Veritas: Combining Expert Opinions without Labeled Data

Sharath R. Cholleti, Sally A. Goldman*
Department of Computer Science and Engineering
Washington University
St. Louis, MO 63130 USA
sharath, sg@cs.wustl.edu

Avrim Blum
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA, 15213 USA
avrim@cs.cmu.edu

David G. Politte, Steven Don
Electronic Radiology Laboratory
Mallinckrodt Institute of Radiology
Washington University School of Medicine
St. Louis, MO 63110 USA
politted, dons@mir.wustl.edu

Abstract

We consider a variation of the problem of combining expert opinions for the situation in which there is no ground truth to use for training. Even though we don't have labeled data, the goal of this work is quite different from an unsupervised learning problem in which the goal is to cluster the data into different groups. Our work is motivated by the application of segmenting a lung nodule in a computed tomography (CT) scan of the human chest. The lack of a gold standard of truth is a critical problem in medical imaging. A variety of experts, both human and computer algorithms, are available that can mark which voxels are part of a nodule. The question is, how to combine these expert opinions to estimate the unknown ground truth. We present the Veritas algorithm that predicts the underlying label using the knowledge in the expert opinions even without the benefit of any labeled data for training. We evaluate Veritas using artificial data and real CT images to which a synthetic nodule has been added, providing a known ground truth.

1. Introduction

The lack of a gold standard of truth for medical images is a critical problem limiting computer aided detection, automated image analysis, and change determination. Establishing ground truth for radiological results or quantitative analyses performed on medical images is exceptionally difficult [2, 3, 13, 15]. Consider the task of segmenting a lung nodule in computed tomography (CT) images of the human chest. Both human experts (radiologists) and computer algorithms are available to mark which voxels are part of a nodule. The accuracy of an expert opinion varies greatly depending on the specific feature and image type, and both inter- and intra-observer variance can be unacceptably high [11]. The question is how to combine these expert image segmentations to estimate the unknown ground truth (the real nodule). While it is easy for human observers to identify and label relevant features in images of natural scenes with high accuracy, the same does not hold for medical images. Currently, expert opinion as to the existence and location of image features and their relevance for a specific diagnosis provide the "gold standard" of truth for medical images. In the CT scans of the human chest, there is no labeled data where we know the nodule exactly. All we have are the CT scans and different expert opinions as to the location and extent of the nodule. While in some instances, correlative clinical findings, such as pathology results and patient outcomes, may be used as substantiating evidence, this is not ground truth. Unlike similar problems of combining experts that have been considered in machine learning [6], for this problem the lack of any accurately labeled training data is a significant obstacle we must overcome.

We present the Veritas algorithm that applies confidence-rated boosting [14] in a novel way to combine expert opinions when there is no labeled data for training. Though this can be applied in different domains where there is a need to combine expert opinions, we investigate it specifically using a medical image segmentation problem. In Section 5, we discuss a variety of other applications for which Veritas

*S. Goldman is currently on leave at Google, Inc.

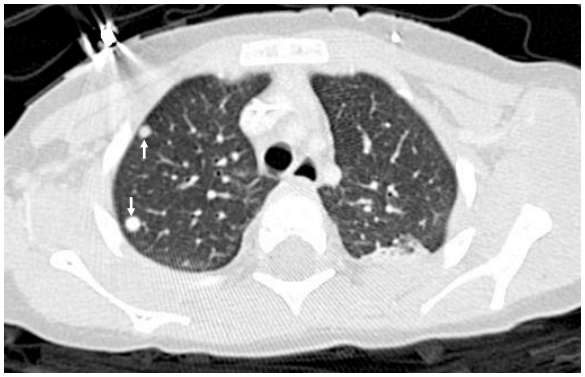

Figure 1. A sample of one of the CT scans that we use in our experimental evaluation. This particular CT slice of the chest shows both a real lung nodule (top arrow) and synthetic lung nodule (bottom arrow).

could be applied. We evaluate Veritas using both artificial data and real CT images in which synthetic nodules have been added that provide a ground truth. Figure 1 shows one of these data sets with both a real nodule and a synthetic nodule marked. Observe that the synthetic nodule is visually similar to the real nodule. Currently the STAPLE algorithm [15] is the standard approach used for developing a model of truth in this context. We compare Veritas with STAPLE, as well as other natural benchmarks.

In Section 2, we describe STAPLE and its drawbacks, and the background for our approach. We describe Veritas in Section 3, and its application to medical image data in Section 4. Conclusions and future work are given in Section 5.

2. Background and Our Approach

In the context of medical image segmentation, the state-of-the-art method for the problem of building a model of truth from the opinions of experts without any ground truth is the STAPLE algorithm [15]. STAPLE aims to simultaneously learn a sensitivity and specificity for each expert as well as the true segmentation, and uses the EM algorithm [8] together with Markov random fields [10] in the following manner. First an estimate for the hidden true segmentation is made (e.g., using the consensus of the experts). This estimated true segmentation can then be used by the "E"-step to estimate a specificity and sensitivity for each expert. For the "M"-step a Markov random field is used as a way to incorporate both the estimated specificity and sensitivity of the experts, as well as a smoothness constraint that penalizes two neighboring pixels for having different values.

The approach used by STAPLE is very natural: combine expert opinions using some weighted vote, but also incorporate spatial constraints. However, it has the limitation of requiring a generative model and a good estimate of the model parameters (e.g., sensitivity and specificity of the experts). As discussed in the paper presenting STAPLE [15], different initial parameters can produce very different hypothesized segmentations. Incorrect parameters could lead to bad estimates, and there is currently no easy way to find good initial parameters except by using prior knowledge about their approximate values.

Another limitation of STAPLE is that the constraint of spatial homogeneity is hard-coded into the edge weights of the hidden Markov model, and is not dependent on the experts' segmentations being considered. Also, STAPLE is designed for a Boolean ground truth and is not well suited for making real-valued predictions. However, in CT images, the lung nodules can have real values indicating the nodule density or that a nodule occupies a fraction of the pixel in the discretized CT image.

Our approach is instead motivated by work in semi-supervised learning — in particular, co-training [5] and co-boosting [7] — as well as work in computational learning theory on learning from random classification noise [1]. The idea of our method is that we will hold out one expert to use as the "label" for each pixel, and then use confidence-rated boosting [14] to learn a good hypothesis for predicting that label based on the predictions of the other experts. Confidence-rated boosting is a generalization of AdaBoost [9] with confidence-rated predictions where the weak learners can abstain (zero confidence) from predicting for some inputs. Specifically, we use the predictions of the other experts on not only the current pixel but also on various combinations of pixels in the surrounding region. For example, if we use expert m as the label, then the space of weak hypotheses for boosting would include prediction rules such as: "if expert 1 predicts '$+$' on at least 7 of the 9 pixels in the local region, then predict '$+$'," or "if experts 1 and 2 both predict '$+$' on at least 6 of the 9 pixels, then predict '$+$'." The idea here is twofold: first, rather than hard-code in beliefs about the strength and form of spatial constraints, we want to allow confidence-rated boosting to select which of these are most important to the problem based on the actual data. Second, if we can model the held-out expert's predictions as corresponding to the true label corrupted by *random classification noise*, then optimizing error rate over the noisy labels has been shown theoretically to optimize error rate with respect to the true labels as well [12]. In particular, the error rate of the predictor trained on the noisy labels can be much lower than that of the noisy labels themselves. Finally, we repeat this process for each expert as the hold-out label, and then combine the outputs of the resulting predictors.

Veritas$(E = \{E_1, \ldots, E_m\})$
  for each expert $E_i \in E$
    $\mathcal{H}_i$ = CreateWeakHypotheses $(E - E_i)$
    P_i = ConfidenceBoost $(\mathcal{H}_i, E_i)$
  Result = Combine $(E, P = \{P_1, \ldots, P_m\})$

Figure 2. Overview of Veritas.

Collins and Singer [7] also consider boosting in a multi-view setting, developing the CoBoost semi-supervised learning algorithm. Unlike CoBoost, however, we do not explicitly optimize for agreement among the m predictors $h_1, \ldots, h_m$ produced in this process (one for each held-out expert). This is because we have overlap in the feature space used for training each predictor, and thus an agreement-based objective among $h_1, \ldots, h_m$ could cause the algorithms to simply focus on common inputs. Instead, we use the expert predictions as labels, with the hope that, as in the simplified version of co-training analyzed by Blum and Mitchell [5], some of the experts will make mistakes that are independent of the mistakes of the other experts, and thus minimizing error with respect to those predictions will minimize error with respect to the ground truth.

3. Veritas Algorithm

In this section we present the Veritas algorithm. Our goal is to create a model of truth that combines the opinions of multiple experts (both human and machine). When the quality of the experts' predictions vary significantly, then a simple consensus model is insufficient.

An overview of the Veritas (truth telling) algorithm is shown in Figure 2. Veritas learns a hypothesis treating one of the m expert segmentations as the label for all pixels. This process is repeated m times, once for each expert segmentation being treated as the label. Each such execution will create a different hypothesis for the ground truth, which are then combined to create the final prediction. Once an expert is selected to be treated as the label, the $m - 1$ other expert segmentations are used to create the feature predictions that form the weak learners for confidence-rated boosting [14]. In this the experts can abstain by saying "I don't know." This is similar to the specialist model of Blum [4]. These predictions are then combined to form a final opinion for the example.

We now describe the details of Veritas for our specific application of lung nodule segmentation in CT images. For each image, Veritas is provided with m expert segmentations, where such a segmentation provides a classification (0 or 1) for each pixel in 2-dimensional data, and for each voxel in 3-dimensional data. Each pixel/voxel will form one example. In particular, to predict the label of the pixel/voxel

CreateWeakHypotheses (E)
/ as we apply, E has $m-1$ experts */*
  for each $E_k \in E$
    generate $\mathcal{S}_k$ */* set contains $2^9 - 1$ subsets of pixels*/*
    set $\mathcal{H}_k$ to empty */* weak hypotheses using E_k */*
  set $\mathcal{H}_{single}, \mathcal{H}_{pairs}, \mathcal{H}_{majority}$ to empty
  / create single expert weak hypotheses */*
  for each $\mathcal{S}_k$, where $k \in \{1, \ldots, |E|\}$
    for each $s \in \mathcal{S}_k$
      / all possible thresholds based on size of s */*
      for (thr = 0; thr< $|s|$; thr=thr++)
        / sum pixel values in s; note $s \in \{0,1\}^{|s|}$ */*
        append the following 4 weak hypotheses to $\mathcal{H}_k$
          if ($\sum s >$ thr) predict 1, else predict 0
          if ($\sum s >$ thr) predict -1, else predict 0
          if ($\sum s >$ thr) predict 0, else predict 1
          if ($\sum s >$ thr) predict 0, else predict -1
    append $\mathcal{H}_k$ to $\mathcal{H}_{single}$
  / create paired experts weak hypotheses*/*
  for ($p = 0; p < |E|; p =$ p++)
    for ($q = p + 1; q < |E|; q =$ q++)
      / combine i^{th} hypothesis from $\mathcal{H}_p$ and $\mathcal{H}_q$ */*
      for ($i = 0; i < |\mathcal{H}_1|; i =$ i++)
        $\mathcal{H}^i_{p,q} = Sign(\mathcal{H}^i_p + \mathcal{H}^i_q)$
        append $\mathcal{H}^i_{p,q}$ to $\mathcal{H}_{pairs}$
  / create majority experts weak hypotheses */*
  for ($i = 0; i < |\mathcal{H}_1|; i =$ i++)
    / combine i^{th} hypothesis from all $\mathcal{H}_k$ */*
    $\mathcal{H}^i_{majority} = Sign(\sum_{k=1}^{|E|} \mathcal{H}^i_k - |E|/2)$
    append $\mathcal{H}^i_{majority}$ to $\mathcal{H}_{majority}$
  / return the set of all weak hypotheses */*
  return $\mathcal{H}_{single} \cup \mathcal{H}_{pairs} \cup \mathcal{H}_{majority}$

Figure 3. Algorithm to create weak hypotheses.

at location ℓ, we apply confidence-rated boosting to a set of weak hypotheses created using a subset of pixels in the neighborhood of ℓ within a single expert, using pairs of experts, and using a majority among all the experts. We have selected these weak hypotheses since they incorporate spatial relationships. Let $\mathcal{S}$ be the set containing all $2^9 - 1$ non-empty subsets of pixels defined by pixel location ℓ and its 8 neighbors. For computational reasons, we consider neighbors from the 2-dimensional slice in a 3-dimensional image. There would be 26 neighbors if the 3-dimensional neighborhood is considered.

We now describe how the weak hypotheses are created. Pseudo code is given in Figure 3.

- *Single expert weak hypotheses:* For each $s \in \mathcal{S}$ and expert segmentation E_k, we introduce four types of weak hypotheses: (1) a weak hypothesis $f^1_s(E_k)$ that predicts 1 if the sum of all the pixels in s is *greater* than a threshold τ (for $0 \leq \tau \leq |s| - 1$), otherwise pre-

ConfidenceBoost $(\mathcal{H}_i, E_i)$
/* E_i^ℓ *provides the label for pixel location* ℓ */
/* $\mathcal{H}_i^\ell$ *is the set of weak hypotheses predictions* */
Run Confidence-rated boosting where weak hypotheses can abstain using the data set: $\langle \mathcal{H}_i^1, E_i^1 \rangle, \langle \mathcal{H}_i^2, E_i^2 \rangle, \ldots$
Using the hypothesis learned, for all ℓ, predict the label P_i^ℓ just using $\mathcal{H}_i^\ell$ without E_i^ℓ
Return P_i

Figure 4. Confidence Boost.

Combine $(E, P = \{P_1, \ldots, P_m\})$
/* *unweighted version* */
for each location ℓ
truth$_\ell = \sum_{k=1}^{m} P_k^\ell / m$

Combine $(E, P = \{P_1, \ldots, P_m\})$ /* *weighted version* */
/* *weight is based on accuracy of* P_k *compared to* E_k */
/* $\forall i,k \in \{1, \ldots, m\}$, *and pixel locations* ℓ */
$W_k = i$ when $\sum_\ell |P_k^\ell - E_k^\ell|$ is the i^{th} highest of m diffs
for each location ℓ
truth$_\ell = \sum_{k=1}^{m} (W_k P_k^\ell) / \sum_{k=1}^{m} W_k$

Figure 5. Combining expert segmentations.

dicts 0 ("I don't know"); (2) a weak hypothesis $f_s^2(E_k)$ that makes the prediction (-1 or 0) complementary to $f_s^1(E_k)$; (3) a weak hypothesis $f_s^3(E_k)$ that predicts 1 if the sum of all the pixels in s is *less than or equal to* a threshold τ (for $0 \leq \tau \leq |s| - 1$), and otherwise predicts 0; (4) a weak hypothesis $f_s^4(E_k)$ that makes the prediction (-1 or 0) complementary to $f_s^3(E_k)$. These weak hypotheses capture spatial relationships within an image.

- *Majority experts weak hypotheses:* For each set $s \in \mathcal{S}$, we introduce a weak hypothesis that predicts based on the majority of the weak hypotheses $f_s^i(E_1), \ldots, f_s^i(E_{m-1})$. These weak hypotheses capture an overall consensus (and its complement) for each pixel set.

- *Paired experts weak hypotheses:* Finally, for each $s \in \mathcal{S}$ and for all distinct pairs E_k, E_j among the $m - 1$ experts, we introduce the weak hypothesis $f_s^i(E_k) \wedge f_s^i(E_j)$. These weak hypotheses capture the importance, in some situations, of when two experts agree with one another.

The *Sign* function used by the *CreateWeakHypotheses* method is defined as:

$$Sign(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \end{cases}$$

Using these weak hypotheses, *ConfidenceBoost* (Figure 4) is executed to generate a hypothesis based on each expert as the label. If the boosting generated hypothesis can predict the label expert well then this expert does not contribute much knowledge to that of the other experts. During the combining phase (Figure 5) of the weighted version of Veritas such a hypothesis is weighted lower. The hypothesis with the lowest accuracy is given the highest weight as this expert is contributing knowledge about the ground truth that other experts do not have. At present, the hypotheses are simply given the integer weights $1, \ldots, m$.

Currently, for the 3-dimensional data, we apply this algorithm to each of the 2-dimensional slices. An area of future work is to introduce a limited set of features (to reduce computation) that involve the corresponding voxels across the slices.

4. Data and Experiments

We evaluate Veritas using both artificial data and real CT data. Though the experts assign either 1 or 0 (part of the nodule or not, respectively) for each voxel of the CT image, the ground truth is not Boolean since a nodule can partially occupy a voxel or the density of the nodule is low in that voxel (corresponding to a fractional ground truth value).

We use the average squared loss among all pixels in which there is disagreement among the original m experts as our criterion for comparing Veritas, STAPLE, and two baseline algorithms. We have selected this measure, rather than the loss over the entire image, since the nodule(s) to be isolated on the real CT images are small, and the experts all agree on the vast majority of the pixels in the image. Thus, we want to focus on the interesting portion of the data.

STAPLE assigns a fractional value to each voxel but its predictions are generally close to 0 or 1 as it was designed with the assumption that the ground truth is Boolean. In contrast, Veritas is designed to work with non-Boolean ground truth values. Though we believe average squared loss compared with ground truth is the best measure, to give a fair comparison to STAPLE, we also use an absolute loss measure where the ground truth and also the output from STAPLE and Veritas are rounded to 0 or 1 before taking the absolute difference. Again, we average the loss over all the voxels where there is disagreement among the original m experts.

4.1. Real CT Data

With Institutional Review Board (IRB) approval we used de-identified spiral CT data collected from a patient using standard pediatric chest protocols on a Siemens Sensation 16 scanner. Images were reconstructed with a voxel size 0.3867 mm in the x- and y-directions (axial), and with a

slice spacing of 1mm in the z-direction (cranial-caudal). Three synthetic nodules have been added to the CT scan of the patient. Figure 1 shows a slice from the patient that includes a real nodule (which do vary in size) and a synthetic nodule.

Significant effort has been made to ensure that the synthetic nodules are similar in form to the real nodules. But at present our model is based on simple lobulated lung nodules which consist of overlapping spheres of constant intensity. We are working on modeling more complicated spiculated lung nodules. De-identified images were modified for the purposes of this study by inserting two or three small overlapping spheres with diameter from 3.5 to 5 mm. In particular, the locations, sizes, and densities (-10 to 100 Hounsfield units) of the nodules were chosen by a pediatric radiologist. Partial volume effects and residual image blur were accounted for by creating the constant-intensity nodules in a grid of sub-voxels smaller than the image voxels, and then blurring with a Gaussian point-spread function with a full-width at half-maximum (FWHM) of 1 mm in the x- and y-directions and a FWHM of 2 mm in the z-direction. The blurred nodules were then down sampled to the original image resolution and added to the CT images. For the 3-dimensional CT data with 193 slices, the image includes over 50,000,000 voxels where 1036, 564 and 1115 voxels are part of the three nodules.

For the CT scans each expert was provided data in DICOM format, the standard for data transfer of medical images, and was asked to segment the synthetic nodules using their preferred methodology. The only constraint placed upon the experts was that they not alter the spatial resolution of the images. For this preliminary data, three segmentations were obtained using seed based region growing techniques, another obtained using an edge detection algorithm, and one obtained using a manual tracing method. Software programs utilized were Analyze (Biomedical Imaging Resource, Rochester, MN), Mimics (Materialise US, Ann Arbor, MI), and ImageJ (National Institutes of Health, USA). Any voxel that was a part of a segmented nodule was assigned a binary 1 and any voxel not part of a segmented nodule was assigned a binary 0. Two experienced experts, each with a minimum of 15 years experience in segmenting CT image data, segmented the data twice using different segmentation algorithms (4 expert segmentations). The other expert was a novice, trained by an experienced expert, who did a manual tracing of the image data.

All of the parameters of the synthetic nodules (location, diameter, and density) are known exactly and constitute "truth" for our empirical results. To convert the continuous nodules into the discretized truth, we define the true label for each voxel to be the fraction of the volume of that voxel that is part of the synthetic nodule (prior to the application of the blurring function). We created three data sets for which ground truth is known – one for each of the three synthetic nodules.

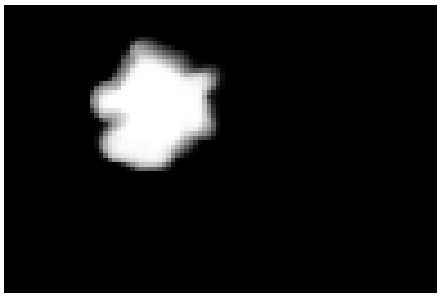

Figure 6. The ground truth for our artificial data.

4.2. Artificial Data

In addition to real CT data, we created artificial data as follows. First we created a 2-dimensional artificial nodule. (See Figure 6.) As in the real CT data it is blurred around the edges of the nodule. Then we created eight different experts, shown in Figure 7, that are intended to simulate the kinds of ways in which experts classify the data.

- Experts E_1 and E_2 represent radiologists with significant experience that provide very accurate segmentations. E_1 is created by hand marking the nodule as best possible, and E_2 is obtained by defining any pixel that has intensity of 128 (of 255) or higher to be part of the nodule.

- Expert E_3 represents an expert that tends to treat boundary pixels as part of the nodule. It is obtained by defining any pixel with intensity of 50 (of 255) or higher to be part of the nodule.

- Experts E_4, E_5, and E_6 represent a variety of expert segmentations that tend to treat boundary pixels as not part of the nodule. Expert E_4 only considers a pixel to be part of the nodule if it is a non-blurred portion of the nodule. E_5 is obtained by marking the boundaries, by hand, in a way that any blurred areas are not part of the nodule, and E_6 is obtained by only defining pixels of maximum intensity of 255 to be part of the nodule.

- Experts E_7 and E_8 represent novice experts whose segmentations are fairly simple in form. Segmentations E_7 is a circle roughly around the true nodule, and expert E_8 marks a rectangle roughly around the true nodule.

The advantage of using such artificial data is that we can vary it in controlled ways to help understand the strengths and limitations of our new algorithm. Similar to the real

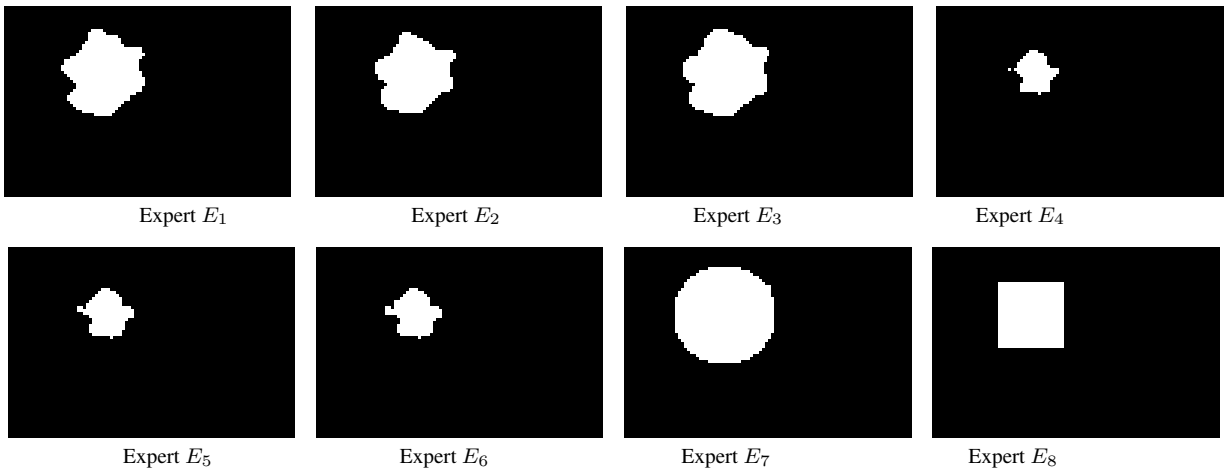

Figure 7. The eight expert segmentations used for our experiments with the artificial data.

data, when the nodule is partly overlapping a pixel, we define the ground truth as the fraction of the area of the pixel covered by the nodule.

4.3. Results

On all data sets we compare the Veritas algorithm to STAPLE, and to two baseline algorithms. The first baseline we use is the "consensus" baseline, which is obtained by using the majority vote of the m experts as the prediction. The second baseline we introduce is the "average" baseline, which predicts according to the average over the expert segmentations. For example, if there are five experts, one that indicates that voxel v is part of a nodule and four that indicate that voxel v is not part of a nodule, then the consensus baseline would predict 0 (not in the nodule) for v, whereas the average baseline would predict 0.2. For STAPLE we use the implementation from the National Library of Medicine Insight Segmentation and Registration Toolkit (ITK) [16] and run it with the default parameters.

Now we formally define the squared and absolute loss measures that we use in our work. Let A_i be the value of i^{th} voxel according to algorithm A, and T_i be the truth value of the i^{th} voxel. Let the total number of voxels where not all experts agree with each other be n, and the foreground (maximum) value of the voxel be M. Squared and absolute loss for an algorithm A is given by:

$$\text{Squared Loss} = \sum_{i=1}^{n} (A_i - T_i)^2 / (n * M)$$

$$\text{Absolute Loss} = \sum_{i=1}^{n} |\text{round}(A_i) - \text{round}(T_i)| / (n * M)$$

Since different types of images can have different ranges of values, division by M normalizes the maximum to 1.

Table 1 compares the unweighted and weighted versions of Veritas. Since the weighted Veritas performs better than the unweighted Veritas for both artificial and real CT data, and with both loss measures, for the remainder of this section we only consider the weighted version of Veritas.

Next we present our results on the artificial data set in Table 2. All the algorithms are run using all subsets of five of the eight experts ($E_1, \ldots, E_8$), subsets of six, subsets of seven experts and also when all the eight experts are used. All the results are averaged over those 93 experiments $\left(\binom{8}{5} + \binom{8}{6} + \binom{8}{7} + \binom{8}{8}\right)$.

Using the squared loss measure, Veritas has an average loss of 0.0389 per voxel compared to the ground truth. Whereas STAPLE has an average loss of 0.0858, baseline average 0.0635 and consensus 0.1475. Veritas outperforms STAPLE by an average of 54.66%, outperforms the average baseline by 38.74%, and outperforms the consensus baseline by 73.63%. Since ground truth is fractional, and STAPLE mostly predicts closer to 0 or 1, its performance is worse than the baseline average method which predicts fractional values.

Using the absolute loss measure which is more suitable to STAPLE (0.1242), it performs much better than baseline average (0.1834). Even using this measure, Veritas (0.0972) performs better than STAPLE by 21.74%.

Table 3 reports the average results from the three data sets we created from the real CT scans of a patient. For each data set we have five expert segmentations. All the algorithms are run using all subsets of four experts, and with all five experts. That gives 6 experiments for each data set, and a total of 18 experiments for all three data sets combined.

Using squared loss, while there are noticeable improvements over STAPLE (40.19%) and the consensus baseline (52.67%), on the real CT data sets the average baseline performs at a similar level as Veritas. We believe this is caused

Data Set	Loss type	Weighted Veritas loss	Unweighted Veritas loss
Artificial	Squared	$.0389 \pm .0071$	$.0619 \pm .0081$
Artificial	Absolute	$.0972 \pm .0237$	$.1647 \pm .0423$
Real CT	Squared	$.0753 \pm .0068$	$.0793 \pm .0058$
Real CT	Absolute	$.2030 \pm .0241$	$.2076 \pm .0266$

Table 1. Average loss with 95% confidence interval for Veritas with weighted and unweighted combining of experts.

Loss type	Veritas loss	STAPLE loss	Veritas percent improv.	Average Baseline loss	Veritas percent improv.	Consensus Baseline loss	Veritas percent improv
Squared	$.0389 \pm .0071$	$.0858 \pm .0047$	54.66	$.0635 \pm .0075$	38.74	$.1475 \pm .0301$	73.63
Absolute	$.0972 \pm .0237$	$.1242 \pm .0095$	21.74	$.1834 \pm .0412$	47.00	$.1834 \pm .0412$	47.00

Table 2. This table shows the average loss with 95% confidence interval for artificial data (93 experiments).

by several factors. First, the synthetic nodules have simple shapes (made of multiple overlapping spheres). Second, the segmentation errors were not necessarily independent of each other (for this experiment two experts each performed two segmentations and three of the segmentations were performed using similar segmentation techniques). Third, all the segmentations were highly accurate allowing the baseline average method to perform quite well, thus there is very little room for improvement with such high quality expert segmentations. These limitations are not found in clinical trials as lung nodules have many sizes, shapes, textures, locations, and attachments to surrounding structures and accurately and precisely measuring them remains a very challenging problem. Clinical trials are usually multi-center with distributed experts and a single expert would not generate two different expert segmentations of the same data, so independent segmentation errors would be expected as opposed to our simple trial. We are currently working on modeling larger lung nodules with more complicated geometry, texture, and locations to better represent the range of nodules found in clinical practice. We will repeat these experiments on these data sets as they become available. As seen with the artificial data, we believe that Veritas will outperform the average baseline once the expert segmentations naturally vary more from the underlying ground truth.

Using the absolute loss measure, STAPLE does better than Veritas (16.67%) and the average baseline. This behavior occurs since whenever the label expert is predicted with high accuracy using weak hypotheses constructed from the other $m-1$ experts, the performance of Veritas suffers. This mainly happens whenever the label expert is similar to one or more other experts used to create the weak hypotheses. When the weak hypotheses from the similar experts can predict the label expert well, the label expert is not really contributing any new knowledge to the group regarding the ground truth. Since our real data is very simple, the segmentations are very similar. We believe more complex lesions (which are common in real medical images) would lead to different segmentations and the performance of Veritas relative to STAPLE and the two baselines would be closer to that seen with the artificial data. The main motivation of our work is to learn the ground truth when there are different expert opinions. If all the experts are very similar (and all very accurate), there is really no need for Veritas or STAPLE – so it is the more complex lesions that are really of interest.

5. Conclusions and Future Directions

We have presented the Veritas (truth telling) algorithm to combine expert opinions when there is no labeled data for training. This is a very different problem compared to others in the machine learning literature. We have shown that Veritas compares favorably to STAPLE and two baseline algorithms for both the artificial data, and the real CT image data (under the squared loss) to which a synthetic nodule was added.

There are many directions for future work. Soon, we will obtain more data sets with more realistic and complicated synthetic nodules, and also with a larger variety of expert segmentations that will enable us to perform more extensive empirical evaluation for real medical images.

In addition, we are exploring a variety of ways to improve our algorithm by introducing features that capture the relationship between corresponding voxels from neighboring slices (using 3-dimensional neighbors instead of just 2-dimensional), and weighting the m hypotheses obtained when treating each expert as the ground truth more precisely. For example, when two learned hypotheses are very close in their accuracies, giving those hypotheses the same weight might be better.

Although the use of a Markov random field in STAPLE hard codes the constraints in regards to spatial homogene-

Loss type	Veritas loss	STAPLE loss	Veritas percent improv.	Average Baseline loss	Veritas percent improv.	Consensus Baseline loss	Veritas percent improv
Squared	$.0753 \pm .0068$	$.1259 \pm .0108$	40.19	$.0798 \pm .0052$	5.64	$.1591 \pm .0208$	52.67
Absolute	$.2030 \pm .0241$	$.1740 \pm .0173$	-16.67	$.2349 \pm .0317$	13.58	$.2349 \pm .0317$	13.58

Table 3. This table shows the average loss with 95% confidence interval for real CT data (18 experiments).

ity, it limits the ability to learn other spatial relationships. We plan to experiment with using Markov random fields as a post-processing step, or as a mechanism to weight the predictions obtained with each expert serving as the label, so that the ground truth obtained has been smoothed. Another research direction is to incorporate some domain knowledge into Veritas by using features from the raw data along with features constructed from the expert opinions to learn which experts are better under different situations.

Some other possible applications of Veritas are an unsupervised robot navigation system that has multiple obstacle detectors, or a spam detector with multiple algorithms running, each looking at different kinds of information. More generally, any setting where we have several different already trained algorithms that make predictions and we want to combine their opinions without access to any labeled data could be a potential application. Even though getting the ground truth is not as difficult in robot navigation or spam detections problems, it is possible that different algorithms are trained on different data sets before being acquired from a variety of sources.

Acknowledgments

We thank Fred Prior, Paul Commean, Kirk Smith, Bruce Whiting, Jason Fritts, Lawrence Tarbox, Mary Wolfsberger, Barry Brunsden, and Robert Schapire for comments, discussions and the segmentations. We gratefully acknowledge support from the Department of Radiology at the Washington University School of Medicine, Dr. R. Gilbert Jost, Chairman.

References

[1] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[2] S. G. Armato, M. F. McNitt-Gray, A. P. Reeves, C. R. Meyer, G. McLennan, D. R. Aberle, E. A. Kazerooni, H. MacMahon, E. J. van Beek, D. Yankelevitz, E. A. Hoffman, C. I. Henschke, R. Y. Roberts, M. S. Brown, R. M. Engelmann, R. C. Pais, C. W. Piker, D. Qing, M. Kocherginsky, B. Y. Croft, and L. P. Clarke. The lung image database consortium (LIDC): An evaluation of radiologist variability in the identification of lung nodules on CT scans. *Academic Radiology*, 14(11):1409–1421, Nov. 2007.

[3] S. G. Armato, R. Y. Roberts, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, G. McLennan, R. M. Engelmann, P. H. Bland, D. R. Aberle, E. A. Kazerooni, H. MacMahon, E. J. van Beek, D. Yankelevitz, B. Y. Croft, and L. P. Clarke. The lung image database consortium (LIDC): Ensuring the integrity of expert-defined "truth". *Academic Radiology*, 14(12):1455–1463, Dec. 2007.

[4] A. Blum. Empirical support for Winnow and weighted-majority algorithms: Results on a calendar scheduling domain. *Machine Learning*, 26(1):5–23, 1997.

[5] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 92–100, New York, NY, 1998. ACM Press.

[6] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, 2006.

[7] M. Collins and Y. Singer. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 100–111, 1999.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B39:1–38, 1977.

[9] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer System Sciences*, 55(1):119–139, 1997.

[10] D. Greig, B. Porteous, and H. Seheult. Exact maximum *a posteriori* estimation for binary images. *Journal of the Royal Statistics Society, B*, 51:271–279, 1989.

[11] C. C. Jaffe. Lecture to American college of radiology. *IEEE Transactions on Medical Imaging*, 1(4):226–229, 1982.

[12] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, Cambridge, Massachusetts, 1994.

[13] F. W. Prior, B. J. Erickson, and L. Tarbox. Open source software projects of the caBIG in vivo imaging workspace software special interest group. *Journal of Digital Imaging*, 20(1):94–100, Nov. 2007.

[14] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

[15] S. K. Warfield, K. H. Zhou, and W. M. Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[16] T. Yoo, M. Ackerman, W. Lorensen, W. Schroeder, V. Chalana, S. Aylward, D. Metaxes, and R. Whitaker. Engineering and algorithm design for an image processing API: A technical report on ITK - the Insight Toolkit, 2002.