SOFTSTAPLE: TRUTH AND PERFORMANCE-LEVEL ESTIMATION FROM PROBABILISTIC SEGMENTATIONS

Neil I. Weisenfeld† and Simon K. Warfield

Computational Radiology Laboratory
Children’s Hospital Boston
Harvard Medical School
Boston, MA, USA

ABSTRACT
We introduce here a new algorithm, called softSTAPLE, for computing estimates of segmentation generator performance and a reference standard segmentation from a collection of probabilistic segmentations of an image. These tasks have previously been investigated for segmentations with discrete label values, but few techniques exploit the information available in probabilistic segmentations. Our new method may be used to evaluate classification algorithms, to fuse “weak” classifiers in a performance-weighted fashion, or to combine the results of a previous fusion of manual segmentations in a hierarchical manner. We describe and validate our new algorithm, and compare its performance to other techniques in two applications with “real-world” data.

Index Terms— segmentation, classification, validation, classifier fusion

1. BACKGROUND

Segmentations of medical images are often assessed by comparison with hand-drawn results. Manually labeling a high-resolution MR data set is not only tedious and time consuming, but also suffers from large inter- and intra-rater variability [1]. Warfield et al. [2] developed an algorithm called STAPLE in order to facilitate the comparison of different segmentation generators for images without a known reference standard. This Expectation-Maximization [3] algorithm simultaneously estimates a reference standard segmentation and the performance of raters attempting to segment an image. Performance, in the binary case, is characterized by estimating the fraction of correct positive (sensitivity) and correct negative (specificity) classifications.

Rohlfing utilized STAPLE for the fusion of segmentations warped to a target subject by non-rigid registration [4] and we recently utilized STAPLE for the fusion of results from several statistical classifiers [5]. One simple strategy for fusing classifiers is to count the number of “votes” that each label receives at each voxel and to take the most prevalent. This approach does not consider the relative performance quality of each classifier, which can be estimated, even in the absence of known ground truth, by considering systematic errors across the image. STAPLE [2] addresses precisely this need and Rohlfing [4] demonstrated the superiority of this approach to vote counting for classifier fusion.

The STAPLE algorithm cannot, however, directly estimate a reference standard from probabilistic segmentations where the probability of each class label is known at each voxel, rather than the class label itself. Such segmentations are the output of statistical classification algorithms or manual ratings specified as confidence levels. In order to utilize STAPLE to fuse probabilistic segmentations, we can turn probabilities into labelmaps by choosing the most probable label at each voxel. Doing so discards information about uncertainty, however, and for certain applications this loss may be unacceptable. For instance, several authors have introduced probabilistic segmentation algorithms designed to provide estimates of tissue segmentations in the presence of partial volume averaging [6, 7]. Turning these probabilistic segmentations into discrete labels would discard critical information.

We introduce here a new algorithm, called softSTAPLE, for computing estimates of segmentation generator performance and a reference standard segmentation from a collection of probabilistic segmentations of an image. This new algorithm enables direct, performance-weighted fusion of probabilistic classifier outputs, without the loss of information about uncertainty or partial volume occupancy due to the conversion of probabilistic classifications to “hard” labels. Furthermore, our algorithm can be used to combine repeated observations of segmentations from a group of domain experts in a hierarchical manner, by first estimating a probabilistic reference standard for each rater, and then combining these into a consensus reference standard across all raters. Such hierarchical analysis enables improved assessment of both within and between rater variability.
2. METHODOLOGY

We assume that each rater is attempting to segment the same gray level image \( G \) and we consider a set of segmentations of this image, one by each of \( J \) raters. We assume that these segmentations are conditionally independent, indicating that each rater is attempting to segment the same underlying image, but working independently of one another. Let \( D_j \) be a hidden random variable representing a segmentation of image \( G \) by rater \( j \), where \( j \in \{1 \ldots J\} \), and \( p(D_j|G) \) be the observed probabilistic segmentation that is available as input to the algorithm. We use \( i \) to index voxels, so \( D_{ij} \) represents a single voxel in a segmentation, \( D_j \) denotes all the voxels of a segmentation, and \( D \) is the collection of all segmentations \( \{D_1 \ldots D_J\} \).

How well each rater performs the segmentation task is described by a matrix \( \theta_j \) of performance parameters. In the case of a binary segmentation, the diagonal elements of the \( 2 \times 2 \) matrix \( \theta_j \) are the specificity and sensitivity of rater \( j \) and the off-diagonal elements are the false-positive and false-negative rates. For arbitrary numbers of labels, we simply add rows and columns such that:

\[
\theta_{j,s's'} = Pr\{D_j = s'|T = s\} = p(D_j = s'|T = s) \tag{1}
\]

In this way, the on-diagonal elements represent rates of agreement or correct segmentation, and the off-diagonal elements represent each possible type of error. We use \( \Theta \) to represent the collection of all \( \theta_j \).

We observe probabilistic segmentations of \( G \) which we denote \( p(D|G) \). If we had observed the hard segmentations \( D \), and we also knew the reference standard segmentation \( T \), then it would be straightforward to generate a maximum-likelihood estimate of rater performance \( \Theta \) by counting the rates of detection of each label in the image for each rater \( j \). If \( D \) and \( T \) were observed, the complete data log likelihood could be used to estimate \( \Theta \):

\[
\arg \max_{\Theta} \ln p(D, T, G|\Theta) \tag{2}
\]

Since \( D \) and \( T \) are unobserved, we have developed an Expectation-Maximization (EM) [3] algorithm for estimating \( \Theta \). The EM algorithm was developed for the solution of missing data maximum-likelihood problems, in cases where the conditional probability distribution for the missing data, given the observed data, is either known or can be estimated. In our application, we are supplied with a distribution for the missing segmentations \( D \), these are the probabilistic input segmentations, and we can estimate a distribution for the reference standard \( T \).

EM operates by iterating between estimating the conditional expectation of the complete-data log likelihood function, called \( Q \) in the EM literature, and estimating the unknown parameters by maximization:

\[
Q(\Theta|\Theta^t) = E \left[ \ln p(D, T, G|\Theta)|G, \Theta^t \right] = \sum_{D,T} \ln p(D, T, G|\Theta)p(D, T|G, \Theta^t) = \sum_D p(D|G) \sum_T \ln p(D, T|G, \Theta)p(T|D, \Theta^t)
\]

where \( \Theta^t \) is the value of \( \Theta \) at the previous iteration. Equation 2 utilizes the observation that under the proposed model, \( D \) does not depend on \( \Theta^t \), and the combination \( (D, \Theta^t) \) is equally informative about \( T \) as \( (D, \Theta^t, G) \).

At each iteration, \( \arg \max_{\Theta} Q(\Theta|\Theta^t) \) is computed and we can disregard terms that do not vary with \( \Theta \) and therefore do not change the maxima:

\[
\arg \max_{\Theta} Q(\Theta|\Theta^t) = \arg \max_{\Theta} \sum_D p(D|G) \sum_T \ln p(D, T|G, \Theta)p(T|D, \Theta^t) = \arg \max_{\Theta} \sum_D p(D|G) \sum_T \ln p(T|D, \Theta)p(T|D, \Theta^t)
\]

We now demonstrate how this procedure can be used to estimate a consensus segmentation and performance parameters.

2.1. Expectation (E) step

Since we have not directly observed \( D \), but our input segmentations are a distribution \( p(D|G) \) for \( D \), we must compute the expectation over all possible combinations of \( D \). Given a fixed number of raters \( J \) and a fixed number of labels \( S \), there are \( S^J \) combinations of decisions that could be made at each voxel. We choose a particular enumeration for these combinations and use a single index, \( c \), to represent a unique configuration of possible rater decisions. In order to calculate the expectation over \( D \), we will enumerate all such combinations.

We also need to calculate the expectation over the true, unknown segmentation \( T \) (our reference standard), and we estimate the probability of \( T \) at each voxel \( i \). Given a particular configuration \( c \) from the set of all possible rater decisions, \( D_c^i \) is the ensemble of decisions by all raters at voxel \( i \) for the chosen configuration. Assuming voxelwise independence, the conditional probability of the reference standard segmentation at a voxel \( i \) is:

\[
W_{si}^c = p(T_i = s|D_c^i, \Theta^t) = \frac{\prod_{j=1}^{J} p(D_{ij}^c|T_i = s, \theta_j^t)p(T_i = s)}{\sum_{s'} \prod_{j=1}^{J} p(D_{ij}^c|T_i = s', \theta_j^t)p(T_i = s')} \tag{4}
\]
This results from Bayes’ rule and the conditional independence of the raters and is equivalent to the result in [2], with actual rater decisions $D_{ij}$ in [2] replaced by the decisions $D_{ij}^c$ specified by the hypothetical configuration $c$ being considered.

### 2.2. Maximization (M) step

Given an estimated reference standard $W_{si}^c$ for each possible configuration of rater decisions, we can generate a maximum-likelihood estimate of each rater’s performance. Setting the derivative of Equation 3 with respect to $\Theta$ to zero and using the method of Lagrange multipliers to enforce the constraint that $\sum_{n'} \theta_{jn'n} = 1$, we find that the performance parameters maximizing Equation 3 are:

$$\hat{\theta}_{jns} = \frac{\sum_c \sum_{D_{ij}^c = s'} p(D_{ij}^c|G) W_{si}^c}{\sum_c \sum_{s'} \sum_{D_{ij}^c} p(D_{ij}^c|G) W_{si}^c}$$

(5)

where $p(D_{ij}^c|G) = \prod_j p(D_{ij}^c|G)$.

### 2.3. Estimated Reference Standard Segmentation

Once the E and M steps have iterated to convergence, we can estimate the reference standard $T$. Equation 4 estimates this value for a particular possible configuration of rater decisions. We estimate the overall reference standard $T$ as the expected value of Equation 4 computed over all possible rater decisions. If we enumerate all possible configurations and $D_{ij}^c$ is the decision of rater $j$ for configuration $c$ at voxel $i$, then we calculate:

$$W_{si} = E_c [W_{si}^c] = \frac{\sum_c \prod_j p(D_{ij}^c|G) W_{si}^c}{\sum_s' \sum_c \prod_j p(D_{ij}^c|G) W_{si}^c}$$

(6)

### 3. VALIDATION AND EXPERIMENTS

In order to validate the algorithm we have conducted experiments with synthetic data. For this purpose, we assumed a true, underlying segmentation as shown at left in Figure 1 and randomly sampled from this in order to generate test data with a certain rate of error. In the first experiment, ten synthetic segmentations were generated with a known sensitivity of 0.90 and specificity of 0.95. The softSTAPLE algorithm was used to estimate a true, underlying segmentation and provide estimates of rater sensitivity and specificity. The sensitivities were estimated as $0.90 \pm 0.001$ (mean $\pm$ standard deviation) and the specificities as $0.95 \pm 0.001$. An example random segmentation is shown (middle) in Figure 1 along with the softSTAPLE estimated reference segmentation (right) and the true segmentation (left). When generating synthetic data, we can simulate a given level of performance by either modulating the error rate voxel-by-voxel or by “softening” the probabilities within a given voxel. In this first experiment, the errors were generated as hard probabilities of either zero or one. The next two experiments utilize soft segmentations.

In a second validation experiment, a similar set of synthetic data was constructed, but in a hierarchical manner. Eight random samples from the same true segmentation as in Experiment 1 were generated with a sensitivity/specificity of 0.95/0.90. Each of these was then, in turn, sampled eight times with a sensitivity/specificity of 0.97/0.97. In so doing, we simulated eight different raters each making eight repeat attempts at segmenting some unknown, true image. The simulated raters have an intra-rater performance of 0.95/0.90, and an intra-rater performance of 0.97/0.97.
Fig. 3. A T2-weighted pelvic MRI from a single female research subject (left), the softSTAPLE consensus from 3 repeated segmentations by rater 1 (left-middle) and rater 2 (right-middle). At (right) is the softSTAPLE consensus from three raters.

<table>
<thead>
<tr>
<th>subject</th>
<th>mean (stddev)</th>
<th>voting</th>
<th>sST</th>
<th>mrfsST</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.952(0.008)</td>
<td>0.955</td>
<td>0.960</td>
<td>0.964</td>
</tr>
<tr>
<td>2</td>
<td>0.975(0.009)</td>
<td>0.975</td>
<td>0.983</td>
<td>0.984</td>
</tr>
<tr>
<td>3</td>
<td>0.939(0.016)</td>
<td>0.951</td>
<td>0.951</td>
<td>0.959</td>
</tr>
<tr>
<td>4</td>
<td>0.967(0.002)</td>
<td>0.967</td>
<td>0.969</td>
<td>0.970</td>
</tr>
<tr>
<td>5</td>
<td>0.978(0.003)</td>
<td>0.980</td>
<td>0.981</td>
<td>0.987</td>
</tr>
<tr>
<td>6</td>
<td>0.928(0.004)</td>
<td>0.931</td>
<td>0.929</td>
<td>0.939</td>
</tr>
<tr>
<td>7</td>
<td>0.915(0.010)</td>
<td>0.921</td>
<td>0.931</td>
<td>0.937</td>
</tr>
<tr>
<td>8</td>
<td>0.966(0.012)</td>
<td>0.971</td>
<td>0.972</td>
<td>0.976</td>
</tr>
</tbody>
</table>

Table 1. Performance of individual versus fused classifiers. Fused classifiers exhibited higher performance than the mean of individual classifiers. In six out of eight cases, softSTAPLE (sST) was an improvement over vote counting and was equal to vote counting in one of the remaining two. The addition of a Markov Random Field model (mrfsST) improved classification accuracy in all subjects.

Table 2. Mean predictive value for reference standards computed using simple vote counting, STAPLE, and hierarchical softSTAPLE. For each of ten subjects, a reference standard was computed and then compared to each of the individual input segmentations. The results are shown as mean predictive value for each method and each subject. The STAPLE reference standard was more like the input segmentations than simple vote counting, and the hierarchical softSTAPLE reference standard was had the highest average predictive value for each subject.
to brain segmentation and we pose this as a binary classification problem using registered T1-weighted and T2-weighted MR images of the brain from eight subjects. We approach this problem by training a supervised classifier using non-rigidly registered template segmentations. In order to create template images, and in order to have a measure of ground truth, each of our eight subjects’ MRIs were hand segmented by an expert.

To each subject to be analyzed, we registered the other seven subjects and resampled the hand drawn segmentations. T1-weighted MRIs were aligned using mutual-information based registration and an affine transform model. The hand-drawn segmentations were resampled into the space of the subject under study and then randomly sampled in order to generate training data for a supervised classifier. Each registered template image led to a different training set for the classifier and therefore a different resulting intensity-based segmentation of the data. The seven segmentations of each subject were combined using softSTAPLE and the individual classifier results, as well as the fused classification, were combined with the hand-drawn segmentation for the subject under study. Dice overlap measurements [8] were made comparing each result with the hand-drawn segmentations. The results are shown in Table 1. The fused classifiers always outperformed the average of the individual classifiers, generally performing at the top of the group. The softSTAPLE classifier outperformed voting in six of eight cases and tied for one of the remaining two. Modeling spatial homogeneity through the addition of a Markov Random Field universally improved the result in this application.

As a second real-world application, we considered repeated manual validation segmentations of pelvic structures by three expert human raters and sought to objectively assess the quality of the reference standard generated using three methods: simple vote counting, STAPLE, and softSTAPLE. Since a reference standard is meant to estimate ground truth from a collection of input segmentations, we evaluated the predictive value of each input segmentation with respect to the resulting reference standard. The ability to fuse probabilistic segmentations using softSTAPLE allowed us to generate a separate probabilistic consensus segmentation and performance estimates for each expert, based on his or her own repeated segmentations, and then to generate an overall consensus segmentation from the group.

Figure 3 shows an example pelvic MRI from a single subject and generated reference standards for two of three raters and the overall consensus reference standard. Each of the three raters segmented each subject’s image three times, and we generated first an intra-rater probabilistic reference standard and then fused these into an inter-rater reference standard. We also computed a simple vote counting reference standard from each input segmentation, as well as a conventional STAPLE reference standard. Each reference standard was then compared to each input segmentation and results are reported as average predictive value for each subject/method combination. The hierarchical softSTAPLE reference standards provided the best representation of the input segmentations for each subject, followed by STAPLE, and then simple vote counting (Table 2). Since the input in this final experiment comprised discrete segmentations, we did not evaluate the use of softSTAPLE in a non-hierarchical manner, as the results would be precisely the same as those obtained by STAPLE. The separate intra- and inter-rater hierarchical analysis is made possible by the new algorithm.

4. DISCUSSION AND CONCLUSION

We have validated and demonstrated a new Expectation-Maximization algorithm for estimating an unknown reference standard and rater performance from a group of probabilistic segmentations. Using synthetic data, the solutions closely agreed with the synthetic input, even in the hierarchical case. We further demonstrated the utility of the algorithm for classifier fusion. The fusion of several classifiers provided a result dramatically better than the average of those classifiers. Finally, we evaluated reference standard quality using repeated real-world segmentations and demonstrated that our algorithm generates a reference standard most like the input segmentations.

Our new algorithm, which we call softSTAPLE, is an extension of the STAPLE algorithm by Warfield, et al.[2]. The key observation made in the derivation of the associated equations is that in the presence of probabilistic segmentations one must consider every possible combination of rater decisions along with their associated probabilities. Indeed, Equations 4 and 5 reduce to those published in [2] when a single configuration of rater decisions is considered to have probability 1 and all others to have probability zero. This is the situation when “hard,” rather than probabilistic, segmentations are used as input.

The independence assumptions of softSTAPLE are the same as those of the original STAPLE algorithm. We assume that the raters are conditionally independent, given the true underlying segmentation and the rater’s performance parameters [2]. Note that this does not preclude systematic errors. Each rater may make systematic errors, and (soft)STAPLE will generate an accurate consensus as long as either the raters do not make the same systematic error or the prior \( p(T_i = s) \) in Equation 4) enables the identification of the systematic error. The prior can be local or global and can be generated externally, such as from an atlas, or estimated from the input data. When the prior is estimated voxelwise from the input data, the STAPLE result with zero iterations is equal to majority voting.

softSTAPLE enables probabilistic label fusion, which is computationally expensive compared to “hard” label fusion. Given \( S \) possible labels at each voxel and \( J \) raters, the number of combinations that must be considered is \( S^J \), so softSTA-
PLE requires $S^4$ as much computation as the original STAPLE. For this reason, we’ve limited the scope of the validation here to two dimensional images and a relatively small number of raters. Since the vast majority of the space of possible configurations will, in practical cases, have zero or near-zero probability, we believe that it will be possible to significantly accelerate the algorithm, even for a large number of labels and raters. Future work will include such an accelerated algorithm.

5. ACKNOWLEDGMENTS

This investigation was supported in part by a research grant from CIMIT, grants RG 3478A2/2 and RG 4032A1/1 from the NMSS, and by NIH grants R01HD047730, R03EB008680, and R01RR021885.

6. REFERENCES


