33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Probabilistic Lung Nodule Classification with Belief Decision Trees

Dmitriy Zinovev, Jonathan Feigenbaum, Jacob Furst, and Daniela Raicu

Abstract— **In reading Computed Tomography (CT) scans with potentially malignant lung nodules, radiologists make use of high level information (semantic characteristics) in their analysis. Computer-Aided Diagnostic Characterization (CADc) systems can assist radiologists by offering a "second opinion" - predicting these semantic characteristics for lung nodules. In this work, we propose a way of predicting the distribution of radiologists' opinions using a multiple-label classification algorithm based on belief decision trees using the National Cancer Institute (NCI) Lung Image Database Consortium (LIDC) dataset, which includes semantic annotations by up to four human radiologists for each one of the 914 nodules. Furthermore, we evaluate our multiple-label results using a novel distance-threshold curve technique - and, measuring the area under this curve, obtain 69% performance on the validation subset. We conclude that multiple-label classification algorithms are an appropriate method of representing the diagnoses of multiple radiologists on lung CT scans when ground truth is unavailable.**

I. Introduction

LUNG cancer is the most prevalent cause of cancer-related deaths in the human population today. Effective treatment often relies on early detection of the disease, which is done by analyzing suspect computed tomography (CT) scans of lungs. Analysis of the size change of suspected tumors – known as lung nodules – and the inspection of their visual characteristics help diagnose the patient.

When radiologist examines the series of computed tomography scans, the aim is to provide a physician with a set of recommendations that will help the physician to make a correct diagnosis. To improve the usefulness and completeness of these recommendations computer-aided diagnosis (CAD) systems have been designed. They provide a "second opinion" to the radiologist, which may help to increase the efficiency of the diagnosis process as well as reduce the rate of false positive diagnoses while maintaining an acceptably low rate of false negative diagnoses at the same time.

The most current findings in this area support and extend the need for creating reference standard data sets that can provide the ground truth for computer-aided diagnosis systems. One such dataset is the Lung Image Database Consortium (LIDC) [1] – a diverse and growing collection of CT scans analyzed by four radiologists. Each radiologist provided a contour for the nodule or nodules present in the scan, as well as a set of characteristics for the nodule as a whole (cross sections of the same nodule are generally present on multiple CT scans). These characteristics are lobulation, malignancy, margin, sphericity, spiculation, subtlety, and texture. Each characteristic received a rating on a scale from one to five.

While the LIDC provides a common framework for training and evaluating CAD algorithms, there are several challenges that the LIDC data presents including the lack of ground truth and the variability among multiple observers as there was no forced consensus among radiologists when assigning ratings for each characteristic (Fig1). Furthermore, the number of nodules on which there was agreement among radiologists was small. These challenges presented by the LIDC data represent the problems encountered in the medical diagnostic process and open new avenues of applying non-traditional machine learning approaches to the medical imaging decision process.

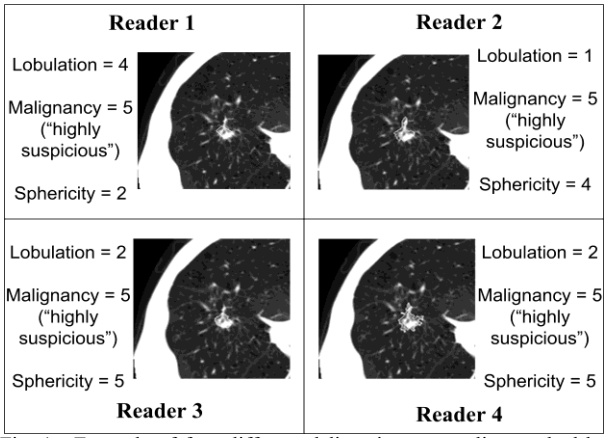

Fig. 1. Example of four different delineations on a slice marked by four different radiologists.

One of the approaches for addressing disagreement issues in the medical imaging area before training the classification model is to artificially force a diagnosis consensus by finding the mean or mode of the characteristics as in the work by Muramatsu et al. [2]. However, the potential drawback of the consensus approach is that the aggregation of individual interpretations produces loss of important information in some cases.

In this paper, we take a further step of classifying lung nodules by considering distributions of ratings (labels) instead of single ratings generated either through consensus or creating multiple instances (each one with a single rating) per nodule. We propose to consolidate the four ratings'

Manuscript received April 5, 2011.

D. Zinovev is with DePaul University, Chicago, IL 60604 USA (corresponding author to provide phone: 630-639-3417; fax: 312-362-6116; e-mail: dzinovev@cdm.depaul.edu).

J. Feigenbaum is with DePaul University, Chicago, IL 60604 USA (e-mail: jfeigenbaum@bus.illinois.edu).

D. Raicu is with DePaul University, Chicago, IL 60604 USA (e-mail: dstan@cdm.depaul.edu).

J.Furst is with DePaul University, Chicago, IL 60604 USA (e-mail: jfurst@cdm.depaul.edu).

interpretations available for each nodule (given the four radiologists interpreting the nodule) into a probability distribution, where each rating ('1', '2'…'5') receives a probability based on the proportion of radiologists who selected that rating. This probability distribution of the ratings becomes the multiple-label and is associated with a single set of image features calculated for the largest of the four outlines provided by the radiologists. In this way, we represent each nodule as a single instance during the training while taking full advantage of all the information available in the radiologists' ratings, instead of discarding the variations in the interpretation process. Having the output of classification system as a multiple-label will be beneficial for the "second opinion" aspect of a CAD process in a sense that radiologist will be provided with not only a single class decision, but with a probability distribution which can provide a radiologist with more insights about the difficulty of a certain diagnosis.

The rest of the paper is organized as follows: Section II discusses the related work in the area of multi-class and uncertain classification; Section III describes the details of our classification and performance evaluation techniques; Section IV presents the evaluation results, and Section V summarizes our presented work and describes possible avenues for future work.

II. Related Work

In a classification task, an instance is a case observation that has to be assigned a label. A label can be either a class membership value or probability of a particular instance where a class defines a certain group that the instance can be a member of. In situations where instances are assigned to multiple classes, then those classification tasks are divided in two categories: multi-label classification and multiple-label classification.

The multi-label classification task is applicable in the situations when the instance can be a member of several non-mutually exclusive classes simultaneously. The examples of such tasks (of video, gene and image classification) are described in [3] - [5].

The multiple-label classification task defined by Jin et al. [6] is similar to the multi-label classification task in a sense that the instance can be a member of several classes at the same time, but differs from it by the fact that only one of these class memberships is correct. Situations in which such a classification task is applicable usually arise due to the presence of multiple observers who do not agree with each other. Multiple studies were conducted to determine whether the presence of multiple observers might be beneficial for solving the classification task. Snow et al [7] examined the problem of dealing with noisy labelers and claimed that the presence of multiple observers can be beneficial for annotation task even when level of expertise of those observers is generally low. Sheng et al [8] performed the study aimed at solving the similar task, but employed the active learning strategies to train the classification model. Results have shown the advantage of combining multiple noisy labelers over the single labeling technique. Kanefsky et al [9] created and tested a system intended for manual annotation of craters on images of Mars. After collecting annotations from multiple volunteer labelers they were combined using weighted clustering technique. Since the study was at the preliminary stage, the obtained results were mostly visual, not quantitative. Raykar et al [10] examined a classification problem at which not only labelers were of different quality, but ground truth was unavailable hidden from the algorithm and used only for evaluation purposes. The performance of created classification model was compared to the performance of a model learned on actual ground truth and the difference in performance was insignificant.

Bjanger and Denœux [11] proposed to modify a traditional decision tree algorithm by defining the impurity measure for each node with respect to a class membership probability distribution of an instance as opposed to single class membership. While Bjanger and Denœux [11] proposed an adaptation of decision trees classifier to the multiple-label problem for classifying uncertain two-class label instances of EEG data (classification approach has shown error rate of 0.34) , Vannoorenberghe and Denœux [12] extended the approach to give it the ability to handle uncertain multiple-class label instances. They proposed to combine trees produced by splitting a multiple-label classification problem into multiple two-class classification problems (one vs. rest classification approach) and demonstrated the approach in the context of classifying data concerning acoustic emission testing of pressure vessels. Authors proposed the evaluation criterion that allowed them to perform the evaluation on instances whose labels were a belief functions rather than single ratings. The reported results showed 0.59, 0.6, 0.59 and 0.57 values for this evaluation criterion, on different datasets correspondingly, which demonstrated a slight improvement over the classification with non-probabilistic labels and suggested that the combination of expert information improves the classification results. The output of such classifier is another basic belief assignment (BBA) that can be evaluated against the original uncertain label using various metrics such as simple accuracy as in [13] or loss function proposed in [12].

Another approach for solving the multiple-label problem using artificial neural networks was proposed by Denœux [14] and by Quost and Denœux [13] using the Dempster Shaefer theory [15]. Their approach consisted in combining uncertain output labels produced by multiple weak classifiers for identifying different types of waveforms in sleep EEG data. The approach proposed by the authors demonstrated an error rate of 13.4. For every classification case the authors performed minimization of mean squared differences between the classifier outputs and target values making a decision on whether the instance was classified correctly or not.

In this paper we propose to handle the variability in the radiologists' interpretation by solving a multiple-label classification problem based on Jin et al. [6] approach. Furthermore, we propose to use belief decision trees [16] to predict class membership probability distribution for each

nodule and also introduce an evaluation metric capable of assessing the performance of such a probabilistic classification algorithm. The same approach that we will apply for any of the seven semantic characterizations classification problems will be applied for the other six characteristics as well.

III. METHODOLOGY

A. LIDC Dataset

The publicly available LIDC database (downloadable through the National Cancer Institute's Imaging Archive web site - http://ncia.nci.nih.gov/) provides the image data, the radiologists' nodule outlines, and the radiologists' subjective ratings of nodule characteristics (with respect to lobulation, malignancy, margin, sphericity, spiculation, subtlety, and texture) for this study. The LIDC database currently contains complete thoracic CT studies for 399 patients acquired over different periods of time and with various scanners. Each study can contain several nodules of a different size; therefore, there may be a different number of slices associated with a particular nodule. Each slice associated with a nodule could contain up to 4 different outlines of this nodule marked by 4 different radiologists. Each radiologist independently rates 7 semantic characteristics of a nodule which produces 4 different semantic labels associated with it. Ground truth for the semantic ratings of lung nodules is not available for LIDC dataset, therefore ratings supplied by radiologists have to be used for training the classification system and evaluating the results.

For each nodule greater than 5×5 pixels (around 3×3 mm) - nodules smaller than this would not have yielded meaningful texture data – we calculate a set of 63 two-dimensional (2D), low-level image features from four categories: shape features, texture features, intensity features, and size. Although each nodule is present in a sequence of slices, in this study we are considering only the slice in which the nodule has the largest area with respect to the outlines provided by up to four radiologists who annotated the corresponding nodule. Therefore, only the largest outline is considered for feature extraction as the most representative. After completion of the feature extraction process, we create a vector representation of every nodule which consists of 63 image features and 7 radiologist annotations. More details on the feature calculations and the rating values for each semantic characteristic are provided in [17].

B. Belief Decision Trees

In this paper we chose to adopt the decision tree based classification approach proposed by Elouedi et al. [19] that is able to handle data instances with uncertain labels. Classification is performed in a manner similar to the one of regular decision trees. On every node, the instance that is currently being classified is redirected to the right or the left child of the node depending on the value of the attribute corresponding to this node. The process is repeated until the instance reaches the leaf node, which has a class membership probability distribution or a basic belief assignment (BBA) associated with it. This BBA is considered to be the newly predicted label of a classified instance. The main difference lies in the way a tree is constructed. At every node of the tree, starting with the root, the algorithm attempts to perform a split based on every attribute/feature existing in the dataset. Out of all constructed splits it determines the best (the selection measure will be defined further) one and uses it for growing the tree further. Every node is associated with a BBA that is constructed by the average of the BBAs of all training cases that reached that node. The newly created node is considered to be a leaf if one of the stopping criteria is reached: 1) there is only one instance that reached this node; 2) all BBAs of the instances which reached the node are equal; 3) all the available attributes/features are split; or 4) the gain ratio of all possible further splits is less than or equal to 0.

In order to define a best split, the algorithm performs the following steps:

First, algorithm computes the pignistic probability (probability calculated from a belief) of instance I_j for each possible class C_i for every instance in the dataset by:

$$BetP^{\Theta}\{I_j\}\{C_i\} = \sum_{C_i \in C \subseteq \Theta} \frac{1}{|C|} \frac{m^{\Theta}\{I_j\}(C)}{1 - m^{\Theta}\{I_j\}(0)}, \forall C_i \in \Theta \quad (1)$$

Where C is a belief mass that C_i is a member of Θ, Θ is a set of all possible classes and $m^{\Theta}\{I_j\}(C)$ is a probability associated with the corresponding belief mass C and $m^{\Theta}\{I_j\}(0)$ is a probability associated with the belief mass of instance not being a member of any class from available pool of classes. Due to the fact that all BBAs in the LIDC dataset are singletons meaning that each radiologist have to pick one class and one class only when assigning the rating to a nodule, the pignistic probability of instance I_j for class C_i is the ratio of observers who assigned the instance to a given class to the total number of observers for that instance (equation 2).

$$BetP^{\Theta}\{I_j\}\{C_i\} = \frac{\lambda_i}{\sum_{l=1}^{5} \lambda_l} \quad (2)$$

(where $\lambda_l=\{0,1,2,3,4\}$ is rater count for every class i rated on a scale from 1 to 5)

Second, the algorithm computes the average pignistic probability function $BetP^{\Theta}\{S\}$ over the set of S instances present in the subset that reached the node to get the average probability on each class:

$$BetP^{\Theta}\{S\}\{C_i\} = \frac{1}{|S|} \sum_{C_i \in C \subseteq \Theta} BetP^{\Theta}\{I_j\}\{C_i\} \quad (3)$$

Third, it computes the entropy of average pignistic probabilities in S:

$$Info(S) = -\sum_{i=1}^{n} BetP^{\Theta}\{S\}\{C_i\} * log_2 BetP^{\Theta}\{S\}\{C_i\} \quad (4)$$

where n is a number of possible classes.

For every attribute/feature, the algorithm collects the subset S_V^A made with the cases having V as a value for the attribute A, compute pignistic probability $BetP^{\Theta}\{S_V^A\}$ for each v of attribute A. Finally algorithm computes $Info_A(S)$ for every attribute as:

$$Info_A(S) = \sum_V \frac{|S_V^A|}{|S|} Info \quad (S_V^A) \tag{5}$$

Where S_V^A is calculated using equation (4).

To calculate goodness of split, the algorithm computes the information gain:

$$Gain(S,A) = Info(S) - Info_A(S) \tag{6}$$

and the gain ratio:

$$Gain\ Ratio(S,A) = \frac{Gain(S,A)}{Split\ Info(S,A)} \tag{7}$$

Where $Split\ Info(S,A)$ is calculated as:

$$Split\ Info \quad (S,A) = \sum_V \frac{|S_V|}{|S|} * log_2 \frac{|S_V|}{|S|} \tag{8}$$

The attribute/feature that produced the largest value of gain ratio is used for the split.

There were several modifications that we made to the original algorithm proposed in [16]. While the approach described by Elouedi et al [16] assumes a categorical nature of the attributes, attributes present in LIDC dataset are continuous. We modified the algorithm to work with continuous attributes by setting the threshold on attribute value that will divide a set of instances into the subset. In order to choose an appropriate threshold, we employed the approach proposed by Quinlan [18]. The approach extracts a separate threshold from every distinct pair of values in the sorted set of attribute values and uses described gain ratio maximization criteria to determine the most suitable one.

We also noticed, while examining the produced classification model, that the Gain Ratio splitting criteria in the case of the LIDC dataset tends to favor very unbalanced splits, assigning a very small ratio of training instances (as small as stopping rules allow) to one of the node's children at every case. As a result the produced trees contained large number of terminal nodes, often equal to the number of training instances, and were over fitted. In order to avoid this we decided to use information gain instead of gain ratio as a splitting criterion.

As the last change we modified one of the stopping rules setting the smallest number of instances that can reach any non-terminal node in a tree to 10 and setting the smallest number of instances that can reach the terminal node to 5. The optimal number of instances that can reach non-terminal node was determined empirically as a compromise between complexity of the classification model and training dataset cross-validation performance. The maximum number of instances at terminal node was fixed to the half of number of instances at its parent node to avoid the unbalanced final splits. This change has also been done to avoid over fitting of the classification model.

C. Performance Evaluation

When evaluating a classification system that utilizes a probability distribution of ratings or classes as an input, and outputs a probability distribution of class membership, evaluation methods beyond accuracy should be used to better capture performance of the system. We propose the idea of a distance curve, in a similar vein to a ROC (receiver operator characteristic) curve [19], to assess the performance of multiple-label classification approach. We were not able to construct ROC curve for the results that we obtained since the definitions of true positive rate and false positive rate are not directly applicable to the multiple-label classification task.

The distance curve is defined as follows:

Let L be a sequence of instance labels, $L= [L_1,L_2,\ldots L_j \ldots L_N]$ where N is the number of instances and each L_j is a discrete probability density function over the label set λ.

Similarly, let P be a sequence of predicted labels, $P = [P_1,P_2,\ldots P_j \ldots P_N]$ where each P_j is discrete probability density function over the label set ε.

Let D be a normalized distance function defined on the instance/prediction pairs, $D(L_j,P_j) \in [0,1]$. We define the distance-threshold curve as

$$\frac{\sum_{j=1}^{N}[D(L_j,P_j)\leq x]}{N}, \tag{9}$$

where x, threshold value for the distance, is defined from 0 to 1, and the [] are Iverson brackets, which equal 1 when the statement inside the brackets is true and 0 otherwise. It can be seen that values of the curve itself are between 0 and 1 and that the curve is monotonically increasing.

We define the area under the distance-threshold curve simply as

$$\int_o^1 \frac{\sum_{j=1}^{N}[D(L_j,P_j)\leq x]}{N}\, dx \tag{10}$$

To generate the curve, we varied the thresholds of distance between the distributions for the classification to be considered "accurate." For example, if we looked for nodules that have a normalized distance of 0, with 0 being a threshold value, between the input and output distributions, we would find little to none. As we increase the distance we find more and more nodules within that threshold. With a normalized distance threshold of 1 between distributions, all the nodules would be considered correct or accurate. Once the curve is generated, the area under the distance threshold curve (AuC_{dt}) was used as the metric for comparison. For this study, we used the Jeffrey Divergence distance metric [20] to generate the D distance function for formula (9), since this distance metric proved to be numerically stable, symmetric and robust with respect to noise [21].

IV. RESULTS

The dataset used for training and testing of the belief decision trees contained 914 instances (1 instance per nodule). The multiple-label of every instance was constructed as class membership probability distribution, where each class probability was calculated as the ratio of radiologists who assigned the nodule to a given class (rating) to the total number of radiologists for that nodule. The set of attributes for an instance was generated from the largest (with respect to the area) outline available for a given nodule.

To build a classification model, we divided the dataset into 90% for training and testing, and 10% for validation subsets in such a way that the nodule distributions of validation subsets mimic the nodule distributions of the original dataset with respect to radiologist agreement and the number of radiologists who rated the nodule. The Belief Decision Tree classification model was constructed for each of the seven semantic characteristics using 10-fold cross validation on the 90% of the data; the model was further validated on 10% validation subset. The distance-threshold curve (AuC_{dt}) and accuracy (ACC) were calculated to evaluate and compare the classification model (Table I). Given the definition of accuracy stands for deterministic labels, we evaluated it by considering the consensus on assigned (majority rating) and predicted probabilistic label (maximum probability).

When analyzing the results, we noticed that belief decision trees demonstrated highest performance on those semantic characteristics for which a highly dominant rating exists. Therefore, in order to determine the impact of a rating distribution's shape (dominated by a rating or not) on classification accuracy, the two subsets of correctly classified (CC) and misclassified (MC) instances were examined independently.

TABLE I
Evaluation of Belief Decision Trees Classification Technique with Respect to Distance-threshold Curve (AuCdt) and Accuracy Metrics (ACC)

	Training subset (90% of instances)		**Testing subset (10% of instances)**	
Characteristic	AuC_{dt} **(%)**	ACC (%)	AuC_{dt} **(%)**	ACC (%)
Lobulation	79.97	69.62	74.46	58.24
Malignancy	73.10	61.58	64.16	49.45
Margin	70.51	61.92	63.72	48.91
Sphericity	60.28	45.93	63.14	37.36
Spiculation	82.05	74.33	76.61	71.74
Subtlety	70.86	60.51	61.67	37.36
Texture	81.94	81.87	76.87	77.17
Average	**74.10**	**65.11**	**68.66**	**54.32**

For the three semantic characteristics with significant increase in the performance by using belief decision trees (spiculation, lobulation, texture) we noticed that belief decision trees accurately predicted the majority of instances with dominant rating. A summary of these findings is reported in Table II; the analysis is provided on the training set given the low number of ratings from each class for the testing data.

An impact of distribution of the ratings on classification performance of decision trees is caused by the way the classification model provides its final probabilistic decision. The instance label is used to calculate the average pignistic probability function (average across 5 classes) which is then used for calculating the entropy of the set and determining the goodness of split for a particular node. Every node in a belief decision tree has a probability distribution associated with it which is calculated by averaging the probability distributions (uncertain labels) of instances that reach that node during the training phase. At the classification step, a classified instance is assigned the probability distribution of a leaf node that it reaches. It is clear that since all predicted labels are produced by averaging the subset of assigned instance labels and there exists a rating which is highly dominant across all 5, there will be fair amount of predicted uncertain labels with the given rating also being dominant. Due to the way accuracy is assessed for every case (mode vs. mode) the model will perform well for instances with dominant ratings.

TABLE II
Misclassification Rate of Belief Decision Tree Classification Approach on Instances with Dominant and Non-dominant Ratings; CCDR Stands for Correctly Classified Dominant Ratings and MCDR for Misclassified Dominant Ratings; (similar abbreviations are used for the non-dominant abbreviations (CCNDR and MCNDR))

% instances per characteristic	**CCDR**	**MCDR**	**CCNDR**	**MCNDR**
Lobulation	67.19%	7.05%	10.57%	15.19%
Spiculation	76.28%	4.38%	4.50%	14.84%
Texture	68.73%	1.22%	9.25%	20.80%

V. CONCLUSIONS

In this paper we adapted and evaluated a multiple-label belief decision tree classification algorithm. We determined that, in certain situations the algorithm demonstrates higher performance than in a general scenario. We learned that these situations correspond to distributions of ratings that are dominant by one rating (unimodal distributions) and therefore, it is possible, by examining the data, to make a decision whether the use of technique is appropriate. While most of the research results for multiple-label classification in the current literature are presented on synthetic data, we demonstrated the multiple-label approach using a real medical dataset. Furthermore, we evaluated the performance using both the standard accuracy measure and the area under a distance-threshold curve – the analog of ROC curves for probabilistic outputs.

In terms of future work, we plan to expand this work as follows: first, we will include 3D image features in addition to the current 2D features; second, we will look at combining radiologists outlines using p-maps approaches instead of considering just the largest outline; third, we will

investigate the use of other base classifiers as a way to improve performance, and lastly, we will look at incorporating belief classifiers into various ensemble learning techniques to take advantage of their classification capabilities.

REFERENCES

[1] S. G. Armato III, et al., "Lung Image Database Consortium: developing a resource for the medical imaging research community," *Radiology,* vol. 232, 2004, pp. 739-748.

[2] C. Muramatsu, Q. Li, K. Suzuki, et al., "Investigation of psychophysical measure for evaluation of similar images for mammographic masses: Preliminary results," *Med Phys,* vol. 32, 2005, pp. 2295-2304.

[3] T. Li, C. Zhang, S. Zhu, "Empirical studies on multi-label classification," *Proc. ICTAI*, 2006, pp. 86–92.

[4] N. Ghamrawi, A. McCallum, "Collective multi-label classification," *Proc ACM CIKM*, 2005, pp. 195-200.

[5] C. Ramachandran, R. Malik, X. Jin, J. Gao, K. Nahrstedt, J. Han, "VideoMule: a consensus learning approach to multi-label classification from noisy user-generated videos," *Proc. 17th ACM Int. Conf. on multimedia*, 2009, pp. 721-724.

[6] R. Jin, Z. Ghahramani, "Learning with multiple labels," *Proc NIPS*, 2002, pp. 897–904.

[7] R. Snow, B. O'Connor, D. Jurafsky, A. Y Ng, "Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks," *Proc. Conf. on Empirical Methods in Natural Language Processing*, 2008 Jan, pp. 254-263.

[8] V. Sheng, F. Provost, P. Ipeirotis, "Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers," *Proc. 14th ACM SIGKDD Int. Conf. on Knowledge discovery and Data Mining*, 2008, pp. 614-622.

[9] B. Kanefsky, NG. Barlow, VC. Gulick, "Can distributed volunteers accomplish massive data analysis task?," *Lunar Planet. Sci. Conf.*, 2001, [CD-ROM] 31:1272.

[10] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, L. Moy, "Supervised learning from multiple experts: whom to trust when everyone lies a bit," *Proc. 26th Annual Int. Conf. on Machine Learning*, 2009, pp. 889-896.

[11] M. S. Bjanger, T. Denœux, "Induction of decision trees from partially classified data using belief functions," *Master thesis at University of Compiegne*. 2000

[12] P. Vannoorenberghe, T. Denoeux, "Handling Uncertain Labels in Multiclass Problems using Belief Decision Trees," *Proc. Int. Conf. on Processing and Management of Uncertainty*, vol. 3, 2002 July, pp. 1916-1926.

[13] B. Quost, T. Denœux, "Learning from data with uncertain labels by boosting credal classifiers," *In Proc. 1st ACM SIGKDD workshop on knowledge discovery from uncertain data*, 2009 June, pp. 38-47.

[14] T. Denœux, "A Neural Network Classifier Based on Dempster-Shafer Theory," *IEEE Trans SMCA,* part A, 2000, pp. 131-150.

[15] G. Shafer, *A mathematical theory of evidence*, 1976; Princeton University Press, Princeton, NJ.

[16] Z. Elouedi, K. Mellouli, P. Smets, "Belief decision trees: Theoretical foundations," *Int J Approx Reason,* vol. 28, 2001, pp. 91–124.

[17] D. Zinovev, D. Raicu, J. Furst, S. G. Armato III, "Predicting radiological panel opinions using a panel of machine learning classifiers," Algorithms Journal, vol. 2, 2009, pp. 1473-1502.

[18] J. R. Quinlan, "Improved Use of Continuous Attributes in C4.5," *J Artif Intell Res,* vol. 4, 1996, pp 77-90.

[19] K. A. Spackman, "Signal detection theory: Valuable tools for evaluating inductive learning," *Proc. 6th Int. Workshop on Machine Learning*, 1989, pp. 160–163.

[20] H. Liu, D. Song, S. Rüger, R. Hu, V. Uren, "Comparing dissimilarity measures for content-based image retrieval," *Proc. 4th Asia Inf. Ret. Conf. on Information Retrieval Technology*, 2008 Jan, pp. 44-50.

[21] J. Puzicha, T. Hofmann, J. M. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," *Proc. IEEE CVPR*, 1997 Jun, pp. 267-272.