Building confidence and credibility into CAD with belief decision trees

Rachael N. Affenit[2], Erik R. Barns[1], Jacob D. Furst[1], Alexander Rasin[1], Daniela S. Raicu[1]

[1] College of Computing and Digital Media
DePaul University
Chicago, IL, US

[2] College of Science
Illinois Institute of Technology
Chicago, IL, US

ABSTRACT

To improve on existing CAD systems, we first examined Belief Decision Trees with an iterative classification approach to provide a vector of predicted probabilistic labels for each case in a balanced version of the Lung Image Database Consortium (LIDC) dataset. We then applied conformal prediction to these results in order to analyze the reliability of the predictions (which was then used in the implementation of iterative classification), and typicality of each case. Probabilistic predictions given with levels of confidence and credibility can build a smarter CAD system that provides more contextual information to the radiologist.

Keywords: *Computer-aided diagnosis, belief decision tree, iterative classification, conformal prediction, reliability, confidence, credibility*

INTRODUCTION

Computer-aided diagnosis (CAD) systems provide radiologists with supplemental diagnostic information for use when analyzing patient CT or X-Ray images. These systems reduce the work required to assess an image by quantitatively analyzing images and predicting qualitative characteristics of the case. To improve on current CAD systems, we investigate strategies to produce better prediction performance and analysis of the quality of a CAD prediction when handling uncertain data. We first examine Belief Decision Trees with an iterative classification approach to provide a vector of predicted probabilistic labels for each case in a balanced version of the Lung Image Database Consortium (LIDC) dataset. We also propose applying conformal prediction to these results in order to analyze the reliability of the predictions, and typicality of each case. Probabilistic predictions given with levels of confidence and credibility can build a smarter CAD system that provides more contextual information to the radiologist. The Belief Decision Tree itself as well as conformal prediction have been implemented with 43.5% testing accuracy and 0.64 Area under the ROC Curve (AUC) on our mean-balanced dataset with 800 cases (514 proper training, 86 calibration, 200 testing).

PURPOSE

CAD systems currently have a relatively low adoption rate in clinical setting, in part because radiologists do not necessarily trust the results due to the lack of supporting contextual information. Without giving clinicians a window into the CAD system's reasoning, or estimates of how well it can predict a particular case, radiologists may come to distrust its results as they observe more misclassified cases [1]. A more comprehensive statistical and contextual output for each case may help build their trust in these types of systems, and give clinicians a better understanding of how well a prediction fits each case. By providing probabilistic labels for potential diagnoses, along with self-evaluation features like case reliability and typicality, we hope to dispel some of the doubt in the potential of these systems.

METHODS

The LIDC dataset [2] contains between one and four radiologist ratings of cancer malignancy for each case, on a scale of 1 (benign) to 5 (malignant). We are working with a mean-balanced subset of this data, which includes only those cases where all four radiologists identified and rated a nodule. The data were balanced between mode ratings by significantly under sampling the label 3 (uncertain) cases, as there were far more of these than the rest, and slightly over sampling cases with other labels. Unfortunately, in this setting a radiologist rating is not akin to a ground truth in machine learning, and the four radiologists agree on a consensus label in only 25% of these cases [3]. This lack of consensus can introduce unwanted bias when classifying new nodules. In order to deal with this uncertainty, we have implemented an algorithm based on a probabilistic classifier called Belief Decision Tree (BDT). It is an adaption of a decision tree (DT) classifier that uses belief function theory to better handle uncertainty.

In a similar vein to a decision tree, a BDT classifies an LIDC case by comparing calculated image feature values to the chosen threshold values to determine which path in the tree the case should follow. When a case reaches a leaf node, it can be assigned a Basic Belief Assignment (BBA) associated with this node as a method of classifying that case. This BBA is a set of probabilities for each of the five classification labels, and represents the average BBA of all cases in the training set that reached this node [4]. The training case BBA's were created using the radiologist ratings from the dataset; for example, ratings of 2, 3, 4, and 4 would yield the BBA [0 .25 .25 .5 0] for five label distribution. Typically the process of calculating these probabilities is much more involved for a belief decision tree, but the LIDC dataset has a few special qualities that allow us to use this method. For one, every radiologist can only pick one label for each

case, and for another, we have no way to gauge the uncertainty of said choice. This allows us to eliminate union labels (label1 U label2) and pure uncertainty labels in our BBA, resulting in a simple probability distribution [5].

The biggest difference between a decision tree and a BDT occurs during tree construction. When deciding whether and how a node should split, a BDT calculates the pignistic probabilities of each class for every case in the dataset (which becomes our BBA), and averages the probabilities of all the cases that reach each node in the tree. The average pignistic probabilities of the parent and child nodes can then be used to calculate the information gain of splitting, using each possible feature and threshold value in the dataset. It then computes the gain ratio, which controls for the size of the child subsets and rewards equally distributed splits, and chooses the feature and threshold that achieved the maximum gain ratio for the split. One can determine whether a node in a BDT is a leaf node if it meets one of four stopping criterion: the maximum information gain of splitting was 0, there is no split that can be made which will result in acceptable numbers of cases at the parent and child nodes (given by n_p and n_c parameters), all of the BBA's at the node are equivalent, or all features have already been used to split [4].

We would like to incorporate an iterative classification approach with these BDTs in order to reinforce their ability to handle the uncertainty in radiologist labels and achieve a reasonable consensus [6, 7]. In this project, we propose the use of iterative classification with four belief decision trees, each incorporating an increasing number of radiologist ratings. After a new case is classified by each tree, the BBA output would be compared to the best previous BBA. If it was better, this would become the best overall BBA. This process provides four different possible BBA's for each case. We can then determine the optimal BBA for the case based on their respective confidence and credibility scores.

Finally, we are implementing Conformal Prediction (CP) in our BDT to produce measures of confidence and credibility for each CAD probability distribution. CP begins as a typical classification problem: the dataset is divided into a training and a testing set, and from the training set we derive a calibration set, which is used to facilitate conformal prediction. The calibration set is classified using the BDT produced by the training set. Using the conformity function given in equation 1 to determine conformity scores for each case (which correlate with case typicality), we then build a set of calibration conformity values. Positive conformity scores represent more typical cases, whereas negative scores represent more atypical cases. Johansson et.al. [8] defines calibration conformity as:

$$\alpha_i = p_i^Y - max_{j=1\ldots C: j\neq Y}\, p_i^j \qquad \text{Eq. 1}$$

In equation 1, α_i is the conformity score for the ith case, p_i^Y is the probability of the actual classification of the case, and $max_{j=1\ldots C: j\neq Y}\, p_i^j$ is the maximum probability of all of the remaining labels in the case. After these calibration conformity values have been calculated, the testing set is run through the classifier to find the predicted labels for the testing cases. With these predicted labels, we can compute the testing conformity of each case using equation 2 below. Testing cases are not associated with a true label, and therefore we must calculate a conformity score for each possible label, as defined by Johansson et.al. [8]:

$$a_i^k = p_i^k - max_{j=1\ldots C: j\neq ck} p_i^j \qquad \text{Eq. 2}$$

Shown in equation 2, a_i^k is the conformity score for class k in the ith case. p_i^k is the probability of class label k and $max_{j=1\ldots C: j\neq ck} p_i^j$ is the maximum probability from the remaining class labels for that case. Utilizing the calibration and testing conformity scores, we can calculate the p-values of the testing cases. This allows us to transform case conformity into our measures of reliability, confidence and credibility. To calculate the p-values, we compare each of the testing conformity scores for a case to the set of calibration conformity scores. It represents the ratio of conformity scores in the calibration set that are less than or equal to the conformity score of that label to the total number of instances in the calibration set. Using this p-value, we can calculate confidence and credibility for each prediction as both an output for the radiologist to consider, and as a method of choosing the best BBA during iterative classification. The p-value as defined by Johansson et.al. [8] is:

$$P_{ik} = \frac{\left|\left\{j=1..l \;\&ik: a_j \leq a_i^k\right\}\right|}{l+1} \qquad \text{Eq. 3}$$

Equation 3 defines the p-value, $P_{ik,}$ for class k of the ith case. P_{ik} is shown as a_j, the number of calibration conformity scores that are less than or equal to the case conformity score a_j^i, over the number of calibration conformity scores. This produces a vector of p-values for a case, which can be used to compute confidence (Cf_i) as one minus the second highest probability where p_j^i is the vector of p-values, and credibility (Cr_j) as the maximum probability of p_{ij}, as defined by Johansson et.al. [8]:

$$Cf_i = 1 - secondMax_{j=1\ldots k}\, p_j^i \qquad \text{Eq. 4}$$
$$Cr_j = max_{j=1\ldots k}\, p_{ij} \qquad \text{Eq. 5}$$

PRELIMINARY RESULTS

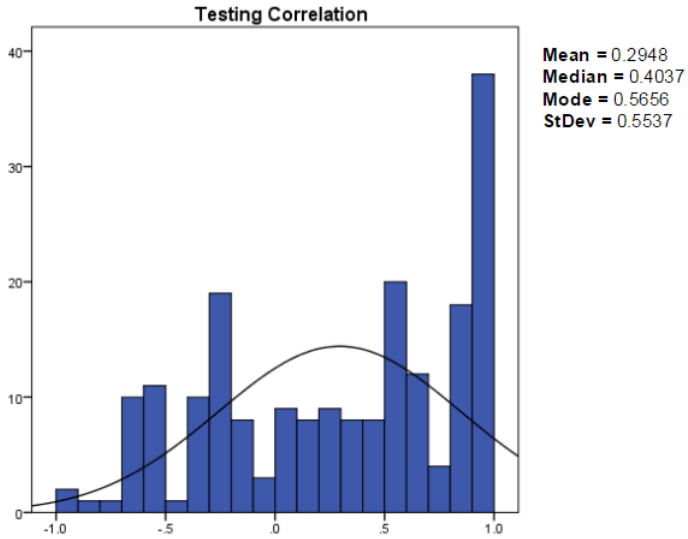
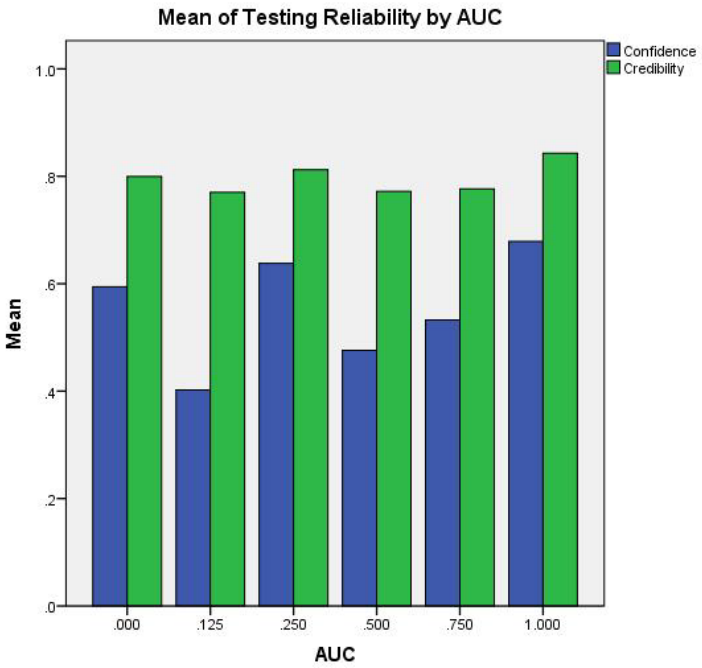


Figure 1. *Histogram of Pearson correlation coefficient values between actual and predicted distributions for the conformal predictive BDT on the testing set with np = 23, nc = 11, and d = 25.*

Our BDT with conformal prediction achieves 43.5% testing accuracy (20% is random classification) and 0.64 AUC (0.5 is random classification) on our mean-balanced dataset with 800 cases (514 proper training, 86 calibration, 200 testing). Its optimal settings were found to be $n_p = 23$, $n_c = 11$, and $d_{max} = 25$, where n_p represents the minimum number of cases a parent node must have to split, n_c represents the minimum number of cases a child node must have for its parent to split, and d_{max} represents the maximum tree depth. Accuracy was defined by taking the maximum probability label from the actual and predicted label distributions, and assuming these were the actual and predicted labels respectively. The accuracy values obtained are likely influenced by the small size of our initial dataset, and the necessity of splitting it further into three even smaller subsets. Accuracy may also be lowered by forcing a consensus out of intentionally uncertain probability distributions, in which case accuracy is likely not the best measure of performance. To improve the accuracy of our tree, we will use k-fold cross validation with k = 4 for further testing [9].

To compare the actual and predicted distributions without forcing consensus, we currently calculate the Pearson correlation between the actual and predicted distributions, shown in Figure 1, which helps us assess the BDT's ability to predict the correct distribution. Pearson correlation for our sample size of 200 testing pairs is significant with $p < 0.05$ at any correlation above 0.139, meaning that 62% of our test instances have a statistically significant correlation between actual and predicted distributions.



np = 23, nc = 11, d = 25 on mean-balanced dataset as compared with the ROC AUC score between the actual and predicted vectors of the case.

We plan to implement the Area under a Distance Threshold Curve (AUCdt) [10] measure to better compare the distributions. The iterative classification step has been implemented, and can be added in at any time, but we chose to remove it until our implementation of conformal prediction is ready, as the process depends on the confidence output from conformal prediction to choose a final BBA for a case.

As shown in Figure 2, our mean CP results show a nonlinear variability for confidence with respect to AUC, and little variation in credibility. We will investigate this phenomenon further to determine the relationships between our performance and reliability measures.

INNOVATION & CONCLUSIONS

By incorporating conformal prediction into the decision-making process of an iterative probabilistic classifier, we are proposing a CAD system that will provide more informative probabilistic predictions for new cases, which would include measures of confidence and credibility for those predictions. In so doing, we aim to generate results that give clinicians a better idea of the context surrounding the predictions, and build their trust in the viability CAD tools for clinical use.

REFERENCES

[1] Drukker, K., Sennett, C., & Giger, M. (2009). Automated Method for Improving System Performance of Computer-Aided Diagnosis in Breast Ultrasound. *IEEE Transactions on Medical Imaging IEEE Trans. Med. Imaging*, *28*(1), 122-128.

[2] Iii, S. G., Mclennan, G., Mcnitt-Gray, M. F., Meyer, C. R., Yankelevitz, D., Aberle, D. R., . . . Clarke, L. P. (2004). Lung Image Database Consortium: Developing a Resource for the Medical Imaging Research Community1. *Radiology*, *232*(3), 739-748.

[3] Ochs, R., Kim, H. J., Angel, E., Panknin, C., Mcnitt-Gray, M., & Brown, M. (2007). Forming a reference standard from LIDC data: Impact of reader agreement on reported CAD performance. *Medical Imaging 2007: Computer-Aided Diagnosis*.

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

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

[6] Ji, M., Han, J., & Danilevsky, M. (2011). Ranking-based classification of heterogeneous information networks. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11*.

[7] Whitehill, J., Wu, T., Bergsma, J., Movellan, J. R., & Ruvolo, P. L. (2009). Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. *Neural Information Processing Systems (NIPS)*, (22).

[8] Johansson, U., Bostrom, H., & Lofstrom, T. (2013). Conformal Prediction Using Decision Trees. *2013 IEEE 13th International Conference on Data Mining*.

[9] Fushiki, T. (2009). Estimation of prediction error by using K-fold cross-validation. Statistics and Computing Stat Comput, 21(2), 137-146.

[10] Williams, S., Harris, M., Furst, J., & Raicu, D. (2013). Area under the Distance Threshold Curve as an Evaluation Measure for Probabilistic Classifiers. *Machine Learning and Data Mining in Pattern Recognition Lecture Notes in Computer Science*, 644-657.