A COMPARATIVE STUDY ON TWO STRATEGIES FOR DISTRIBUTED CLASSIFICATION

A Thesis

Presented to the Faculty of the Department of Mathematical Sciences

Middle Tennessee State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Mathematical Sciences

by

Honglan Xu

May 2018

Thesis Committee:

Qiang Wu, Chair

Don Hong

Yeqian Liu

Lisa Green

Copyright © 2018, Honglan Xu

ACKNOWLEDGMENTS

I would first like to thank my thesis advisor, Dr. Qiang Wu, who gave guiding opinions and recommendations on the research direction of my thesis. During the process of writing the thesis, he promptly pointed out the difficulties and doubts and put forward many useful suggestions for improvement. I would also like to thank all the professors and students in the program of Mathematics and Actuarial Science in the Department of Mathematical Sciences at Middle Tennessee State University who helped me to perfect my algorithms.

ABSTRACT

Distributed learning is an effective tool to process big data. An easy and effective distributed learning approach is the divide and conquer method. It first partitions the whole data set into multiple subsets. A base learning algorithm is then applied to each subset. Finally, the results from these subsets are coupled together. In the classification setting, many classification algorithms can be used in the second stage. Typical ones include the logistic regression and support vector machines. For the third stage, both voting and averaging can be used as the coupling strategies. In this thesis, empirical studies are done to thoroughly compare the effectiveness of these two coupling strategies. Averaging is found to be more effective in most scenarios.

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER 1: INTRODUCTION 1

1.1 Outline of This Thesis 2

CHAPTER 2: BINARY CLASSIFICATION 4

2.1 Problem Overview 4

2.2 Basic Classification Algorithms 5

2.2.1 Logistic Regression 5

2.2.2 Support Vector Machine 6

CHAPTER 3: DISTRIBUTED LEARNING 9

3.1 Voting 9

3.2 Averaging 10

CHAPTER 4: DATA 11

4.1 Data Collection 11

4.2 Description of Data Sets 12

4.2.1 Default of Credit Card Clients Data Set 12

4.2.2 APS Failure at Scania Trucks Data Set 12

4.2.3 Epileptic Seizure Recognition Data Set 12

4.2.4 Turkiye Student Evaluation Data Set 13

4.2.5 Wireless Indoor Localization Data Set 13

4.2.6 Wilt Data Set 13

4.3 Data Processing 13
4.4 Data Separation 14

CHAPTER 5: RESULTS AND CONCLUSION 16

5.1 Results 16
5.2 Conclusion 20

BIBLIOGRAPHY 21

LIST OF TABLES

1 Description Of 6 Data Sets. 11

2 Classification Accuracy Of Logistic Regression With $k = 7$ And 30% Data Used For Training. 16

3 Classification Accuracy Of SVM With $k = 7$ And 30% Data Used For Training. 17

4 Classification Accuracy Of Logistic Regression With $k = 7$ And 70% Used For Training. 17

5 Classification Accuracy Of SVM With $k = 7$ And 70% Used For Training. 18

6 Classification Accuracy Of Logistic Regression With $k = 21$ And 70% Data Used For Training. 18

7 Classification Accuracy Of SVM With $k = 21$ And 70% Data Used For Training. 19

LIST OF FIGURES

1 *The Outline Of This Thesis.* 3

2 *Optimal Hyperplane.* 7

3 *Data Separation.* 14

4 *Effect Of Number Of Blocks On Accuracy In Wireless1 Data Set.* 19

5 *Effect Of Number Of Blocks On Accuracy In APS Failure Data Set.* 20

CHAPTER 1

INTRODUCTION

Classification is a critical research field in machine learning, data mining, and pattern recognition. Its objective is to produce a classification function or classification model based on these features of the data set to classify the information properly and efficiently. The classification model can map unknown samples into a given category. It is common knowledge that both classification and regression can be used for prediction. Unlike regression methods, the output of the classification is a discrete category value, but the output of the regression is a continuous or ordinal value.

Many problems in real life can be converted into classification problems. In a binary classification problem, we try to predict whether the result belongs to one of two classes, such as true or false. For example, a classification model can be constructed by customer classification to perform risk evaluation on bank credit card business; an important feature in current marketing is to emphasize customer segmentation. Other classification applications exist as well, such as image recognition technology and automatic text classification techniques in search engines.

When the researchers are faced with the classification of the data set, they usually apply their desired “best” classifier. This expectation is determined by their knowledge of the available classifiers. Different classification algorithms will produce different classifiers, and the quality of the classifiers will directly affect the accuracy of the classification results and the efficiency of machine learning. Therefore, when categorizing large-scale mass data, it is crucial to choose the most appropriate classification algorithm.

Recently, with the rapid development of the internet and the continuous expansion of network information and data information, effective use of this rich data information has become one of the focuses of information technology.

Recent research on classification algorithms mainly focuses on the following two aspects: First, the traditional classification algorithms or combinations are applied to develop various application systems. Second, the traditional classification algorithms are improved.

Distributed learning is an effective tool to handle big data and received considerable attention recently. In classification setting, the big data is first divided into several subsets. A base classification algorithm is then applied on each subset. Finally, these classifiers are coupled together to produce the final classifier, which will be used to classify new data. This thesis will analyze and compare two distributed classification strategies in depth, summarize their respective advantages and disadvantages and applicable situations, and provide a reference for future application of distribution learning in binary classification problems.

1.1 Outline of This Thesis

In this thesis, my goal is to compare two different decision strategies for distributed classification. I will describe the classification problems and provide the judging criteria briefly in Chapter 2. The selection of the classification algorithm at every disjointed data subsets is vital to produce a new function. Detailed introductions to Logistic Regression and Support Vector Machine algorithms will be given.

Chapter 3 provides a brief introduction to distributed learning. I will explain two different discrimination strategies for distributed classification, namely, averaging and voting.

In Chapter 4, I will introduce the data that will be used for the comparative study. Pre-processing and experiment settings will be described.

Chapter 5 discusses the results of those two strategies, exploring the accuracy and interpretation which will lead to any conclusion. Figure 1 indicates the flow chart of my study.

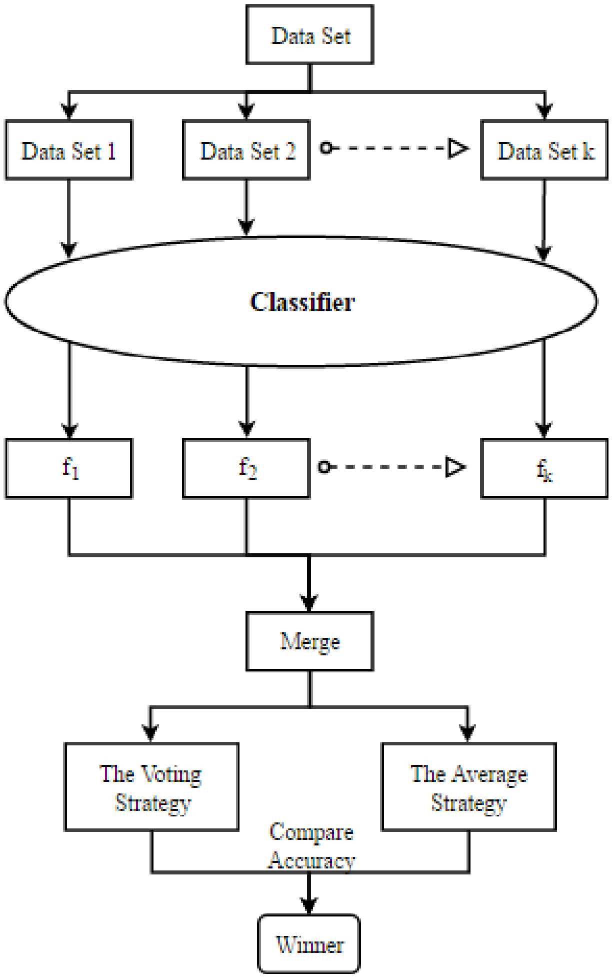

Figure 1: *The Outline Of This Thesis.*

CHAPTER 2

BINARY CLASSIFICATION

2.1 Problem Overview

In this paper, with respect to the classification problem, the input is a p-dimensional vector, which is actually what we call "feature." And we use F to represent the p-dimensional feature space [1]. Here, we limit ourselves to binary classification problems. In general, we identify the two classes with the symbols $(+)$ and $(-)$. A training set of a number of patterns $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ with known class labels $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ is given. A classification algorithm allows us to use the training set to capture a discriminant function $\mathbf{f}(\mathbf{x})$, which is usually a real valued objective function of an input pattern $\mathbf{x}$.

A discriminant function that is simple weighted sums of the training patterns is called a linear decision function. That is,

$$\mathbf{f}(\mathbf{x}) = b + \theta^T \mathbf{x}, \tag{1}$$

where b is the bias of linear function, θ^T is the weight vector of features and $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$ is a vector of p features.

Usually in the field of machine learning, $\mathbf{y}_i \in \{-1, 1\}$. A new sample can be classified by the sign of the discriminant function in this case. This convention is used by a support vector machine algorithm where the decision function is

$$\mathbf{f}(\mathbf{x}) > 0, \implies \mathbf{y} = 1 \implies \mathbf{x} \in class(+), \tag{2}$$

$$\mathbf{f}(\mathbf{x}) < 0, \implies \mathbf{y} = -1 \implies \mathbf{x} \in class(-). \tag{3}$$

However, people in the statistical field prefer that $\mathbf{y}_i \in \{0, 1\}$. For instance, in a logistic regression algorithm, a function $h_\theta(\mathbf{x})$ is learned, which represents the estimation for $P(\mathbf{y} = 1|\mathbf{x})$. The classification for the new sample can be made by

$$h_\theta(\mathbf{x}) > 0.5 \implies \mathbf{y} = 1 \implies \mathbf{x} \in class(+), \tag{4}$$

$$h_\theta(\mathbf{x}) < 0.5 \implies \mathbf{y} = 0 \implies \mathbf{x} \in class(-), \tag{5}$$

$$h_\theta(\mathbf{x}) = 0.5, \text{decision boundary.} \tag{6}$$

This is equivalent to using the sign of the function

$$\mathbf{f}(\mathbf{x}) = h_\theta(\mathbf{x}) - 0.5. \tag{7}$$

2.2 Basic Classification Algorithms

2.2.1 Logistic Regression

Logistic regression is one of the most common and effective classification methods. It has been widely used in image processing, text classification, and semantic recognition.

A nature of this algorithm is that its output value is always between 0 and 1. Linear logistic regression is based on the theory of linear regression, and then adds sigmoid functions. The linear regression formula is as follows

$$z = b + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p = b + \theta^T \mathbf{x}. \tag{8}$$

We define a hypothesis function $h_\theta(\mathbf{x})$. In binary classification problems, the logistics regression model is

$$h_\theta(\mathbf{x}) = g(\theta^T \mathbf{x}) = \frac{1}{1 + e^{-(b+\theta^T \mathbf{x})}}, \tag{9}$$

$$\text{where } g(z) = \frac{1}{1 + e^{-z}}. \tag{10}$$

Let $h_\theta(\mathbf{x})$ represent the probability that a pattern x belongs to class(+). Then

$$P(\mathbf{y} = 1|\mathbf{x}; \theta) = h_\theta(\mathbf{x}), \tag{11}$$

$$P(\mathbf{y} = 0|\mathbf{x}; \theta) = 1 - h_\theta(\mathbf{x}). \tag{12}$$

The function $g(z) = \frac{1}{1+e^{-z}}$ is often called the sigmoid or logistic function. It is an S-shaped function that squashes the value of $b + \theta^T\mathbf{x}$ into the range [0,1] so that we interpret $h_\theta(\mathbf{x})$ as a probability [7]. The parameters $b, \theta_1, \theta_2, \ldots, \theta_p$ are estimated using maximum likelihood estimation. The log likelihood function is given by:

$$l(\theta) = \sum_{i=1}^{n} [\mathbf{y_i} ln[h_\theta(\mathbf{x_i})] + (1 - \mathbf{y_i}) ln[1 - h_\theta(\mathbf{x_i})]] \tag{13}$$

The classification for a new sample $\mathbf{x}$ can be made by

$$\mathbf{y} = \begin{cases} 0 & h_\theta(\mathbf{x}) < 0.5 \\ \\ 1 & h_\theta(\mathbf{x}) > 0.5. \end{cases} \tag{14}$$

2.2.2 Support Vector Machine

The Support Vector Machine (SVM) was invented by Boser (1992) and Vapnik (1998). Compared with Logistic Regression, SVM is clearer and more powerful in the experiment of complex nonlinear classification problems.

The Support vector machine (SVM) is a supervised machine learning algorithm. Based on limited sample information, it strives to find the best balance between model complexity and learning ability. Due to its excellent generalization performance, it has been widely used in the field of pattern recognition and regression analysis.

In binary classification problems, we plot each data item as a point in p-dimensional space (where p is the number of features) with the value of each feature being the value of a particular coordinate [6]. Then, we perform classification by finding the hyperplane that differentiates the two classes distinctly.

Linear SVM is a special linear discriminant classifier. A data set will be “linearly separable” if a linear decision function can separate it without bias [1]. If the training is linearly separable, a linear SVM is a classifier which has a maximum margin. The decision boundary (a hyperplane in high dimensional case or a straight line in 2-dimensional case) is positioned to leave the largest possible margin on either side [1].

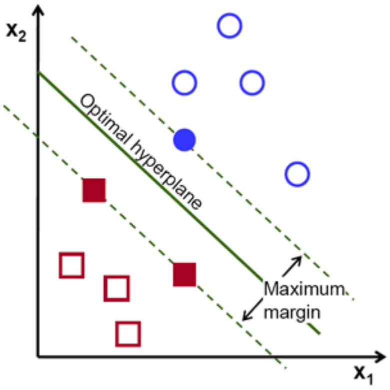

Figure 2: *Optimal Hyperplane.*

The optimal hyperplane is selected by the maximum margin principle, through picking the boundary which maximizes the distance between those two classes. The distance between these two hyper-planes is $\frac{2}{||\theta||}$. The optimal hyperplane is illustrated in Figure 2 [24].

The hyperplane can be expressed by its normal vector θ and a bias b as follows,

$$b + \theta^T \mathbf{x} = \mathbf{0}. \tag{15}$$

It produces the corresponding discriminant function

$$\mathbf{y} = sgn(b + \theta^T \mathbf{x}). \tag{16}$$

where sgn is the sign function extracting the sign of real numbers. The optimization problem to solve linear SVM is

$$\underset{\theta}{min} \frac{1}{2} \sum_{j=1}^{p} \theta_j^2 \tag{17}$$

$$\text{s.t. } b + \theta^T \mathbf{x} > 0, \text{if } \mathbf{y} = 1; \tag{18}$$

$$b + \theta^T \mathbf{x} < 0, \text{if } \mathbf{y} = -1. \tag{19}$$

When the data is not linearly separable, errors are unavoidable. A regularization parameter, referred to as “Cost,” is necessary to trade off the errors and the margin between the two classes. Tuning this cost parameter, as well as any kernel parameters, becomes vital to avoid overfitting or underfitting problems.

CHAPTER 3

DISTRIBUTED LEARNING

A distributed system is a system consisting of a set of computer nodes that communicate through the network and work together to accomplish common tasks. The emergence of distributed systems is because cheap, common machines can complete the computing and storage tasks that a single computer cannot. Its purpose is to use more machines to handle more data. It is crucial to understand that when the processing power of a single node cannot meet the increasing tasks of computing and storage, the hardware upgrade (adding memory, adding disks, and using a better CPU) is too costly, and further optimization is not available, we need to consider distributed systems.

Nowadays, technology has entered the era of big data. Distributed learning is an effective method to process big data. The divide and conquer approach is a simple but effective way to implement distributed learning. It first partitions a data set into multiple disjointed data subsets. Then, a base learning algorithm is applied to each subset. Finally, the results from all subsets are coupled together to make a prediction.

In this thesis, I focus on distributed classification problems. For the second stage, many classification algorithms can be used. I will use logistic regression and support vector machine, which have been discussed in Chapter 2.

For the third stage, I will consider two coupling strategies and compare them empirically.

3.1 Voting

We use distributed learning to randomly divide the training set into k blocks. Through the logistic regression and support vector machine classifiers mentioned in Chapter 2, we can get an objective function on each subset, $\hat{\mathbf{f}}_\mathbf{1}(\mathbf{x}), \hat{\mathbf{f}}_\mathbf{2}(\mathbf{x}), \hat{\mathbf{f}}_\mathbf{3}(\mathbf{x}), \ldots, \hat{\mathbf{f}}_\mathbf{k}(\mathbf{x})$. These functions are real valued and can be used to label new data points. So each new data point is labeled k times, which is regarded as k votes.

By voting strategy, the category of a new sample point is the label that received more votes. If the number of positive signs during these k functions exceeds the negative signs, this sample belongs to class(+); otherwise, it belongs to class(−). Mathematically, the formula is

$$\mathbf{y} = sgn(sgn(\mathbf{\hat{f}_1}(\mathbf{x})) + sgn(\mathbf{\hat{f}_2}(\mathbf{x})) + \ldots + sgn(\mathbf{\hat{f}_k}(\mathbf{x})), \tag{20}$$

where if logistic regression is used,

$$\mathbf{\hat{f}_i}(\mathbf{x}) = \hat{h}_\theta^i(\mathbf{x}) - 0.5,$$

and if linear SVM is used,

$$\mathbf{\hat{f}_i}(\mathbf{x}) = \hat{\theta}^T\mathbf{x} + b.$$

For instance, we divide the training set into seven parts. If the symbolic results of a new sample point in these seven objective classifiers are $+, -, +, +, +, +, -$, then the number of positive signs exceeds the negative. So, this new sample point belongs to class(+) by using voting strategy.

3.2 Averaging

By averaging strategy, we will compare the average value of all k objective functions learnt from the k subsets. The decision on new data points will be made based on the average value.

When the logistic regression algorithm is used, we get the particular output function, $\hat{h}_\theta^1(\mathbf{x}), \hat{h}_\theta^2(\mathbf{x}), \ldots, \hat{h}_\theta^k(\mathbf{x})$. We find the mean of $\hat{h}_\theta^1(\mathbf{x}), \hat{h}_\theta^2(\mathbf{x}), \ldots,$ and $\hat{h}_\theta^k(\mathbf{x})$, denoted by $\overline{\hat{h}_\theta(\mathbf{x})}$. If $\overline{\hat{h}_\theta(\mathbf{x})} > 0.5$, then $\mathbf{x} \in class(+)$; If $\overline{\hat{h}_\theta(\mathbf{x})} < 0.5$, then $\mathbf{x} \in class(-)$.

When SVM is used, we collect these k output discriminant functions, $\mathbf{\hat{f}_1}(\mathbf{x}), \mathbf{\hat{f}_2}(\mathbf{x}), \ldots,$ and $\mathbf{\hat{f}_k}(\mathbf{x})$, and then average them. We use $\overline{\mathbf{\hat{f}}(\mathbf{x})}$ to represent the average value. If $\overline{\mathbf{\hat{f}}(\mathbf{x})} > 0$, then $\mathbf{x} \in class(+)$; If $\overline{\mathbf{\hat{f}}(\mathbf{x})} < 0$, then $\mathbf{x} \in class(-)$.

CHAPTER 4

DATA

4.1 Data Collection

For this study, the data sets used come from the UC Irvine Machine Learning Repository, which is the most widely used database in machine learning literature. As of February 2018, there are 313 data sets which can be used for classification tasks on the UCI website. Table 1 shows the number of observations and the number of features of each data set.

We select 6 different representative data sets to develop the strategy comparison. These are Default of Credit Card Clients Data Set, APS Failure at Scania Trucks Data Set, Wireless Indoor Localization Data Set, Turkiye Student Evaluation Data Set, Wilt Data Set and Epileptic Seizure Data Set. These data sets cover several different fields, including finance, communications, education, medical treatment, and transportation. This makes the comparative study representative and convincing.

Table 1: Description Of 6 Data Sets.

Data Set	#Sample	#Feature
Credit Card	30,000	23
APS Failure	60,000	170
Epileptic	9,200	178
Student Eva	5,046	32
Wireless	2,000	7
Wilt	4,889	5

As the formats of all the data sets are different, they are first converted into cvs-format using EXCEL. The Credit Card, APS Failure, Studen Evaluation and Wilt are binary problems by treating the largest class as 1 and the rest as -1. As for the Epileptic Seizure Data Set and Wireless Indoor Location Data Set, these are multiclass problems. I created two binary classification subtasks for each data set. In total, I have 8 binary classification tasks.

4.2 Description of Data Sets

4.2.1 Default of Credit Card Clients Data Set

This data set displays the customers' default payments in Taiwan where the binary result of this classification is credible or not credible clients [13]. There are 30,000 credit card customers' personal information, including 23 influencing factors such as amount of the given credit, education, gender, age, marital status, history of past payment, etc.

4.2.2 APS Failure at Scania Trucks Data Set

The data was collected from heavy Scania Trucks in everyday usage [19]. An air pressure system (APS) that generates pressurized air can be used for various functions in the truck, such as braking and shifting. The positive class of this data is that the truck failure is due to APS, and the negative class is the failure that has nothing to do with APS. There are 60,000 failure incidents and 170 factors. Some information is missing, replaced by NA.

4.2.3 Epileptic Seizure Recognition Data Set

There are five classes in this data set. All subjects belonging to classes 2, 3, 4, and 5 are individuals who do not have epileptic seizures, while the individuals in class 1 actually have epileptic seizures [23]. Since I limit my study to binary classification problems, I consider two subproblems, the class 1 versus class 2, and the class 1 versus class 3. Both tasks involve a data set of 4,600 observations and 178 variables.

4.2.4 Turkiye Student Evaluation Data Set

This data set includes 5,046 useful student evaluation scores given by Gazi University in Ankara. There are 32 attributes, such as “the Instructor came prepared for classes”, the number of times the student has taken this course, level of difficulty of the course as perceived by the student, etc [22].

4.2.5 Wireless Indoor Localization Data Set

The data set is collected to perform experiments on how to use wifi signal strength to determine indoor location. The seven variables are obtained by observing the wifi signal strength on smartphones [21]. The decision variable is one of the four indoor rooms. I only consider two binary classification subtasks, the room 1 versus room 2, and the room 3 versus room 4.

4.2.6 Wilt Data Set

The data set consists of image segments generated by a segmented raster image. The Quickbird multi-special image band information and texture information from the pan image band are presented in these clips [20]. It has a total of 4,889 information collection points. The five variables are GLCM mean texture (Pan band), Mean green value, Mean red value, Mean NIR value, and Standard deviation (Pan band). The class w represents diseased trees, and the class n means all other land cover.

4.3 Data Processing

In general, a knowledge discovery process consists of the following steps: data cleaning, data intergation, data selection, data transformation, data mining, pattern evaluation and knowledge presentation [17]. In this study, data cleaning, data selection and data transformation occur.

To have robust models, we should make sure that the data sets are robust as well. We know that some models have different assumptions of the predictor data and may need to be pre-processed. Here, I need the predictors to be centered and scaled since most of the predictors follow normal distributions.

At the same time, I would like to delete the missing values. Since APS Failure at Scania Trucks Data Set has plenty of predictors, I use Principal Component Analysis (PCA) to select important variables and reduce dimensions from 178 to 142. To get real estimates of performance, all data transformations ought to be involved within the cross validation. After that, we can use this model to predict new sample data.

4.4 Data Separation

The data used to build the most optimal model usually comes from multiple data sets. If we spend too much data in training, it will not be beneficial to obtain a believable evaluation of predictive performance. Meanwhile, the model is likely to over-fit. However, putting too much data in testing cannot get a successful assessment of model parameters.

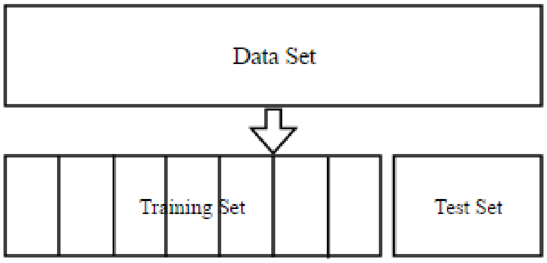

Figure 3: *Data Separation.*

In this paper, I split data into training and test data sets (Shown in Figure 3). The training set is used to fit the parameters of the model. The hyperparameters of the models are selected by cross-validation.

I randomly divide the training data into m disjointed blocks of approximately equal size. The final result is based on the hold-out predictions. The data in the test data set can be used to create an unbiased evaluation of a final model obtained from the training data set.

In this study, in order to compare the two different distributed classification strategies and obtain more widely applied results, we will do the following different situations: (A) There are more testing data than training data. In this case, testing data accounts for 70% of the total for all eight tasks. (B) The training data is more than testing data. I use 70% of the data for the training set and 30% for the testing. (C) The number of groups in distributed learning varies. In this case, I still use 70% for training and 30% for the testing. But, I will partition the training set into k subsets, with k varying from 7 to 21.

CHAPTER 5

RESULTS AND CONCLUSION

5.1 Results

The overall accuracy is used to evaluate the performance of classification. We may quickly judge which strategy prevails by comparing the mean and standard deviation of the accuracy of the two strategies. If the average of the accuracy of one of the strategies is significantly higher than the other and the variance is small, then this strategy must be better than the other. For a more sophisticated comparison, we adopt the statistical t-test. Then, the p-value will be calculated to examine the significance of the difference.

Let μ_1 be the mean accuracy of the Voting Strategy after 20 experiments and μ_2 the mean accuracy of the Averaging Strategy. We test if averaging is better than voting. If the *p-value* is very small, the Averaging Strategy is better than the Voting. Otherwise, these two strategies are considered not significantly different.

Table 2: Classification Accuracy Of Logistic Regression With $k = 7$ And 30% Data Used For Training.

Data Set	Accuracy		p-value
	Voting	Averaging	
Credit Card	79.51% (0.804%)	80.72% (0.168%)	0.01412
APS Failure	98.18% (0.128%)	98.58% (0.264%)	1.41e-7
Epileptic Seizure1	58.53% (1.472%)	60.23% (0.702%)	0.00912
Epileptic Seizure2	49.90% (0.493%)	50.43% (0.845%)	0.01303
Student Eva	70.61% (1.389%)	69.03% (0.945%)	0.85421
Wireless1	69.56% (1.454%)	70.95% (0.243%)	0.00154
Wireless2	70.85% (1.358%)	72.04% (0.980%)	0.01247
Wilt	96.80% (0.213%)	96.91% (0.197%)	0.05391

Table 3: Classification Accuracy Of SVM With $k = 7$ And 30% Data Used For Training.

Data Set	Accuracy		p-value
	Voting	Averaging	
Credit Card	65.64% (1.445%)	67.25% (0.895%)	0.02541
APS Failure	87.21% (0.174%)	88.16% (0.356%)	2.67e-6
Epileptic Seizure1	50.62% (2.354%)	51.83% (1.852%)	0.01024
Epileptic Seizure2	47.56% (0.503%)	48.32% (0.635%)	0.01238
Student Eva	63.31% (1.145%)	63.45% (1.405%)	0.65221
Wireless1	60.42% (1.024%)	61.46% (0.825%)	0.04231
Wireless2	59.86% (1.562%)	60.47% (1.052%)	0.02408
Wilt	79.43% (1.475%)	80.74% (2.627%)	0.03333

In Table 2 and 3, we use 30% data for training and 70% for testing. The number of subsets is set as 7. As the average of accuracy of the Averaging Strategy for all data sets except Student Eva is higher than the Voting Strategy, we can say that the Average Strategy is generally better than the Voting Strategy in this scenario.

Table 4: Classification Accuracy Of Logistic Regression With $k = 7$ And 70% Used For Training.

Data Set	Accuracy		p-value
	Voting	Averaging	
Credit Card	85.13% (0.625%)	86.79% (0.738%)	0.02314
APS Failure	98.25% (0.564%)	98.91% (0.321%)	0.00012
Epileptic Seizure1	60.05% (1.037%)	60.86% (0.814%)	0.01252
Epileptic Seizure2	57.85% (0.587%)	58.37% (1.420%)	0.00125
Student Eva	73.16% (1.044%)	74.25% (0.756%)	0.74511
Wireless1	70.21% (1.412%)	73.05% (0.123%)	0.00142
Wireless2	73.12% (1.457%)	74.08% (0.653%)	0.01324
Wilt	96.68% (0.541%)	97.02% (0.406%)	0.06452

Table 5: Classification Accuracy Of SVM With $k = 7$ And 70% Used For Training.

Data Set	Accuracy		p-value
	Voting	Averaging	
Credit Card	71.05% (0.973%)	74.48% (2.726%)	0.03234
APS Failure	88.35% (0.728%)	89.92% (0.234%)	0.00008
Epileptic Seizure1	55.83% (1.232%)	56.930% (1.328%)	0.03143
Epileptic Seizure2	55.62% (0.683%)	56.034% (0.390%)	0.03623
Student Eva	68.34% (1.035%)	69.38% (1.521%)	0.48422
Wireless1	69.33% (1.432%)	70.87% (1.072%)	0.03562
Wireless2	67.89% (2.001%)	70.09% (1.243%)	0.01375
Wilt	87.45% (0.654%)	88.93% (1.334%)	0.09287

In Table 4 and 5, we also let the number of blocks be 7, but the 70% data are used for the training. As expected, the accuracy is significantly better than before. The Averaging Strategy is still found to be better than the Voting Strategy, indicating the superiority of averaging strategy is irrelevant to the size of training set.

Table 6: Classification Accuracy Of Logistic Regression With $k = 21$ And 70% Data Used For Training.

Data Set	Accuracy		p-value
	Voting	Averaging	
Credit Card	82.13% (0.728%)	83.91% (0.235%)	0.03156
APS Failure	97.05% (0.635%)	97.68% (0.334%)	0.01245
Epileptic Seizure1	56.03% (1.235%)	57.88% (0.937%)	0.03248
Epileptic Seizure2	57.36% (0.673%)	58.24% (0.653%)	0.02469
Student Eva	69.46% (0.768%)	70.37% (0.878%)	0.56319
Wireless1	68.38% (0.869%)	69.45% (0.463%)	0.00246
Wireless2	67.98% (1.027%)	70.24% (0.994%)	0.01924
Wilt	90.78% (1.238%)	92.09% (1.026%)	0.07329

Table 7: Classification Accuracy Of SVM With $k = 21$ And 70% Data Used For Training.

Data Set	Accuracy		p-value
	Voting	Averaging	
Credit Card	70.98% (0.876%)	72.09% (1.256%)	0.02345
APS Failure	86.83% (0.889%)	88.31% (0.653%)	0.00021
Epileptic Seizure1	51.09% (1.023%)	52.40% (1.129%)	0.06289
Epileptic Seizure2	50.69% (0.997%)	51.97% (0.785%)	0.05273
Student Eva	65.83% (1.145%)	66.04% (1.445%)	0.33456
Wireless1	66.03% (1.028%)	67.23% (1.034%)	0.05382
Wireless2	65.85% (1.372%)	67.29% (1.342%)	0.02341
Wilt	83.81% (0.732%)	84.39% (1.038%)	0.10284

In Table 6 and 7, we use 70% for training but set the number of subsets as 21. So in each subset, there are less data. Though, we see a similar results: the Averaging Strategy is better than the Voting Strategy. At the same time, we observe that the accuracy on the test set is slightly reduced.

I take out two of the most stable data sets, Wireless Indoor Localization Data Set1 and APS Failure at Scania Trucks Data Set, for further comparative analysis.

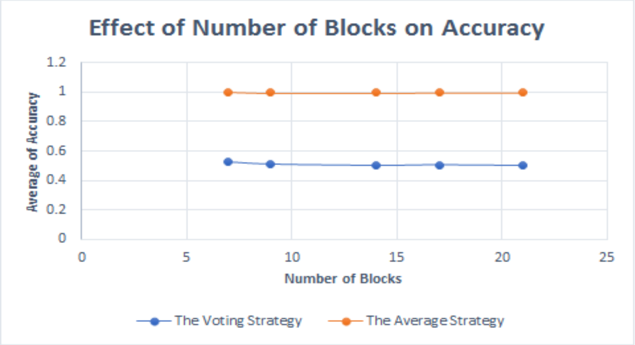

Figure 4: *Effect Of Number Of Blocks On Accuracy In Wireless1 Data Set.*

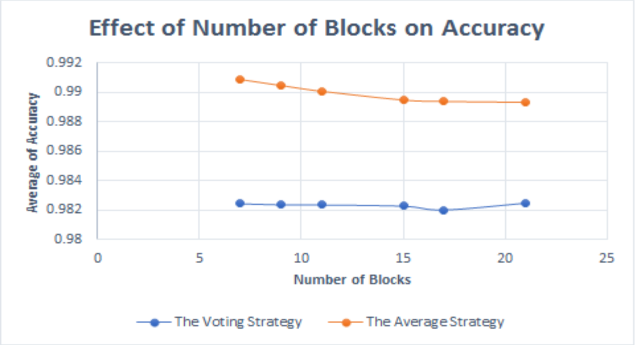

Figure 5: *Effect Of Number Of Blocks On Accuracy In APS Failure Data Set.*

There is a weak negative relationship shown in Figure 4 and 5 between the accuracy of the strategy and the number of blocks in distributed learning.

5.2 Conclusion

By the comparison of the accuracy of two strategies, we can conclude that Averaging is generally slightly better than Voting, and therefore is preferred for distributed classification problems. The accuracy of the strategy is negatively correlated with the number of subsets in distributed learning. This is intuitivly expected and compatible with the study of distributed regression where the optimal prediction is shown to be achieved only when the number of subsets is not too large.

BIBLIOGRAPHY

[1] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning*, **46**(2002), 389-422.

[2] T. S. Furey, N. Cristinanini, N. Duffy, D. W. Bednarski, M. Schummer, D. Haussler, Support Vector Machine Classification and Validation of Cancer Tissue Samples using Microarray Expression Data, *Bioinformatics*, **10**(2000), 906-914.

[3] M. Fernandez-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, **15**(2014), 3133-3181.

[4] L. Hemphill, Distributed Learning: Definitions We Can Use, *[http://eds-courses.ucsd.edu//tep290/fa05/final-papers/sdlcFinalLibby.pdf]*

[5] M. Jekel, S. Fiedler, A. Glckner, Diagnostic Task Selection for Strategy Classification in Judgment and Decision Making: Theory, Validation, and Implementation in R, *Judgment and Decision Making*, **8**(2011), pp. 782-799.

[6] S. Ray, Understanding Support Vector Machine algorithm from examples (along with code), *[https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/]*

[7] C. Shalizi, Chapter 12 Logistic Regression, *[http://www.stat.cmu.edu/ cshalizi/uADA/12/lectures/ch12.pdf]*

[8] A.K. Jain, R.P.W. Duin, J. Mao, Statistical Pattern Recognition: A Review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(1999), pp. 4-37.

[9] J. Rosenblatt, B. Nadler, On the Optimality of Averaging in Distributed Statistical Learning, *Information and Inference*, **5**: 4(2016), 379-404.

[10] S. Lin, X. Guo, D. Zhou, Distributed Learning with Regularized Least Squares, *Journal of Machine Learning Research*(2015), 1-28.

[11] Y. Zhang, J. Duchi, M. Wainwright, Divide and Conquer Kernel Ridge Regression, *Journal of Machine Learning Research*, **30**(2013), 1-26.

[12] D. Peteiro-Barral, B. Guijarro-Berdias, A survey of methods for distributed machine learning, *Prog Artif Intell*, **2**(2013), 1-11.

[13] I. C. Yeh, C. H. Lien, The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, *Expert Systems with Applications*, **36**: 2(2009), 2473-2480.

[14] G. Oreki, S. Oreki, An experimental comparison of classification algorithm performances for highly imbalanced datasets, $[https://bib.irb.hr/datoteka/717193.CECIIS_2014_Oreski_Oreski.pdf]$

[15] S. Batsuuri, J. Ahn, J. Ko, Steel surface defects detection and classification using SIFT and voting strategy, *International Journal of Software Engineering and Its Applications*, **6**: 2(2012), 161-166.

[16] B. Umamaheswararao, P. Seetharamaiah, S. Phanikumar, An Incorporated Voting Strategy on Majority and Score-based Fuzzy Voting Algorithms for Safety-Critical Systems, *International Journal of Computer Applications*, **98**: 4(2014), 0975-8887.

[17] J. Kuo, An Automatic Library Data Classification System Using Layer Structure and Voting Strategy, *ICADL*, **8839**(2014), 279-287.

[18] M. Kuhn, Building Predictive Models in R Using the caret Package, *Journal of Statistical Software*, **28**: 5(2008), 1-26.

[19] E.C. Ozan, E. Riabchenko, S. Kiranyaz, M. Gabbouj, An Optimized k-NN Approach for Classification on Imbalanced Datasets with Missing Data, *Advances in Intelligent Data Analysis XV*, **9897**(2016), 387-392.

[20] B. Johnson, R. Tateishi, N. Hoan, User Localization in an Indoor Environment Using Fuzzy Hybrid of Particle Swarm Optimization and Gravitational Search Algorithm with Neural Networks, *in Proceedings of Sixth International Conference on Soft Computing for Problem Solving*, **34**: 20(2017), 286-295.

[21] J. Rohra, B. Perumal, S. J. Narayanan, P. Thakur, R. B. Bhatt, A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees, *International Journal of Remote Sensing*, **34**: 20(2013), 6969-6982.

[22] G. Gunduz, E. Fokoue, UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science(2013).

[23] RG. Andrzejak, K. Lehnertz, C. Rieke, F. Mormann, P. David, CE. Elger, Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state, *Phys. Rev. E*, **64**(2001).

[24] F. I. Garca, Introduction to Support Vector Machines,
https://docs.opencv.org/2.4/doc/tutorials/tutorials.html(2012).