**Reducing Annotation Cost and Uncertainty in Computer-Aided Diagnosis through Selective Iterative Classification**

Amelia Riely, Kyle Sablan, Xiaotao Fang, Jacob Furst, and Daniela Raicu

**Abstract:** Advances in medical imaging technology have expanded the need for Computer-Aided Diagnosis (CAD) systems to assist radiologists in analyzing image data. However, where there is no ground truth diagnosis, experts in CAD systems often turn to multiple annotations to approximate a reference truth. Especially in the medical field, each of these annotations is expensive, and may not necessarily add to the quality of the classification model. This paper outlines a *selective iterative* CAD system to minimize the number of annotations required to obtain desirable accuracy, by requesting additional labels only for cases that could benefit from them. This approach simultaneously increases accuracy over a non-selective system, since it reduces noise and uncertainty in the labels. In the case of the LIDC dataset, our results show that using only 46.91% of the original label set, we can achieve 81.17% accuracy, as opposed to only 70.86% when 100% labels are used.

**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 find 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, it requires label data from radiologists representing their diagnosis interpretations. However, annotations from multiple radiologists are very costly, both in time and money, and are often uncertain because of the subjectivity of the interpretation process and the variability among annotators. We propose an approach that minimizes the number annotations needed to correctly classify or diagnose a case, and therefore, benefiting cost and label uncertainty reduction.

**Method(s)**

*A. Lung Image Database Consortium (LIDC)*

The LIDC dataset (available at <http://ncia.nci.nih.ogv>) used in this study is a diverse collection of Computed Tomography (CT) scans interpreted by up to four radiologists [2]. 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 including the disagreement among radiologists, 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 make the LIDC data a good candidate to demonstrate our iterative selective approach for CAD systems. While in this study we focus only on the malignancy ratings, the same approach can be applied for the other semantic characteristics. Further, to analyze the variability among annotators, we consider only the 810 nodules out of the 2,600 nodules for which all four radiologists provided a rating for malignancy.

*B. Feature Extraction*

Based on our previous work for CAD systems [3], we extracted 64 low-level image features for each nodule instance delineated by the largest radiologist outline across all slices in which the nodule appeared. Eight shape features included circularity, roughness, elongation, compactness, eccentricity, solidity, extent, and standard deviation of the radial distance. Seven size features included area, ConvexArea, perimeter, ConvexPerimeter, EquivDiameter, MajorAxisLength, MinorAxisLength. Intensity features were minimum, maximum, mean, and standard deviation of the gray-level intensity for the segmented image and its background. Texture features included five Markov features and twenty-four Gabor features.

Each nodule instance also has four malignancy ratings associated with it in the range from 1 to 5, where 1 is “*Highly Unlikely,”* 3 is *“Indeterminate,”* and 5 is *“Highly Suspicious*.” However, since the ratings 2 and 4 were underrepresented in the dataset, we rescaled the labels from a 1-5 rating to a 1-3 rating, where the original 1 & 2 became 1, the original 3 becomes 2, and the original 4 & 5 become 3. This created a more even distribution of ratings and also decreased the complexity of the multi-label problem.

*C. Classification Models*

Our classification method involved the creation of two sets of four decision trees (given the four annotators). We used Classification and Regression Tree (CRT/CART) decision trees introduced by Breiman [4] because of their simplicity and because they make no assumptions about the distribution of the data. To build the model, the data was split into training (60%), testing (30%) and validation (10%). The validation dataset was kept separate from other data during the entire process and used for evaluation to simulate how completely new data will be classified by the iterative approach. The process was repeated for twenty trials to take into account the variation in the accuracy as well as with different label permutations to eliminate the process of getting additional labels and eliminate any potential label bias.

*C.1 Selective Iterative Classification*

We call our approach selective iterative classification because the four classification models, *Mk* (*k* = 1:4, 4 being the number of annotators) we create are built off of each other so that the labels used to create one model depend on which cases were misclassified in the previous model. For example, the first classification model, *M1*, is trained using only one randomly selected label among the four labels provided by LIDC radiologists as the reference truth. The label for an instance *I*, , is selected randomly out of the set of all available labels for that case, , and removed from the set to simulate the process of acquiring a new annotation from an expert. If the instance *I* is misclassified at the by the first model, the label used for the second model, , will be the mode of label from the first iteration and a new label selected out of the set of all available labels. If the instance, is not misclassified by the first model, then the reference truth used to create the second model, , will not change from the reference truth of the first model, .

*C.2 Non-Selective Iterative Classification*

We compared our proposed approach with a non-selective iterative approach [3] in which additional labels are required iteratively for all cases instead of just the miss-classified ones. Thus, in the general case of *p annotators,* the number of used labels *t*, is always equal to *k*, the number of the iteration and all instances will receive their maximum number of labels, *p*, by the end of the algorithm.

**Results**

In Table I, we report the accuracy in predicting malignancy for the validation set at each iteration for both iterative approaches above. Accuracy was calculated as the mean accuracy over all 20 trials. The reported label count is for the most representative trial, and represents the total number of labels used to form any given model (other trials show a similar distribution). Marginal benefit is calculated as change in accuracy between iterations. Cost-based marginal benefit was calculated as marginal benefit over the increase in label count. The parameters for the decision trees that were used to report the results from Table I were empirically determined as 250 for the minimum case per parent and 58 for the minimum per child.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Approach** | **Iteration** | **Accuracy** | **Margin of Error** | **Marginal Benefit** | **Cumulative Label Count** | **Cost-Based Marginal Benefit** |
| *Selective* | 1 | 52.59% | 2.35% |  | 810 |  |
|  | 2 | 70.62% | 2.65% | 18.03% | 1165 | 0.051% |
|  | 3 | 77.35% | 2.45% | 6.73% | 1366 | 0.033% |
|  | 4 | 81.17% | 2.99% | 3.82% | 1520 | 0.025% |
| *Non-Selective* | 1 | 52.59% | 2.35% |  | 810 |  |
|  | 2 | 66.23% | 3.26% | 13.64% | 1620 | 0.017% |
|  | 3 | 65.80% | 2.50% | -0.43% | 2430 | -0.001% |
|  | 4 | 70.86% | 2.23% | 5.06% | 3240 | 0.006% |

Table I: Performance comparisons for validation set.

As one would expect, in both approaches the marginal benefit of additional labels decreases in the later iterations. Further, the selective approach achieved much higher accuracy even in early iterations than the non-selective approach, using far fewer labels. If each label corresponds to a fix price of acquisition, the selective approach after four iterations cost 46.91% of what the non-selective approach cost at the same point and achieved 10.31% higher accuracy. This gain in accuracy is related to the reduction in label uncertainty that occurs when labels are judged as adequate or inadequate based on other labels that may be more certain and are linked to a large set of image features.

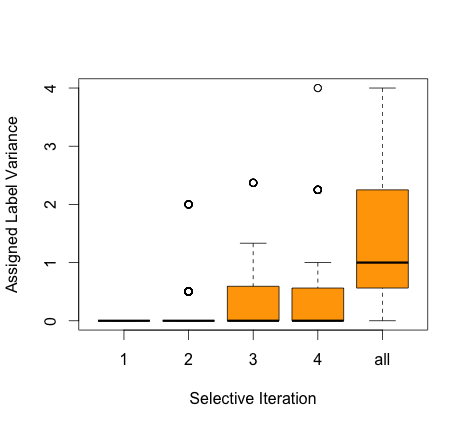
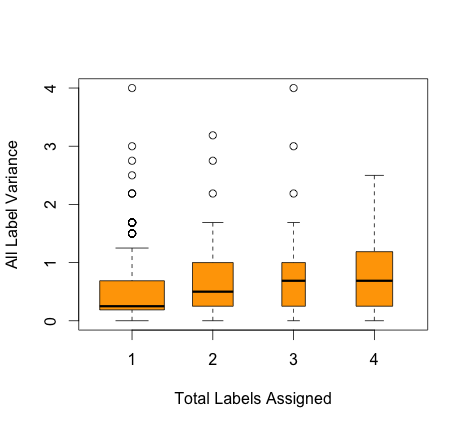
Figure I: **left)** Label variances for cases receiving one label, two labels, three labels, and four labels respectively; **right)** Label variances at each selective iteration and for all possible labels provided by radiologists.

Figure I.left shows a boxplot of the variances per case of all radiologist labels, separated by the total number of labels required at the final iteration of the selective algorithm. A higher variance indicates more radiologist disagreement and a harder to diagnose case. Cases that required more labels from the algorithm had higher variances on average than cases that required fewer labels. This tendency to agree with radiologists on which cases are harder to diagnose makes the selective iterative algorithm a strong candidate for a future easy vs. hard differentiation as described in [3].

We can further quantify the reduction in uncertainty by looking at the variance of the labels used in each iteration of the selective approach in comparison with the variance of all four labels, which would be used in a traditional all-label consensus approach. Figure I.right shows that the variance even at the fourth iteration of the selective iterative model is much lower overall than the variance of all radiologist labels, which have a high degree of uncertainty. This u ncertainty from the labels is transferred to the classification model and results in lower accuracy numbers, as seen in Table I. With the iterative selective approach, we use the image features from all instances to determine whether the image features support the diagnosis of the current label. If they do not agree, we ask for another opinion until the model and the label agree under the assumption that the current label has a high degree of uncertainty. If they do agree, no further labels are requested, avoiding the possibility of acquiring new labels that increase cost and only add uncertainty, unless the instance is misclassified in a later iteration.

**New or breakthrough work to be presented**

While there are some studies that present iterative approaches for other domains such as optical character recognition [5], violence risk assessment [6], and ranking of networks [7], this study is the first study to look at iterative cost reduction and uncertainty when selectively obtaining annotations for building computer-aided diagnosis systems. Furthermore, all related work from the other domains assume knowledge about annotator quality, as well as a ground truth, both of which are unavailble with the LIDC and many other medical datasets.

**Conclusions**

From our preliminary results, our selective iterative classification method used less than half of the labels of a traditional consensus based CAD system, as in [3], and achieved significantly better accuracy than the non-selective iterative approach. Selectively allocating additional labels using our algorithm reduces cost and uncertainty in CAD systems.

**References**

[1] *Centers for Disease Control and Prevention*. Centers for Disease Control and Prevention, 15 May 2014. Web. 29 July 2014. <<http://www.cdc.gov/cancer/lung/>>.

[2] Armato SG III, McLennan G, Bidaut L, et al: *The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans*. Med Physics, 38: 915-931, 2011.

[3] Zamacona, Jose R., Alexander Rasin, Jacob D. Furst, and Daniela S. Raicu. *Reducing Classification Cost through Strategic Annotation Assignment*. Proc. of 2013 IEEE 13th ICDM, DMBIH Workshop, USA, Dallas, TX. N.p.: IEEE, 2013. 287-94.

[4] Breiman, L., Friedman, J.H., Olshen, R., Stone and C.J. *Classification and Regression. Tree* Wadsworth & *Brooks/Cole Advanced Books & Software, Pacific California.* 1984.

[5] Trapeznikov, Kirill, and Venkatesh Saligrama. *Supervised Sequential Classification Under Budget Constraints*. Proc. of 16th International Conference on Artificial Intelligence and Statistics (AISTATS), USA, Scottsdale, AZ. Vol. XX. 2013.

[6] Monahan, J. *Developing a Clinically Useful Actuarial Tool for Assessing Violence Risk*. The British Journal of Psychiatry 176.4 (2000): 312-19.

[7] M. Ji, J. Han, and M. Danilevsky. *Ranking-based classification of heterogeneous information networks*. In Proc. 2011 ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD’11), San Diego, CA, August 2011.