Applying Boosting Algorithm for Improving Diagnosis of Interstitial Lung Diseases

Jason Yang | jason2@stanford.edu

BACKGROUND

Interstitial Lung Disease (ILD) is a group of irreversible lung pathologies presented as progressive scarring of lung tissue around the air sacs, which causes lung stiffness. This scarring reduces the ability to breathe and may possibly lead to life-threatening complications, such as pulmonary hypertension and respiratory failure[1]. In order to diagnose ILDs, physicians turn to high-resolution computed tomography (HRCT) scans to identify abnormal lung texture patterns[2]. Currently, radiologists comb through hundreds, or even thousands, of image slices in one patient's HRCT scan to find any histological markers of ILDs. Additionally, lung texture patterns are often difficult to distinguish, resulting in low diagnostic accuracy. Because of these reasons, there is motivation for computers to assist radiologists in diagnosing ILDs.

DATA SOURCE

This project utilized the collection of high resolution CT scans from the MedGIFT project at the University of Geneva, Switzerland[3]. The dataset contains three-dimensional grayscale regions of the lung tissue from 128 patients, who had at least one of the 13 histological diagnoses of ILD. Each region of interest was annotated with one of 17 different lung texture classes. This publicly available dataset is extremely useful for this project because of the high data quality, as multiple radiologists were consulted for a consensus of opinion whether a region contains healthy or pathological lung tissue.

IMAGE PROCESSING AND FEATURE EXTRACTION

Preprocessing

For each slide of CT scan with an annotation, a bounding box was found that encompassed the whole region of interest. This bounding box was expanded, such that the dimensions of the box were multiples of 32 pixels. The box was then split into nonoverlapping 32x32 pixel image patches. Each patch was examined to ensure that there is more than 80% overlap with the lung and more than 70% overlap with the region of interest.

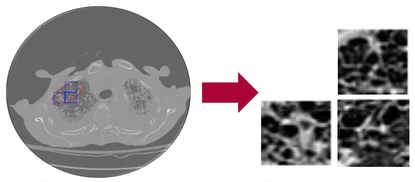


Figure 1. Example of Image Patch Extraction from HRCT Scan. Region of Interest in red. Valid image patches in blue.

Rigid transformations, such as rotations and reflections, were performed on each image class that had less than the class with the largest count of images. New images were randomly added to each class so that each texture class had the same number of images.

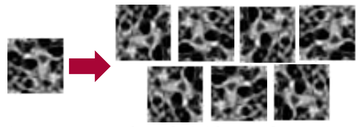


Figure 2. Examples of Rigid Transformations. Each image patch produced 7 new images.

Gabor Filter

Gabor filter bank is used for texture analysis and edge detection of images. Frequency and orientation representations are similar to how the human visual system works.

Passing the 32x32 image patches through the Gabor filter bank yielded 96 features each[4].

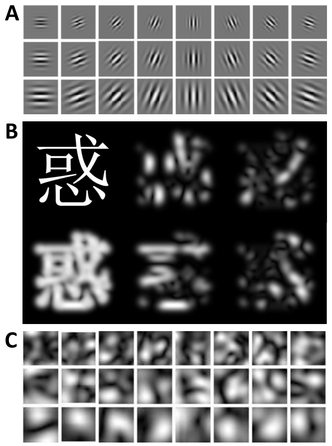


Figure 3. Representations of Gabor Filter Bank. A) Image of the 3-scale, 8-orientation Gabor filter. B) The right images are filtered at 4 different orientations and the left images are the original picture and the superposition of the right images. C) Example of a filtered image patch.

FEATURE SELECTION AND MODEL EXECUTION

Principal Component Analysis (PCA)[5]

PCA uses orthogonal transformation to convert the set of features into linearly uncorrelated variables in order to select the features that explain nearly all of the variance in the dataset. The cutoff was placed at features that explained more than 0.1% of the variance. Of the 96 features, 72 were selected to be inserted in the boosting algorithm.

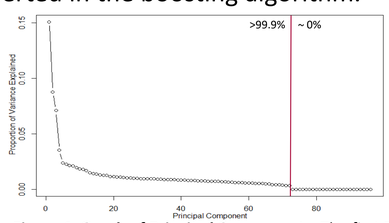
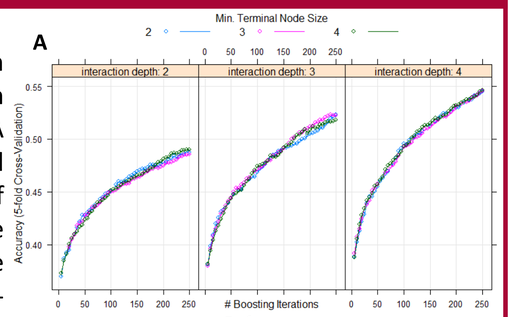


Figure 4. Graph of Principal Components. The first 72 features explains more than 99.9% of the variance. There was a drop off of variance after the 72th feature.

Gradient Boosting Model (GBM)

GBM uses an ensemble of decision trees to generate a model based on the features inputted after the PCA filter. The model was trained and five-fold cross-validated on 80% of the total image patches. The parameters were being tuned in the range of 5-250 decision trees with 2-4 terminal nodes and 2-4 interaction depth (splits) per tree. The final model used 250 trees, 4 terminal nodes, and 4 interaction depth (splits) per tree. The algorithm produced probabilities of each test image having each label. The highest probability was taken to be the predicted label of the test image.



B

Reference

Predicted	Fibrosis	Ground Glass	Healthy	Micro-nodules	PCP
Fibrosis	357	89	89	78	64
Ground Glass	103	359	91	41	119
Healthy	94	75	363	140	37
Micro-nodules	113	29	166	480	20
PCP	80	181	30	15	528

Figure 5. Results of GBM. A) Training accuracy of varying parameters of GBM. B) Confusion matrix of the test data.

DISCUSSION

The final Gradient Boosting Model predicted with 55.7% accuracy on a held out test set. There was concern for the model overfitting the data because of the number of trees in the model; however, the training and test accuracy were both very similar. This accuracy is unacceptable in the medical field in predicting health. Physicians and radiologist would need a very accurate predictor assisting them in practice. This result was unexpected, but at least performed better than randomly predictions, which would have had an accuracy of 20%. The features extracted from a Gabor filter bank might not have been enough for the algorithm to achieve accuracy.

FUTURE DIRECTIONS

One of the improvements to this project could be a more robust feature selection step. Sequential feature search was attempted in this project, but yielded negligible results. Additionally, other features can also be extracted from the images, such as histogram of oriented gradients. Another improvement could be the use of a more powerful machine learning algorithm. There recently has been an approach that utilizes ensemble learning with a neural network. Applying that approach to this project could produce much better accuracy in the diagnoses of interstitial lung disease.

REFERENCES

1. "Interstitial lung disease," *Mayoclinic*, Jun. 2015. [Online]. Accessed: Nov. 11, 2016.
2. B. Elicker, et al. "High-resolution computed tomography patterns of diffuse interstitial lung disease with clinical and pathological correlation," *Journal Brasileiro de Pneumologia*, vol. 34, no. 9, pp. 714-744, Sep. 2008.
3. A. Depeursinge, et al. "Building a reference multimedia database for interstitial lung diseases," *Computerized Medical Imaging and Graphics*, vol. 36, no. 3, pp. 227–238, Apr. 2012.
4. M. Haghighatm, et al. "CloudID: Trustworthy cloud-based and cross-enterprise biometric identification," *Expert Systems with Applications*, vol. 42, no. 21, pp. 7905–7916, Nov. 2015.
5. "Practical guide to principal component analysis (PCA) in R & python," Analytics Vidhya, 2016. [Online]. Accessed: Dec. 9, 2016.