocol is available. We show that the asymmetry is informative and the direction of measurement needs to be chosen carefully. We also show that a point set similarity measure is robust across different studies, and outperforms state-of-the-art results on a multi-center brain tissue segmentation task.

***Index Terms*—** Transfer learning, similarity measure, asymmetry, tissue segmentation

1. INTRODUCTION

Manual biomedical image segmentation is time-consuming, and thus in recent years a lot of advances have been made to automate this process. Because of its good performance, supervised voxelwise classification [1, 2, 3, 4], where manually labeled images are used to train supervised classifiers, has been used successfully in many applications. However, supervised classifiers need large amounts of labeled source data that is representative of the target data in order to be successful. In multi-center studies or longitudinal studies where different scanning protocols are used, these requirements are often not fulfilled, leading to poor performances.

Fortunately, *transfer learning* [5] techniques can be employed in order to deal with the differences between source and target data. Such approaches frequently rely on a small amount of labeled target data ([2, 6, 7], to name a few), or can be unsupervised with respect to the target [1, 8], which is favorable for tasks where annotation is costly. In the latter case, typically the transfer is achieved by weighing the training samples such that the differences between training and target data are minimized. However, this means that adding novel scanners or scanning protocols is time-consuming, because the classifiers need to be retrained.

We propose to approach voxelwise classification by similarity-weighted ensembles. The classifiers are trained *only once*, each on a different source image. For a target image, the classifier outputs are fused by weighted averaging, where the weights are determined by the similarity of the source image and the target image. Similarity-weighting of ensembles is shown to outperform a classifier trained on all data [9], but unfortunately, the definition of similarity in this case is supervised, i.e., labeled target data needs to be available. We investigate several similarity measures that do not have this requirement. We show that an unsupervised weighting scheme based on an asymmetric point set distance is able to achieve good performances. On a brain tissue segmentation dataset, our method outperforms state-of-the-art, while being computationally efficient, and readily extendable to novel target images.

1.1. Related Work

[3] propose a classifier-based approach to multi-atlas segmentation, where each atlas is encoded by a random forest, therefore leading to a technique called atlas forest (AF). At test time, each forest outputs posterior probabilities, which are averaged across all forests. In later papers [4, 10], target-specific atlas selection of AF is investigated. Our approach is conceptually similar to AF in that we construct individual classifiers for each training image. Our approach is different because it focuss on images with different feature distributions. Because [4] uses class probabilities based on a model of intensities of all images as additional inputs, differences in feature distributions would introduce additional class over-

lap. Furthermore, these inputs would change if more data becomes available, and retraining the classifiers would be necessary. Similarly, [10] requires transforming the training data to determine the similarities of the images, and including additional source domains would require full retraining.

[11, 1] propose a classifier-based approach for brain tissue segmentation, but in a setting where the training images originate from different scanning protocols and the feature distributions are therefore different. The approach overcomes these differences by weighting the training images such that a divergence, such as Kullback-Leibler (KL), of the training and test distributions is minimized. These image weights are then used to weight the samples before training a support vector machine (SVM). We consider the same setting, but our strategy is different: we train the classifiers only once, independent of the test image, and determine only classifier weights at test time. Our approach therefore does not need re-training for each target image.

While classifiers that are adapted to the test data have been investigated in different communities, the question of how the similarity definition affects the result is underexplored, in particular if the data is from different distributions. The novelty of our contribution lies in the comparison of different unsupervised metrics, which allow for on-the-fly handling of additional training or testing data.

2. SIMILARITY-WEIGHTED ENSEMBLES

We use training data from M different images, which may be acquired with different scanning protocols or from different studies. The m-th image is represented by N_m training samples $(\mathbf{x}_i^m, y_i^m)$, where $\mathbf{x}_i^m \in \mathbb{R}^n$ is a feature vector describing the local image structure around the voxel, and $y_i^m \in \{1, 2, \ldots, C\}$ is a class label provided by manual segmentation. The goal is to learn a classifier $F : \mathbb{R}^n \to \{1, 2, \ldots, C\}$ that can predict the labels y_i^z of the N_z voxels $\mathbf{x}_i^z$ of the target image. Without loss of generality we focus on a classifier $F^y : \mathbb{R}^n \to [0, 1]$ that may have posterior probabilities as outputs, indicating the probability of a sample belonging to class y, and which can easily be extended to a multi-class classifier by one-vs-one or one-vs-all approaches.

We propose to use an ensemble of M classifiers, where each base classifier $\{f_1, \ldots, f_M\}$ is trained on samples from a different image. Herein f_m^y stands for the m-th classifier's output for class y. The final ensemble F is determined by a weighted average of the posterior probabilities, i.e. $F(\mathbf{x}_i^z) = \sum_{m=1}^{M} w_{mz} f_m(\mathbf{x}_i^z))$. The weights are determined per pair of training image and target image, and are inversely related to a dissimilarity d_{mz} between the images:

$$w_{mz} = (d_{max} - d_{mz})^p / \sum_{m=1}^{M} (d_{max} - d_{mz})^p \qquad (1)$$

where $d_{max} = \max_m\{d_{mz}\}$ and p is a parameter that influences the scaling of the weights (with high p, highly dissimilar images are downweighed more). In what follows, we examine four measures to define d_{mz}, which can be seen as supervised (i.e., requiring labeled data from the target domain) and unsupervised.

Supervised Similarity. We use the mean square error (MSE) of the posterior probabilities, because it distinguishes between classifiers that are slightly or very inaccurate. We denote this ensemble by F^{sup}, the corresponding dissimilarity is defined as:

$$d_{mz}^{sup} = \sum_{(\mathbf{x}_i^z, y_i^z)} (1 - f_m^y(\mathbf{x}_i^z))^2. \qquad (2)$$

Clustering Similarity. In the absence of labels $\{y_i^z\}$, we estimate the target labels using a clustering procedure. In MRI brain tissue segmentation, the tissue classes can be determined by examining the average intensity within each cluster. Therefore we can define d_{mz}^{clu} by performing an unsupervised, 3-class clustering and substituting the clustering labels c_i^z into (2), i.e. computing the MSE over the pairs $(\mathbf{x}_i^z, c_i^z)$. We denote this ensemble by F^{clu}.

Distribution Similarity. The clustering approach depends on the classifier and clustering algorithm used. We also propose a classifier-independent approach, where the assumption is that if the probability density functions (PDF) of the source image $P_m(\mathbf{x})$ and target image $P_z(\mathbf{x})$ are similar, that the labeling functions $P_m(y|\mathbf{x})$ and $P_z(y|\mathbf{x})$ are also similar. We propose to evaluate the similarity of the PDFs with the Kullback-Leibler divergence, following the approach in [1]. The dissimilarities are proportional to the term

$$d_{mz}^{div} = -\frac{1}{N_z} \sum_{i=1}^{N_z} \log P_m(\mathbf{x}_i^z) \qquad (3)$$

where $P_m(\mathbf{x})$ is determined by kernel density estimation (KDE) on the samples $\{\mathbf{x}_i^m\}$. Note that in [1], these weights are used for weighting the samples prior to training the classifier, while we weight the classifier outputs. We denote this ensemble by F^{div}.

Point Set Similarity. Rather than viewing the voxels of each image as a distribution, we can view them as a discrete point set or *bag*. Both the advantage and the disadvantage of this approach is that KDE can be omitted: on the one hand, there is no need to choose a kernel width, on the other hand, outliers which would have been smoothed out by KDE may now greatly influence the results. A dissimilarity that characterizes such bags well even in high-dimensional situations [12] and is related to the Hausdorff distance is defined as:

$$d_{mz}^{bag} = \frac{1}{N_z} \sum_{i=1}^{N_z} \min_j ||\mathbf{x}_i^z - \mathbf{x}_j^m||^2. \qquad (4)$$

In other words, each voxel's feature vector in the target image is matched with its nearest neighbor in the feature space of the source image, and these nearest neighbor distances are averaged. We denote this ensemble by F^{bag}.

Asymmetry. All proposed measures are asymmetric. However, we are only allowed to compute both asymmetric versions for d^{bag} and d^{div} because unlike d^{clu} and d^{sup} they do not require labels. In (3) and (4), we are matching the target samples to the source samples. This ensures that for all samples in the target data, the classifier has been trained on similar source samples. Our hypothesis is that an ensemble with similarities based on matching target samples to the source samples will outperform an ensembles with similarities in the opposite direction.

3. EXPERIMENTS

3.1. Data

We use the brain tissue segmentation dataset from [1, 11], which includes 56 manually segmented MR brain images. The images originate from four sources (two from [13] and two from [14]), where some sources used multiple scanners, and from two different MR scanning sequences (T1 and HASTE-Odd, a sequence that resembles inversed T1). Subjects included both healthy young adults and elderly.

Initial normalization of images included bias-field correction, inversion of HASTE-Odd images, and rescaling the voxel intensities by [4,96]-th percentile range matching inside the manually annotated brain mask. We extract 100K voxels per image. The voxels are labeled as WM (white matter), GM (grey matter) or CSF (cerebrospinal fluid). Each voxel is described by 13 features: intensity, {intensity, gradient magnitude, absolute value of Laplacian of intensity} each after convolution with a Gaussian kernel with $\sigma = 1, 2, 3$ mm^3, and the 3D position of the voxel normalized for the size of the brain. Fig. 1 shows a 2D embedding of some voxels from sources 1 and 4, produced by the dimensionality reduction technique t-SNE [15]. Note the area in the top left where clusters of CSF voxels from the two images are quite dissimilar.

3.2. Experimental Setup

We train the 56 image classifiers only once, each on 10K samples from an image. We then perform four experiments. In each experiment, three studies are used as the source data, and the fourth one is the target data. For each target image, we use 10K samples to determine the weights, and 100K samples to evaluate the classifiers.

We determine the supervised weights for F^{sup} using the test image labels in order to show the best possible performance of the ensemble. All other weights are determined without using label information. For F^{clu}, we use the k-Means algorithm with $k = 3$. For scaling the weights, we use $p = 10$ to enlarge the differences between the worst and best classifiers, but in practice this parameter could be set by leave-one-source-out cross-validation on the training set. All experiments are performed with a random forest classifier (RF). We chose RF because of its inherent ability to handle multi-class problems, speed and reduced need of parameter optimization, but other classifiers could be easily used instead. We use 100 trees and otherwise default parameters (https://code.google.com/p/randomforest-matlab/).

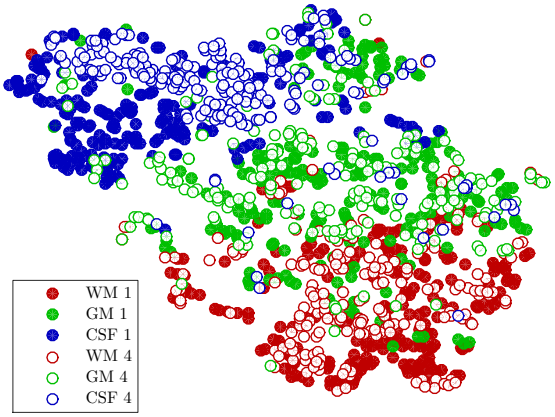

Fig. 1. Embedding of 600 voxels, where half are uniformly sampled from a source 1 image, and half are uniformly sampled from a source 4 image.

	Method				
	f^{all}	F^{uni}	F^{bag}	SPM8	WSVM
1	**9.5 (2.3)**	19.1 (1.0)	**11.5 (4.2)**	12.6 (2.0)	20.3 (4.9)
2	13.1 (1.1)	24.5 (1.2)	12.8 (2.6)	**10.0 (2.5)**	16.7 (2.6)
3	22.2 (2.7)	11.6 (1.3)	**11.5 (3.9)**	20.8 (3.4)	**10.6 (1.2)**
4	26.7 (8.4)	23.7 (7.6)	**16.3 (6.7)**	24.6 (2.1)	**16.2 (6.6)**
all	20.5 (8.2)	19.5 (7.3)	**13.5 (5.3)**	18.9 (6.4)	14.9 (5.4)

Table 1. Classification errors (mean and standard deviation, in %) of methods, unsupervised w.r.t. the target data. Last row shows average over all 56 images. Bold = best or not significantly worse (paired t-test, $\alpha < 0.05$) than best.

3.3. Results

The error rates of the different weight strategies are shown in Fig. 2. The much better performances of F^{sup} demonstrate that with suitable weights, very good performances are attainable. Note that F^{sup} is an oracle since it uses the target labels, and is only presented in order to get an impression of the best possible performances. For example, these results demonstrate that study 4 has two very atypical images, which cannot possibly be classified well.

Out of the unsupervised similarities, F^{clu} performs very poorly. F^{div} and F^{bag} perform better, and F^{bag} gives the best results overall. The asymmetric versions F^{bag} and F^{div} show similar trends. As we hypothesized, measuring the similarity

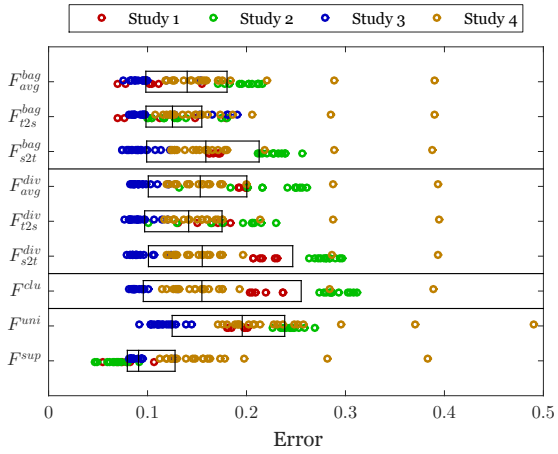

Fig. 2. Errors of different weighting strategies (rows). For the unsupervised measures, both directions and the symmetrized versions are shown. Images from different studies are indicated by color. Each plot shows the 25th, 50th and 75th quantiles of all image performances per method.

from the target to the source (*t2s*) samples, as in F^{bag}_{t2s} and F^{div}_{t2s}, outperforms the opposite direction. In the *t2s* case, a high weight assigned to a classifier means that for all samples in the target image, the classifier has seen similar samples (if such samples are present) during training. The *s2t* direction does not enforce this, so even with high similarity, a classifier could have no information about large regions of the target feature space. Interestingly, symmetrizing the similarity does not solve this problem, and only deteriorates the results. We therefore use the asymmetric version, F^{bag}_{t2s}, in further experiments.

We compare the average classification errors of F^{bag}_{t2s} to several baselines, which also do not use labeled target data, in Table 1. This outperforms uniform weights (F^{uni}) and, in 3 out of 4 cases, training a single classifier f^{all} on all the data. We also compare to the state-of-the-art brain tissue segmentation tool SPM8 [16], and the weighted SVM [1], which weights the training images by minimizing the KL divergence between training and test data, and trains an SVM. Comparing to these methods, our approach is the only one which provides reasonable results for all the four studies. When averaging over all the images, F^{bag}_{t2s} with a random forest classifier is significantly better than the other approaches.

4. DISCUSSION

We used random forests with default parameters throughout all experiments and did not investigate the effect of parameter selection on the results. The performances might improve if such parameters are set leave-one-source cross-validation on the training data. However, this means we would need to retrain the classifiers based on the target data. Furthermore, at test time (large number of voxels in practice) RFs are very efficient both in terms of time and storage.

An intriguing result is the effect of asymmetry on the similarity measures. Our results show that it is better to measure the similarity of the target data to the source data, and not the other way round. Symmetrizing the similarity also negatively affects performance. This suggests that in various similarity measures, the focus should be on the atypical samples in the target data - those which contribute to the from-target similarity. This could explain the slightly lower performances of the KL divergence, which smoothes out such outliers.

It would be interesting to investigate more similarity measures, which are unsupervised with respect to the target data. One possibility is STAPLE [17], which implicitly weights different candidate segmentations, thus following the approach of [4]. However, [18] show that with a large number of candidate segmentations, STAPLE's behaviour is similar to averaging, which would not be appropriate if the majority of segmented images are very different from the target image.

We applied our approach on brain tissue segmentation. However, two out of three similarity measures (including the best performing measure) do not use any prior knowledge about brain tissue. As such, our approach is not restricted to brain tissue segmentation, and can be applied to other tasks where the training and test data are expected to come from different distributions. Furthermore, all the investigated measures can be extended to be spatially dependent, such that the weights are not determined for whole images, but for local image patches.

5. CONCLUSIONS

In this study we proposed an ensemble approach to transfer learning. We investigated different similarity measures for a weighted combination of classifiers, each trained on a source domain that is dissimilar to the test data from the target domain. We showed that weighing the classifiers this way outperforms training a classifier on all the data, or assigning uniform weights for classifiers trained on different sources. An asymmetric bag similarity measure based on averaging the nearest neighbor distances between the feature vectors describing the voxels of the source and target images, performed best. The proposed ensemble can effectively combine heterogenous training data. The classifiers do not need retraining to segment novel target images. We therefore believe our approach will be useful for longitudinal or multi-center studies in which multiple protocols are used, and in clinical practice.

Acknowledgements. This research was performed as part of the research project "Transfer learning in biomedical image analysis" which is financed by the Netherlands Organization for Scientific Research (NWO) grant no. 639.022.010.

6. REFERENCES

[1] Annegreet Van Opbroek, M Arfan Ikram, Meike W Vernooij, and Marleen De Bruijne, "A transfer-learning approach to image segmentation across scanners by maximizing distribution similarity," in *MLMI*, pp. 49–56. Springer, 2013.

[2] Annegreet Van Opbroek, M Arfan Ikram, Meike W Vernooij, and Marleen De Bruijne, "Transfer learning improves supervised image segmentation across imaging protocols," *TMI*, 2014.

[3] D Zikic, B Glocker, and A Criminisi, "Encoding atlases by randomized classification forests for efficient multi-atlas label propagation," *MedIA*, vol. 18, no. 8, pp. 1262–1273, 2014.

[4] Darko Zikic, Ben Glocker, and Antonio Criminisi, "Classifier-based multi-atlas label propagation with test-specific atlas weighting for correspondence-free scenarios," in *Medical Computer Vision: Algorithms for Big Data*, pp. 116–124. Springer, 2014.

[5] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[6] Carlos Becker, C Christoudias, and Pascal Fua, "Domain adaptation for microscopy imaging," *TMI, in press*, 2014.

[7] Bo Cheng, Daoqiang Zhang, and Dinggang Shen, "Domain transfer learning for MCI conversion prediction," in *MICCAI*, pp. 82–90. Springer, 2012.

[8] Tobias Heimann, Peter Mountney, Matthias John, and Razvan Ionasec, "Real-time ultrasound transducer localization in fluoroscopy images by transfer learning from synthetic training data," *MedIA*, vol. 18, no. 8, pp. 1320–1328, 2014.

[9] Haixun Wang, Wei Fan, Philip S Yu, and Jiawei Han, "Mining concept-drifting data streams using ensemble classifiers," in *Knowledge Discovery and Data mining*. ACM, 2003, pp. 226–235.

[10] Herve Lombaert, Darko Zikic, Antonio Criminisi, and Nicholas Ayache, "Laplacian forests: Semantic image segmentation by guided bagging," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*, pp. 496–504. Springer, 2014.

[11] Annegreet van Opbroek, Meike W Vernooij, M Arfan Ikram, and Marleen de Bruijne, "Weighting training images by maximizing distribution similarity for supervised segmentation across scanners," *Medical image analysis*, vol. 24, no. 1, pp. 245–254, 2015.

[12] Veronika Cheplygina, David M J Tax, and Marco Loog, "Multiple instance learning with bag dissimilarities," *Pattern Recognition*, vol. 48, no. 1, pp. 264–275, 2015.

[13] M Arfan Ikram, Aad van der Lugt, Wiro J Niessen, Gabriel P Krestin, Peter J Koudstaal, Albert Hofman, Monique MB Breteler, and Meike W Vernooij, "The Rotterdam Scan Study: design and update up to 2012," *European journal of epidemiology*, vol. 26, no. 10, pp. 811–824, 2011.

[14] "Internet brain segmentation repository," `http://www.nitrc.org/projects/ibsr`.

[15] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, pp. 85, 2008.

[16] John Ashburner and Karl J Friston, "Unified segmentation," *Neuroimage*, vol. 26, no. 3, pp. 839–851, 2005.

[17] Simon K Warfield, Kelly H Zou, and William M Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation," *TMI*, vol. 23, no. 7, pp. 903–921, 2004.

[18] Koen Van Leemput and Mert R Sabuncu, "A cautionary analysis of STAPLE using direct inference of segmentation truth," in *MICCAI*, pp. 398–406. Springer, 2014.