Learning from Small Amounts of Labeled Data in a Brain Tumor Classification Task

Michael Goetz
Medical Image Computing (MIC)
German Cancer Research Center (DKFZ)
Heidelberg, Germany
m.goetz@dkfz.de

Christian Weber
Medical Image Computing (MIC)
German Cancer Research Center (DKFZ)
Heidelberg, Germany
ch.weber@dkfz.de

Bram Stieltjes
Department of Radiology
University Hospital Basel
Basel, Switzerland
bram.stieltjes@usb.ch

Klaus Maier-Hein
Medical Image Computing (MIC)
German Cancer Research Center (DKFZ)
Heidelberg, Germany
K.Maier-Hein@dkfz.de

Abstract

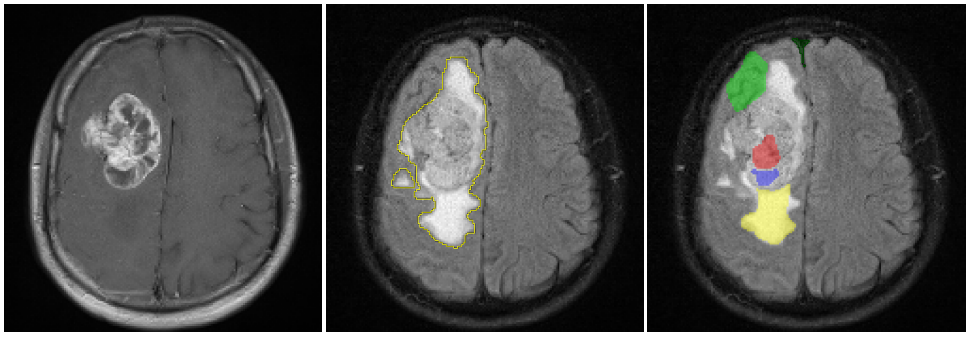
Current learning-based brain tumor classification methods show good performance but require large datasets of manually annotated training examples. Since image acquisition hardware and setup vary from clinic to clinic, training has to be repeated and the required time-consuming labeling effort limits a wider applicability of these approaches in clinical routine. We propose an approach that allows labelling of only small and unabiguous parts of the training data. Domain adaptation is applied to correct for the induced sampling error. We validated our approach using multimodal MR-scans of 19 patients and showed that our approach reduces the labeling time significantly while giving results that closely match those from a fully annotated training set. This is an important step towards bringing automatic tumor segmentation into clinical routine.

1 Introduction

Manual segmentation of tumors in the context of treatment planning or therapy control is time-consuming and error-prone. It often requires the simultaneous consideration of complex imaging features and partial volume effects (blurry and unclear borders) in multiple 3D images. Mazzara *et al.*, for example, reported that it takes between 20 min and 1 hour to label a 3D-MR-scan that contains a malignant glioma – the most common primary brain tumor [1]. They also reported an intra-rater and inter-rater volume variability of $20 \pm 15\%$ and $28 \pm 12\%$ respectively. Menze *et al.* reported that they needed about 4 hour to label a single training patient [2]. An automated segmentation can reduce the work-load while giving more consistent segmentations [1].

Automated machine learning-based methods were previously shown to successfully learn glioma appearance from training databases [3–8]. The tumor in new images is then segmented by predicting the label of each voxel separately. This step integrates multiple sources of information such as different modalities (e.g. different magnetic resonance imaging (MRI)-protocols which are commonly available in clinical routine), derived features, or brain atlas-based information.

One common drawback of these approaches is that the training and labelling has to be repeated if the clinical setup changes. MRI-images have a high variability depending on the scanner type, sequence, and configuration, so for optimal performance each clinic has to create a unique training base reflecting their setting. The above mentioned problems of this tedious process often render



(a) (b) (c)

Figure 1: **a**: An exemplary slice of a $T1_w$-MR image of the brain of a patient with glioma tumor. **b**: The corresponding slice of a T2-Flair image. The tumor is outlined in yellow. **c**: The T2-Flair with the small segmentations. Green is healthy - red, blue and yellow are tumor. Note that border regions are included in all labels.

learning-based methods inadequate for clinical use. To alleviate these problems, Verma *et al.* [9] avoided full manual segmentations and learned from nearly complete segmentations excluding ambiguous areas. We hypothesize that this introduces a domain adaption problem since the test and train data are generated from different distributions. In this work, we also exclusively label nonambiguous regions in only small fractions of the data. We correct the resulting sampling selection bias using a domain adaptation technique that assumes a covariate shift.

2 Method

Labeling only small parts of the image leads to a sampling bias. Some areas are over-represented while others are under-represented. This is true for the tissue classes as well as for the feature distribution. We assume that the small segmentations are representative for the labels, i.e. the likelihood P for a label y and a given feature vector x was assumed to be the same in the complete and the small segmentation. Only the likelihoods for given feature vectors were assumed to be different:

$$P_{\text{Small}}(y \mid x) = P_{\text{Complete}}(y \mid x)$$
$$P_{\text{Small}}(x) \neq P_{\text{Complete}}(x) \quad .$$

We therefore assumed a covariate shift within the training data and corrected it by weighting all samples as suggested by Shimodaira [10] with

$$w(x) = \left(\frac{P_{\text{Complete}}(x)}{P_{\text{Small}}(x)} \right)^{\lambda} \quad .$$

There are several ways to estimate the correction factor w. Since the tissue appearance is learned voxel-wise the training-base is rather large. A single 3D-MR scan contains usually more than 100.000 voxels. We decided to use a logistic regression classifier (LRC) to calculate w because it was previously successfully used [11] and we found that it works fast on large data sets. We trained a LRC that predicts if a voxel is in the complete or the small segmentation. According to Sugiyama and Kawanabe [12] w can be estimated with the trained LRC-parameters $\theta(x)$ by

$$w(x) = \left(c \cdot \exp\left(\theta(x)\right)\right)^{\lambda} \quad .$$

We calculated the voxel-weights for each image separately instead for all images at once. This is important since the tissue appearance differs greatly between different MR images. Consequently we set $c = 1$. The sum of all weights for a single image match the number of voxels instead of the number of voxels from the small segmentation as it would be if we used c from [12]. This is

important in order to ensure that size of the small segmentation does not influence the importance of an image during training. We set $\lambda = 1$ as we found this to give the best result.

For the classifier we chose random forests since they have previously shown good performance in brain tumor segmentation [3, 4]. The noise sensitivity was reduced by limiting the tree depth as suggested in [13] and the weights were incorporated by extending the Gini Impurity. Instead of estimating the label probability by the number of elements with this label the sum of all weights corresponding to this label is used:

$$I(V) = 1 - \sum_{y_c \in Y} \left[\frac{1}{\sum w_i} \cdot \sum_{y_j = y_c} w_j \right]^2 .$$

3 Experiments and Results

The evaluation of the proposed method was carried out using 19 patients with malignant gliomas. Each patient had 16 different MR images, including T1 with contrast enhancement, T2, T2 Flair, and MR-Diffusion-tensor-imaging derived maps. The feature vector of a voxel contains the intensity of the 16 images at the corresponding positions after a MR histogram normalization step.

Trained experts created both, a complete tumor segmentation with two classes (healthy and tumorous) and small segmentations with 5 classes (fluid, healthy brain, edema, active tumor, and necrosis). They performed multiple refinement steps for the full segmentation to increase the quality of these segmentations. To compare the results we fused the 5 labels of the small segmentations into two labels which match those of the full segmentations.

Contrary to a complete segmentation for which all slices (usually between 40 and 50) of an image are labeled, the small segmentations are usually located only in a single slice. (The labeled regions within these slices is small.) The small regions were drawn in locations that the expert evaluated as being representative. Figure 1 shows an example of both a complete- and a set of small segmentations.

Using these segmentations, we ran leave-one-out experiments. Excluding one patient from the training base we trained 3 different classifiers using the remaining 18 patients. The first two classifiers are based on the small segmentations - one with and one without domain adaptation. The third classifier is based on 0.5% random samples from the complete segmentations which is roughly the area covered by the small segmentations.

3.1 Timing analysis

The time required for the different steps is listed Table 1. It shows a significant reduction of the time necessary for the creation of the training base. Since it takes less than 5 minutes to generate the labels for a single patient it is possible to label all patients within 2 hours. Thus, a radiologist may label patients prospectively during daily routine. This allows for a continuous growth of the training base and fast adaptation to changes in the imaging protocol.

Table 1: Durations of different tasks

Method	Labeling	Training	Prediction
Small Seg.	< 5 min	12.4 ± 1.1 sec	45.7 ± 4.3 sec
Small Seg. with DA	< 5 min	63.8 ± 14.4 sec	74.4 ± 8.3 sec
Complete	> 240 min	46.9 ± 1.1 sec	$149.3.4 \pm 16.3$ sec

The training time is minimal if small segmentations are used and no DA is performed during the training. Due to the estimation of w, the training takes significantly longer if DA is used. Since the training is fully automated this is usually not a problem. It is also worth noting that the prediction times for the small segmentation based classifiers are only half of those of the complete segmentation based classifiers. This is important for interactive applications where a fast response is important.

3.2 Quality analysis

The evaluation of the produced predictions is based on the DICE-score [14]. The prediction for each patient is compared to the manually created complete segmentation and the results are given in Fig. 2. Compared to a complete segmentation trained classifier the small segmentation trained classifier shows a significant[1] ($p = .008$) drop in the segmentation quality. We think that there are two reasons for this drop. First the small segmentations contain less information than the complete segmentation, therefore the classifier is less general. A second reason for this drop is the sampling bias within the training data.

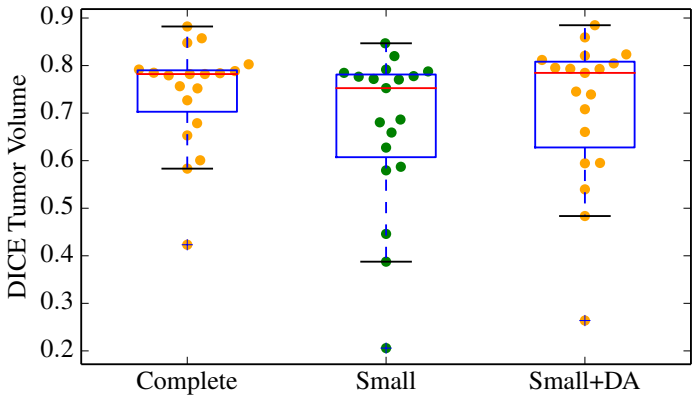


Figure 2: Leave-one-out results of the classifier trained on 3 different training bases. The DICE score is calculated against the manually created complete segmentation.

As expected the correction of the sampling bias improves the results. A classifier with DA gives significantly ($p = .015$) better results than one trained without domain adaptation. The DA results are comparable to the results obtained from a classifier trained on the whole segmentation. There is no significant ($p = .10$) difference between the two results although the complete segmentation-trained classifier seems to have a better generalization. This difference could be reduced by adding more patients to the training base, which is now much easier than extending the complete segmentation training base.

4 Conclusion

We showed that domain adaptation allows training classifiers for tumor segmentation on partially labeled data. It reduces the sampling error made during the creation of the training base and the so-trained classifiers perform similar to classifier trained with complete segmentations. This is an important step towards including automatic brain tumor segmentation in clinical routine since it allows creating a custom training base in reasonable time. Further research needs to evaluate the choice of the weight estimation algorithm and to validate the effect of an extended training base on small segmentation-trained classifier.

Acknowledgments

This work was carried out with the support of the German Research Foundation (DFG) as part of project I04, SFB/TRR 125 Cognition-Guided Surgery.

We like to thank Franciszek Binczyck, Joanna Polanska, Rafal Tarnawski, and Barbara Bobek-Billewicz from the 'Silesian University of Technology' and the 'Maria Skodowska-Curie Memorial Cancer Center and Institute of Oncology' in Gliwice, Poland for providing the data.

[1] Significance is tested with paired t-tests

References

[1] G. P. Mazzara, R. P. Velthuizen, J. L. Pearlman, H. M. Greenberg, and H. Wagner, "Brain tumor target volume determination for radiation treatment planning through automated mri segmentation," *International Journal of Radiation Oncology* Biology* Physics*, 2004.

[2] B. Menze, et al., and K. Van Leemput, "The multimodal brain tumor image segmentation challenge (BraTS)." *Submitted to IEEE Transactions on medical imaging*, 2014.

[3] D. Zikic, B. Glocker, E. Konukoglu, A. Criminisi, C. Demiralp, J. Shotton, O. Thomas, T. Das, R. Jena, and S. Price, "Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel MR," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, 2012.

[4] S. Bauer, T. Fejes, J. Slotboom, R. Wiest, L.-P. Nolte, and M. Reyes, "Segmentation of brain tumor images based on integrated hierarchical classification and regularization," in *MICCAI BraTS Workshop: Miccai Society*, 2012.

[5] S. Bauer and L.-P. Nolte, "Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, 2011.

[6] P. Buendia, T. Taylor, M. Ryan, and N. John, "A grouping artificial immune network for segmentation of tumor images," *MICCAI BraTS Workshop: Miccai Society*, 2013.

[7] S. Doyle, F. Vasseur, M. Dojat, and F. Forbes, "Fully automatic brain tumor segmentation from multiple mr sequences using hidden markov fields and variational em," 2013.

[8] S. Bauer, R. Wiest, L.-P. Nolte, and M. Reyes, "A survey of mri-based medical image analysis for brain tumor studies," *Physics in medicine and biology*, 2013.

[9] R. Verma, E. I. Zacharaki, Y. Ou, H. Cai, S. Chawla, S.-K. Lee, E. R. Melhem, R. Wolf, and C. Davatzikos, "Multiparametric tissue characterization of brain neoplasms and their recurrence using pattern classification of MR images," *Academic Radiology*, 2008.

[10] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of statistical planning and inference*, 2000.

[11] T. Heimann, P. Mountney, M. John, and R. Ionasec, "Real-time ultrasound transducer localization in fluoroscopy images by transfer learning from synthetic training data," *Medical image analysis*, 2014.

[12] M. Sugiyama and M. Kawanabe, *Machine learning in non-stationary environments: introduction to covariate shift adaptation*. MIT Press, 2012.

[13] A. Criminisi and J. Shotton, Eds., *Decision Forests for Computer Vision and Medical Image Analysis*. Springer London.

[14] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, 1945.