UCLA

UCLA Electronic Theses and Dissertations

Title

Unbalanced Data Classification Using Support Vector Machines with Active Learning on Scleroderma Lung Disease Patterns

Permalink

<https://escholarship.org/uc/item/56j388nz>

Author

Lee, Jongyoon

Publication Date

2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

**Unbalanced Data Classification Using Support Vector
Machines with Active Learning on Scleroderma Lung
Disease Patterns**

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Jongyoon Lee

2013

© Copyright by
Jongyoon Lee
2013

ABSTRACT OF THE THESIS

Unbalanced Data Classification Using Support Vector Machines with Active Learning on Scleroderma Lung Disease Patterns

by

Jongyoon Lee

Master of Science in Statistics

University of California, Los Angeles, 2013

Professor Ying Nian Wu, Chair

Unbalanced data classification has been a long standing issue in the field of Medical Vision Science and Pattern Recognition. In our research, we introduced the methods of Support Vector Machines (SVM) with Active Learning (AL) to improve prediction of unbalanced classes of Scleroderma Lung Disease, Lung Fibrosis (LF) and Honeycomb (HC). Four different SVM with AL approaches are proposed: 1) random sampling to select the initial pool to begin the AL algorithm; 2) doubling the training instances of Honeycomb to reduce the imbalance ratio before the AL algorithm; 3) a balanced pool with equal number Honeycomb instances and Lung Fibrosis instances; 4) a balanced pool of Honeycomb and Lung Fibrosis and implements balanced sampling throughout the AL algorithm. Grid pixel data of Lung Fibrosis and Honeycomb was extracted from computed tomography (CT) images of 71 patients from 13 clinical centers around the United States to produce a training set of 348 HC and 3009 LF instances and a test set of 291 HC and 2665 LF. From our research, SVM with AL using balanced sampling compared to random sampling increased the sensitivity of HC by 0.56 (0.175 vs. 0.735) and 0.47 (0.23 vs. 0.70) for the original and de-noised dataset respectively. These results show that by implementing SVM with Active Learning paired with balanced sampling, we can improve the prediction performances of unbalanced data classification.

The thesis of Jongyoon Lee is approved.

Nicolas Christou

Janet Sinsheimer

Grace Hyun Jung Kim

Ying Nian Wu, Committee Chair

University of California, Los Angeles

2013

To my mother . . .

To my father . . .

To my brother.

TABLE OF CONTENTS

1	Introduction	1
2	Methods	4
2.1	Summary	4
2.2	Support Vector Machines (SVM)	4
2.2.1	Support Vector Machines with Active Learning (AL)	6
2.3	Four SVM with AL Methods	7
2.4	Feature Selection	8
2.4.1	Logistic Regression Backwards Stepwise Selection	8
3	Data	10
3.1	Scleroderma Lung Study (SLS)	10
3.2	Image Denoising	11
3.3	Texture Features	11
4	Results	14
4.1	Original Dataset	14
4.1.1	Logistic Regression Backwards Stepwise Selection	14
4.1.2	Results of Method 1 on Original Training Dataset	14
4.1.3	Results of Method 2 on Original Training Dataset	15
4.1.4	Results of Method 3 on Original Training Dataset	16
4.1.5	Results of Method 4 on Original Training Dataset	17
4.1.6	G-means Performance	18
4.1.7	Original Test Dataset Results	18

4.2	Denoised Dataset	19
4.2.1	Logistic Regression Backwards Stepwise Selection	19
4.2.2	Results of Method 1-4 on the Denoised Training Dataset	20
4.2.3	G-means Performance	21
4.2.4	Denoised Test Dataset Results	22
5	Discussion	23
6	Conclusions & Future Work	25
	References	28

LIST OF FIGURES

1.1	CT scans of Normal Lung (NL), Lung Fibrosis (LF), and Honeycomb (HC) .	2
3.1	Denoised Images of Normal Lung (NL), Lung Fibrosis (LF), and Honeycomb (HC)	11
4.1	G-means Performance of Method 1-4 on Original Training Dataset	18
4.2	G-means Performance of Method 1-4 on Denoised Training Dataset	21

LIST OF TABLES

3.1	Summary of Training and Test Sets	10
3.2	Summary of Statistical Texture Features	12
3.3	Summary of Co-occurrence Matrices & Acquisition Length Parameters Texture Features	13
4.1	Results of Method 1 on Original Training Dataset	15
4.2	Results of Method 2 on Original Training Dataset	16
4.3	Results of Method 3 on Original Training Dataset	16
4.4	Results of Method 4 on Original Training Dataset	17
4.5	Results of Method 1-4 on Original Test Dataset	19
4.6	Results of Method 1-4 on Denoised Training Dataset	20
4.7	Results of Method 1-4 on Denoised Test Dataset	22

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my thesis advisors, Dr. Grace H.J. Kim, Dr. YingNian Wu, Dr. Janet Sinsheimer, Dr. Nicolas Christou for their support and guidance throughout my graduate studies. I also would like to thank MedQIA, UCLA Department of Radiology, UCLA Center for Applied Statistics, and R01 HL072424 from the National Institutes of Health for sharing their data and funding and making this research project possible.

CHAPTER 1

Introduction

Class imbalance has been an ongoing issue the field of Machine Learning and Classification. As the availability of raw data grows at an exponential rate, there is a great deal of motivation in many engineering and science fields to construct artificially intelligent learning systems to sort through the immense quantities of data. However, as data availability increases, it is common to be faced with complex unbalanced data structures during classification processes. Many solutions have been proposed in previous research to address the class imbalance problem, such as oversampling and undersampling, cost-sensitive weighting sampling, and kernel-based and active learning methods [1]. In more recent studies, empirical evidence suggest machine learning techniques with an active learning framework can achieve higher and more efficient prediction performances on significantly unbalanced data [2].

In this paper, we present the applications of machine learning with active learning to address the class imbalance problem of Interstitial Lung Disease (ILD) of Scleroderma in the field of Radiological Sciences. Interstitial Lung Disease refers to the collection of lung disorders characterized by abnormal tissue growth (fibrosis) and/or inflammation [3]. Using high resolution computed tomography (CT) scans of lungs, radiologists can detect one common and one less common category of ILD patterns: Lung Fibrosis (LF) and Honeycomb (HC). Lung Fibrosis is identified by large amounts of visible tissue scarring in the CT images, while Honeycomb is characterized by honeycomb-shaped cysts that thicken the walls of the lung (Figure 1.1).

Computerized classification techniques have been introduced in ILD research to aid radiologists in the efforts to classify large amounts of CT image data [4]. However, these

Figure 1.1: CT scans of Normal Lung (NL), Lung Fibrosis (LF), and Honeycomb (HC)

techniques still require further development when building an artificially intelligent learner due to the class imbalance problem, such as between Lung Fibrosis and Honeycomb. Due to the rarity of honeycomb patterns in the lung, recent studies have shown an imbalance ratio of 10 to 1 between Lung Fibrosis and Honeycomb categories [5].

Prior research suggests machine learning models called Support Vector Machines (SVM) paired with active learning can offer a fast and computationally attractive way to classify unbalanced data categories without sacrificing prediction performance [2]. Our research examines the capabilities of SVM with active learning to investigate the unbalanced data classification problem between Lung Fibrosis and Honeycomb. Four different approaches to SVM with active learning are proposed and assessed using classification rates of a training

and test set, and also visually through comparisons of G-means plots.

CHAPTER 2

Methods

2.1 Summary

Support Vector Machines (SVM) have been shown to perform well in medical imaging studies [9, 10], more specifically in the field of pattern recognition of interstitial lung disease [11]. In addition, we applied this method not only due to its popularity and efficiency, but also due to its strong regard for unbalanced data problems using active learning properties [2]. With Active Learning (AL), Support Vector Machines are able to iteratively extract the most informative training instances and carefully build a classifier using a more balanced training sample. Using these ideas, we ran four variations of SVM with AL to assess their capabilities to accurately predict both rare and overall classifications. In this section, we also discuss feature selection methods to select the most important texture features as predictor variables to generate a more parsimonious classification model. All analyses are run using R programming language.

2.2 Support Vector Machines (SVM)

Support Vector Machines (SVM) [12] are among the most popular supervised learning methods due to their simplicity and efficiency in classification and prediction. The foundation of a Support Vector Machine model is an optimal hyperplane that linearly separates the data into two categories by the largest margins. This method can also be extended to situations where the original data is not linearly separable, by implementing a kernel function that maps the input data into a higher dimensional new space called the feature space. The

hyperplane is obtained by minimizing the following objective function:

$$\begin{aligned} \min_{w, b, \xi_i} \quad & \frac{1}{2} \mathbf{w} \cdot \mathbf{w}^T + C \sum_{i=1}^N \xi_i \\ \text{subject to} \quad & \begin{cases} \forall i \ y_i(\mathbf{w}^T \Phi(x_i) - b) \geq 1 - \xi_i \\ \forall i \ \xi_i \geq 0 \end{cases} \end{aligned}$$

where $\mathbf{w}$ is the norm of the hyperplane, b is the offset, y_i are the labels, $\Phi(\cdot)$ is the mapping from the input space to the feature space, and ξ_i are the slack variables that permit the non-separable case by allowing misclassification of training instances. The objective function is often solved by optimizing the dual cost function:

$$\begin{aligned} \max W(\alpha) \equiv \quad & \sum_{i=1}^N \alpha_i - \frac{1}{2} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{subject to} \quad & \begin{cases} \forall i \ 0 \leq \alpha_i \leq C \\ \sum_{i=1}^N \alpha_i y_i = 0 \end{cases} \end{aligned}$$

where y_i are the labels, $\Phi(\cdot)$ is the mapping function from the input space to the feature space, $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$ is the kernel matrix, and the α_i 's are the Lagrange multipliers which are non-zero only for the training instances which fall on the margin.

The instances that fall on the margins that separate the data are called Support Vectors. Due to this relationship between the Support Vectors and the margins, only the Support Vectors have an effect on the solution of the Support Vector Machine. This fact leads to the idea that instances have different levels of informativeness, depending on their distance to the hyperplane [2].

However, we have not only chosen to implement Support Vector Machines due to its high prediction performance and efficiency, but because of its approach to dealing with classification on unbalanced data. We applied a variation of Support Vector Machines called Support Vector Machines with Active Learning.

2.2.1 Support Vector Machines with Active Learning (AL)

Support Vector Machines with Active Learning are developed using the idea that not all instances are equally important when creating a Support Vector Machine model [2]. The most informative instances are believed to be the ones closest to the hyperplane. The farther an instance is from the hyperplane, the more the classifier is confident about its classification prediction, thus provides no additional information for the SVM. Hence, the closer an instance is to the hyperplane, the more informative is that instance to build an accurate classifier. In Ertekin’s article, she uses this method to address the unbalanced ratio problem with the idea that the unbalanced ratio of instances within the margin is much smaller than the unbalanced ratio of the entire dataset. A classification technique that samples from the instances in the margin will likely obtain a more balanced class distribution than of the entire input dataset.

The basic steps of SVM with active learning is to run an SVM on an initial small pool of training instances, select the closest instance to the hyperplane from the remaining pool of instances, include the selected instance in the small pool, and train the SVM again. This process iterates until there is a convergence in the number of Support Vectors or the G-means performance. This means that the inclusion of additional training instances will not improve the Support Vector Machine classifier.

In our research, we used the 59 trick to determine the sample size to select the most informative instances for each iteration [13]. The 59 trick states that a sample size of 59 is the optimal sample size to select one instance from it, based on the condition that this selected instance has the shortest distance from the hyperplane. This stems from the fact that in order to obtain an instance that has a 95% probability that it is the top 5% closest instances to the hyperplane, the random sample will need to be of size $\lceil \log 0.05 / \log 0.95 \rceil = 59$. Previous literature suggests that this algorithm not only reduces training time but it also does not decrease the quality of the classifier.

2.3 Four SVM with AL Methods

Using this idea of informativeness and active learning, we investigated and assessed four variations of SVM with AL. These methods incorporate the concepts and techniques of active learning in slightly different ways.

1) The first method begins the Active Learning algorithm by generating an initial SVM pool by randomly sampling 300 instances from the training set. Then the Active Learning algorithm is run using the 59 trick and the classifier is generated.

2) The second method implements weighted sampling by doubling the number of Honeycomb instances in the training set to decrease the imbalance ratio before sampling the initial pool. The initial SVM pool is generated by randomly sampling 300 training instances, and the Active Learning algorithm is run using the 59 trick to generate the classifier. Due to the intensity of the overall imbalance ratio of 10:1, weighting the Honeycomb instances higher to reduce the imbalance ratio to 4:1 will aid the classifier in strengthening its prediction capabilities for the rare category.

3) The third method begins the Active Learning algorithm by generating an initial SVM pool by randomly sampling 150 training instances from the Honeycomb category and 150 training instances from the Lung Fibrosis category to create a balanced initial SVM pool. Then the Active Learning algorithm is run using the 59 trick and a classifier is generated. This technique is derived from the idea of aiding the SVM classifier towards the balance before the algorithm begins.

4) The fourth method begins the Active Learning algorithm by generating an initial SVM pool by randomly sampling 150 training instances from the Honeycomb category and 150 training instances from the Lung Fibrosis category to create a balanced initial SVM pool. Then, the Active Learning algorithm is run using a stratified sampling technique to add

a constraint of maintaining the 1:1 balanced ratio throughout the entire AL algorithm to generate the classifier.

Performances of these four methods are assessed by reviewing the sensitivity, specificity, and accuracy rates of the training and testing, set as well as referring to the G-means plots for visual comparison.

2.4 Feature Selection

Feature selection, also known as variable selection, is the process of selecting the important and relevant variables from the dataset to construct a more parsimonious model for analysis and prediction. With parsimony, models will be simpler and will have better prediction performances because the classifiers determined by the models will not be overly dependent on the composition of the training set. In our research, we used Logistic Regression Backwards Stepwise Selection to identify the statistically significant variables relevant to predicting the lung disease patterns as accurately as possible. We chose this method due to its simplicity and ease of implementation. The variables selected from this selection process will be used to generate classification models using Support Vector Machines with Active Learning.

2.4.1 Logistic Regression Backwards Stepwise Selection

Logistic Regression Stepwise Selection is an automatic feature selection method under the branch of Logistic Regression Models [14]. This feature selection process involves selecting the most important variables in a dataset that will accurately predict a categorical, in most cases binary, response variable. This is appropriate for our research because our interest is finding a classifier to accurately to predict the classification of two Scleroderma Lung Disease types, Honeycomb and Lung Fibrosis. Stepwise selection is also advantageous in that it focuses on reducing multicollinearity among the variables. This is a key point in building our classifier due to the large number of predictor variables that we have to start with in our

lung disease classification study.

There are three approaches to this selection process: forward, backward, and bi-directional. In our case, we have decided to use Backwards Stepwise Selection, which is a technique in which all the variables are included in the beginning of the variable selection process and removed iteratively depending on the comparisons of the Akaike information criterion (AIC) values. Each statistical model has an AIC value which measures the goodness of fit of that model, relative to the other models in question. The smaller the value, the better the model. Thus, the model with the smallest AIC value will be chosen at the end of the selection process.

CHAPTER 3

Data

3.1 Scleroderma Lung Study (SLS)

The data were collected during a multicenter National Institutes of Health-sponsored randomized, controlled trial investigating the effects of cyclophosphamide versus placebo on patients admitted for Interstitial Lung Disease [5]. This data collection was held in 2000 across 13 clinical centers throughout the United States. Baseline thoracic high-resolution computed tomography (CT) scans were taken of each patients total lung capacity. Within this study, 50 ROIs (46 LF, 4 HC) were extracted from 38 selected patients to compose the training set. An independent set of 48 ROIs (44 LF, 4 HC) from 33 patients were sampled to compose the test set. Note that the selection of ROIs from each patient ranged from 1 to 11. Within each ROI, grid sampling of 4 x 4 pixel squares was implemented to generate pixel instances [5]. Using this technique, a total of 3,557 pixels (3,009 LF, 348 HC) were extracted to produce the training set and 3,001 (2,665 LF, 291 HC) instances for the test set. From these datasets, we observed an imbalance ratio of about 10:1 (LF:HC). Table 3.1 summarizes the study data.

	Patients	ROI	Grid Sample Pixels
Training Set	38	50 (46 LF, 4 HC)	3,557 (3,009 LF, 348 HC)
Test Set	33	48 (44 LF, 4 HC)	3,001 (2,665 LF, 291 HC)

Table 3.1: Summary of Training and Test Sets

3.2 Image Denoising

Image noise has been a common issue in the field of pattern recognition and computer vision. Image noise are random, unwanted signals that affect the clarity of the picture and hinder the amount of information that can be retrieved from it [6]. In our study, images are denoised using a modified variation of Aujols algorithm with an added constraint in the parameters (Figure 3.1) [5]. We will present the results of our investigations using both the original and denoised datasets.

Figure 3.1: Denoised Images of Normal Lung (NL), Lung Fibrosis (LF), and Honeycomb (HC)

3.3 Texture Features

A total of fifty-six texture features were investigated in the SLS study for both the original and denoised datasets, including 9 statistical features, 27 texture features from co-occurrence

matrix, and 20 acquisition length texture features [5, 7, 8]. Among these features, the original dataset was reduced to 45 texture features due to missing data. Similarly, the denoised dataset was reduced to 38 features. Table 3.2 and 3.3 summarizes the texture features of the SLS study dataset.

Texture Features	Dimension
Statistical Features	Mean
	Standard Deviation
	Skewness
	Kurtosis
	Median
	Mode
	Mean32
	Energy
	Entropy
	Ratio Highest
	Ratio Highest Lowest

Table 3.2: Summary of Statistical Texture Features

Texture Features	Dimension
Co-occurrence matrices 4 direction (0, 45, 90, 135), average, and range	Angular second moment
	contrast
	correlation
	difference entropy
	difference mean
	difference variance
	entropy
	inverse difference moment
	information measures of correlation-a
	information measures of correlation-b
	maximal correlation coefficient
	sum average
	sum entropy
	sum variance
	variance
Acquisition Length parameters 4 direction (0, 45, 90, 135), average, and range	short run emphasis
	long run emphasis
	low grey-level run emphasis
	high grey-level run emphasis
	short run low grey-level emphasis
	short run high grey-level emphasis
	long run low grey-level emphasis
	long run high grey-level emphasis
	run length non-uniformity
	grey level non-uniformity
	run percentage

Table 3.3: Summary of Co-occurrence Matrices & Acquisition Length Parameters Texture Features

CHAPTER 4

Results

4.1 Original Dataset

4.1.1 Logistic Regression Backwards Stepwise Selection

We first conducted Logistic Regression backwards stepwise selection to determine the most significant variables in predicting classification of Lung Disease, Honeycomb and Lung Fibrosis. After the initial implementation of logistic regression variable selection, we then incrementally removed variables in the logistic regression model that had a p-value greater than 0.05. We reduced the number of predictor variables from 45 to 17. By reducing the number of variables in the model, we can construct a more parsimonious model, and reduce the chances of overfitting the training dataset. Details and descriptions of these variables can be found in the data section. We will conduct the remaining of our analyses with these selected variables.

4.1.2 Results of Method 1 on Original Training Dataset

We began our classification analyses by presenting a model using simply Support Vector Machines with Active Learning with an initial pool of 300 randomly sampled instances. This will act as a baseline to compare the efficiency of our three methods of interest.

We ran SVM with Active Learning with an initial pool of 300 randomly selected training instances, and ran the Active Learning algorithm for a total of 350 iterations of our Active Learning algorithm. The performance metrics of the model are presented in Table 4.1 at different iterations across the algorithm. Due to the fact that Honeycomb is the rare event,

we indicated Honeycomb Lung Disease at the positive and Lung Fibrosis as the negative event. Thus, the sensitivity refers to the accuracy of the Honeycomb classification and specificity refers to the accurate classification rate of Lung Fibrosis.

	start	100th iteration	200th iteration	350th iteration
sensitivity	0.679 (19/28)	0.738 (59/80)	0.636 (77/121)	0.457 (80/175)
specificity	1.000 (272/272)	0.997 (319/320)	0.997 (378/379)	1.000 (475/475)
accuracy	0.970 (291/300)	0.945 (378/400)	0.910 (455/500)	0.854 (555/650)

Table 4.1: Results of Method 1 on Original Training Dataset

From the table, we can see that when we randomly sampled 300 training instances to generate the initial pool, we sampled 272 Lung Fibrosis instances and 28 Honeycomb instances. This closely reflects the imbalance ratio of the entire dataset (around 10:1). Before the start of the Active Learning algorithm, we observed a sensitivity of 0.679 and a specificity of 1.0 using 77 support vectors in our model. As the algorithm continues to run, sensitivity gradually begins to decrease and specificity stabilizes around 1.0. At the end of all the iterations, the final SVM model classifier, shown in the last column of Table 4.1, shows a sensitivity of 0.457 and an overall accuracy of 0.854. It also represents a smaller imbalance ratio of about 3:1 (175/475) relative to the overall imbalance ratio. This supports idea that the ratio within the margins of the SVM hyperplane is more balanced than the ratio of the entire dataset.

4.1.3 Results of Method 2 on Original Training Dataset

For method 2, we investigate the efficiency of weighted sampling by artificially doubling the number of Honeycomb instances in the training dataset before running the Active Learning algorithm. The results of the model performance for this method are provided in Table 4.2. Doubling the number of Honeycomb instances in the overall training dataset increased the total sample size to 3705, with 696 Honeycomb instances and 3009 Lung Fibrosis instances.

From this method, we generated an initial pool of 46 Honeycomb instances and 254

	start	100th iteration	200th iteration	350th iteration
sensitivity	0.826 (38/46)	0.907 (97/107)	0.821 (128/156)	0.902 (211/234)
specificity	0.980 (249/254)	0.922 (270/293)	0.936 (323/344)	0.897 (373/416)
accuracy	0.957 (287/300)	0.918 (367/400)	0.900 (450/500)	0.898 (584/650)

Table 4.2: Results of Method 2 on Original Training Dataset

Lung Fibrosis instances, producing an initial imbalance ratio between the two Lung Disease categories of about 5:1. This was a much better improvement from the overall imbalance ratio of 10:1.

The model began with a sensitivity of 0.826 and a specificity of 0.980 before starting the Active Learning algorithm. This is a drastic increase in sensitivity, compared to the sensitivity that we began with in method 1 (0.679). After running the Active Learning algorithm for 350 iterations, the final SVM model classifier produced a sensitivity of .902, a specificity value of .897, and an overall accuracy of .898. These values imply stabilized sensitivity and specificity rates across the Active Learning algorithm, which we will also observe in the G-means plots latter in the section. From comparing these results to Table 4.1 of method 1, we observed the sensitivity rate increased from 0.457 to 0.902.

4.1.4 Results of Method 3 on Original Training Dataset

In method 3, the efficiency of beginning the Active Learning algorithm using a balanced initial pool of 150 randomly selected Honeycomb instances and 150 randomly selected Lung Fibrosis instances was investigated. The performance metric results of this method are provided in Table 4.3 below.

	start	100th iteration	200th iteration	350th iteration
sensitivity	0.987 (148/150)	0.792 (133/168)	0.649 (120/185)	0.505 (108/214)
specificity	0.940 (141/150)	0.987 (299/232)	1.000 (315/315)	1.000 (436/436)
accuracy	0.963 (289/300)	0.905 (362/400)	0.870 (435/500)	0.837 (544/650)

Table 4.3: Results of Method 3 on Original Training Dataset

The balanced initial pool provided high percentages for sensitivity (0.987) while maintaining a high specificity rate (0.963) before beginning the Active Learning algorithm. After running the algorithm for 350 iterations, we observed a sensitivity rate of 0.505 and a specificity rate of 1.0. Similar to method 1, the sensitivity value slowly diminishes as the number of iterations increase. Specificity maintained at a high rate of 1.0 throughout the algorithm which produced an accuracy rate of 0.837. Although we start with a balanced initial pool, along the algorithm, the SVM goes back to the nature of the imbalance.

4.1.5 Results of Method 4 on Original Training Dataset

Method 4 investigates using a balanced initial pool and an Active Learning algorithm that maintains the balance throughout all the iterations. The results are provided in Table 4.4 below. Note in this method, we only ran 200 iterations. Due to the nature of method 4 sampling one training instance from each category per iteration, the size of the SVM pool grows faster than the other three methods. Due to this and the total number of Honeycomb training instances being 348, the number of iterations was reduced from 350 to 200. This ended with a final SVM pool only 50 training instances shorter than the other SVM methods.

	start	50th iteration	100th iteration	200th iteration
sensitivity	0.987 (148/150)	0.985 (197/200)	0.980 (245/250)	0.977 (293/300)
specificity	0.940 (141/150)	0.935 (187/200)	0.924 (231/250)	0.930 (279/300)
accuracy	0.963 (289/300)	0.960 (384/400)	0.952 (476/500)	0.953 (572/600)

Table 4.4: Results of Method 4 on Original Training Dataset

Method 4 produced high sensitivity and specificity rates before and throughout the entire algorithm, similar to method 2. The iteration began with a balanced pool with high sensitivity (0.987) and specificity (0.940) and maintained these values across 200 iterations. The final SVM model classifier showed a sensitivity rate of 0.977 and a high specificity rate of .930 to produce an overall accuracy of 0.953. Out of all four methods, Method 4 produced the highest rates of these performance metrics using the training dataset.

Figure 4.1: G-means Performance of Method 1-4 on Original Training Dataset

4.1.6 G-means Performance

The G-means performances of the four SVM classification models are given in Figure 4.1. The means and error bounds of the G-mean values are plotted against the iterations of the AL algorithm. Note, the measurements of the plots are scaled consistently across the four plots and provide a red horizontal line indicating the 0.9 value. Method 1 shows erratic and decreasing G-means performance across the AL algorithm, which reflects the performance metrics provided in Table 4.1. Method 2, although it maintains a relatively high G-means performance, presents rigged error bounds. Method 3, although it begins with a high G-means value at the beginning of the algorithm, shows a gradually decreasing G-means performance as the number of iterations increase. This reflects that although we start with a balanced initial pool for the AL algorithm, the SVM tends to revert back to the nature of the imbalance between the two categories. This is also shown in the performance metrics of the training dataset in Table 4.3. Method 4 provides not only the highest G-means performance above the 0.9 line across the iterations, but it also the most stable SVM model with narrow error bounds. These figures suggest Method 4 presents the best prediction performances simultaneously in sensitivity and specificity among the four methods.

4.1.7 Original Test Dataset Results

Results conducting predictions on the test dataset using the four SVM model classifiers are produced in Table 4.5.

	Method 1	Method 2	Method 3	Method 4
sensitivity	0.175 (51/291)	0.340 (99/291)	0.251 (73/291)	0.735 (214/291)
specificity	0.996 (2655/2665)	0.979 (2609/2665)	0.992 (2645/2665)	0.941 (2507/2665)
accuracy	0.915 (2706/2956)	0.916 (2708/2956)	0.919 (2718/2956)	0.921 (2721/2956)

Table 4.5: Results of Method 1-4 on Original Test Dataset

Using the method 1 produced the poorest results in terms of correctly classifying the rare Lung Disease, Honeycomb. The sensitivity value was 0.175 and the accuracy value was 0.915, but this is largely due to a pull from the high specificity rate of 0.996. Method 2 using doubled Honeycomb instances produced a higher sensitivity value in comparison, but was still a fairly low value of 0.34, correctly predicting 99 out of 291 Honeycomb instances. Similar to method 1, specificity and accuracy values were high due to the prediction precision of Lung Fibrosis, 0.979 and 0.916 respectively. Method 3 produced low classification rates, although sensitivity rates were higher for method 3 compared to method 1 by .075. Method 4 produced the best results by far, in terms of classification of Honeycomb and overall classification with a sensitivity of 0.735 and an accuracy of 0.921. Sensitivity increased by 0.56 compared to method 1. However, we observed the lowest specificity rate of 0.941 among the methods, but it is fairly high correctly predicting 2507 out of 2665 Lung Fibrosis instances.

4.2 Denoised Dataset

4.2.1 Logistic Regression Backwards Stepwise Selection

For the de-noised dataset, when we ran logistic regression variable selection, we reduced the number of predictor variables from 38 variables to 19, using the same selection scheme as we did for the original dataset. First, we applied logistic regression backwards stepwise selection, and then incrementally removed variables in the logistic regression model according to largest p-value. At the end of the selection process, we kept only the statistically significant predictors that remained in the final model (p-value ≤ 0.05). Details and descriptions of these variables can be found in the data section. The SVM models will be built solely on these

selected variables.

4.2.2 Results of Method 1-4 on the Denoised Training Dataset

		start	100th iteration	200th iteration	350th iteration
Method 1	sensitivity	0.840 (21/25)	0.867 (65/75)	0.658 (75/114)	0.531 (85/160)
	specificity	1.000 (275/275)	0.978 (318/325)	0.995 (384/386)	0.998 (489/490)
	accuracy	0.987 (296/300)	0.958 (383/400)	0.918 (459/500)	0.883 (574/650)
Method 2	sensitivity	0.898 (44/49)	0.702 (66/94)	0.885 (131/148)	0.883 (196/222)
	specificity	0.996 (250/251)	0.990 (303/306)	0.932 (328/352)	0.925 (396/428)
	accuracy	0.980 (294/300)	0.922 (369/400)	0.918 (459/500)	0.911 (592/650)
Method 3	sensitivity	0.980 (147/150)	0.807 (138/171)	0.682 (131/192)	0.524 (120/229)
	specificity	0.940 (141/150)	0.991 (227/229)	1.000 (308/308)	1.000 (421/421)
	accuracy	0.960 (288/300)	0.912 (365/400)	0.878 (439/500)	0.832 (541/650)
		start	50th iteration	100th iteration	200th iteration
Method 4	sensitivity	0.973 (146/150)	0.980 (196/200)	0.988 (247/250)	0.990 (297/300)
	specificity	0.933 (140/150)	0.945 (189/200)	0.948 (237/250)	0.943 (283/300)
	accuracy	0.953 (286/300)	0.962 (385/400)	0.968 (484/500)	0.967 (580/600)

Table 4.6: Results of Method 1-4 on Denoised Training Dataset

Table 4.6 is a report of the performance metrics across all four methods on the denoised training dataset. We observed similar patterns to that of the original dataset. In Method 1, we note that the random initial pool begins with a ratio that reflects the overall training dataset of about 10:1. Then, as the algorithm iterates, sensitivity decreases and specificity is maintained around 1.0. Method 2, which uses weighted sampling, maintains a high sensitivity rate around .90 throughout the AL algorithm. However, specificity gradually decreases but stays fairly high above .90. In method 3, similar to the results of the original dataset, the balanced initial pool shows very high performance metrics but the SVM returns to the nature of the imbalance as the iterations are run. Method 4 shows the best results compared to all the methods in terms of prediction performance on the denoised training dataset. The SVM model for method 4 maintained high sensitivity and specificity values from the

beginning to end of the AL algorithm, around .98 and .94 respectively. This also led to the highest accuracy value of the final SVM model classifier of 0.967 compared to the other SVM methods. This leads up to believe maintaining a balanced sample when building an SVM classifier will provide the best prediction performances even on an unbalanced dataset.

4.2.3 G-means Performance

Figure 4.2: G-means Performance of Method 1-4 on Denoised Training Dataset

We investigated the G-means performances of the four SVM methods on the denoised training dataset in Figure 4.2. In Figure 4.2, the measurements are scaled consistently across all four plots with a red horizontal line indicating the 0.9 value. These plots show similar results to the G-means performance of the original training dataset. Method 2 and method 4 maintain high g-mean scores relative to method 1 and method 3. Method 1 not only shows a gradually decreasing G-means performance, but it also shows large error bounds around the averages of the G-mean values. Method 3 begins with a very high G-means value at the beginning of the AL algorithm but shows a decrease in G-means performance as the number of iterations increase. This is similar to method 3 on the original training set, shown in Figure 4.1. However, in the denoised training set, we also observe an increase in the width of the error bounds as the number of iterations increase. This suggests a very unstable SVM model. Method 4 provides the best G-means performance maintaining high values above 0.9 throughout the entire AL algorithm. This suggests method 4 will best classify Honeycomb and Lung Fibrosis and provide high sensitivity and specificity values simultaneously. These results on the denoised dataset agree with the results found from the original dataset.

4.2.4 Denoised Test Dataset Results

	Method 1	Method 2	Method 3	Method 4
sensitivity	0.230 (67/291)	0.426 (124/291)	0.340 (99/291)	0.701 (204/291)
specificity	0.991 (2640/2665)	0.968 (2581/2665)	0.986 (2629/2665)	0.917 (2445/2665)
accuracy	0.916 (2707/2956)	0.915 (2705/2956)	0.923 (2728/2956)	0.896 (2649/2956)

Table 4.7: Results of Method 1-4 on Denoised Test Dataset

Results of the prediction performances of the four SVM model classifiers are given in Table 4.7. These results also closely reflected the results of the SVM classifiers on the original dataset. Method 1 and 3 produced the lowest prediction rates. Method 2, which reduced the imbalance from 10:1 to 4:1 while building the classifier, produced a sensitivity rate of 0.426, correctly predicting 124 out of 291 Honeycomb instances. Accuracy was also high around 0.923 but this was largely due to the high specificity value of 0.986, correctly predicting 2629 out of 2665 Lung Fibrosis instances. Method 4 produced the highest sensitivity rate by far of 0.701. However, using this method sacrificed the prediction performances of Lung Fibrosis with a specificity rate of 0.917. This produced an overall accuracy of a little less than 0.90 (0.896). All in all, from reviewing the prediction performance on both the training and test set, we suggest method 4 is optimal in classifying the rare category Honeycomb out of the four methods of interest.

CHAPTER 5

Discussion

According to the prediction outcomes of the original dataset, we found Method 4 provided the highest G-means performances and highest test sensitivity values. By incorporating the technique of maintaining a balanced ratio in the training instances while building the SVM model, Method 4 maintained stable and high G-means performance well above 0.9 throughout the entire AL algorithm, thus providing high sensitivity and specificity values simultaneously. This also could be seen in the performance metrics of the test set. Using this technique of maintaining a balanced ratio caused the sensitivity of the original test set to increase by 0.56, from 0.175 to 0.735, compared to the baseline, method 1. Method 1 and method 3 showed decreasing prediction performances for the rare category, Honeycomb, in the training dataset, which reflected onto the prediction rates of the test set with low sensitivity values, 0.175 and 0.251 respectively. This demonstrated the idea that due to the intensity of the imbalance ratio, the SVM tends to return to the nature of the 10:1 imbalance. Method 2 although showed high levels of the G-means performance for the training dataset, showed only a sensitivity of 0.340 in the original test dataset, which was lower than random guess.

The findings from the denoised dataset also agreed with the results from the original dataset. We observed narrow error bounds and high G-means values throughout the entire AL algorithm for method 4, showing stabilized prediction performance. This ultimately provided a sensitivity value of 0.701 on the test set, which was 0.47 higher than of method 1 (0.230). The other methods also showed similar patterns to those provided using the original dataset. Method 1 and method 3 showed gradually decreasing performance metrics across the AL algorithm with a test sensitivity of 0.230 and 0.340. Method 2 showed unstable G-

means performance with rigged error bounds, and a test sensitivity larger than method 1 and method 2, but still a rather low value of 0.426. Due to these close similarities of the results using the original and denoised dataset, our studies imply that the imbalance classification issue in ILD research should aim classification efforts more on the machine learning methods of the rare categories, than the quality of the image themselves.

However, one caveat of our research is that although the fourth method produced the best results in terms of predicting Honeycomb Lung classification compared to the other methods, we still observed fairly low sensitivity values for both the original and denoised test sets (0.735 and 0.701 respectively). This is largely due to the violation of the independence assumption of the Support Vector Machine model. In our case, because the instances were multiple pixels extracted from the same limited number of ROI samples, many of the pixels were not independent from one another. This could have affected the outcome of the predictive power of our SVM models. Further research should be conducted to investigate this issue and delve into more analysis on the technique of reducing the imbalance ratio by maintaining a balanced training dataset while building the SVM model.

CHAPTER 6

Conclusions & Future Work

Unbalanced data classification has been a long standing issue in the field of Medical Vision Science and Classification. In our research in particular, we were interested in how we could apply Machine Learning methods to accurately distinguish unbalanced Scleroderma Lung Diseases categories: one common pattern, Lung Fibrosis, and one uncommon, Honeycomb.

Ertekin shed light on a technique using Support Vector Machines with Active Learning for accurate classification of unbalanced data in her article [2]. A technique was built based on the idea that the ratio between two unbalanced categories is less unbalanced within the margins of the SVM hyperplane than the ratio of the entire data. Thus, she implemented an AL sampling algorithm to only sample the most informative training instances closest to the hyperplane to actively build the SVM classifier as new incoming data were presented.

Using this information, we implemented and assessed four methods of SVM with Active Learning for our study. Method 1, which we referred to as our baseline, drew a random sample of training instances to initiate the SVM Active Learning algorithm. Method 2 applied weighted sampling by artificially doubling the number of the rare Lung Disease category, Honeycomb, to lower the imbalance ratio before beginning the AL algorithm. Method 3 equally sampled from the Honeycomb Lung and Lung Fibrosis instances to construct an initial SVM pool before starting the AL algorithm. Method 4 initiated the AL algorithm with a balanced pool and maintained the balance between the two categories throughout the entire AL algorithm. Lung image data collected from 13 clinical centers throughout the United States collected in 2000 were used to evaluate these methods.

In our results, we found that maintaining a balanced ratio in the sampled training instances provided the highest prediction performances for the rare lung category. Results

from Method 4 and method 2, which also reduced the imbalance in the training dataset, suggests it is useful to try and reduce and even eliminate the imbalance ratio of the training set while building the classifier to provide more predictive power for the more rare categories. In our case, we eliminated the imbalance between the two categories while building the SVM model by adding a constraint of keeping the 1:1 balance in the AL algorithm while adding new training instances. The results from the original dataset and the de-noised dataset agreed in conjunction to these conclusions. We found the method 4 provided the highest test sensitivity values above 0.70 for both the original and denoised dataset. However, the performance of accurately predicting Lung Fibrosis was sacrificed, with test specificity rates decreasing to 0.941 and 0.917 in the original and denoised dataset respectively, compared to the test specificity of the other methods, ranging from around 0.97 to 0.99. However, due to the significant increase in the prediction rates of the rare category, Honeycomb, tied with fairly high specificity rates above 0.90 in the test set, we believe maintaining a balanced ratio while building the SVM model is the optimal technique to address the unbalanced data classification issue.

Limitations of our research include finding a more efficient and reliable way of selecting the significant predictors to use in our classification models. Although logistic regression is a classic method for variable selection, we are interested in investigating advanced techniques in the field of Machine Learning or other methods such as Spike and Slab regression that not only produce more optimal models to better assess overall significance, but also increases computational efficiency. Another issue we came across was the limitations of our software. Ertekin used an online tool called LASVM to implement SVM with Active Learning. Because we did not have access to this programming instrument, we ran all of our analyses using the provided SVM functions in R. Lastly, our biggest concern was dealing with the violation of the independence assumption of Support Vector Machines. Because of this violation, the prediction performance of our SVM models suffered. We would like to find a Machine Learning system that deals with dependent training instances or find a way to extract independent pixels from the CT image data.

In future research, we hope to further investigate the cause and effects of reducing the

imbalance ratio of the training set while building classifiers to improve the classifications of the more rare categories. More emphasis can be placed on strengthening the classification of the rare categories, such as Honeycomb Lung, due to the already existing high classification rates of the more common categories, such as Lung Fibrosis. We would also like to extend this research into classifying a broader range of lung disease patterns, such as two types of Ground-glass patterns and including a category for normal non-disease lungs. This paper provided key analyses in addressing the unbalanced data classification issue that can be extended to a more complete set of lung disease classifications in future studies.

REFERENCES

- [1] He, H., Garcia, E. (2009). Learning from Imbalanced Data. *IEEE Trans. Knowledge and Data Eng.* 21:1263-1284.
- [2] Ertekin, S., Huang, J., Bottou, L., Giles, L. (2007). Learning on the Border: Active Learning in Imbalanced Data Classification. *In Proc. ACM Conf. on Information and Knowledge Management (CIKM 07)*.
- [3] Olson AL, et al. Interstitial lung disease. *Schraufnagel DE, et al. Breathing in America: Diseases, Progress and Hope. New York, N.Y.: American Thoracic Society* <http://www.thoracic.org/education/breathing-in-america.pdf>. Accessed April 20, 2013.
- [4] Chabat, F., Yang G.Z., Hansell, D., (2003) Obstructive Lung Diseases: Texture Classification for Differentiation at CT. *Radiology* 228:871-877.
- [5] Kim, H.J., Li, G., Gjertson, D., Elashoff, R., Shah, S., Ochs, R., Vasunilashorn, F., Abtin, F., Brown, M., Goldin, J. (2008). Classification of Parenchymal Abnormality in Scleroderma Lung Using a Novel Approach to Denoise Images Collected via a Multicenter Study. *Acad Radio* 15:1004-1016.
- [6] Aujol, J., Chambolle, A. (2005). Dual Norm and Image Decomposition Models. *Int J Comput Vision* 63:85-104
- [7] Sonka, M., Hlavac, V., Boyle, R. (1993). *Image Processing, Analysis and Machine Vision*. London, England: Chapman & Hall.
- [8] Haralick, R., Shanmugam, K., (1973). Texture Features for Image Classification. *IEEE Trans Syst Man Cybernetics* 3:610-621.
- [9] Maglogiannis, I.G., Zafiroopoulos, E.P. (2004) Characterization of digital medical images utilizing support vector machines. *BMC Med Decis Mak* 4:4-32.
- [10] Chapelle, O., Haffner, P., Vapnik, V. (1999) Support Vector Machines for Histogram Based Image Classification. *IEEE Trans Neural Networks* 10:1055-1064.
- [11] Xu, Y., van Beek, E.J., Hwanjo, Y., et al. (2006) Computer-aided Classification of Interstitial Lung Diseases via MDCT: 3D Adaptive Multiple Feature Method (3D AMFM). *Acad Radio* 13:969-978.
- [12] Vapnik, V. (1995) *The Nature of Statistical Learning Theory*. Springer, New York.
- [13] Smola, A.J., Scholkopf, B., Sparse Greedy Matrix Approximation for Machine Learning. *In Proc. of 17th Int. Conference on Machine Learning (ICML)*.
- [14] Agresti, A. (2007) *An Introduction to Categorical Data Analysis*. Wiley, New Jersey.