Semi-Supervised and Active Learning for Automatic Segmentation of Crohn's Disease

Dwarikanath Mahapatra[1,⋆], Peter J. Schüffler[1], Jeroen A.W. Tielbeek[2], Franciscus M. Vos[2,3], and Joachim M. Buhmann[1]

[1] Department of Computer Science, ETH Zurich, Switzerland
dwarikanath.mahapatra@inf.ethz.ch
[2] Department of Radiology, Academic Medical Center, The Netherlands
[3] Quantitative Imaging Group, Delft University of Technology, The Netherlands

Abstract. Our proposed method combines semi supervised learning (SSL) and active learning (AL) for automatic detection and segmentation of Crohn's disease (CD) from abdominal magnetic resonance (MR) images. Random forest (RF) classifiers are used due to fast SSL classification and capacity to interpret learned knowledge. Query samples for AL are selected by a novel information density weighted approach using context information, semantic knowledge and labeling uncertainty. Experimental results show that our proposed method combines the advantages of SSL and AL, and with fewer samples achieves higher classification and segmentation accuracy over fully supervised methods.

1 Introduction

Crohn's Disease (CD) has emerged as an important challenge for the medical imaging community [11,9]. CD is a type of inflammatory bowel disease (IBD) affecting the digestive tract leading to abdominal pain, diarrhea, weight loss, fatigue and anemia. Colonoscopy, the current reference standard for diagnosis, is painful, invasive and poses the risk of bowel perforation. Thus non-invasive imaging techniques like Magnetic resonance imaging (MRI) are being explored to assess extent and severity of CD [9].

Vos et al. in [11] proposed a pipeline for automatic detection and diagnosis of CD from MRI while in [8] we explored different image features and *fully supervised* classifiers for classifying CD samples. However a reliable fully supervised classifier requires sufficient number of samples of different degrees of disease severity. Semi-supervised learning (SSL) [3] and active learning (AL) [10,6] methods have been used to overcome the limitations of insufficient labeled samples in medical applications.

We propose a novel method for automatic detection and segmentation of CD tissues from abdominal MRI that combines the principles of SSL and AL. Our method is denoted as SS-AL. SSL makes use of a few labeled voxels and many unlabeled voxels to classify unseen voxels and generate their probability maps.

⋆ Corresponding author.

K. Mori et al. (Eds.): MICCAI 2013, Part II, LNCS 8150, pp. 214–221, 2013.

The AL part of our method selects those voxels whose labels are likely to lead to maximum improvement in classifier performance. The final classification is used to segment CD affected tissues using semantic knowledge and graph cuts. We use random forest (RF) classifiers [2] which is popular in medical applications [6] because: 1) recent work [4] has showed the advantages of RF over other classifiers for SSL classification in terms of accuracy and speed; and 2) the knowledge from the trained data can be used to design different strategies for query sample selection and segmentation.

This paper has the following contributions: 1) a framework combining AL and SSL for classification of Crohn's disease tissues; 2) a novel query sample selection strategy is presented which makes use of learned semantic knowledge from RF classifiers, and context information from the neighborhood; and 3) semantic knowledge is used to design a novel smoothness cost for graph cut segmentation. We describe our method in Section 2, present our results in Section 3 and conclude with Section 4.

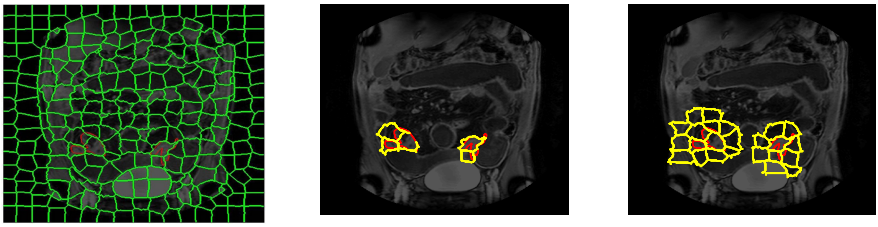
2 Methods

Overview of Our Method. Our proposed framework has the following steps: 1) volume of interest (VOI) detection from initial annotations using supervoxel segmentation and SSL classification; 2) generating probability maps for each voxel within VOI using the annotations in 1) and SSL; 3) querying labels of most informative samples; 4)using new labels to retrain classifier and generate new probability maps; and 5) alternating between steps 3),4) till convergence.

2.1 Initial VOI Detection by SSL

A radiologist identifies diseased and normal regions on an MR T-1 test volume. Subsequently the volume is segmented into supervoxels (SVs) using intensity values and the method in [1]. We calculate the mean, variance, skewness and kurtosis of intensity, texture and mean 3D curvature values of every supervoxel. The texture maps are calculated for each slice of the supervoxel using $2D$ Gabor filters oriented at $0°, 45°, 90°, 135°$ at the original scale. Thus each supervoxel gives a 24 dimensional feature vector.

The extracted features extracted are used to classify all SVs as "diseased" or "normal" using RFs for SSL ($RF - SSL1$). RFs for SSL have the advantage that unlabeled samples are classified in one step without iterative retraining of the classifier. When the fraction of diseased tissues in a SV is low it may escape detection. We observe that diseased SVs are clustered together in groups of more than two. To get the final VOI we include all adjoining SVs of those classified as "diseased". Although this increases computation time, it ensures that in most cases we don't miss any diseased regions. Figure 1 shows an example where the above step includes SVs that were missed in the initial classification.



(a) (b) (c)

Fig. 1. (a) Supervoxels (green) and ground truth diseased region (red);(b) diseased supervoxels (in yellow);(c) final VOI after label change of neighboring supervoxels

2.2 SSL-AL Based Classification of VOI Voxels

All voxels within the designated VOI are analysed by a second set of RF SSL classifiers ($RF-SSL2$). For every VOI voxel the higher order features described in Section 2.1 are calculated over its $30 \times 30 \times 3$ neighborhood, and $RF-SSL2$ gives probabilities of each voxel being diseased or normal.

The querying strategy chooses the test sample whose label would result in maximum improvement of classifier performance. Common query strategies include uncertainty sampling [7], query-by-committee (QBC) [5] and density weighting [10]. The first two methods select samples for which the classifier is least certain although uninformative samples may be selected, e.g., if they are lying on the classifier boundary or situated in low-density regions. To overcome this problem density weighted methods weigh samples according to their informativeness.

Query Sample Selection Strategy: The information density framework [10] chooses samples which are uncertain and inhabit dense regions of the input space. The informativeness of sample x is

$$Inf(x^*) = \arg\max_x \left(\phi(x) + \alpha \times \sum_{u=1}^{N} \text{sim}\,(x, x^u) \right). \quad (1)$$

ϕ is the uncertainty of x, sim is the similarity between x and x^u (all other unlabeled samples in N), and α is a parameter that weighs the relative contribution of ϕ and sim. $N = 35 \times 35 \times 3$ is a neighborhood around x (instead of the entire unlabeled set U) to reduce computational complexity. Note that in Eqn. 1 we do not normalize by the number of samples in N. In a high-density region this ensures that the informativeness of x is higher, while in a sparsely populated region x's informativeness is low. The uncertainty of x is its entropy given by

$$\phi(x) = -\sum_{\widehat{y}} P((\widehat{y}|x) \log P((\widehat{y}|x)\,, \quad (2)$$

where $\widehat{y}$ indicates all possible labels (in this case two) for x, and $P((\widehat{y}|x)$ is calculated by $RF-SSL2$. Higher entropy indicates greater uncertainty.

$RF-SSL2$ provides semantic information of the relative importance of each element of the feature vector in classification. Although the feature vector has 24

elements there are in effect 3 types of features - intensity, texture and curvature. We aggregate the importance values of each feature type and divide by their sum to get a set of normalized importance measures. Let the normalized importance measures of intensity, texture and curvature be $w_I, w_T, w_C (w_I + w_T + w_C = 1)$. The similarity between two samples after incorporating *semantic information* is

$$sim(x, x^N) = w_I \left\| x_I - x_I^N \right\| + w_T \left\| x_T - x_T^N \right\| + w_C \left\| x_C - x_C^N \right\|. \quad (3)$$

x_I, x_T, x_C are, respectively, the intensity, texture and curvature components of the feature vector of sample x and superscript N indicates unlabeled samples in neighborhood N. $\|.\|$ indicates the mean square distance of the two vectors. The weights typically take the following range of values $w_I = 0.19 - 0.24, w_T = 0.31 - 0.35, w_C = 0.41 - 0.47$ in different iterations. A constant weight excludes semantic information, whose importance is analyzed in Section 3.2.

Context information from N also contributes to a sample's informativeness. Medical images have inherent context information because the relative arrangement of organs in the human body is constant. The proximity of an unlabeled sample to a labeled sample gives an idea about its possible label. An unlabeled sample close to a labeled sample is assigned lower importance than one far away. This ensures that during query time, the radiologist provides maximum information to the classifier. If the radiologist were to annotate samples close to an already labeled sample it *does not* lead to significant information gain. α incorporates context information and is given by

$$\alpha = \min \left(\left\| x - x^L \right\| \right). \quad (4)$$

where x^L denotes all the labeled samples, and $\|.\|$ denotes the Euclidean distance based on voxel co-ordinates.

Stopping Criteria: After every classification we determine the distribution of probability values and compare with those of the previous iteration using the Student $t-$test. $p < 0.05$ denotes statistically different classification in the current iteration. However, $p > 0.05$ indicates minor difference in the distributions. If $p > 0.05$ for two consecutive iterations there is no further querying of labels.

2.3 Graph Cut Segmentation

The probability maps by $RF - SSL2$ are used as penalty costs in a second order MRF energy function given by

$$E(L) = \sum_{s \in P} D(L_s) + \lambda \sum_{(s,t) \in N_s} V(L_s, L_t), \quad (5)$$

where P denotes the set of pixels; N_s is the 8 neighbors of pixel s; L_s is hte label of s and L is the set of labels for all s. λ determines the relative contribution of penalty cost (D) and smoothness cost (V). $D(L_s)$ is the negative log-likelihood of probabilities given by

$$D(L_s) = -\log \left(Pr(L_s) + \epsilon \right), \quad (6)$$

where Pr is the likelihood (or probabilities) obtained using $RF-SSL2$ and $\epsilon = 0.00001$ is a very small value to ensure that the cost is a real number.

Semantic Information for Smoothness Cost: w_I, w_T, w_C obtained after final $RF-SSL2$ classification incorporate *semantic information* in the smoothness cost. V is given by

$$V(L_s, L_t) = \left\{ \begin{array}{ll} w_I V_I + w_T V_T + w_C V_C, & L_s \neq L_t, \\ 0 & L_s = L_t. \end{array} \right\} \tag{7}$$

where V_I, V_T, V_C are the individual contributions to the smoothness by intensity, texture and curvature.w_I, w_T, w_C have been defined previously. V_I is defined as

$$V_I(L_s, L_t) = e^{-\frac{(I_s - I_t)^2}{2\sigma^2}} \cdot \frac{1}{\|s-t\|}, \tag{8}$$

I is the intensity. V_T and V_C are similarly defined using texture and curvature instead of intensity. Graph cuts was preferred over other methods because: 1) it can find a global minimum for binary labeled segmentation problems; and 2) facilitates easy integration of probabilistic outputs from $RF-SSL2$.

3 Experiments and Results

$T1$ weighted MR images were acquired from 26 patients diagnosed with CD (mean age 37 ± 19.4 years, 16 females) using a 3-T scanner (Intera, Philips), having 16-channel torso phased array body coil. The images had pixel spacing of $1.02\times\ 1.02\times\ 2$ mm, and matrix size of $400\times400\times100$ voxels. Diseased and normal regions were manually annotated in all patients which served as the ground truth for our segmentation algorithm. In total there were 673 diseased regions and 629 normal regions. Our experiments were performed on a system running MATLAB on a Core2 quad core 2.66 GHz CPU having 4 GB RAM. $\lambda = 0.02$ set after cross validation using an entirely different dataset of 5 patients.

3.1 Classification Performances of SSL and AL

Results in this section demonstrate: 1) SSL gives better classification performance than FSL; and 2) our query selection strategy based on distance to nearest labeled samples is better than constant α. A certain percentage of labeled samples (N_L) were used to train a fully supervised RF classifier ($RF-FSL$) and classify the remaining samples. The same N_L and unlabeled data are classified using $RF-SSL2$ with the results summarized in the first part of Table 1. Note that for a fixed N_L, $RF-SSL2$'s and $RF-FSL$'s performance is the average of 10 runs with randomly drawn N_L samples.

For $N_L < 50$ $RF-SSL2$ performs significantly better than $RF-FSL$ as the number of labeled samples are not sufficient to train a reliable $RF-FSL$. On the other hand $RF-SSL2$ exploits the information from the unlabeled samples to obtain high accuracy and sensitivity. With $N_L \geq 50$, the difference between

Table 1. Comparative classification performances between $RF-SSL2$ and $RF-FSL$, and between α' and adaptive α for $RF-SSL2$. *Sen*-sensitivity (percentage of correctly classified diseased samples), *Spe*-specificity (percentage of correctly classified normal samples), *Acc*-overall accuracy.

	N_L										$N_L = 25$
	5		10		25		50		75		$\alpha' = 3$
	SSL	FSL	SSL	FSL	SSL	FSL	SSL	FSL	SSL	FSL	
Sen	71.3	64.2	77.4	70.1	85.4	80.2	91.5	89.2	93.6	92.1	81.3
Spe	69.2	62.8	74.6	69.8	83.1	78.9	89.8	88.1	93.1	91.6	79.7
Acc	70.4	63.7	76.2	70.1	84.4	79.8	90.7	88.7	93.4	91.8	80.6

$RF-SSL2$ and $RF-FSL$ is less because the unlabeled samples do not give much extra information for $RF-SSL2$ to exploit and significantly outperform $RF-FSL$. However, $RF-SSL2$ continues to outperform $RF-FSL$ for all N_L. This clearly establishes the superiority of SSL classification over fully supervised methods. Compared to our *fully supervised* method in [8] we are able to achieve higher sensitivity and accuracy by combining SSL and AL.

Recall that α in Eqn. 1 is the Euclidean distance of an unlabeled sample to the nearest labeled sample, For $N_L = 25$, we varied the value of α from 0 to 10 (denoted by α') and cross validation shows best classification results (by $RF-SSL2$) for $\alpha' = 3$. The results are shown in the last column of Table 1. The classification performance is lower than our adaptive α (first part of Table 1 for $N_L = 25$). Since a constant α discards contextual information from the nearest labeled sample, it has lower performance than our adaptive α in Eqn. 1.

3.2 SS-AL Based Graph Cut Segmentation

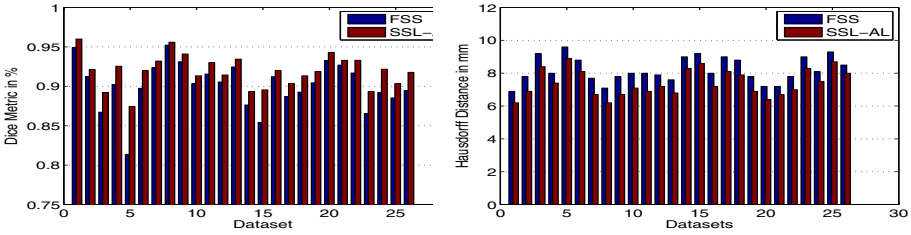
Next we show that combining SSL and AL leads to superior segmentation over a fully supervised method. This is due to more accurate probability maps by $RF-SSL2$. We start with labels of 1 region each from diseased and normal regions, and requests for sample labels till convergence. Number of queries for each slice varies from $2-4$ depending upon the size of the diseased area. In each query the expert annotates a small region around the queried sample providing labels of $12-25$ pixels per labeling. We compare our method ($SS-AL$) with a fully supervised segmentation (FSS) method using a leave-one-out approach; a RF is trained on 25 patients and the 26th patient is segmented by generating probability maps, using them as cost functions and formulating the smoothness cost using semantic information. The VOI for both methods is the same as obtained from $RF-SSL1$.

The algorithm segmentations are compared with manual segmentations using Dice Metric (DM) and Hausdorff distance(HD) and our results are summarized in Table 2. Figure 2 shows the DM and HD values of individual patients using our method. $SS-AL$ gives better segmentations for all patients and achieves significantly higher DM values in many cases where FSS gives low values.

Table 2 presents results for: 1) $SS-AL_{nV}$: $SS-AL$ without semantic context in V; $w_I = w_T = w_C = 0.33$; 2) $SS-AL_{nV_I}$: $SS-AL$ with $w_I = 0$; 3)

Table 2. Quantitative measures for segmentation accuracy. DM- Dice Metric in %, HD-Hausdorff distance in mm, Time-computation time in minutes.

	$SS-AL$	FSS	$SS-AL_{nV}$	$SS-AL_{nV_I}$	$SS-AL_{nV_T}$	$SS-AL_{nV_C}$	$SS-AL_{nV_{TC}}$
DM	92.1	90.3	88.1	88.3	88.0	87.8	87.0
HD	6.8	7.3	8.9	8.8	9.2	9.3	9.8
Time(min)	51	43	50	49	51	52	50



(a) (b)

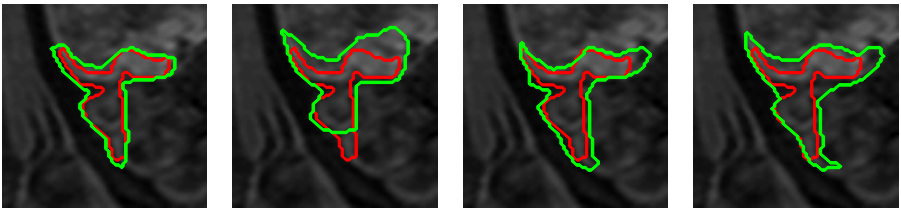
Fig. 2. (a) Dice Metric and (b) Hausdorff distance measures for individual datasets

$SS-AL_{nV_T}$: $SS-AL$ with $w_T = 0$; 4) $SS-AL_{nV_C}$: $SS-AL$ with $w_C = 0$; 5) $SS-AL_{nV_{TC}}$: $SS-AL$ with $w_I = 1, w_T = 0, w_C = 0$ which is a conventional smoothness cost based on intensity features. For $2, 3, 4$ above the weights are normalized by the sum of the two values. We do not compare with other methods as there are no segmentation methods specific to CD.

$SS-AL$ performs the best among all methods followed by FSS and $SS-AL_{nV_I}$. $SS-AL_{nV_{TC}}$ gives the worst performance because it imposes smoothness constraints using only intensity information. If we exclude any information from $SS-AL$ the performance is worse than FSS which indicates the importance of semantic information and individual features for our method. Figure 3 shows segmentation results for $SS-AL$, FSS, $SS-AL_{nV}$, $SS-AL_{nV_C}$ on Patient 21 with the diseased regions cropped for clarity. FSS gave poor results in this case because of weak boundary between diseased and surrounding normal regions. But $SS-AL$ is able to achieve high DM values. The segmentation results are consistent with the values observed in Table 2, clearly indicating the superiority of the combination of active and semi supervised learning.

4 Conclusion

We have proposed a framework combining the principles of SSL and AL for automatic detection and segmentation of CD from abdominal MRI. RF classifiers provide quick SSL classification, probabilistic outputs of test data and exploitation of semantic information. A novel query selection strategy incorporates contextual information and does not require user defined parameters. Experimental results show our method achieves greater classification accuracy, higher DM and lower HD values than fully supervised approaches with equal number of training samples. Thus SSL-AL is very relevant in this scenario to derive maximum benefit from few labeled samples and insufficient training data.



(a) (b) (c) (d)

Fig. 3. Segmentation results for Patient 21: (a) $SS-AL$ ($DM = 0.93$)(b) FSS $DM = 0.85$; (c) $SS-AL_{nV}$ $DM = 0.89$;(d) $SS-AL_{nV_C}$ $DM = 0.87$. Manual segmentation is shown in red while algorithm segmentations are shown in green. DM values are for entire volume.

Acknowledgement. This research was partly funded from the European Communitys Seventh Framework Programme (FP7/2007-2013): the VIGOR++ Project (grant agreement nr. 270379).

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Patt. Anal. Mach. Intell. 34(11), 2274–2282 (2012)
2. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
3. Chapelle, O., Scholkopf, B., Zien, A.: Semi-Supervised Learning. MIT Press, Cambridge (2006)
4. Criminisi, A., Shotton, J.: Decision Forests for Computer Vision and Medical Image Analysis. Springer (2013)
5. Freund, Y., Seung, H., Samir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Mach. Learn. 28(2), 133–168 (1997)
6. Iglesias, J.E., Konukoglu, E., Montillo, A., Tu, Z., Criminisi, A.: Combining generative and discriminative models for semantic segmentation of CT scans via active learning. In: Székely, G., Hahn, H.K. (eds.) IPMI 2011. LNCS, vol. 6801, pp. 25–36. Springer, Heidelberg (2011)
7. Lewis, D., Catlett, J.: Heterogenous uncertainty sampling for supervised learning. In: ICML, pp. 148–156 (1994)
8. Mahapatra, D., Schüffler, P.J., Tielbeek, J., Buhmann, J.M., Vos, F.M.: A supervised learning based approach to detect crohn's disease in abdominal mr volumes. In: Proc. MICCAI-ABD, pp. 97–106 (2012)
9. Rimola, J., Rodriguez, S., Garcia Bosch, O., et al.: Magnetic resonance for assessment of disease activity and severity in ileocolonic Crohn's disease. Gut. 58, 1113–1120 (2009)
10. Settles, B.: Active learning literature survey. Tech. Rep. 1648, University of Wisconsin-Madison (January 2010)
11. Vos, F.M., et al.: Computational modeling for assessment of IBD: to be or not to be? In: Proc. IEEE EMBC, pp. 3974–3977 (2012)