Multi-level Classification of Emphysema in HRCT Lung Images Using Delegated Classifiers

Mithun Prasad[1] and Arcot Sowmya[2]

[1] Cedars-Sinai Medical Center,
8700 Beverly Blvd.
Los Angeles, CA 90048
[2] School of Computer Science and Engineering,
University of New South Wales,
Sydney, NSW, 2052, Australia
mithunp@cse.unsw.edu.au, sowmya@cse.unsw.edu.au

Abstract. Emphysema is a common chronic respiratory disorder characterized by the destruction of lung tissue. It is a progressive disease where the early stages are characterized by diffuse appearance of small air spaces and later stages exhibit large air spaces called bullae. A bullous region is a sharply demarcated region of emphysema. In this paper, we show that an automated texture-based system based on delegated classifiers is capable of achieving multiple levels of emphysema extraction in High Resolution Computed Tomography (HRCT) images. The key idea of delegation is that a cautious classifier makes predictions that meet a minimum level of confidence, and delegates the difficult or uncertain predictions to a more specialized classifier. In this paper, we design a two-step scenario where a first classifier chooses the examples to classify on and delegates the more difficult examples to a second classifier. We compare this technique to well known emphysema classification techniques and ensemble methods such as bagging and boosting. Comparison of the results shows that the techniques presented here are more accurate. From a medical standpoint, the classifiers built at different iterations appear to show an interesting correlation with different levels of emphysema.

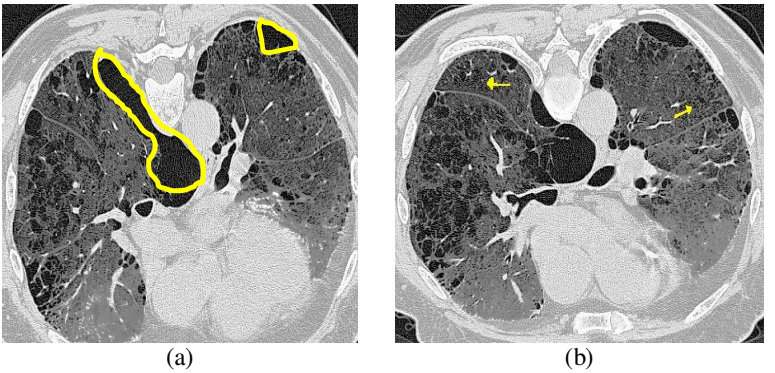
Keywords: CT, emphysema, texture, delegated classifiers.

1 Introduction

High Resolution Computer Tomography (HRCT) is a valuable imaging modality for assessing diffuse lung diseases and in particular, emphysema. The automated analysis of HRCT scans poses difficult problems, because the radiographic patterns observed are often varied and subtle. Emphysema diagnosis by radiologists is often based on visual recognition of imaging patterns augmented by anatomical knowledge. Emphysema is a common chronic respiratory disorder characterised by the destruction of lung tissue and is often reflected as areas of low attenuation in CT images [1].

Emphysema regions are typically small in the early stages, but become larger and involve the lung more diffusely over time. Large air spaces called bullae may

D. Metaxas et al. (Eds.): MICCAI 2008, Part I, LNCS 5241, pp. 59–66, 2008.



(a) (b)

Fig. 1. 1(a) A typical HRCT scan containing bullous emphysema. 1(b). A typical HRCT scan containing diffuse regions of emphysema.

develop, particularly in the later stages. Bullous emphysema is histologically defined as the presence of emphysematous areas with complete destruction of lung tissue. Classification of bullae is useful to evaluate patients as candidates for surgery. Figure 1 visually presents examples of bullous and diffuse regions of emphysema.

The techniques utilized in this paper are intended to automate the recognition process and assist radiologists in the diagnosis of emphysema by providing accurate measures of severity across each HRCT scan. This is achieved by using the idea of delegated classifiers. In automated emphysema detection in lung images, a common technique called "Density Mask" is applied simply to threshold the image [1]. However, a fixed threshold yields unsatisfactory results when the degree of emphysema is low. Computerised techniques for classifying emphysema have been explored [2-4] using texture and machine learning approaches with reports of reasonable accuracy. However, very few techniques that distinguish the type of emphsyema have been reported [4].

The proposed system is based on delegated classifiers [5]. Delegation can be summarized by the motto: *let others do the things you cannot do well*. We use the notion of a cautious classifier which only classifies the examples for which its predictions have high confidence, leaving the other examples to another classifier. Delegated classifiers have been successfully developed and tested on "artificial" datasets from the UCI repository [5]. However, application of delegated classifiers to vision problems has not been addressed to our knowledge. In this work, we also show that delegation is capable of classifying different levels of diagnosis automatically. The levels range from the larger set of diffuse and bullous regions, to bullous regions only.

2 Methods

2.1 Texture Feature Extraction

In this work, textural features are used to characterize emphysema. We extract textural features using two main steps: automatic segmentation of the lungs and

feature extraction. The lungs in the image are initially located and extracted. A suite of classical image processing techniques is used to segment lung regions using in house software [6]. This is quite a straightforward approach where the different morphological operations performed to segment lung regions include dilation, erosion and thresholding. The percentage area occupied by lung regions in the whole image is used to decide whether the image is of interest. A percentage value of less than 6 is considered unacceptable.

In our application, feature extraction is primarily based on texture as emphysema is a finding that can be well characterized by texture. A feature vector is defined as a set of textural parameters calculated on a small neighborhood of 12x12 pixels surrounding each image point belonging to the lung region. Window sizes less than 12x12 do not provide uniformity of disease patterns and window sizes larger than 12x12 are computationally expensive. The textural parameters used in the experiments are based on the following methods:

1. moments of gray level histogram of a local area
2. gray level co-occurrence matrix method (GLCMM)
3. gray level run length matrix method (GLRLMM)
4. gray level difference method (GLDM)

The GLCMM, one of the well known texture analysis methods, estimates image properties related to second-order statistics. Each entry (i,j) in GLCM corresponds to the number of occurrences of the pair of gray levels i and j at a distance d apart at an angle θ in original image. The configurations of the co-occurrence matrix used in our experiments include $1 \leq d \leq 2$ and $0 \leq \theta \leq 90, \pm 45$ since these values are sufficient to cover uniformity of disease features. The GLRLMM is based on computing the number of gray level runs of various lengths in different directions. Each element of the GLRLM (i,j) specifies the estimated number of times a picture contains a run of length j, for gray level i, in the direction of angle θ. Three grey level run length matrices, where $0 \leq \theta \leq 360$ in steps of 45, are used in our experiments. The full range of θ provides greater uniformity among the various disease features used in our experiments. GLDM is concerned with the spatial gray-level distribution and spatial dependence among the gray levels in a local area. The features extracted from the methods are displayed in Table 1; some features have multiple values, as discussed above.

Table 1. Textural Features

Moments of Histogram	GLCMM	GLRLM	GLDM
Mean	Energy	Short Emphasis	Mean
SD	Entropy	Long run emphasis	Contrast
Variance	Homogeneity	Gray level uniformity	Entropy
Energy	Contrast	Primitive length uniformity	SD
Entropy		Primitive percentage	Variance

2.2 Delegation

The idea of delegation revolves around two main issues [5]. Firstly, one needs to determine a threshold or a rule to decide when to apply the first classifier and when to

delegate to the second one. Secondly, one needs to determine good techniques to generate classifiers that perform better than the first one for the examples that the first one has delegated. We use a probabilistic classifier to estimate the reliability. The second issue is addressed by specializing the second classifier on examples for which the first classifier behaves worst. This is achieved by training the second classifier on solely the examples rejected by the first classifier. There are two main advantages of delegation fold. Firstly, since the first classifier is holding part of the examples, the second classifier has fewer examples for training. This results in a more efficient process than other ensemble techniques. Secondly, the resulting overall classifier is a decision list whose decisions can be traced and understood, unlike in comparable techiques.

2.2.1 Cautious Classifier

In many application areas, a classifier that abstains from making a prediction when it is not sure of being able to make the right decision is preferable over a greedy classifier that always makes a classification. A cautious classifier is one that gives predictions for the subset of inputs for which it is more confident (that may still be right or wrong) but abstains for the rest of its inputs. In other words, a cautious classifier is a partial function.

Any classifier that can reliably estimate class probabilities or the reliability of each prediction, can be converted to a cautious classifier. For a classifier f we can consider the associated functions $f_{CLASS}(e)$, $f_{CONF}(e)$ and $f_{PROBc}(e)$ (for each class c from a total of C classes). The function $f_{CLASS}(e)$ returns the class assigned by classifier f to example e, $f_{PROBc}(e)$ returns the probability of class c for example e, and $f_{CONF}(e)$ returns the highest probability among all classes for example e. Unless stated otherwise, we assume that $f_{CLASS}(e) = \arg\max_c f_{PROBc}(e)$ and $f_{CONF}(e) = \max_c \{f_{PROBc}(e)\}$.

Given these definitions, a cautious classifier *f* can be obtained from a soft classifier using a confidence threshold.

Decision Rule for a cautious classifier with threshold τ:

If $f_{CONF}(e) > \tau$ *then predict* $f_{CLASS}(e)$ *else abstain*

A soft classifier will be converted into a good cautious classifier if the reliabilities are well estimated, as achieved by, for instance, a good class probability estimator, or, for binary problems, a good ranker. It is the idea of completing the cautious classifiers that leads to the concept of *delegation*. If a cautious classifier f^1 decides that it is not competent to classify an example with good confidence, but wants to complete the work, then it can delegate the example to another classifier. If we had this second classifier, denoted by f^2, and a confidence threshold τ, then the delegating rule is as follows:

Decision Rule for a delegating classifier with threshold τ:

If $f^1_{CONF}(e) > \tau$ *then predict* $f^1_{CLASS}(e)$ *else predict* $f^2_{CLASS}(e)$

A common way to obtain classifier f^2 is to train it only on the training examples for which f^1 has low confidence. In this way, the second classifier will be specialized for these examples. More formally, given a training set *Tr*, a soft classifier f and a confidence threshold τ, we divide this set into two data sets $Tr_f^{>} = \{e \in Tr : f_{CONF}(e) > \tau\}$ and $Tr_f^{\leq} = Tr - Tr_f^{>}$. In other words, we can refer to $Tr_f^{>}$ as the "retained" or "high confidence" examples and $Tr_f^{\leq}$ as the "delegated" or "low-confidence" examples. In this work, we use the same threshold for training and for prediction. The approach taken here is that a classifier retains a fixed percentage of the examples. For instance, we may stipulate that the first classifier should retain 70% of the most highly ranked examples, delegating the rest to the second classifier. This technique is known as *Global Absolute Percentage*. In the case of imbalanced datasets, a technique known as *Stratified Absolute Percentage* may be used, where the decision rule can be modified to incorporate a different threshold τ_c for each class *c* as shown below. In the case of stratified absolute percentage, if we denote Tr_c as the examples in Tr of class *c*, the retained examples in this case are:

$$Tr_f^{>} = \{e \in Tr_c : c = f_{CONF}(e) \wedge f_{CLASS}(e) > \tau_c\}$$

Decision Rule for a delegating classifier with threshold $\tau_1, \tau_2, \ldots, \tau_c$:

If $f^1_{CONF}(e) > \tau_c$ *then predict* $f^1_{CLASS}(e)$ else predict $f^2_{CLASS}(e)$ where $c = f^1_{CLASS}(e)$

3 Experimental Results and Discussion

The accuracy measure is used to evaluate the performance of our algorithm. Accuracy is defined as the percentage of correctly classified examples (which includes both positive and negative examples). The emphysema regions were manually characterized in consultation with a board-certified radiologist. In the experiments, we use naive Bayesian classifier to perform classification at both levels in the delegated framework. Naive Bayes estimates the probability of class membership based on Bayes rule [7]. The delegated classifier was trained on 13 HRCT scans (chosen randomly from 8 subjects) and evaluated on a separate labelled test set comprising 60 HRCT scans (randomly chosen from 9 subjects), and the visual results are presented in Figure 2. It can be seen in 2(d) that the second classifier identifies more regions of emphysema than the first one precisely on the low-confidence examples delegated by the first classifier. This is the key to the overall improvement achieved by the delegated classifier. It is also worth noting that the low confidence examples in the HRCT scans correspond to diffuse regions of emphysema. The output of the "density mask" algorithm (Figure 2(b)) shows that a lot of noise is picked up along with the emphysema regions.

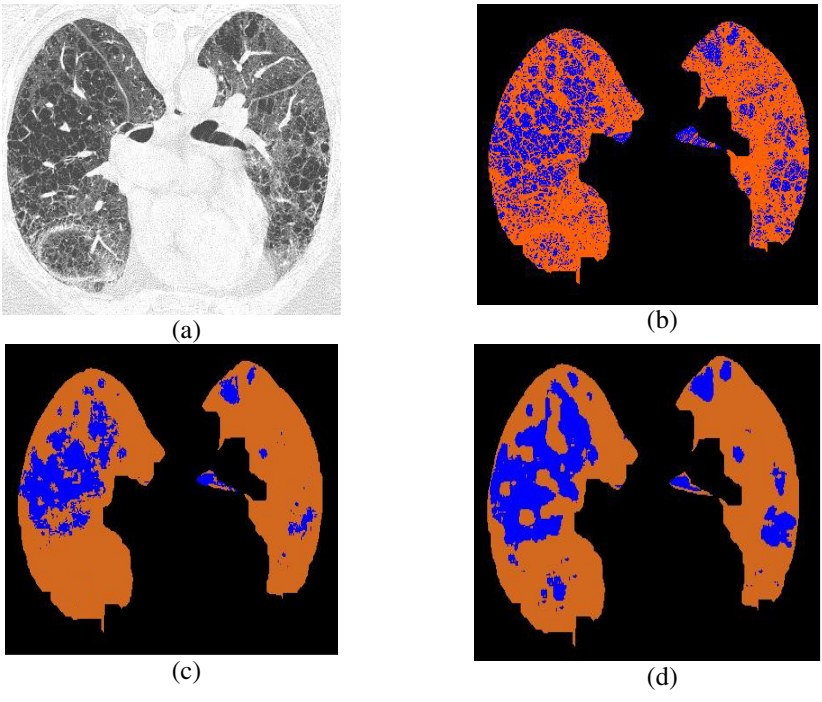


Fig. 2. 2(a) contains the original image where the dark regions correspond to emphysema. 2(b) is the output of "Density Mask" where the blue regions are emphysema. 2(c) and 2(d) correspond to the output of the first classifier and the overall classifier in the delegated framework respectively. The first classifier identifies mostly the "more confident" or the bullous regions whereas very diffuse regions of emphysema can be classified using the second classifier in the delegated framework.

Additionally, Table 2 demonstrates that accuracy is not very sensitive to the percentage of examples retained, although it seems that 60% is a good compromise. Lower percentages would mean most of the work is left to the second classifier, which is then very similar to the first one and not specialized sufficiently to improve the results. A high retention lowers the influence of the second classifier and may perhaps lead to its overfitting. However, the model appears to be robust in the sense that, with different configurations, the mean accuracy is never worse than for a single classifier (0%).

Table 2. Performance of delegated classifiers with different global absolute percentage (GAP) and stratified absolute percentage (SAP)

	0%	10%	20%	30%	40%	50%	60%	70%	80%
GAP	89	93	93	93	93	92	93	93	93
SAF	89	92	93	91	93	92	93	92	92

In addition, comparison was also performed on other well known techniques that have been explored for emphysema detection as can be seen in Table 3. The accuracy is higher for the delegated classifier (global absolute percentage of 20% used). Intuitively, the delegation forces the classifier to focus on the weak examples using a new decision rule. The "Density mask" identifies a large amount of emphysema that is mostly "not correct" as shown in figure 2.

Table 3. Comparison of average accuracy of the delegated classifier with well known emphysema classification techniques

	Average Accuracy
Density Mask	95
Seeded K-means	88
ICA – C4.5	84
ICA – Naive Bayes	84
Error Backpropagation	84
Support vector machine	86
Delegated classifier, GAP of 20%	94

Finally, we compare delegating (GAP of 20%) with two ensemble techniques, namely boosting and bagging [8], using naive Bayes as the base classifiers. As can be seen in Table 4, the delegated classifier is better than the ensemble methods in terms of average accuracy. Delegated classifiers are more efficient than classical ensemble methods, because each subsequent classifier is learned using fewer examples than the previous one. In the delegated framework, predictions of classifiers are not combined. Each instance is classified by a single classifier. This does not degrade the comprehensibility of the model as ensembles do. Additionally, comparison of the results in Table 4 was made using a mixed-effects linear model [9] and we found a significant difference between our method and the other techniques ($p<0.001$).

Table 4. Comparison between delegation and ensemble techniques. Bagging and boosting were performed with 10 and 20 iterations.

	Average Accuracy
Bagging – 10	87
Boosting – 10	88
Bagging – 20	86
Boosting – 20	89
Delegated classifier, GAP of 20%	94

4 Conclusions

In this paper, an approach to perform multi-level diagnosis of emphysema detection based on delegated classifiers has been presented. Delegation only makes predictions that meet a minimum level of confidence and delegates to another classifier otherwise. Results have been compared against "density mask", a standard approach used for emphysema detection in medical image analysis. In general, the density mask

method has been known to mark more pixels as emphysematous than warranted, and it has been speculated that many marked pixels do not represent true emphysema. The results of the method proposed here appear to confirm this. Other well known computerized techniques used for classification of emphysema have also been used for comparison and the results show that by using a specialized classifier, classification accuracy can be improved. A system that is capable of differentiating the appearance of emphysema regions has been successfully reported in this paper, which would help experts in the medical setting to analyze the progressive nature of the disease. In the future, we plan to investigate the use of different base classifiers at each delegation stage. At different levels of the delegation chain, different classifiers can be used. We also plan to investigate the utility of delegated classifiers in multi-class classification tasks within the HRCT setting.

Acknowledgement

This research was partially supported by the Australian Research Council through a Linkage grant (2002-2004), with Medical Imaging Australasia as clinical and Philips Medical Systems as industrial partners.

References

1. Kinsella, M., Mueller, N.L., Abboud, R.T., Morrison, N.J., DyBuncio, A.: Quantification of emphysema by computed tomography using a density mask program and correlation with pulmonary function tests 97, 315–321 (1990)
2. Friman, O., Borga, M., Lundberg, M., Tylén, U., Knutsson, H.: Recognizing emphysema - A Neural Network Approach. In: Proceedings of 16th International Conference on Pattern Recognition (August 2002)
3. Prasad, M., Sowmya, A., Koch, I.: Feature Subset Selection using ICA for classifying Emphysema in HRCT images. In: Proc. of international conference on intelligent sensors, sensor networks and information processing, Melbourne, Australia (2004)
4. Prasad, M., Sowmya, A., Wilson, P.: Multi-level Classification of Emphysema in HRCT lung images. Pattern Analysis and Applications Journal (in press, 2007)
5. Ferri, C., Flach, P., Hernandez-Orallo, J.: Delegating classifiers. In: Proceedings of 21th International Conference on Machine Learning, pp. 106–110. Omni press, Alberta (2004)
6. Chiu, P.T., Sowmya, A.: Lung Boundary Detection and Low Level feature extraction and analysis from HRCT images. In: VISIM: Information Retrieval and Exploration of Large Medical Image Collections (October 2001)
7. Mitchell, T.: Machine Learning. McGraw Hill, New York (1997)
8. Freund, Y., Schapire, R.E.: A short introduction to boosting. Journal of Japanese Society for Artificial Intelligence 14(5), 771–780 (1999)
9. Laird, N.M., Ware, J.H.: Random Effects Models for Longitudinal Data. Biometrics 38, 963–974 (1982)