An Evaluation of Consensus Techniques for Diagnostic Interpretation

Jake Sauter[a], Victoria LaBarre[b], Jacob Furst[c] and Daniela Raicu[c]

[a] SUNY Oswego, 7060 NY-104, Oswego, NY 13126, USA
[b] McLennan Community College, 1400 College Dr, Waco, TX 76708, USA
[c] Intelligent Multimedia Processing Laboratory, DePaul University, College of Computing and Digital Media, 243 S. Wabash Avenue, Chicago, IL 60604, USA

Abstract

Most computer-aided diagnosis systems use low-level image features extracted directly from image content to train and test machine learning classifiers for diagnostic label prediction. When the ground truth for the diagnostic labels is not available, reference truth is generated from the experts' diagnostic interpretations of the image/region of interest. In this paper, we evaluate three consensus approaches (median, mode, and mean) that are typically used to encode the variability in the experts' labeling of the medical data. Given that the NIH/NCI Lung Image Database Consortium (LIDC) data provides interpretations for lung nodules by up to four radiologists, we leverage the LIDC data to evaluate and compare these consensus approaches when creating computer-aided diagnosis systems for lung nodules. First, low-level image features of nodules are extracted and paired with their radiologists' semantic ratings (1= "most likely benign",…, 5="most likely malignant"); second, machine learning multi-class classifiers that handle deterministic labels (decision trees) and probabilistic vector of labels (belief decision trees) are built to predict the lung nodules' semantic ratings. We show that the mean-based consensus generates the most robust classifier overall when compared to the median- and mode-based consensus.

Description of Purpose

Lung cancer is the leading cause of cancer death and the second most common cancer among both men and women[1], causing a critical need to explore ways to diagnose malignant lung nodules early on. Computer-aided diagnosis (CAD) systems have been developed to assist radiologists by providing an additional opinion. To create these CAD systems, we require label data from radiologists representing their diagnosis interpretations. However, annotations from multiple radiologists are often uncertain because of the subjectivity of the interpretation process and the variability among annotators. We propose a study of different approaches to reaching a consensus label that minimizes false positive and false negative rates and maximizes true negative and true positive rates. The results of this study will show that, when building CAD systems with uncertain diagnostic interpretation, it is important to evaluate different strategies for encoding and predicting the diagnostic label.

Methods

A. The Lung Image Database Consortium (LIDC) Data

The LIDC dataset[2] (available at http://ncia.nci.nih.gov) is a diverse collection of Computed Tomography (CT) scans interpreted by up to four radiologists. Each radiologist outlined a boundary for the nodule or nodules present in the scan, as well as provided ratings on various semantic characteristics (such as malignancy, texture, margin, spiculation, and lobulation) for the nodule as a whole. The challenges of the LIDC data include the disagreement among radiologist annotators, the multi-class label (for example, there are five ratings for malignancy rather than the traditional two class problem - malignant versus benign), and lack of ground truth. These challenges make the LIDC data a good candidate to study behavioral changes of classifiers when the reference truth is calculated in various ways. While in this study we focus only on the malignancy ratings, the same approach can be applied to the other semantic characteristics (e.g. degree of spiculation, lobulation, and texture). Further, to analyze the variability among annotators, we consider only the 809 nodules out of the 2,600 nodules for which all four radiologists provided a rating for malignancy.

B. Low-level Image Feature and Consensus-based Label Extraction

Based on our previous work[3], we extracted 64 low-level image features for each nodule instance delineated by the radiologists' largest outline across all slices in which the nodule appeared[4]. These image features encode the lung nodule intensity, size, shape, and texture.

The malignancy label has five ratings (1= "most likely benign",…, 3="indeterminate",…, 5="most likely malignant) which makes the classification a multi-class problem. Furthermore, given the variability in the radiologists' interpretation, for the same nodule, there could be multiple ratings. Therefore, we encode the malignancy label using a vote vector which is further converted into a probabilistic label vector (PLV) which represents the probability distribution over the five possible class ratings of malignancy for that specific nodule.

Three methods of calculating a zero-initialized PLV, denoted as *P*, from a vote vector *V* were tested, what we call mean, median and mode consensus techniques. These techniques generate a consensus label *C*, which is then converted into *P*. The mean vote vector consensus technique is applied by first arriving at *C* by the standard definition of mean. This *C* is converted to a probability distribution over the five possible classes. In a similar fashion, the median- and mode-based consensus label *C* are created for the same nodule diagnostic interpretation. For example, if the radiologists' rating for a certain nodule were *V*= [2, 2, 3, 4], the mean of this vote vector is *C*=2.75, leading to a *P* vector defined as [0, .25, .75, 0, 0] with a higher probability assigned to 3 than 2 given that the mean is closer to 3 than to 2. Similarly for median and mode conversion of *V*, *C* = 2.5 and 2, respectively, which will further produce P = [0, .5, .5, 0, 0] for the median label and P= [0, 1, 0, 0, 0] for the mode label.

We name our approach of encoding the uncertainty of the label *consensus PLV* (Figure 1) because it first creates the consensus and then it encodes it as a probabilistic vector of labels. This is different from previous approaches[5], including ours, in which we used a *direct PLV approach* in which the vector V is converted to P directly without considering consensus. For example, the direct PLV approach will convert V into P based on the probabilities of each rating (*P*= [0, .5, .25, .25, 0]).

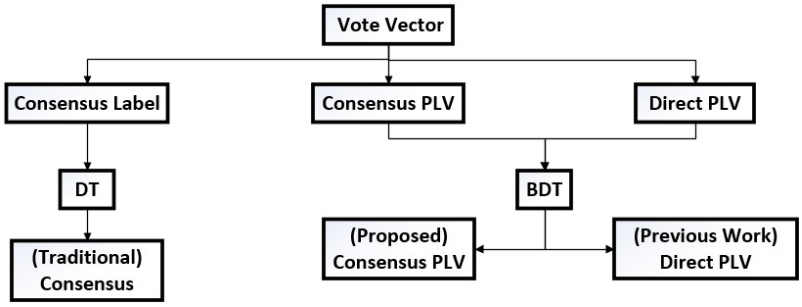

Figure 1: Diagram of the different methods to evaluate consensus-based classifiers

C. Classification Models

We analyze the impact of encoding the reference truth through consensus-based approaches using Belief Decision Tree (BDTs). BDTs inherently model uncertainty through predictions being probabilistic distributions over the possible classes. This, however, requires extra processing because vote vector *V* is converted to consensus label *C*, then into probabilistic vector *P* for a target label to train the BDTs, and then the predicted *P* is converted back to predicted *C* for classification performance assessment. Therefore, we compare the results for the belief decision trees with those obtained using simple decision trees[6] trained and tested on the consensus-based label C (Figure 1).

The construction of the BDTs[6] is similar to the traditional decision trees with the exception that the decision to split a node is based on averaging the pignistic probabilities (called basic belief assignment (BBA)) for all the cases that reach that node of the tree. The average pignistic probabilities of the parent and child nodes can then be used to calculate the information gain to decide the split with respect to each possible feature and threshold value in the dataset. Given that the information gain split criteria can produce heavily unbalanced trees, we propose to use the gain ratio which controls for the size of the child subsets and rewards equally distributed splits. The low-level image feature that achieves the maximum gain ratio for the corresponding split is then selected at that node to perform the partitioning of the cases. Furthermore, to determine whether a node in a BDT is a leaf node, four stopping criteria were used: the maximum gain ratio of splitting is 0, there is no split that can be made which will result in acceptable numbers of cases at the parent and child nodes, all the BBA's at the node are equivalent, or all features have already been used to split[6].

Results

In Figure 2, true positive (TP), correctly classified indeterminate (TI), true negative (TN), false positive (FP), incorrectly classified indeterminate (FI), false negative (FN) rates are compared across consensus-based techniques and classification models. The grouping of bars on the left represents the actual negative cases and the grouping on the right represents the actual positive cases. Within the groupings, spans at the top of the bars indicating results from the DTs can be found on the left and results from the BDTs can be found on the right, noted by a span on the top of the respective bars. The red regions of the bars represent incorrectly predicted category, grey regions represent indeterminate predicted category (predicted as a rating of 3), and green regions represent correctly predicted category. Included at the right of these two groupings are rates calculated from directly comparing the distribution of the BDT without forming a consensus label, to compare how forming a consensus affects these rates.

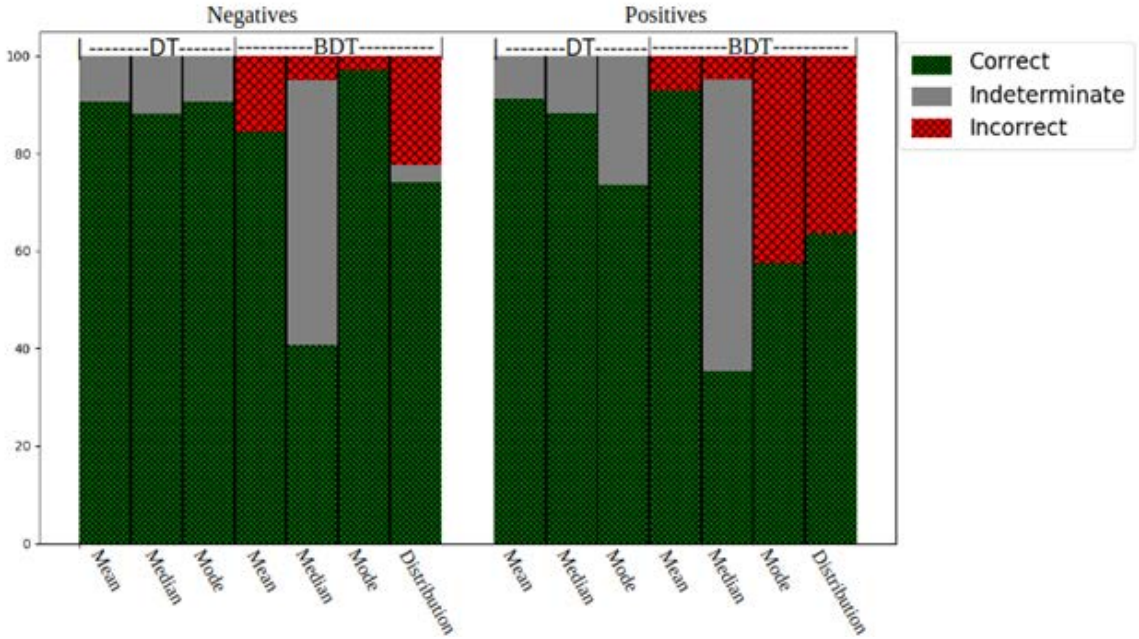

Figure 2: True Positive/Negative and False Positive/Negative Rates, includes Indeterminate Rate

It is important to specify that labels predicted as a three (neither malignant nor benign) are calculated separately as the "Indeterminate Rate". This means that the total true positive/negative or false positive/negative rates do not sum to a total of 100 percent.

New or breakthrough work to be presented

In some studies, choosing a consensus of a list of votes is necessary but the method of consensus has not been deeply explored[7]. This study is the first to perform a comparison of classifiers' behavioral changes with use of different consensus techniques to form a reference truth. This study also investigates the comparison of a classifier with a single class output against a classifier that outputs a probabilistic distribution over possible classes.

Conclusions

After comparing results from a classifier with and without consensus integrated in the label, we found that a consensus produced more robust results due to lower FP/FN rates and higher TP/TN rates overall. Mean-based consensus has the highest ratio of positive responses to negative responses among the three consensus techniques over both classifiers. We also found that the model that was trained on a consensus PLV performed better that the model trained on a direct PLV, due to the fact that the prediction indicating the correct single label was our qualifier for accuracy. We should note that comparing distributions not only allows data to be retained, but also allows for more intuitive evaluation of outputs, though a tradeoff of robustness to information loss is present. This tradeoff is especially important to consider when lacking the ground truth. Future research into belief decision trees and their probabilistic vector outputs will be pursued, possibly utilizing AUCdt[8] to aid in the evaluation of a distribution-based classifier rather than consensus-based classifier.

This work is not being submitted for presentation or publication elsewhere.

References

[1] American Cancer Society. "Key Statistics for Lung Cancer." American Cancer Society. American Cancer Society, 5 Jan. 2017. Web. 17 July 2017.

[2] Armato, S. G., et al. (2004). Lung Image Database Consortium: Developing a Resource for the Medical Imaging Research Community. Radiology, 232(3), 739-748.

[3] Zinovev, D., J. Feigenbaum, J. Furst, and D. Raicu. "Probabilistic lung nodule classification with belief decision trees." *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2011): n. pag. Web.

[4] Riely, Amelia, et al. "Reducing Annotation Cost and Uncertainty in Computer-Aided Diagnosis through Selective Iterative Classification." Medical Imaging 2015: Computer-Aided Diagnosis | MI15 | SPIE Proceedings | SPIE, International Society for Optics and Photonics, 20 Mar. 2015, proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=2211245.

[5] Affenit, Rachael N., et al. "Building Confidence and Credibility into CAD with Belief Decision Trees." Medical Imaging 2017: Computer-Aided Diagnosis | MI17 | SPIE Proceedings | SPIE, International Society for Optics and Photonics, 3 Mar. 2017, proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=2609070.

[6]Elouedi, Z., Mellouli, K., & Smets, P. (2001). Belief decision trees: Theoretical foundations. International Journal of Approximate Reasoning, 28(2-3), 91-124.

[7] Filho, Antonio Oseas de Carvalho, et al. "Computer-Aided Diagnosis of Lung Nodules in Computed Tomography by Using Phylogenetic Diversity, Genetic Algorithm, and SVM." SpringerLink, Springer International Publishing, 19 May 2017, link.springer.com/article/10.1007/s10278-017-9973-6.

[8]Williams, Sydney, et al. "Area under the Distance Threshold Curve as an Evaluation Measure for Probabilistic Classifiers." SpringerLink, Springer, Berlin, Heidelberg, 19 July 2013, link.springer.com/chapter/10.1007/978-3-642-39712-7_49.