Intelligent Discrimination of Growing Areas based on Near-Infrared Spectra

by

Lin Xie

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Master of Applied Science
in
Engineering

Guelph, Ontario, Canada

© Lin Xie, April, 2012
ABSTRACT

Intelligent Discrimination of Growing Areas based on Near-Infrared Spectra

Lin Xie
University of Guelph, 2012

Advisor:
Professor Simon X. Yang

The tobacco growing area is an important aspect for the consistency of cigarette aroma and the control of cigarette quality. The fragrance of tobacco leaves would be different for climates and planting environments, such as soil and rainfalls. Accurately discriminating tobacco growing areas is very important to maintain the specifications of cigarettes. In this thesis, the relationship between tobacco near-infrared (NIR) spectra and growing areas is studied. Soft computing models and statistical classifiers are established, and the performance of the developed classifiers are compared in the prediction accuracy and in evaluations derived from confusion matrix. An artificial neural network (ANN) classifier and a statistical model are firstly developed. The best prediction accuracy of ANN model reaches to 79.3% in 226 training samples and 78.7% in 66 testing samples, respectively, which are 2.2% and 4.5% higher than the best results of the conventional statistical model in training (77.1%) and in testing (74.2%), respectively. A support vector machine (SVM) model is proposed to investigate the characteristics of growing areas based on risk error minimum, and produces a higher classification accuracy than ANN model does, demonstrating the effectiveness and robustness of SVM model. In addition, a genetic algorithm (GA) optimized SVM (GA-SVM) model is proposed for taking the influence of the interaction of individual inputs on the performance of classifiers into account. With the application of GA, the sensitive input subset is identified and used to discrimination models. The simulation results demonstrate that the GA-SVM model has the best performance among the other developed models, and the model complexity is simplified, which is shown by requiring fewer inputs to achieve the equivalent prediction accuracy. The GA-SVM classifier is preferred for solving multi-category problems.
Dedication

To my dear parents for their endless support and love.
Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Simon X. Yang, for his invaluable guidance, sustained support, and encouragement during my graduate study in University of Guelph. His knowledge, patience, vision, and valuable feedback to my research have provided me with lifetime benefits.

I am very grateful to Dr. Wenjie Pan for his suggestions to my research work and thesis, especially the kindly support with experiment dataset.

I would also like to thank my advisory committee member Prof. Fantahun M. Defersha and the examination committee chair Prof. Gregori for their valuable comments on the thesis drafts as well as my presentation.

The School of Engineering and the ARIS Lab offer me with the enjoyable environment for studying and working. I greatly appreciate all the help and support come from all the faculty, staff, my classmates and colleagues.

Last but not least, I want to take this opportunity to express my utmost thank to my husband, Kai Zeng, for his love and support during my graduate study and thesis writing period. Also, I am grateful for my dear parents’ strong encouragement all the time.


## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>viii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>x</td>
</tr>
<tr>
<td>List of Symbols</td>
<td>xii</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>xv</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Problem Statement</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Database</td>
<td>3</td>
</tr>
<tr>
<td>1.3 The Proposed Approaches</td>
<td>4</td>
</tr>
<tr>
<td>1.4 Contribution</td>
<td>5</td>
</tr>
<tr>
<td>1.5 Organization of This Thesis</td>
<td>6</td>
</tr>
<tr>
<td>2 Literature Review</td>
<td>8</td>
</tr>
<tr>
<td>2.1 Tobacco Classification</td>
<td>8</td>
</tr>
<tr>
<td>2.2 Near Infrared Spectroscopy and Preprocessing</td>
<td>10</td>
</tr>
<tr>
<td>2.3 Methods for Classification</td>
<td>14</td>
</tr>
<tr>
<td>2.3.1 Statistic Models</td>
<td>14</td>
</tr>
<tr>
<td>2.3.2 Artificial Neural Networks</td>
<td>15</td>
</tr>
<tr>
<td>2.3.3 Support Vector Machine</td>
<td>19</td>
</tr>
<tr>
<td>2.4 Genetic Algorithms</td>
<td>20</td>
</tr>
</tbody>
</table>
3 The Proposed Neural Network Classifier

3.1 Introduction ....................................................... 23
3.2 The Proposed Methods for Growing Areas Discrimination .......... 24
  3.2.1 De-noising: Savitzky-Golay Algorithm .......................... 24
  3.2.2 Feature Extraction Method: Principal Component Analysis .... 27
  3.2.3 A Neural Network Classifier for Tobacco Growing Area Discrimination ....................................................... 29
  3.2.4 A Statistical Method for Tobacco Growing Area Discrimination: Mahalanobis Distance Model ............................... 32
  3.2.5 Model Evaluations .............................................. 33
3.3 Experimental Results ................................................. 35
  3.3.1 Results of De-noising ............................................ 36
  3.3.2 Results of Feature Extraction ..................................... 36
  3.3.3 Performances ..................................................... 39
3.4 Discussion ............................................................ 44
3.5 Summary ............................................................. 47

4 The Proposed Support Vector Machine Classifier

4.1 The Proposed Method .................................................. 48
  4.1.1 The Support Vector Machine Algorithm .......................... 49
  4.1.2 Typical Kernel Functions ........................................ 52
  4.1.3 Multi-class Support Vector Machine .............................. 53
4.2 Experimental Results ................................................... 53
4.3 Discussion ............................................................. 55
  4.3.1 Selection of Preprocessing Method ............................... 55
  4.3.2 Selection of Kernel Function ..................................... 57
  4.3.3 Selection of Parameters for SVM with RBF Kernel ............. 57
4.4 Summary ............................................................. 60

5 The Proposed Genetic Algorithm Optimized SVM Classifier

5.1 Introduction ........................................................... 61
5.2 Genetic Algorithm Optimized Support Vector Machine Approaches 63
5.3 Experimental Results ................................................. 67
5.4 Discussion ............................................................... 72
   5.4.1 GA-SVM Model Setting .......................................... 72
   5.4.2 Population Size ................................................... 72
   5.4.3 Genetic Operators ................................................ 74
5.5 Summary ............................................................... 76

6 Conclusions and Future Work ........................................ 77

References ........................................................................ 80

A The First 25 PCs of the Preprocessed Data and the Raw Data 92
List of Tables

1.1 The constitution of dataset. .................................................. 3

3.1 The confusion matrix .......................................................... 34

3.2 The prediction accuracy of ANN the MD models for training set with raw spectra and three pro-processed spectra. ......................... 42

3.3 The prediction accuracy of the ANN and MD models for tobacco NIR spectra de-noised using SG smoothing method on testing ............... 43

3.4 The identification capability of ANN and MD for each class are analyzed in terms of $\gamma_{TP}$, $\gamma_{TN}$, $\gamma_{PP}$, and $\gamma$ based on testing set with SG smoothed tobacco NIR spectra. ........................................ 44

4.1 The most commonly used kernel function for SVM. ...................... 53

4.2 The performance of SVM model for tobacco growing area discrimina-
tion. $P_a$ is the prediction accuracy. ......................................... 54

4.3 The identification capability of SVM classifier and ANN model for each class are analyzed in terms of $\gamma_{TP}$, $\gamma_{TN}$, $\gamma_{PP}$, and $\gamma$ based on testing set with SG smoothed tobacco NIR spectra. ......................... 55

4.4 The prediction accuracy of the SVM model for tobacco growing area discrimination using the raw data and the data preprocessed by different de-noising methods. ................................. 56

4.5 The prediction accuracy when the SVM model respectively adopts the commonly used kernel function along with different number of inputs for the SG smoothed data. ................................. 58
4.6 The parameters selected by grid-search and the performance of SVM model for different number of inputs. ........................................ 59

5.1 The amount of information for each PC ranging from 1 to 25 using the SG smoothing processed tobacco NIR spectra. .......................... 62

5.2 The best individual PCs are selected using GA-SVM model based on SG smoothed tobacco NIR spectra. $N_f$ is the number of inputs .... 68

5.3 The prediction accuracy of GA-SVM and SVM for different number of inputs, and the corresponding information amount based on testing set with SG smoothed tobacco NIR spectra. $N_f$ is the number of inputs; $P_a$ is prediction accuracy; and $I_m$ is the amount of information. .... 69

5.4 The identification capability of GA-SVM for each class are analyzed in terms of $\gamma_{TP}$, $\gamma_{TN}$, $\gamma_{PP}$, and $\gamma$ based on testing set with tobacco NIR spectra using SG smoothing method. The input number is 14; Class 1: the north of Guizhou Province in China; Class 2: the middle of Guizhou Province in China; Class 3: the northwest of Guizhou Province in China; Class 4: the southwest of Guizhou Province in China; $\gamma_{TP}$: true positive rate; $\gamma_{TN}$: specificity or true negative rate; $\gamma_{PP}$: positive predictive rate; and $\gamma$: F1 score. ............................... 71

A.1 The first 25 PCs of the tobacco NIR spectra, which are preprocessed by SG smoothing and PCA method. ................................. 93

A.2 The first 25 PCs of the raw tobacco NIR spectra, which are only preprocessed by PCA method. ................................. 95
List of Figures

1.1 The distribution of the tobacco planting areas in Guizhou Province of China. .................................................... 4

2.1 The specific rules for tobacco quality grading that is adopted in the cigarette industry. .................................................. 10

2.2 The range of wavelength and the divisions of the infrared light. . . . . 11

2.3 The general structure of an ANN. ........................................ 17

3.1 The flow diagram of the proposed classifier for tobacco growing area discrimination. .................................................. 25

3.2 The structure of the proposed three-layer ANN model for tobacco growing area discrimination. ........................................ 30

3.3 The raw and de-noised tobacco NIR spectra. (a) Raw tobacco NIR spectra; (b) tobacco NIR spectra de-noised using SG smoothing method; (c) tobacco NIR spectra de-noised using SG first-order derivative method; (d) tobacco NIR spectra de-noised using SG second-order derivative method. .................................................. 37

3.4 PCA of raw and de-noised tobacco NIR spectra. (a) PCA feature extraction of raw data; (b) PCA feature extraction of spectra de-noised using SG smoothing method; (c) PCA feature extraction of spectra de-noised using SG first-order derivative method; (d) PCA feature extraction of spectra de-noised using SG second-order derivative method. 38
3.5 The distribution of scores of the first principal component (PC) and second PC for four classes. (a) on the raw data; (b) on the spectra de-noised using SG smoothing method; (c) on the spectra de-noised using SG first-order derivative method; (d) on the spectra de-noised using SG second-order derivative method. 40

3.6 The selection for number of hidden nodes (HNs). (a) The discrimination accuracy for training and validation along with the change of HNs; (b) The mean absolute squared error along with the change of HNs. 45

3.7 The absolute squared error along with the increase of epochs, when HNs is 11 and number of inputs is 14. 46

4.1 The idea of SVM algorithm to transfer nonlinear separable problem into linear separable one. 50

4.2 The SVM algorithm deals with some instances that are still nonlinear separable in feature space. 52

5.1 The framework of the proposed GA-SVM classifier for tobacco growing areas. 64

5.2 The corresponding elements for genetic algorithm and the Darwinian principle of survival of the fittest. 65

5.3 The flow diagram of GA to implement the procedure of evolution. 66

5.4 The best fitness and average fitness during the training using GA-SVM classifier. 70

5.5 The best fitness along with different population size for GA-SVM classifier on training. 73

5.6 The best fitness along with different crossover rate for GA-SVM classifier on training. 74

5.7 The best fitness along with different mutation rate for GA-SVM classifier on training. 75
# List of Symbols

- \( a_0, a_1, a_2 \) Coefficients of quadratic polynomial
- \( b \) Bias
- \( C \) Penalty constant
- \( C_{m \times m} \) Covariance matrix of dataset with zero mean
- \( d \) Value of raw spectrum
- \( D_{n_1,n_2} \) MD for any two samples
- \( e \) Error of the output neuron
- \( E \) The error function
- \( f(\cdot) \) Activation function of hidden neurons
- \( g(\cdot) \) Activation function of output neurons
- \( g(x) \) Decision boundary
- \( h_0 \) Threshold of hidden neurons
- \( h_j \) Output of hidden neuron \( j, j = 1, 2, \cdots, m \)
- \( \mathbb{H} \) Feature space
- \( i,j \) The index of input and hidden neurons for ANN
- \( J \) Matrix for calculating coefficients of quadratic polynomial
- \( k \) Size of the moving window
- \( K(x_i, x_j) \) Inner product kernel function
- \( l \) Total number of training set
- \( m \) Dimension of individual raw sample
- \( n \) Total number of tobacco samples
- \( n_{TP} \) number of true positive predicted samples
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_{FN}$</td>
<td>number of false negative samples</td>
</tr>
<tr>
<td>$n_{FP}$</td>
<td>number of false positive samples</td>
</tr>
<tr>
<td>$n_{TN}$</td>
<td>number of true negative samples</td>
</tr>
<tr>
<td>$p(x)$</td>
<td>Regression value of the quadratic polynomial</td>
</tr>
<tr>
<td>$P_a$</td>
<td>Prediction accuracy of training or testing</td>
</tr>
<tr>
<td>$P_{(cva)}$</td>
<td>Prediction accuracy of validation</td>
</tr>
<tr>
<td>$Q$</td>
<td>Eigenvectors matrix</td>
</tr>
<tr>
<td>$\mathbb{R}$</td>
<td>Input space</td>
</tr>
<tr>
<td>$S$</td>
<td>Covariance matrix for MD method</td>
</tr>
<tr>
<td>$t$</td>
<td>Target output value for ANN</td>
</tr>
<tr>
<td>$v_j$</td>
<td>Weight from hidden neuron $j$ to the output neuron</td>
</tr>
<tr>
<td>$w$</td>
<td>Adjustable slope</td>
</tr>
<tr>
<td>$w_{ij}$</td>
<td>Weight from input $i$ to hidden neuron $i$</td>
</tr>
<tr>
<td>$|w|$</td>
<td>Second norm of $w$</td>
</tr>
<tr>
<td>$x$</td>
<td>A tobacco NIR sample</td>
</tr>
<tr>
<td>$x_i$</td>
<td>The $i$-th input of ANN, $i = 1, 2, \cdots, n$</td>
</tr>
<tr>
<td>$x_i^j$</td>
<td>The $i$th variable of sample $j$</td>
</tr>
<tr>
<td>$X$</td>
<td>whole tobacco NIR sample database</td>
</tr>
<tr>
<td>$X_i$</td>
<td>The $i$-th variable in the whole dataset</td>
</tr>
<tr>
<td>$X_i, X_j$</td>
<td>Sample vectors</td>
</tr>
<tr>
<td>$\bar{X}_i$</td>
<td>Zero mean dataset for PCA</td>
</tr>
<tr>
<td>$y$</td>
<td>Model prediction of the output for developed models</td>
</tr>
<tr>
<td>$y_i$</td>
<td>Target label for sample $x_i$, $i = 1, 2, 3, 4$</td>
</tr>
<tr>
<td>$Y$</td>
<td>Score of principal components</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Learning rate</td>
</tr>
<tr>
<td>$\alpha_0, \alpha_1, p$</td>
<td>Constants of polynomial function</td>
</tr>
<tr>
<td>$\beta_0, \beta_1$</td>
<td>Constant of Sigmoid function</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Distance away the hyperplane</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>F score or $F_1$-score</td>
</tr>
</tbody>
</table>
\( \gamma_{TP} \)  Sensitivity or true positive rate
\( \gamma_{TN} \)  Specificity or true negative rate
\( \gamma_{PP} \)  Precision or positive predictive rate
\( \sigma \)  Width of Gaussian function
\( \mu_i \)  The mean value variable \( i \) in MD model
\( \xi_i \)  Slack variable
\( \lambda, \mu \)  Positive constants
\( \lambda_1, \lambda_2, \ldots, \lambda_n \)  Deviation of principal components
\( \lambda_i \)  Eigenvalues, \( i = 1, 2, \ldots, 2084 \)
List of Abbreviations

ANN  Artificial Neural Network
ASTM American Society for Testing and Materials
BPNN Back Propagation Neural Network
FBP  Fast Back Propagation
FDR  False Discovery Rate
FP   False Positive
FN   False Negative
GA   Genetic Algorithm
GA-SVM Genetic Algorithm Optimized Support Vector Machine
HNs  Number of Hidden Neurons
IFS  Interactive Feature Selection algorithm
KNN  K-Nearest Neighbor
LDA  Linear Discriminant Analysis
MD   Mahalanobis Distance
NIR  Near Infrared
MLR  Multiple Linear Regression
MSC  Multiplicative Scatter Correction
PCA  Principal Component Analysis
PLSR Partial Least Squares Regression
PLS  Partial Least Squares
PPV  Positive Predictive Value
PSO  Particle Swarm Optimization
QDA  Quadratic Discriminant Analysis
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>QP</td>
<td>Quadratic Programming</td>
</tr>
<tr>
<td>RBF</td>
<td>Radial-Basis Function</td>
</tr>
<tr>
<td>SFS</td>
<td>Sequential floating Forward Selection Algorithm</td>
</tr>
<tr>
<td>SIMCA</td>
<td>Soft Independent Modeling of Class Analogy</td>
</tr>
<tr>
<td>SPC</td>
<td>Specificity or True Negative Rate</td>
</tr>
<tr>
<td>SG</td>
<td>Savitzky-Golay</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>SNV</td>
<td>Standard Normal Variate</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
<tr>
<td>TPR</td>
<td>True Positive Rate</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Tobacco is a valuable cash crop worldwide, with universal usage in the cigarette industry. The features of tobacco leaves, such as flavor, aroma, and physical characteristics, greatly influence the quality of cigarettes. High quality cigarettes are expected to provide the end consumers with a positive taste experience, optimum smoke yield, and less harm to health. The economic effect, like the price of cigarettes, is highly determined by the quality and fragrance of tobaccos. These elements are affected by many factors, particularly the climate and cultivation environmental effect, including rainfall, temperature, soil and geographical position (White et al., 1979; Karaivazoglou et al., 2007). In the tobacco industry, tobacco leaves are usually evaluated and distributed for use in accordance with growing areas since tobaccos from the same cultivation geographical area usually have the similar fragrance style (Pandeya et al., 1983). Cigarettes are ranked based on fragrance styles and a reference retail price can be set accordingly. Generally, some special growing areas can produce tobaccos, which are famous and used to produce top level cigarettes because of their fragrance styles. Therefore, an automatic evaluation approach, which is able to obtain reliable and accurate classification results of tobacco growing areas, is highly required for reasons of both economy and efficiency.
1.1 Problem Statement

In the tobacco industry, the growing area evaluation is generally conducted by experts according to their sensory inspections including individual sense of smell, taste and smoke quality. The manual evaluation is time-consuming and laborious. The evaluated results of growing areas highly depend on the experience of experts, and may vary with human emotion as well as the environment outside, such as the temperature and light of tobacco leaves assessment rooms. Due to the involvement of human beings, the tobacco assessment inevitably introduces unpredictable factors and the evaluation results may not be accurate and objective. Besides, it is expensive to train estimation experts for cigarette manufactures.

During the acquisition process, tobacco leaves from different planting areas may be blended together intentionally or unintentionally due to the economic reason. The work of inspecting planting areas needs to be repeated several times before tobacco leaves are actually used to produce cigarettes. However, current tobacco growing area evaluation cannot satisfy the requirement of reproducible evaluation process for the cigarette quality control and supervision.

Essentially, the fragrance of tobacco leaves is the external reflection on aroma, when internal chemical compositions are burnt and smoked (Wojtowicz et al., 2003). Although much has been known about the role of many chemical constituents playing in the tobacco quality. The routine chemical constituents, such as nicotine, nitrogen and sugars, have been recognized to be important to flue-cured tobacco quality. No single chemical quality index has been found to be applicable to classify grades, types, or blends of tobaccos (Gaines and Miles, 1975). Furthermore, there are more than 3000 chemical compounds reported in the literature (Gaines and Miles, 1975). It is almost impossible to analyze such a large number of chemical compounds using conventional chemical analysis methods.
1.2 Database

The proposed approaches are tested on a four-class tobacco near infrared (NIR) spectrum database. A total number of 332 tobacco samples were collected from four areas of Guizhou Province by Guizhou Tobacco Science Research Institute of China in 2008. The locations of Guizhou Province and the four areas are shown in Figure 1.1. Guizhou Province is in the southwest of China, and the four areas are in the north, middle, northwest and southwest of Guizhou Province, respectively. Because of the great ecological environment and the local climate condition, tobaccos cultivated in Guizhou Province are very popular in the cigarette industry. On the other hand, the complex climate makes growing area discrimination extremely hard.

<table>
<thead>
<tr>
<th>1, North</th>
<th>2, Middle</th>
<th>3, northwest</th>
<th>4, Southwest</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>144</td>
<td>70</td>
<td>80</td>
<td>332</td>
</tr>
</tbody>
</table>

The tobacco NIR spectra of the total 332 instances are recorded by Thermo Antaris 2 (Thermo Fisher Scientific Inc. Waltham, USA.) with the spectral resolution of 8 cm\(^{-1}\) and 64 scans in the NIR range of 12000 cm\(^{-1}\) and 3499 cm\(^{-1}\). In this study, three soft computing models and a statistic model are developed to discriminate the four tobacco growing areas. The specific geographical distribution of the four areas is marked in Figure 1.1.

The total dataset is randomly divided into two parts, training set and testing set, to the proportion of 80% and 20%, respectively. The specific constitution of datasets is depicted in Table 1.1. The division proportion of training set and testing set has been indicated to be the proper percentage with respect to the performance of discrimination models (Qin et al., 2008)
1.3 The Proposed Approaches

In this study, soft computing approaches are proposed aiming to developing an objective and effective tobacco cultivation area discriminator. The NIR spectroscopy, a fast, accurate, easy and non-destructive method, is used to conduct tobacco data acquisition. The Savitzky-Golay (SG) smoothing algorithm is employed in order to eliminate noise disturbance from the performance of classifiers. In addition, principal component analysis (PCA) technique is also applied to extract the features of tobacco NIR spectra and to compress the extremely high dimensional tobacco NIR spectra, so that the samples are more suitable to act as inputs for classification models. In order to demonstrate the effectiveness of the proposed classifiers, the model predictions are evaluated by prediction accuracy and the evaluation criteria derived from confusion matrix. Additionally, Mahalanobis distance (MD) discriminator (Ni et al., 2009), which was employed in tobacco growing area classification in the literature is developed as comparison study.

First of all, an artificial neural network (ANN) with fast learning algorithm is
developed to approximate the relationship between tobacco NIR spectra and growing areas. The important parameters, the number of hidden neurons (HNs) and learning rate, are discussed for the proper setting of ANN model.

Secondly, a support vector machine (SVM) model is developed to deal with the nonlinear separable classification problem. The selection of kernel function is discussed in order to improve the performance of SVM model. Other parameters, such as penalty factor and setting for the selected kernel function are also analyzed in details.

Finally, Genetic algorithm (GA) conjunct with SVM (GA-SVM) is applied to investigate the sensitivity of a input subset and to simplify classification models. The parameters of the proposed GA-SVM model are also discussed and analyzed.

1.4 Contribution

Through investigating the tobacco growing area discrimination problem, this thesis possesses following contributions.

(1) Based on NIR spectrum, three soft computing approaches are proposed to solve tobacco growing area discrimination problem automatically and non-destructively. The investigated tobacco NIR data have four classes and are extremely complex. The simulation results demonstrate the feasibility of applying the soft computing models for growing area classification.

(2) The proposed three-layer ANN model adopts the fast learning algorithm to adjust connection weights between adjacent layers, which not only exerts the great capability of nonlinear relationship approximation, but also accelerates the training process.

(3) An SVM classifier with Gaussian kernel is developed to avoid the drawbacks of ANN model, such as convergence speed and local minimum. The effectiveness and efficiency of SVM model is evidenced by its superiority in terms of prediction accuracy.

(4) GA is applied to investigate the interaction effect of features extracted by PCA
method in terms of prediction accuracy. The most sensitive input subset is selected by GA. The model conjunction GA with SVM possesses a better performance than other developed models.

(5) The preprocessing methods, SG de-noising and PCA feature extraction, are introduced and applied. Several de-noising methods are analyzed to select a proper one for tobacco NIR spectra investigation.

1.5 Organization of This Thesis

In this thesis, artificial systems are developed to model the connection or association between the tobacco NIR spectrum data and growing areas. The soft computing models, ANN, SVM and GA-SVM, are studied. The conventional statistical method is developed to comparing the classification performance with that of the soft computing approaches. The SG de-noising methods are employed to eliminate the noise disturbance from the performance of classifiers. The preprocessing methods are presented and discussed. This thesis is organized as follows.

Chapter 2 introduces previous work on tobacco growing area discrimination and the relevant works in the literature. Those approaches include conventional statistic methods and artificial approaches (ANN, SVM and GA-SVM). In addition, data preprocessing methods for NIR spectrum data and the performance assessment for classification models are reviewed and discussed.

Chapter 3 presents the proposed ANN model and discusses the structure and parameters for performance improvement of ANN models. The typical statistic method is also developed for tobacco growing area discrimination. In addition, the SG de-noising methods for the raw tobacco NIR spectra and PCA feature extraction technique are presented and discussed in details.

Chapter 4 describes the proposed SVM pattern classification model with the Gaussian inner product kernel function for the tobacco growing area discrimination application. The proper configurations for SVM model is analyzed and discussed.

Chapter 5 introduces GA for taking the influence of interaction among inputs
into account and for selecting the most sensitive input subset to model the relation
between the tobacco NIR spectrum data and growing areas. Relative configurations,
like population size, crossover rate, and mutation rate, are discussed for optimum
model performance.

In Chapter 6, the conclusions are drawn according to the results of the soft com-
puting models and the statistic model in previous chapters. Additionally, the further
work is recommended as an extension of the present study.
Chapter 2

Literature Review

Tobacco, as the main raw material of cigarettes, belongs to the Nicotiana genus of Nightshade family and mainly grows in warm places, like southern of China, United States, and Cuba. In the cigarette industry, the smoking experience of cigarette highly depends on the quality of tobacco leaves. Tobacco leaves are collected and sorted according to their geographical areas and positions on the whole stalks. Certain geographical areas produce high quality tobacco leaves, which can provide consumers with better taste and less harm to health, and can help cigarette manufactures gain economic benefits. The tobacco from different parts of the stalks has different utility. In the literature, the research about tobacco mainly focuses on growing area discrimination, grading and cigarette brand identification. This thesis focuses on the first aspect.

2.1 Tobacco Classification

For tobacco growing area classification, much research effort has been put into exploring the feasibility of using pattern recognition models. Hana and McClure (1997) employed a back propagation neural network (BPNN) to identify Burley tobacco as growing within the native USA or the outside. In their study, 19 points from tobacco NIR spectra are extracted as the input of the back propagation classifier. The reported results indicated that the BPNN had great performance with the identifica-
tion ratio 89%. This study demonstrated that BPNN was feasible to utilize pattern recognition methods in conjunction with NIR spectrum technique to identify tobacco growing area speedily and nondestructively. However, it did not clearly state how the 19 points were selected from tobacco NIR spectra. Moreover, the tobacco from native and outside of USA may have large difference due to the distinct geometrical distribution of growing area, which reduced the difficulty of this problem. Ni et al. (2009) discriminated Chinese flue-cured tobacco by a Mahalanobia distance (MD) method, which calculated the centroid of classes using training dataset and then identified tobacco growing areas by the shortest Mahalanobis distance to the class centroid. However, only the small range of spectra, of which the absorbency were less than 0.8, were interested for classification model establishment, but there is no guarantee that the selected range does positively contribute to classification. In addition, the method adopted to avoid the overfitting was based on the errors including instrument noise, operation error and instrument precision. The inaccurate estimation of those uncertainties would certainly leads to the performance degradation of this method. Additionally, the discriminated growing areas were based on province, most of which are far away from each others.

For tobacco quality grading, physical features identification, such as the width, length and colours of tobacco leaves, is investigated. Figure 2.1 shows the specific rule adopted by the cigarette industry. Existing automatic grading systems took advantage of machine vision algorithm to extract those physical features from tobacco leaves (Zhang et al., 1997; Yawootti and Kaewtrakulpong, 2005; Zhang and Zhang, 2008).

For cigarette brand discrimination, most of research work combined NIR technique with pattern recognition approaches. Moreira et al. (2009) applied a linear discriminate analysis (LDA) model on the NIR spectra to assess cigarette authenticity. The reported results indicated the model was a promising method whose accuracy was up to 100% for cigarette brands identification. Zhang et al. (2008) employed multi-class SVM and K-nearest neighbor (KNN) algorithm on NIR spectra data for cigarette brand discrimination. They claimed that the proposed approach was low
Figure 2.1: The specific rules for tobacco quality grading that is adopted in the cigarette industry.

cost, time-saving and accurate. Tan et al. (2007) distinguished cigarettes of different brands using an adaboost algorithm and near-infrared spectroscopy. Multivariate chemometric discrimination was also adopted in (Giokas et al., 2011).

2.2 Near Infrared Spectroscopy and Preprocessing

Near infrared (NIR) light is invisible light, and locates between middle infrared and visible light. According to the American Society for Testing and Materials (ASTM), the wavelength of NIR light ranges from around 800nm to 2500nm as shown in Figure 2.2. NIR spectroscopy is a measurement which utilizes the NIR region of the electromagnetic spectrum to analyze the chemical components of tested samples. It supports several options to measure different types of inter atomic bond when they are vibrating. Different molecule has different atomic bond and each of them has their unique spectrum. We can identify different substance in them, such as N-H, O-H and C-H of organic molecules. NIR technique possesses several merits, including that (1) it does not require pretreatment for samples; (2) it is a non-destructive measurement, and the experiment samples won’t be damaged during NIR data acquisition and can be reused, and (3) it can go through a container, which allows samples to be analyzed
in solid or liquid forms. Moreover, NIR technique is fast and costs low. Due to these outstanding advantages, NIR spectrophotometry has become a powerful tool to do qualitative and/or quantitative analysis in various applications.

In food safety and quality control, NIR spectroscopy is applied for food authentication and quality assessment. Existing work includes classifying the pure and the adulterated honey samples using NIR spectroscopy (Toher et al., 2007); assessing egg quality and freshness (Abdel-Nour, 2008); verifying adulteration of alcoholic beverages (whiskey, brandy, rum and vodka) (Pontesa et al., 2006); predicting oil content in instant noodles (Chen et al., 2002); evaluating pectin constituents of Japanese pear (Sirisomboon et al., 2007); identifying fishmeal batches (Cozzolino et al., 2005); investigating quantification of crystalline lactose content in whey permeate powder (Norgaard et al., 2005), and predicting somatic cell count of un-homogenized milk (Pravdov et al., 2001). Cen and He (2007) summarized tens of applications for NIR in aquatic and stock products, dairy products, agricultural products (rice, wheat, soybean, tea and so on), fruits (peach and apple) and vegetables.

Some research work focus on application of disease diagnosis using information
on the chemical composition of tissue provided by NIR, such as detecting human colorectal cancer (Kondepati et al., 2007), aortic atheroscleroses (Moreno et al., 2002), and imaging tumour vascular in living subjects (Cai et al., 2006).

Agricultural goods analysis uses NIR for qualitative and/or quantitative analysis, which covers detecting weed-free and weed-infested areas of a soybean field (Chang et al., 2004), monitoring fruit properties such as maturity, sensory properties and storability non-destructively both prior to harvest and during storage (Kim et al., 2000).

Soil content analysis is another application for NIR. Vasques et al. (2008) utilized NIR technique to estimate soil carbon across large landscapes; Zornoza et al. (2008) predicted various physical, chemical and biochemical properties in Mediterranean soils. (Maleki et al., 2006) predicted the phosphorus content in fresh soil as a preliminary step toward the development of an automated variable rate fertilizer distributor. W. Yoon et al. (2004) applied NIR to distinguish between the manufacture sites of a number of proprietary tablets.

Other applications include pharmaceutical production (Beer et al., 2011; Reich, 2005), gasoline petroleum industry (Balabin and Safieva, 2007; Balabin and Safieva, 2008; Kim et al., 2000). All these research work based on NIR reported promising results.

As a spectral analysis technique, NIR spectra owns great property to combine well with the traditional chemometrics models, like multiple linear regression (MLR) (Vasques et al., 2008), partial least squares regression (PLSR) (Vasques et al., 2008; Abdel-Nour, 2008; Cozzolino et al., 2005), LDA (Cozzolino et al., 2005) and PCA (Cozzolino et al., 2005; Kim et al., 2000). It also works with nonlinear models very well, such as SVM (Chandra et al., 2009), ANN, and KNN, to solve the complex problems. Conjunction with nonlinear models, NIR can be easily applied to predict and assess properties of testing samples.

NIR technique, because of its advantages of fast speed, non-destruction and no sample preparation, has successfully been applied for quantitative and qualitative analysis. However, NIR data acquisition inevitably introduces unexpected noise that
will contaminate the NIR spectra and disturb the characteristic of the subjectives. NIR spectra are the signals responded to NIR capture apparatus. And it is human to operate the process of NIR data acquisition. Therefore, noise from NIR spectrum instrument and human operation has to be considered (Ni et al., 2009). Additionally, complex backgrounds and baselines of NIR spectra make non-concentration-correlated contributions to the NIR spectra data. The influence from the noise, backgrounds and baselines would (1) degrade the signal-to-noise ratio (SNR), (2) introduce pseudo-biases into calibration model, (3) make model performs poor, and (4) lead to unnecessarily complicated analysis models (Xu et al., 2008). Therefore, de-noising raw NIR spectra is indispensably in the model establishment, especially for pattern recognition.

An appropriate NIR de-noising method is important. The selection of proper processing method should depend on the specific NIR data and there is no guidance for it, only by trail. There are many preprocessing techniques applied for NIR spectra to remove the interference, like multiplicative scatter correction (MSC) (Shao et al., 2007; Stordrange et al., 2002; Thennadil and Martin, 2005), standard normal variate (SNV) (Song et al., 2010; Stordrange et al., 2002), wavelet methods (Shao et al., 2007), and so on, where SG smoothing/derivative methods are the most frequently used approaches for eliminating noise and improving NIR determinations.

SG smoothing/derivative methods was firstly proposed by Savitzky and Golay (1964) and has been extensively used for NIR data de-noising. A first-order SG derivative method was employed for flue-cured tobacco NIR spectra to help improve the prediction performance of a simplified K-nearest neighbours classifier (Ni et al., 2009). A second-order SG filter was stated to be better in both of simulated NIR dataset and two measured NIR spectral datasets (water samples and plasma samples) (Xie et al., 2009). Xu et al. (2008) designed an ensemble SG preprocessing method to improve the performance of partial least squares (PLS) in wheat kernels application. SG smoothing/derivative methods work like a low-pass filter and can remove all high frequency noise. They are functionally well for these objectives with little or no priori knowledge, unlike other preprocessing methods.
2.3 Methods for Classification

The performance of a classifier highly depends on the characteristics of the data to be classified. There is no single classifier that works best on all given problems. Various empirical tests have been conducted to compare the performance of classifiers, and to find the characteristics of data. Determining a suitable classifier for a given problem is however still more like an art than a science. There are a bunch of approaches for various classification problems. Comprehensively, there are linear classifiers, like logistic regression, naive Bayes classifier, Fisher’s linear discriminant, and nonlinear models, such as ANN, SVM, and some statistic methods.

2.3.1 Statistic Models

In the literature, multi-variable input is often involved in pattern recognition problems. The statistic models, which function to measure the distance of samples, can deal with pattern recognition problems well. For instance, the KNN algorithm is to measure the Euclidean distance between all stored vectors and the test samples, and to assign the class of the neighbours with the minimum average distance. Mahalanobis distance (MD), proposed by Mahalanobis (1936), is another kind of distance to measure the distance of variables and their relationships. MD focuses on the correlations between variables or features by calculating the inverse of variance-covariance matrix, and measures the similarity of variables of interested data by linking similar population together.

In the past decades, MD has been widely used in supervised pattern recognition methods (Maesschalck et al., 2000). Some researchers directly utilized MD to identify different patterns. DeVries (2005) and Etherington et al. (2009) applied MD to model the distribution of vegetation communities and distribution of species respectively. Xiang et al. (2008) investigated information pairs and classify them into must-link and cannot-link. They also introduced the application of interactive natural image segmentation and face estimation to demonstrate the effectiveness of MD. In addition, MD is successfully applied in distribution prediction of marine organisms (Cameron
et al., 2011), fault identification of electronic product (Kumar et al., 2010), habitat selection studies for the estimation of environmental suitability maps (Calenge et al., 2008), agents control and non-target feeding habitats prediction (Hodges, 2009). There are a large number of research to build global discrimination model based on MD, such as linear discrimination analysis (LDS) model, nonlinear models soft independent modelling of class analogy (SIMCA), quadratic discriminant analysis (QDA), as surveyed by Maesschalck et al. (2000).

There are also a lot research on applying MD and the models built upon MD to deal with NIR spectra in classification problems. Generally, MD works with principal component of NIR data and functions to discriminate the authenticity of cigarettes (Moreira et al., 2009; Tan et al., 2009), to recognize brown core of pears for safely storage (Han et al., 2006), to quantitatively monitoring polymorph conversion in medicine field (Li et al., 2005), and so on.

Even though MD has been applied to pattern recognition in various fields, its problems are quite obvious. Because it takes correlation of the interested data into consideration, inverse of the variance-covariance matrix of data has to be computed. However, the invested data with multi-variable from the real world, especially NIR spectra, contain redundant or irrelevant information. The variance-covariance matrix is easily to be singular and has no inverse matrix, which leads to the failure of applied MD. Therefore, the data which attempt to employ MD, have to be pretreated before computing the MD. Secondly, as a supervised statistic method, the centre of each class has to be calculated from training set. The noise and the non-typical training set may lead to the centre far away from the real centre, thereby predicting the unknown data incorrectly. Furthermore, like other statistic methods, only sufficient number of samples can make MD have good performance. MD is not appropriate for applications with only small sample set.

2.3.2 Artificial Neural Networks

ANN is a powerful tool for nonlinear classification problems. It is inspired by the biological neuron networks, which are identified as groups of neurons that adaptively
perform specific functions. ANN focuses on abstracting this complexity from problems and takes care what may hypothetically matter most from an information processing point of view (Haykin, 1999). ANN, also commonly called neural network (NN), is composed of multilayer artificial neurons. The neurons on different layers are connected with each other and there is no connection between those on the same layer, as shown in Figure 2.3. Multi-layer ANN is famous for its strong ability to approximate nonlinear relationships by training neurons via a supervised manner. It has many hidden layers according to the specific problem, except for the input layer and output layer. A four layers ANN with two hidden layers is shown in Figure 2.3. Generally, a three-layer ANN, which includes an input layer, a hidden layer with nonlinear activation function, and output layer, are most commonly used for complex problems. There are two passes in ANN. The first one is forward pass. The input features or patterns are applied to neurons of the input layer. The effect propagates through the neuron connection between the layers and finally activates the output layer. After input feature forward passing, error back propagation is conducted for adjusting the connection weights between layers by a supervised learning algorithm with the error-correction rule. The actual output is as close as desired output after enough training iterations. Therefore, ANN possesses high self-learning ability and is capable to predict unknown relationship with high accuracy. Theoretically, a three layers ANN can approximate any nonlinear function at arbitrary accuracy, provided sufficient training time and training data. Moreover, there is no need for ANN to know much priori knowledge for complex problems themselves.

A fast learning algorithm was proposed by Karayiannis and Venetsanopoulos (1993) to accelerate network convergence speed, which has been proved that it can work much faster than the existing learning algorithms (Karayiannis and Venetsanopoulos, 1992).

In past decades, ANN has been widely applied for pattern recognition to investigate complex relationship and to predict unknown data. For fault diagnosis, Upendar et al. (2010) proposed ANN model to predict the fault in electric power system; Hu et al. (2007) applied BPNN on fault diagnosis for analog circuits; and Liu et al. (2010)
extended ANN to defects detection in embedded reconfigurable system. ANN worked very well in this field and exhibited promising performance.

In recent decades, ANN is popular to be employed in expert systems in wide applications, particularly, in medical diagnoses. Lin (2008) developed an artificial system based on ANN with three layers, 17 input feature and 5 outputs. It was used to identify the classification patterns of the scoliosis spinal deformity to help the preoperative surgical planning. Mazurowski et al. (2008) investigated the effect of class imbalance in training data for computer-aided medical diagnosis and took real clinical data of breast cancer diagnosis for validation. The experimental results showed that ANN was generally preferable over particle swarm optimization (PSO)
for imbalanced training data, especially with small data sample and large number of features. An ANN model was developed to predict survival/death and growth/no-growth interfaces for Escherichia coli O157:H7 in a mayonnaise-type system (Yu et al., 2006). The model correctly predicted the growth/no-growth in 1810 combinations (99.5%) with 8 false positives and 2 false negatives, and survival/death in 1804 combinations (99.1%) with 13 false positives and 3 false negatives.

A lot of effort have been put into investigating factor analysis and food identification in agricultural field. For instance, the single-component model using ANN (Janes et al., 2004) was applied to model the pork farm odour yields. The authors claimed more accurate and precise odour intensity predictions using ANN than the statistical models. Janes et al. (2004) applied ANN for multiple-component multiple-factor analysis in the complex pork farm odour analysis problem. The multiple-component neural network model provided better performance than the corresponding linear multiple regression model. Zhang et al. (2002) proposed an ANN (a four-layered back-propagation neural network, with 8 and 5 neurons in the first and the second hidden layers, respectively) model for rough rice drying to predict six performance indices: energy consumption, kernel cracking, final moisture content, moisture removal rate, drying intensity and water mass removal rate. Chegini et al. (2008) designed a four-layer BPNN (with 14 and 10 neurons in the first and the second hidden layers, respectively) to investigate the effect of feed flow rate in an orange juice semi-industrial spray dryer, with respect to seven performance indices, namely: residual moisture content of orange juice powder, particles size, bulk density, average time of wet ability, insoluble solids, outlet air temperature and dryer yield. The ANN technology has been shown to be an useful tool to investigate, approximate and predict the physical properties of orange juice powder as well as process parameters of spray dryers.

There are also other applications, which utilize the great advantage of ANN to perform prediction and classification. Lv et al. (2010) constructed a function of intersection for traffic conditions and accidents to provide basis for intersection adaption reconstruction. The reported prediction accuracy was up to 89%. Jiang et al. (2010) adopted a BPNN to classify remote sensing image. Researchers applied ANN to
analyze EMG signal patterns (Chong and Sundaraj, 2009), to track pulverized fuel
during combustion (Xu et al., 2005), to identify vehicle types (Ki and Baik, 2006),
to forecast bankruptcy and credit scoring (Tsai and Wu, 2008), and to predict flood
discharge (Chen et al., 2006).

2.3.3 Support Vector Machine

SVM, originally proposed by Vapnik (1995), is known as a powerful technique for
pattern recognition, classification, regression, and so on. It is particularly designed for
the high dimension and nonlinear separable problems. SVM model performs pattern
recognition through constructing an optimal separating hyperplane (OSH) (for two
classes problems) or a set of OSHs (for multi-classes problems) in high-dimension
space. The OSHs can make the largest distance between classes. The larger the
margin is, the smaller the structural error of the classifier is.

SVM, as a supervised pattern recognition method, has well developed theoretical
foundation (statistical learning theory), and widespread practical applications. It is
appealing more and more attention due to its outstanding advantages: (1) suitable
for small samples set; (2) reasonable mathematic support; (3) good performance as a
nonlinear classifier and effectively avoiding dimensional disaster.

More and more different applications in the real world have been explored by SVM
(Abhisek, 2007), such as image processing (like object, facial expression and hand-
written character recognition), speech processing, data mining (like data classification,
text classification and noise classification and reduction), power systems (fault
classification and load forecasting) and bioinformatics (protein structure prediction,
breast cancer prognosis). In the bioinformatics and medical field, SVM model was
developed to classify older adults from younger adults based on their resting-state
functional connectivity and to characterize reorganization of functional brain net-
works with respect to age (Meier et al., 2012). Chandra et al. (2009) proposed an
SVM model with Gaussian RBF kernel to automatically detect brain tumour from
MRI images. The SVM gave higher precision and lower error rates than the Adaboost
classifier. Yu et al. (2010) presented an SVM method to be an useful alternative to
predict diabetes. Zuba et al. (2012) utilized SVM to effectively identify risk and/or protective factors for high-risk drinking for helping detect and address alcohol misuse in college students. Recently, some researches are investigated to explore the potential of SVM in financial problems. It can be used to suggest better explanatory power, robustness and stability, like bankruptcy prediction (Shin et al., 2005; Min and Lee, 2005), financial distress of Chinese listed companies forecast (Ding et al., 2008) and credit rating analysis (Huang et al., 2004). SVM model has also been applied to biochemical field. For instance, it is used to explore protein interactions (Guo et al., 2008), and to analyze the contribution of chemical factors for classification (Krooshof et al., 2010). Other problems, like classifying remote sensing images (Bazi and Melgani, 2006) and monitoring fault diagnosis (Widodo and Yang, 2007), were investigated and reported promising performance.

2.4 Genetic Algorithms

The GA is a famous heuristic search method, which was proposed by Holland (1975). It is inspired by evolutionary processes of nature and is based on the Darwinian principle of survival of the fittest. GA mimics the whole natural evolution, encodes problems to be solved and conducts fitness of environment evaluation, gene selection, crossover, mutation and inheritance. Finally, the best fit individual is kept after competition. The evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated. Multiple individuals are stochastically selected from the current population (based on their fitness), and modified (recombined and possibly randomly mutated) to form a new population. The new population is then used in the next iteration. Commonly, the iteration is terminated when either a maximum number of generations has been reached, or a satisfactory fitness level has been produced for the population.

Unlike other optimization approaches, GA can efficiently search in large space and perform well for those problems with little priori knowledge (Pudil and Hovovi-
The only link between GA and the problems to be solved is the fitness function, which acts as environment selection function in natural evolution, that is, the predefined fitness function evaluates how fit the solution of GA for the real problems. Furthermore, GA can function well without the mathematical assumptions, which is required in statistical models, such as continuity and differentiability (Chambers, 2000). GA, as a derivative-free stochastic optimization method, can find the global optimal solution but never be trapped by local minimum solution, and reach convergence faster than exhaustive methods. Therefore, GA always draws much more attention to various applications, especially to optimal feature subset selection or feature selection for building robust models.

Many researches are investigated to discover the essential characteristics of complex problems, which are known little by human beings, and to provide better understanding about the data by showing the important features of them. In medical diagnosis and disease detection, bunches of factors, which may be relevant with disease forecast and diagnosis, are involved. Identifying factors, which contribute positive to the disease prediction and diagnosis, would be rather important. Zhang et al. (2005) employed GA for feature selection in conjunction with neural network and statistical classifiers to detect breast cancer based on digital mammography. The appropriate feature subset, suspicious areas containing benign and malignant microcalcifications in digital mammography, was found out by GA. The classifier achieved higher accuracy of prediction. For the Salmonella Typhimurium infection problem, Qin et al. (2008) employed GA for factor subset extraction, and combined with neural and statistical classifiers for Salmonella Typhimurium infection. Finally, 18 risk factors were selected from the total of 91 factors as the inputs of classifiers, which greatly decreases the complexity of classifiers and provides better understanding for Salmonella Typhimurium infection. Alexandre et al. (2007) applied GA to reduce the number of signal-describing features, which feed the automatic classifier. A subset of signal-describing features with lower cardinality was selected to enhance listening comprehension, when the user went from one sound environment to another. Nuclear transients classification also employs GA to identify the most important factors (Zio
The GA also is combined with pattern recognition models to select the most sensitive feature subset to improve the performance of classifiers for various applications. The research (Kim et al., 2006) investigated the three methods (traditional sequential floating forward selection algorithm (SFS), interactive feature selection algorithm (IFS) and GA) to conduct feature selection for emotion recognition, and reported that GA had superior performance than IFS, and similar with SFS in terms of best fitness and average fitness. Moreover, GA for feature selection is used in channel estimation (Ali et al., 2007) and lung nodule diagnosis (Boroczky et al., 2005).
Chapter 3

The Proposed Neural Network Classifier

In this chapter, the tobacco NIR spectra are preprocessed using SG algorithm and PCA for de-noising and feature extraction, respectively. An ANN model is designed for discriminating tobacco growing areas. A statistical model, MD model, is also developed for this problem as comparison, in order to demonstrate the effectiveness of the proposed approach. In addition, the ANN setting is discussed in details for the performance improvement.

3.1 Introduction

The characteristics of tobaccos are particularly influenced by the planting environment and climate. For instance, the soil acidity and nitrogen affect the growth, yield, quality characteristics and chemical constitutions in tobaccos (Karaivazoglou et al., 2007), including total sugar, reducing sugar, nicotine, total nitrogen, protein, and so on. Due to the huge number (>3000) of chemical constituents, it is extremely difficult for existing research to explore the specific influence of individual chemical compounds by quantitative analysis. Thus, how to identify tobacco growing areas is a challenging problem.
3.2 The Proposed Methods for Growing Areas Discrimination

In this study, ANN method is proposed aiming to developing an objective and effective tobacco cultivation area discriminator based on NIR spectrum. The framework of the proposed method is shown in Figure 3.1. Firstly, the SG de-noising method is employed to eliminate noise disturbance from the performance of the classifier. Then, PCA technique is applied to extract features from tobacco NIR spectra as well as to compress high dimensional NIR spectra so that the data are suitable for the ANN model as inputs. In this way, the original high dimensional noisy NIR spectra is transformed into clean low-dimensional PCs, which are much more convenient for building prediction models. Then, the whole tobacco NIR dataset is divided into training set and testing set randomly with proportion of 80% and 20%, respectively. In addition, the structure of the proposed ANN model and de-noising methods are also discussed with respect to discriminating accuracy. In order to demonstrate the effectiveness of the proposed ANN classifier, MD discriminant method (Ni et al., 2009), which was employed in tobacco growing area classification in the literature, are developed for comparison. The developed MD model also follows the same procedure with the ANN model. In order to keep the comparison to be fair, the data exact same with the ANN model are used for MD modeling.

3.2.1 De-noising: Savitzky-Golay Algorithm

Because NIR spectrum data are recorded through spectrum instruments, which are operated by human, it is inevitably contaminated by disturbances, including background noise, instruments error and operation error. The presence of noise will introduce distortion into the data and lead to the hidden of data characteristics. Generally, it is better to pre-process spectra before applied to establish the classification models. For spectroscopic analysis, the SG smoothing and differentiation methods are often used to remove the background noise, to eliminate the spectra differences from
Figure 3.1: The flow diagram of the proposed classifier for tobacco growing area discrimination.
baseline drifts, and to improve the SNR.

The SG method is a polynomial regression algorithm with a local sliding window. It functions de-noising by replacing the original contaminated value with the averaged value, which is determined by the local polynomial regression. The raw spectra distorted by various noise are replaced by the value of the local polynomial regression, which makes the de-noised data are more close to the true signals. Essentially, the SG method approximates the true signals of the raw spectra in the sense of least squares principle. In this study, a quadratic polynomial is adopted, which is represented as

\[ p(x) = a_0 + a_1 x + a_2 x^2, \]  
(3.1)

where \( a_0, a_1 \) and \( a_2 \) are the coefficients of the quadratic polynomial, and are determined by raw spectrum value \( x \) and spectrum wavelength \( d \) in this case; and \( p \) is the regression value of quadratic polynomial (3.1). The size of the sliding window is required to be much larger than the order of the polynomial, and it must be odd (in this case, \( k=121 \), and \( k \gg 2 \)). The coefficients \( a_0, a_1 \) and \( a_2 \) can be obtained by

\[ a = (J^T J)^{-1} d, \]  
(3.2)

where \( a = \{a_0, a_1, a_2\} \); the \( i \)-th row of matrix \( J = \{1, x_i, x_i^2\} \), and \( i = \{1, 2, 3, \ldots 121\} \); and \( d \) is the value of the raw spectrum. Because the rank of Equation (3.2) is 3 and there are 121 equations, the coefficients can be determined in the least-square sense.

For the SG smoothing, the approximated value can be calculated by Equation (3.1), and coefficients \( a = \{a_0, a_1, a_2\} \) are obtained by Equation (3.2). For the SG differentiation at a given point, the first order is the first derivative of the polynomial (3.1) at point \( x = 0 \), i.e., \( a_1 \) in the moving window. And the second order is the second derivative of the polynomial (3.1) at point \( x = 0 \), which results to \( a_2 \). This is because of \( \frac{dp}{dx}|_{x=0} = a_1 \) and \( \frac{d^2p}{dx^2}|_{x=0} = a_2 \) in the local window.

In this study, three de-noising methods, the SG smoothing, first-order and second-order differentiation, are applied respectively in order to find out the appropriate de-noising approach for tobacco NIR spectra, thereby detecting the characteristics of growing areas.
3.2.2 Feature Extraction Method: Principal Component Analysis

PCA method is one of the most popular approach to extract features from high dimensional data in analytical chemistry and spectroscopic analysis. It can extract patterns from high dimensional data and express them by highlighting the similarities and the differences. In this study, PCA method is applied to extract the features from the raw tobacco NIR spectra and from the spectra de-noised using three SG methods, so that the dimensionality of datasets for modelling is reduced. In order to keep as much information as possible, the full spectra range of the raw data from 12000 cm\(^{-1}\) to 3961 cm\(^{-1}\), is adopted for model establishment. They are denoted by 2084 data points as \(x = [x_1 \ x_2 \ ... \ x_m]^T\). The whole tobacco dataset is denoted by an \(m \times n\) array \(X\) as

\[
X = \begin{bmatrix}
x_1^1 & x_1^2 & \cdots & x_1^n \\
x_2^1 & x_2^2 & \cdots & x_2^n \\
\vdots & \vdots & \ddots & \vdots \\
x_m^1 & x_m^2 & \cdots & x_m^n 
\end{bmatrix}, \tag{3.3}
\]

where \(m\) is the dimensionality of the individual raw tobacco spectrum, \(m = 2084\); and \(n\) is the total number of tobacco data, \(n = 332\). The columns of dataset array \(X\) are sample vectors; and the rows are the variables, e.g., the \(i\)-th row is the \(i\)-th variable in the whole dataset, denoted by

\[
X_i = [x_i^1 \ x_i^2 \ ... \ x_i^n]. \tag{3.4}
\]

In order to make the dataset to be suitable to apply PCA, the zero mean dataset are constructed by subtracting the mean from each variable using

\[
\bar{X}_i = X_i - \mu_i = [\bar{x}_i^1 \ \bar{x}_i^2 \ ... \ \bar{x}_i^n], \tag{3.5}
\]

where \(\mu_i\) is the mean of the \(i\)-th variable \(x_i\), and is obtained by

\[
\mu_i = \frac{1}{n} \sum_{j=1}^{n} x_i^j, \tag{3.6}
\]
where \( x_i^j \) is the \( i \)-th variable of the \( j \)-th sample.

The covariance matrix \( C_{m \times m} \) of the dataset with zero mean is

\[
C_{m \times m} = \begin{bmatrix}
\text{cov}(\bar{X}_1, \bar{X}_1) & \text{cov}(\bar{X}_1, \bar{X}_2) & \cdots & \text{cov}(\bar{X}_1, \bar{X}_m) \\
\text{cov}(\bar{X}_2, \bar{X}_1) & \text{cov}(\bar{X}_2, \bar{X}_2) & \cdots & \text{cov}(\bar{X}_2, \bar{X}_m) \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}(\bar{X}_m, \bar{X}_1) & \text{cov}(\bar{X}_m, \bar{X}_2) & \cdots & \text{cov}(\bar{X}_m, \bar{X}_m)
\end{bmatrix},
\tag{3.7}
\]

where the covariance of two variables is calculated by

\[
\text{cov}(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1},
\tag{3.8}
\]

where the covariance matrix \( C_{m \times m} \) is squared and symmetric. Therefore, it is easy to find out the eigenvalues \( \lambda_i \) and eigenvectors matrix \( Q \), which express useful information about dataset and eliminate the redundant information. The PC in eigenvector with the larger value is more important and contains more information or energy of original dataset. The PCs can be calculated as

\[
Y = Q\bar{X}.
\tag{3.9}
\]

The values of PCs are the projection of centralized dataset to the eigenvectors. The eigenvector determines the corresponding PC of the centralized dataset. After this projection, the information about the centralized dataset is concentrated on PCs in the front part of PCA results, and the noise is evenly spread to all eigenvectors’ space. Therefore, PCA method achieves redundant information elimination and dimensionality reduction. The importance of PCs is decided by the eigenvalues of the corresponding eigenvectors. The larger the eigenvalue is, the more important the corresponding PC is than others with small eigenvalues. The less important PCs can be ignored without much influence on applications. Number of PCs should be carefully selected, since using too many PCs for modeling may bring on too much noise, while insufficient number of PCs cannot provide enough spectra information for classification. In this study, different number of PCs is used to establish classification models. The suitable number of PCs is selected based on model prediction accuracy.
3.2.3 A Neural Network Classifier for Tobacco Growing Area Discrimination

An ANN model with the fast learning algorithm is developed in growing area discrimination. The schematic diagram of the proposed method is shown in Figure 3.2. A three-layer structure is adopted, where there are \( n \) input nodes, \( m \) hidden nodes and one output node. The nodes are fully connected with the nodes in the adjacent layers, and there is no connection between nodes within the same layer. All the connection weights and thresholds are initialized in the range of \([-1, 1]\). A hyperbolic tangent sigmoid function is used as the activation function for the hidden layer. For the output layer, a linear scalar function is applied as the activation function. The input node \( i \) of the ANN model is injected into the \( n \)-th PC extracted from tobacco NIR spectra by the PCA technique, and the output corresponds to the tobacco growing area. In this study, four geographical areas are investigated. Theoretically, a three-layer neural network with nonlinear activation function in the hidden layer could approximate any nonlinear function at arbitrary accuracy, provided sufficient training time and training data.

An ANN model consist of two phases, the input signal forward transmission and the error backward propagation. For the input signal forward transmission, the output of the neuron \( j \) in the hidden layer is given by

\[
    h_j = f \left( \sum_{i=0}^{n} w_{ij} x_i \right),
\]

(3.10)

where \( x_i \) is the \( i \)-th input node, and \( x_0 \equiv 1 \); \( n \) is the number of input nodes; \( w_{ij} \) is the connection weight between the input node \( i \) and the hidden node \( j \); and \( f(\cdot) \) is the activation function of the hidden layer, which is required to be a nonlinear function when approximating nonlinear input-output relation. In this work, it is defined to be the tangent function as

\[
    f(a) = \tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}.
\]

(3.11)

To ensure that each input variable provide equal contribution to ANN model, the inputs (PCs) are normalized into the range of \([-1, 1]\). A linear scalar function
Figure 3.2: The structure of the proposed three-layer ANN model for tobacco growing area discrimination, where \(x_i\) is the \(i\)-th input of ANN model; \(h_j\) is the output of the hidden layer; \(w_{ij}\) is the weight between the \(i\)-th input neuron in the input layer and the \(j\)-th neuron in the hidden layer; \(v_j\) is the weight between the \(j\)-th neuron in the hidden layer and the output neuron; \(n\) is the number of the input neurons; \(m\) is the number of the hidden neuron; and \(y\) is the prediction growing area.
$g(a) = a$ is applied to be the activation function of the output layer. The output of the output layer is the sum of products of each hidden layer output with corresponding connection weight, and is given by

$$y = \sum_{j=0}^{m} v_j h_j,$$

(3.12)

where $h_j$ is the output of hidden node $j$, and $h_0 \equiv 1$; $m$ is the number of the hidden nodes; and $v_j$ is the connection weight between hidden node $h_j$ and the output. Normally, all the connection weights and thresholds are randomly initialized in the range of $[-1, 1]$. The output $y$ indicates the tobacco growing area.

For the error backward propagation part, ANN learns from training data and tries to approximate the relationship between inputs and outputs by minimizing error function. Squared error and learning rate are utilized to update weights and thresholds in the sense of least squared error. For the convenience of derivation of the learning algorithm to adjust connection weights and thresholds, the error function is defined as

$$E = \frac{1}{2} e^2 = \frac{1}{2}(t - y)^2,$$

(3.13)

where $e$ is the error of target value and prediction output; $t$ is the target value; and $y$ is the predicted output of the ANN.

The fast learning algorithm, proposed by Karayiannis and Venetsanopoulos (1993), is adopted to minimize the squared error function (3.13), and to adjusts the connection weights and thresholds with noticeably fast speed. For any specific sample, the updates of weights and thresholds are shown as

$$v_j = v_j + \alpha \varepsilon^o h_j, \quad j = 0, 1, \cdots, m,$$

(3.14)

$$w_{ij} = w_{ij} + \alpha \varepsilon^h x_i, \quad i = 0, 1, \cdots, n,$$

(3.15)

where

$$\varepsilon^o = \lambda e + (1 - \lambda) \tanh(\beta e),$$

(3.16)

$$\varepsilon^h_j = (1 - h^2_j) \varepsilon_o v_j,$$

(3.17)

$$\beta = e^{-\mu/E^2},$$

(3.18)
where $\alpha$ is the learning rate, which is an important parameter for ANN model and can determine the learning efficiency in one epoch or iteration. If $\alpha$ is too large, the network will oscillate and is hard to converge. Conversely, if it is too small, a long time will be taken for ANN model to learn and reach the convergence. Currently, there is no normal way to choose the value of $\alpha$, since it depends on the specific applications. Generally, $\alpha$ is a small positive value within the range of $[0, 1]$, and proper value of $\alpha$ still needs to be decided by trial, which is the selection method for $\alpha$ in this study. Parameters $\lambda$ and $\mu$ are two positive constants. The number of hidden nodes is another important parameter for the network. Overfitting and underfitting problems can be caused by the inappropriate number of hidden nodes. More hidden nodes will approximate the input-output relation on training set too accurate to make the network have good generalization property. The typical behavior of ANN in this case is that the prediction accuracy for training set is particularly high but that for testing set is pretty poor. On the other hands, insufficient number of hidden nodes will cause the network not able to learn enough from the training set. Consequentially, the ANN will have poor performance for both training set and testing set. Similar as $\alpha$, there is no good way to decide the number of hidden nodes except for experiments. This study will discuss how the hidden nodes are selected in Section 3.4.

### 3.2.4 A Statistical Method for Tobacco Growing Area Discrimination: Mahalanobis Distance Model

Ni et al. (2009) applied the MD discrimination model to predict tobacco growing area which is described by provincial districts. The performance was evaluated by the prediction accuracies for training set and testing set. Basically, there are two stages for this approach. At first, the training set is used to calculate the centroid of each class. In this study, tobacco spectra are extracted by PCA method, and the first $i$-th PCs are selected for modeling. They are represented as $x = [x_1 \ x_2 \ ... \ x_i]$. The centroid of class $j$ is $\bar{x}_j = [\bar{x}_{1j} \ \bar{x}_{2j} \ \cdots \ \bar{x}_{ij}]$, and $\bar{x}_{ij}$ is calculated by

$$\bar{x}_{ij} = \frac{1}{n} \sum_{m=1}^{n} x_{ijm}, \quad (3.19)$$
where there are \( n \) samples in the group \( j \) for the training set. Secondly, for any given sample, the Mahalanobis distances between the centroid of classes and the samples are calculated. The class with the shortest MD away from the sample is the final predicted output. MD is an useful technique to measure the difference of two measured vectors, in which the elements have different calibration attributes. The calculation of MD between two given objects \( n_i \) and \( n_j \) is

\[
D_{n_1n_2} = (X_{n_1} - X_{n_2})^T S^{-1} (X_{n_1} - X_{n_2}),
\]

(3.20)

where \( X_{n_i} \) and \( X_{n_j} \) are the sample vectors with the same dimensionality; \( S \) is the covariance matrix of dataset; the superscript “-1” is the inverse operator; and the superscript “T” is the transpose operator. Because each PC is orthogonal with each other, \( S \) is a diagonal matrix and the diagonal elements are PCs’ deviation \( \lambda_1, \lambda_2, \ldots, \lambda_n \). Therefore, MD between the sample \( n_1 \) and \( n_2 \) can be written as

\[
D_{n_1n_2} = \sum_{m=1}^{n} \frac{(x_{n_1m} - x_{n_2m})^2}{\lambda_m}.
\]

(3.21)

In this study, the MD between each sample and each class centroid needs to be calculated. For sample \( n_1 \), the MD between \( n_1 \) and class \( j \) can be calculated using

\[
D_{n_1j} = \sum_{m=1}^{n} \frac{(x_{n_1m} - \bar{x}_{jm})^2}{\lambda_m},
\]

(3.22)

where \( j \) represents the class number, and \( j = \{0, 1, 2, 3\} \) in this study. The prediction class is based on the shortest MD rule. The performance is evaluated by prediction accuracies of training set and testing set.

3.2.5 Model Evaluations

The prediction accuracy for the four classes is applied to assess the overall performance of the classification problem, and it is defined as

\[
P_a = \frac{n_r}{N_t},
\]

(3.23)

where \( n_r \) is the number of samples, which are correctly predicted; \( N_t \) is the number of samples for prediction. In this study, \( N_t \) is equal to 266 (80% of the whole samples)
Table 3.1: The confusion matrix. $n_{TP}$ is the number of true positive samples whose known label and predicted output are both positive; $n_{FN}$ is the number of false negative samples whose known label is positive, and predicted output is negative; $n_{FP}$ is the number of false positive samples whose known label is negative, and predicted label is positive; $n_{TN}$ is the number of true negative samples whose known label is negative, and predicted label is negative.

<table>
<thead>
<tr>
<th>Known label</th>
<th>Predicted label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>$n_{TP}$</td>
</tr>
<tr>
<td></td>
<td>$n_{FN}$</td>
</tr>
<tr>
<td>Negative</td>
<td>$n_{FP}$</td>
</tr>
<tr>
<td></td>
<td>$n_{TN}$</td>
</tr>
</tbody>
</table>

and 66 (20% of the whole samples) for the calculation of the training and testing prediction accuracy, respectively.

Because the distribution of dataset classes to be predicted is unknown in the real application, sufficient evaluation should be done for summarizing prediction capability. A two-by-two confusion matrix (or contingency table) (Brown and Davis, 2006) is introduced, which can be used to derive some interested evaluation criteria for classifiers, such as precision, sensitivity, specificity and so on. The confusion matrix is depicted by Table 3.1, in which $n_{TP}$ is the number of true positive samples whose known label and predicted output are both positive; $n_{FN}$ is the number of false negative samples whose known label is positive, and predicted output is negative; $n_{FP}$ is the number of false positive samples whose known label is negative, and predicted label is positive; $n_{TN}$ is the number of true negative samples whose known label is negative, and predicted label is negative.
From the confusion matrix, various evaluation criteria can be constructed as

\[
\gamma_{TP} = \frac{n_{TP}}{n_{TP} + n_{FN}},
\]

\[
\gamma_{TN} = \frac{n_{TN}}{n_{FP} + n_{TN}},
\]

\[
\gamma_{PP} = \frac{n_{TP}}{n_{TP} + n_{FP}},
\]

\[
\gamma = \frac{2n_{TP}}{2n_{TP} + n_{FP} + n_{FN}},
\]

where \( \gamma_{TP} \) is sensitivity or true positive rate, which is the fraction of truly positive events that are correctly predicted to be positive; \( \gamma_{TN} \) is specificity or true negative rate, which is the fraction of truly negative events that are correctly predicted to be negative; \( \gamma_{PP} \) is precision or positive predictive rate, which is the fraction of predicted positive events that are truly positive; \( \gamma \) is also called F1-score, which considers both precision and sensitivity of the test, and can be interpreted as a weighted average of them. The range of all above four evaluation criteria is between 0 (worst score) and 1 (best score).

### 3.3 Experimental Results

In this section, a tobacco NIR spectrum dataset with growing area labels is taken to investigate the feasibility of the proposed approach for cultivation area discrimination. The effectiveness of the proposed ANN classifier is demonstrated in terms of prediction accuracy. The results of preprocessing methods for raw NIR data and the setting of ANN model are discussed in details. All the simulations are implemented on Window Vista operating system using Matlab R2008b, which is running on a laptop with Intel Pentium (R) Dual CPU T3400 2.16GHz 2.17GHz and 2Hz RAM. The statistic toolbox and bio-informatics toolbox are required to facilitate the de-noising process and statistical model. A total of 332 tobacco NIR spectrum samples are used. Tobacco leaves are collected from four areas of Guizhou Province in China, and the NIR spectra are recorded by Thermo Antaris 2 (Thermo Fisher Scientific Inc. Waltham, USA) with the spectral resolution of 8 cm\(^{-1}\) and 64 scans within the near-infrared range of 3499 cm\(^{-1}\) and 12000 cm\(^{-1}\).
3.3.1 Results of De-noising

Considering the inevitable noise introduced by the NIR spectrum instruments, the experimental environment and operational errors, it is necessary to remove noise from raw data before NIR spectra are used for modelling. In this study, SG smoothing, SG first-order, and second-order derivative approaches with a 121-tap sliding window are employed, respectively. The results of three de-noising methods and raw spectra are shown in Figure 3.4, where the absorbance of the raw tobacco NIR spectra are shown in Figure 3.3(a). The spectra de-noised using the the SG smoothing, first-order, and second-order derivative methods are shown in Figure 3.3(b), (c) and (d), respectively. As can be seen, the absorbance is scattered from 0.15 to 0.7 for the raw dataset. The results of SG smoothing keep this range and become more smooth. Meanwhile, de-noised by the first-order derivative, the spectra is more compact and range from [-0.0012, 0.001]. For second-order derivative outcome, the spectra become even more compact and range from [-0.0001, 0.0005]. In the last two cases, the noise is greatly reduced but the important spectral information may have been lost. Therefore, the smoothing method needs to be selected based on its effect on specific classification model, and will be discussed later. In addition, absorbance bands of the resulted spectra treated by the three de-noising methods are highly overlapped, which indicates that tobacco growing area discrimination is a challenging problem.

3.3.2 Results of Feature Extraction

PCA technique is effective to extract the most important information, and to remove the redundant information from high dimensional data. In this study, there are 2084 information points for one spectrum. Because there is no evidence to show which range of spectral wave is sensitive for classification models, the full range of NIR spectra is adopted. After applying PCA, the most important information is kept as PCs, which are sorted by the amount of information carried on them.

Generally, the number of PCs for establishing model is determined by the cumulative contribution rate of PCs, i.e., accumulated scores of PCs. Indeed, the selected
Figure 3.3: The raw and de-noised tobacco NIR spectra. (a) Raw tobacco NIR spectra; (b) tobacco NIR spectra de-noised using SG smoothing method; (c) tobacco NIR spectra de-noised using SG first-order derivative method; (d) tobacco NIR spectra de-noised using SG second-order derivative method.
Figure 3.4: PCA of raw and de-noised tobacco NIR spectra. (a) PCA feature extraction of raw data; (b) PCA feature extraction of spectra de-noised using SG smoothing method; (c) PCA feature extraction of spectra de-noised using SG first-order derivative method; (d) PCA feature extraction of spectra de-noised using SG second-order derivative method.
PCs carry on most of information of the spectrum. The proportional information (called variance explained) carried by each PC for raw data and each of de-noised spectra is depicted in Figure 3.4. The bar is the proportional contribution for each PC, and red line indicates the accumulated score. It can be verified that the first several PCs possess very large proportion of total information. For example, the accumulated score of first three PCs for raw data is 97.88%; the accumulated score of first three PCs for SG smoothed data is 98.36%; the accumulated score of first six PCs for first-order derivative de-noised data is 89.32%, and that of first fifteen PCs for second-order derivative de-noised data is 89.08%.

However, it is not guaranteed that these PCs contain the information that stands for the difference among areas, as the Figure 3.5 illustrated. The joint distributions of first PC and second PC for raw data and three kinds of smoothed spectral are depicted in Figure 3.5(a), (b), (c) and (d), respectively. It can be observed that the instances in the four classes are seriously mixed, and there is no explicit clue to differentiate the four areas directly. It is expect more PCs to find the implicit difference for this problem. In this study, the proper number of PCs is determined by the corresponding discrimination accuracy of the model with a set of trail number of PCs.

In the experiment, a set of trial number of PCs are examined on training and testing dataset to find out the most proper number of PCs for model establishment. However, there is no intention to fix the number of PCs since it is highly related to the selected data.

### 3.3.3 Performances

Three smoothing methods: the SG (1) smoothing, (2) first-order, and (3) second-order derivative, are employed for noise removal, respectively. Without loss of generality, the number of PCs is selected from \{6, 8, 10, 12, 14, 16\} after PCA process. As comparison, the corresponding prediction accuracies of developed ANN models and MD models for training are shown in Table 3.2. It can be observed that for all cases, the ANN models consistently outperform the MD models. When applying 14 PCs, ANN
Figure 3.5: The distribution of scores of the first principal component (PC) and second PC for four classes. (a) on the raw data; (b) on the spectra de-noised using SG smoothing method; (c) on the spectra de-noised using SG first-order derivative method; (d) on the spectra de-noised using SG second-order derivative method.
model reaches to best performance with the prediction accuracy 79.3%, which is 2.2% higher than the best prediction accuracy of MD. For smoothing method, the prediction accuracy of both ANN and MD model on training are the highest, when spectra are de-noised using SG smoothing method. Therefore, the SG smoothing method is selected as de-noising method. For the number of PCs, it can be observed that more PCs could potentially improve the classification results, because the information may be lost when less principal components are selected. However, the prediction accuracy is degraded when 16 PCs are employed, which may be caused by more noise when the number of PCs is increased.

In order to verify the validation of developed tobacco growing area classifiers, the test set that occupies 20% of the original 332 samples is used to examine the two models. After SG smoothing and PCA method, the same number of PCs are tested as the same as that on the training set. The prediction accuracy for ANN and MD models under each input case is shown in Table 3.3. It can still be observed that the prediction accuracy ($P_a$) of ANN model is consistently higher than that of MD regardless of the number of principal components. The improvement of ANN, compared with MD, can be around 4%. The best performance of ANN, $P_a = 78.2\%$, is achieved when the first 14 PCs are chosen as inputs for model construction. On the other hand, the best $P_a$ of MD is 72.7%, when the first 14 PCs are selected. Therefore, it is recommended to use the first 14 PCs to discriminate tobacco growing areas in this study.

Furthermore, the overall prediction accuracy is highly effected by the number of samples in each class. The identification capabilities of ANN and MD models for each class are also analyzed in terms of $\gamma_{TP}$, $\gamma_{TN}$, $\gamma_{PP}$, and $\gamma$ when the number of input is 14. The higher value of these criteria, the better performance of corresponding model. The values of each criteria for all four classes assessed by two models are shown in Table tab:ANN-MD. It can be observed that (1) the ANN is superior than MD for most individual class; (2) the ANN is more suitable to identify the tobacco from the north and southwest of Guizhou province of China than MD classifier; (3) the ANN and MD have similar identification capability for the tobacco from the middle and
Table 3.2: The prediction accuracy of ANN the MD models for training set with raw spectra and three pro-processed spectra. \( n \): the number of principal components as input of models; \( P_a \): prediction accuracy for the training set (in this case) or testing set; ANN1: the ANN model with 6 PCs of tobacco NIR as inputs; ANN2: the ANN model with 8 PCs of tobacco NIR as inputs; ANN3: the ANN model with 10 PCs of tobacco NIR as inputs; MD1: the MD classifier with 6 PCs of tobacco NIR as inputs; MD2: the MD classifier with 8 PCs of tobacco NIR as inputs; MD3: MD classifier with 10 PCs of tobacco NIR as inputs; ANN4: the ANN model with 12 PCs of tobacco NIR as inputs; ANN5: the ANN model with 14 PCs of tobacco NIR as inputs; ANN6: the ANN model with 16 PCs of tobacco NIR as inputs; MD4: the MD classifier with 12 PCs of tobacco NIR as inputs; MD5: the MD classifier with 14 PCs of tobacco NIR as inputs; and MD6: the MD classifier with 16 PCs of tobacco NIR as inputs

<table>
<thead>
<tr>
<th>Training Set</th>
<th>MD1</th>
<th>ANN1</th>
<th>MD2</th>
<th>ANN2</th>
<th>MD3</th>
<th>ANN3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Raw</td>
<td>50</td>
<td>51.5</td>
<td>52</td>
<td>53.03</td>
<td>62.12</td>
<td>62.12</td>
</tr>
<tr>
<td>SG S</td>
<td>62</td>
<td>66.2</td>
<td>65</td>
<td>68.8</td>
<td>67</td>
<td>66.8</td>
</tr>
<tr>
<td>SG 1st</td>
<td>61</td>
<td>63.6</td>
<td>58</td>
<td>65.2</td>
<td>65</td>
<td>68.2</td>
</tr>
<tr>
<td>SG 2nd</td>
<td>45</td>
<td>55.5</td>
<td>48</td>
<td>59.7</td>
<td>52</td>
<td>66.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training Set</th>
<th>MD4</th>
<th>ANN4</th>
<th>MD5</th>
<th>ANN5</th>
<th>MD6</th>
<th>ANN6</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td>12</td>
<td>12</td>
<td>14</td>
<td>14</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Raw</td>
<td>59</td>
<td>56.1</td>
<td>65</td>
<td>66.7</td>
<td>65</td>
<td>65</td>
</tr>
<tr>
<td>SG S</td>
<td>71.4</td>
<td>75.6</td>
<td>77.1</td>
<td>79.3</td>
<td>74.4</td>
<td>74.8</td>
</tr>
<tr>
<td>SG 1st</td>
<td>68.4</td>
<td>71.4</td>
<td>70.6</td>
<td>73.7</td>
<td>71</td>
<td>72.7</td>
</tr>
<tr>
<td>SG 2nd</td>
<td>56</td>
<td>62.4</td>
<td>59</td>
<td>69.7</td>
<td>61</td>
<td>64.3</td>
</tr>
</tbody>
</table>
Table 3.3: The prediction accuracy of the ANN and MD models for tobacco NIR spectra de-noised using SG smoothing method on testing. $n$: the number of principal components as in put of models; $P_a$: prediction accuracy for the training set (in this case) or testing set; ANN1: the ANN model with 6 PCs of tobacco NIR as inputs; ANN2: the ANN model with 8 PCs of tobacco NIR as inputs; ANN3: the ANN model with 10 PCs of tobacco NIR as inputs; MD1: the MD classifier with 6 PCs of tobacco NIR as inputs; MD2: the MD classifier with 8 PCs of tobacco NIR as inputs; MD3: MD classifier with 10 PCs of tobacco NIR as inputs; ANN4: the ANN model with 12 PCs of tobacco NIR as inputs; ANN5: the ANN model with 14 PCs of tobacco NIR as inputs; ANN6: the ANN model with 16 PCs of tobacco NIR as inputs; MD4: the MD classifier with 12 PCs of tobacco NIR as inputs; MD5: the MD classifier with 14 PCs of tobacco NIR as inputs; and MD6: the MD classifier with 16 PCs of tobacco NIR as inputs.

<table>
<thead>
<tr>
<th>Testing Set</th>
<th>MD1</th>
<th>ANN1</th>
<th>MD2</th>
<th>ANN2</th>
<th>MD3</th>
<th>ANN3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>$P_a$(%)</td>
<td>61</td>
<td>63.6</td>
<td>58</td>
<td>65.2</td>
<td>65</td>
<td>68.2</td>
</tr>
<tr>
<td>Testing Set</td>
<td>MD4</td>
<td>ANN4</td>
<td>MD5</td>
<td>ANN5</td>
<td>MD6</td>
<td>ANN6</td>
</tr>
<tr>
<td>$n$</td>
<td>12</td>
<td>12</td>
<td>14</td>
<td>14</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>$P_a$(%)</td>
<td>68</td>
<td>71.2</td>
<td>72.7</td>
<td>78.2</td>
<td>71</td>
<td>74.2</td>
</tr>
</tbody>
</table>
Table 3.4: The identification capability of ANN and MD for each class are analyzed in terms of $\gamma_{TP}$, $\gamma_{TN}$, $\gamma_{PP}$, and $\gamma$ based on testing set with SG smoothed tobacco NIR spectra. Class 1: north region; Class 2: middle region; Class 3: northwest region; Class 4: southwest region; $\gamma_{TP}$: true positive rate; $\gamma_{TN}$: specificity or true negative rate; $\gamma_{PP}$: positive predictive rate; $\gamma$: F1 score;

<table>
<thead>
<tr>
<th></th>
<th>Class 1: North</th>
<th>Class 2: Middle</th>
<th></th>
<th>Class 3: Northwest</th>
<th>Class 4: Southwest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\gamma_{TP}$</td>
<td>$\gamma_{TN}$</td>
<td>$\gamma_{PP}$</td>
<td>$\gamma$</td>
<td>$\gamma_{TP}$</td>
</tr>
<tr>
<td>ANN</td>
<td>0.75</td>
<td>0.98</td>
<td>0.9</td>
<td>0.82</td>
<td>0.6</td>
</tr>
<tr>
<td>MD</td>
<td>0.83</td>
<td>0.94</td>
<td>0.77</td>
<td>0.8</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 3:</td>
<td>$\gamma_{TP}$</td>
<td>$\gamma_{TN}$</td>
<td>$\gamma_{PP}$</td>
<td>$\gamma$</td>
<td>$\gamma_{TP}$</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.87</td>
<td>0.63</td>
<td>0.77</td>
<td>0.88</td>
</tr>
<tr>
<td>Class 4:</td>
<td>$\gamma_{TP}$</td>
<td>$\gamma_{TN}$</td>
<td>$\gamma_{PP}$</td>
<td>$\gamma$</td>
<td>$\gamma_{TP}$</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.88</td>
<td>0.74</td>
<td>0.76</td>
<td>0.76</td>
</tr>
</tbody>
</table>

northwest of Guizhou province.

3.4 Discussion

For the proposed ANN model, the number of hidden neurons and learning rate are two critical parameters, which highly affect the performance of the developed model. A proper setting of those two parameters can improve the prediction accuracy, and otherwise may lead to the underfitting or overfitting of ANN model. For example, insufficient hidden neurons would make ANN model unable to approximate the relationship between complex inputs and desired output, thereby resulting poor prediction. On the contrary, too many hidden neurons not only greatly increase the complexity of ANN model, but also harm to its generalization ability. The similar phenomenon is also applied to learning rate.
Figure 3.6: The selection for number of hidden nodes (HNs). (a) The discrimination accuracy for training and validation along with the change of HNs; (b) The mean absolute squared error along with the change of HNs.
Figure 3.7: The absolute squared error along with the increase of epochs, when HNs is 11 and number of inputs is 14.

Generally, those two parameters are selected empirically. The results for the HNs selection is shown in Figure 3.6. The discrimination accuracy and error on the training set and validation are changed with different HNs. It can be observed that along with the increase of HNs, the prediction accuracy increases and error decreases, especially for the training process. After a certain threshold (11 in this case), both measures for validation become worse if HNs is further increased. This indicates that the ANN enters into overfitting status, irrespective that the training performance is still improved. Thus, 11 is chosen as the HNs in the ANN model. For learning rate, 0.0007 is selected through a set of trial values based on prediction accuracy.

As a conclusion for the experimental results of ANN classifier, the best observed prediction accuracy for training and testing are 79.3% and 78.2%, respectively. The error tendency along with epoch is depicted in Figure 3.7. Along with the increase of epoch, the error decreases sharply before 1000 epochs, followed by slightly and slowly decreasing. Because the studied four-class classification problem is still a challenge for the state-of-the-art classifiers, the reported outcomes fully evidence the effectiveness of the proposed approach.
3.5 Summary

In this chapter, the ANN model are developed for tobacco growing area identification. The prediction accuracies on the training and testing demonstrate the effectiveness of established models. In order to give a comprehensive prediction capability assessment, the confusion matrix is introduced to build the statistic evaluation for each class, which includes $\gamma_{TP}$, $\gamma_{TN}$, $\gamma_{PP}$, and $\gamma$. The experimental results indicate that the ANN classifier outperform the conventional statistical method, MD model. In addition, the pre-processing methods are analyzed based on the prediction accuracy. The SG smoothing method and PCA algorithm are chosen to be the de-noising and feature extraction method in pre-processing stage, respectively. The parameter setting, which related to the number of hidden neurons and learning rate of models, is also discussed to further prove that the developed ANN model can be a promising solution for tobacco growing area discrimination problem.
Chapter 4

The Proposed Support Vector Machine Classifier

SVM, originally proposed by Vapnik (1995), is a powerful technique to be applied for classification, especially for the high-dimensional nonlinear separable problems. As a supervised pattern recognition method, it possesses well developed theoretical foundation (statistical learning theory), and is applicable for widespread practical applications. It is appealing more and more attention due to its outstanding advantages: suitable for small sample set, strong mathematic support, good performance as a nonlinear classifier, and effectively avoiding the curse of dimensionality. It has been employed in various research fields, such as text categorization, computer vision, and bio-informatics.

4.1 The Proposed Method

In this chapter, the SVM model is developed to investigate the implicit relation between growing areas and tobacco NIR spectra. Firstly, the same preprocessing methods are adopted as stated in Section 3.2.1 and Section 3.2.2, including the SG smoothing for noise elimination and PCA for feature extraction. The dimensionality of raw tobacco spectrum data is greatly reduced after PCA process. Most of information for NIR spectrum concentrates on the first several extracted PCs, which are
considered as inputs for the SVM model to predict growing areas of the unknown tobacco samples.

4.1.1 The Support Vector Machine Algorithm

In the last decades, SVM algorithm has been successfully applied to many real-world applications, especially on those with small sample set. The introduction of the kernel function and the structural risk minimization are two critical components for SVM to achieve superior performance in terms of prediction accuracy and good generalization ability. Using kernel function, the nonlinear separable inputs in low-dimensional space can be mapped into high-dimensional feature space, so that they become linear or approximate linear separable. Unlike other classification methods, which aim for minimizing the empirical training error, the SVM makes the structural risk minimized through building optimal hyperplanes or decision planes, which makes the margin of separation between different classes maximized. Therefore, SVM algorithm is able to classify the complex nonlinear separable data with excellent performance.

Originally, SVM is designed for binary pattern recognition problems. However, it is easily to be extended to multi-category problems. Without loss of generalization, the basic binary classification problem is employed to illustrate the essence of SVM algorithm in the following.

Assume \((\vec{x}, y)\) is a set of samples, in which \(\vec{x}_i \in \mathbb{R}^d\) is the input vector, and \(y_i \in \{-1, +1\}(i = 1, 2, \cdots, N)\) is the corresponding class label. A linear decision hyperplane can be formulated as

\[
g(\vec{x}) = \vec{w} \cdot \vec{x} + b, \tag{4.1}
\]

where \(\vec{w}\) is the weight vector; \(b\) is the bias; \("\cdot"\) is the inner product operator; \(g(\vec{x}) = 0\) is the separating hyperplane, and the corresponding classifier is \(\text{sgn}(g(\vec{x}))\).

The Euclidean distance between any instance \(\vec{x}_i\) to the separating hyperplane can be calculated by

\[
\delta = \frac{1}{\|\vec{w}\|} | g(\vec{x}_i) |, \tag{4.2}
\]
Figure 4.1: The idea of SVM algorithm to transfer nonlinear separable problem into linear separable one. There are two classes in this example.
where $\| \vec{w} \|$ is the second norm of vector $\vec{w}$. Therefore, to maximize margin between $g(\vec{x}) = 1$ and $g(\vec{x}) = -1$ is equivalent to minimize $\frac{1}{2} \| \vec{w} \|^2$.

As shown in Figure 4.1(b), the distance between the plane $L_1$ and $L_2$ is the maximum margin and the red plane is the OSH. The red plane in $\mathbb{R}^d$ shown in Figure 4.1(c) corresponds to the OSH in $\mathbb{H}$ shown in Figure 4.1(b). All instances, which locate on the two dashed parallel plane $L_1$ and $L_2$ in $\mathbb{H}$, are called support vectors.

However, it is an idealized scenario that the nonlinear separable data in $\mathbb{R}^d$ are fully transformed into the linear separable in $\mathbb{H}$. After mapped into $\mathbb{H}$, nonlinear separable instances are always encountered, such as the instances “1” to “6” shown in Figure 4.2. Those minor samples lead to the performance degradation of SVM. In order to improve the capability of SVM model, a slack variable $\xi$ is introduced to enlarge the tolerance for those nonlinear separable instances. The $\xi$ slightly reduces the margin between two classes to an acceptable level so that those minor objects can be accounted.

Finally, the OSH can be found by solving the following quadratic programming problem

$$
\min (w, b, \xi) \quad \frac{1}{2} \| \vec{w} \|^2 + C \sum_{i=1}^{N} \xi_i, \tag{4.3}
$$

subject to

$$
y_i[\vec{w}^T \cdot \vec{x}_i + b] \geq 1 - \xi_i,
$$

where $i = \{1, 2, \cdots N\}; C$ is the penalty constant; and $\xi_i$ is the slack variable.

The $\vec{w}$ and $b$ can be solved by finding the solution of dual form of Equation (4.3) using Lagrange multipliers method (Haykin, 1999) as

$$
\min (\alpha_i) \quad \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(\vec{x}_i, \vec{x}_j) - \sum_{i=1}^{N} \alpha_i, \tag{4.4}
$$

subject to

$$
\sum_{i=1}^{N} \alpha_i y_i = 0,
$$

$$
0 \leq \alpha_i \leq C \text{ for } i = 1, \cdots, N,
$$

where $K(\cdot, \cdot)$ is the kernel function, which is designed to compute the inner product in $\mathbb{H}$; and $\vec{x}_j$ is the support vector with corresponding Lagrange coefficient $\alpha_j \neq 0$. 

51
Figure 4.2: The SVM algorithm deals with some instances that are still nonlinear separable in feature space.

4.1.2 Typical Kernel Functions

Because the kernel function typically reflects the essential relation between nonlinearly separable data in $\mathbb{R}^d$ and linearly separable data in $\mathbb{H}$, the selection of it is important to guarantee the good performance of SVM model. The performance of SVM model highly depends on the characteristics of investigated data, which makes it hard to choose a proper kernel by employing an universal criterion. Generally, there are four commonly used kernel functions: linear kernel, polynomial learning machine, radial-basis function (RBF), and sigmoid function, whose details are listed in Table 4.1. Empirically, the RBF kernel function always shows strong competitiveness.
Table 4.1: The most commonly used kernel function for SVM.

<table>
<thead>
<tr>
<th>Kernel function</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>$\vec{x} \cdot \vec{y}$</td>
</tr>
<tr>
<td>Polynomial</td>
<td>$(\alpha_0 \vec{x}^T \cdot \vec{y} + \alpha_1)^p$ $\alpha_0, \alpha_1$ and power $p$</td>
</tr>
<tr>
<td>RBF</td>
<td>$-\frac{|\vec{x}^2 - \vec{y}^2|}{2\sigma^2}$ $\sigma$ width</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>$\tanh(\beta_0 \vec{x}^T \cdot \vec{y} + \beta_1)$ $\beta_0$ and $\beta_1$</td>
</tr>
</tbody>
</table>

4.1.3 Multi-class Support Vector Machine

Because the real-world applications are never limited to binary classification, it is necessary to extend SVM into a multi-class classifier. There are two common methods to achieve the expansion: one-versus-all and one-versus-one. One-versus-all scheme is done by a winner-takes-all strategy. For the one-versus-one expansion, classification is conducted through a max-wins voting strategy. Essentially, the basic idea of both strategies is a combination of binary classifiers and an election process. Therefore, for an $n$-class problem ($n > 2$), there are $n$ binary SVM classifiers to be built when one-versus-all strategy is adopted. While for the one-versus-one extension, $\frac{n(n-1)}{2}$ SVM classifiers need to be trained. In this paper, the one-versus-one strategy is employed to maintain the good classification performance of SVM model.

4.2 Experimental Results

The developed system is implemented and tested on Windows Vista platform using Matlab. The program is executed on the same laptop. LIBSVM (Chang and Lin, 2011), which is an integrated and easy-to-use library for SVM algorithm, is employed to construct the discrimination system. The training and testing dataset are the same as the previous proposed method, which occupy 80% and 20% of the whole dataset, respectively. The SG smoothing is selected as de-noising method. Because SVM model could achieve very high prediction accuracy on the training set, the
Table 4.2: The performance of SVM model for tobacco growing area discrimination. $P_a$ is the prediction accuracy.

<table>
<thead>
<tr>
<th>$n$</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM $P_a$ (%)</td>
<td>60.6</td>
<td>67</td>
<td>72.7</td>
<td>74.2</td>
<td>80.3</td>
<td>78.8</td>
</tr>
<tr>
<td>ANN $P_a$ (%)</td>
<td>63.6</td>
<td>65.2</td>
<td>68.2</td>
<td>71.2</td>
<td>78.2</td>
<td>74.2</td>
</tr>
</tbody>
</table>

performance of it on the testing set is of importance and mainly discussed in this section. Actually, the prediction accuracy on testing set is used as the primary criteria for model performance evaluation.

In order to demonstrate the superior performance of the SVM model, the ANN model is tested in the same scheme as comparison in terms of prediction accuracy ($P_a$). As shown in Table 4.2, irrespective of the number of selected PCs ($n$), the SVM provides higher $P_a$ than that of ANN in most cases. Specifically, the $P_a$ improvements are 1.8%, 4.5%, 3%, 2.1%, and 4.6% when the number of input is 8, 10, 12, 14, and 16, respectively. The best performance of the SVM is achieved when $n = 14$, and the $P_a = 80.3\%$, which is 2.1% higher than that of the ANN model.

In addition to the prediction accuracy, which is the model performance evaluation regarding all four classes, the discrimination effectiveness of the developed model for each individual class is assessed by the four criteria introduced in Chapter 3, including $\gamma_{TP}$, $\gamma_{TN}$, $\gamma_{PP}$, and $\gamma$. The values of those criteria on the testing set are shown in Table 5.4. The first 14 PCs are taken as inputs for the SVM model. It can be observed that the SVM is superior to ANN in most cases, especially for class 2 and 4. For example, for class 2, the $\gamma_{TP}$ of SVM is 0.28 higher than that of ANN. Therefore, SVM model outperforms ANN model in tobacco growing area discrimination with respect to both individual and all classes in most cases.
Table 4.3: The identification capability of SVM classifier and ANN model for each class are analyzed in terms of $\gamma_{TP}$, $\gamma_{TN}$, $\gamma_{PP}$, and $\gamma$ based on testing set with SG smoothed tobacco NIR spectra. Class 1: the north of Guizhou Province in China; Class 2: the middle of Guizhou Province in China; Class 3: the northwest of Guizhou Province in China; Class 4: the southwest of Guizhou Province in China; $\gamma_{TP}$: true positive rate; $\gamma_{TN}$: specificity or true negative rate; $\gamma_{PP}$: positive predictive rate; and $\gamma$: F1 score.

<table>
<thead>
<tr>
<th></th>
<th>Class 1: North</th>
<th>Class 2: Middle</th>
<th>Class 3: Northwest</th>
<th>Class 4: Southwest</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.67 0.98 1 0.76</td>
<td>0.88 0.78 0.75 0.79</td>
<td>SVM 0.75 0.94 0.71 0.75</td>
<td>0.82 1 0.89 0.9</td>
</tr>
<tr>
<td>ANN</td>
<td>0.75 0.98 0.9 0.82</td>
<td>0.6 0.88 0.75 0.67</td>
<td>ANN 1 0.87 0.63 0.77</td>
<td>0.88 0.96 0.88 0.88</td>
</tr>
</tbody>
</table>

4.3 Discussion

The selection of preprocessing method is firstly discussed in this section. Secondly, because the settings of SVM model, such as the penalty constant, kernel function and corresponding parameters, are important for achieving the appealing performance. In this section, the effects of these configurations with respect to the prediction accuracy are discussed, and the corresponding proper values are stated.

4.3.1 Selection of Preprocessing Method

As the same situation as the ANN model in Chapter 3, the noise removal from original tobacco NIR spectrum data is necessary, because the noise will make the character-
Table 4.4: The prediction accuracy of the SVM model for tobacco growing area discrimination using the raw data and the data preprocessed by different de-noising methods. \( n \): the first \( n \)-th PCs of tobacco NIR spectrum data; SG smoothed data: the data using the SG smoothing method; SG 1st OD data: the data using the SG first-order derivative; SG 2nd OD data: the data using the SG second-order derivative; \( P_a \) is the prediction accuracy.

<table>
<thead>
<tr>
<th>( n )</th>
<th>( P_a ) (%)</th>
<th>( P_a ) (%)</th>
<th>( P_a ) (%)</th>
<th>( P_a ) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>the raw data</td>
<td>SG smoothed data</td>
<td>SG 1st OD data</td>
<td>SG 2nd OD data</td>
</tr>
<tr>
<td>6</td>
<td>68.18</td>
<td>74.24</td>
<td>78.80</td>
<td>60.60</td>
</tr>
<tr>
<td>8</td>
<td>63.64</td>
<td>69.70</td>
<td>63.61</td>
<td>50.00</td>
</tr>
<tr>
<td>10</td>
<td>60.61</td>
<td>71.20</td>
<td>72.73</td>
<td>60.60</td>
</tr>
<tr>
<td>12</td>
<td>63.64</td>
<td>80.3</td>
<td>75.80</td>
<td>59.00</td>
</tr>
<tr>
<td>14</td>
<td>71.21</td>
<td>78.79</td>
<td>75.80</td>
<td>65.15</td>
</tr>
<tr>
<td>16</td>
<td>74.24</td>
<td>71.72</td>
<td>81.80</td>
<td>57.60</td>
</tr>
</tbody>
</table>

Statistics of tobacco NIR spectral data hard to identify. In this section, three de-noising methods are studied and the most proper one is selected from: (1) SG smoothing; (2) SG first-order derivative; and (3) SG second-order derivative method. The raw data are tested as comparison. A comprehensive test is conducted, whose results are shown in Table 4.4. Each de-noising method is tested based on the prediction accuracy of different numbers of PCs. It can be seen that SG smoothing method gives better performance than others. Specifically, for the number of inputs range from 6 to 14, the \( P_a \) of SG smoothing method is always the highest among four candidates. It can be seen that SG smoothing performs worse than the raw data and SG 1st order derivative when the number of inputs is 16, which is not preferred due to the high model complexity in this case. Therefore, SG smoothing method is selected as de-noising method for the merit of better performance and its simplicity.
4.3.2 Selection of Kernel Function

One of the advances for SVM algorithm is that the kernel function in the input space can replace the computation of inner product in the feature space. The patterns in the input space can be mapped into the high-dimensional feature space, and transferred to be linear separated. Kernel functions help to construct the OSH in the feature space without knowing its explicit form. The functions, which satisfy the Mercer’s theorem, can act as kernel functions (Haykin, 1999).

Generally, the Gaussian RBF kernel function is selected to construct SVM model for pattern recognition. There are mainly three reasons. (1) Unlike the linear kernel, RBF kernel nonlinearly maps samples into a higher dimensional space, which can handle the case when the relation between class labels and attributes is nonlinear; (2) The number of hyper parameters influences the complexity of constructed model. As the third column in Table 4.1 shows, the RBF kernel only has one parameter to be determined. Compared with three parameters of polynomial kernel and the two parameters of sigmoid kernel, RBF kernel greatly reduces the complexity of the SVM model and saves computational resources; (3) The sigmoid kernel is not valid (not the inner product of two vectors) under some scenarios.

The experimental results verify that RBF kernel is suitable for constructing the SVM classifier for tobacco growing areas. The 5-fold cross-validation prediction accuracy ($P_{cva}$) on the training set are recorded in Table 4.5. Along with the increase of inputs number, the ($P_{cva}$) for different kernel functions are compared. It can be easily observed that the performances of SVM model using RBF kernel are always much better than other three kernels, regardless of the number of inputs. The highest prediction accuracy ($P_{cva}$) of 5-fold cross-validation is 75.9%, compared with linear kernel 69.2%, polynomial kernel 69.9% and sigmoid kernel 69.2%.

4.3.3 Selection of Parameters for SVM with RBF Kernel

The aim of the parameter selection is to identify the most proper parameter pair: the penalty factor ($C$), and the width of RBF kernel function ($\sigma$), so that the SVM model
Table 4.5: The prediction accuracy when the SVM model respectively adopts the commonly used kernel function along with different number of inputs for the SG smoothed data. $P_{cva}$ (%): the prediction accuracy for 5-fold cross-validation on training set.

<table>
<thead>
<tr>
<th>n</th>
<th>$P_{cva}$ (%)</th>
<th>$P_{cva}$ (%)</th>
<th>$P_{cva}$ (%)</th>
<th>$P_{cva}$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>linear kernel</td>
<td>polynomial kernel</td>
<td>RBF kernel</td>
<td>sigmoid kernel</td>
</tr>
<tr>
<td>6</td>
<td>50.8</td>
<td>50.0</td>
<td>52.6</td>
<td>51.9</td>
</tr>
<tr>
<td>8</td>
<td>59.8</td>
<td>54.5</td>
<td>61.3</td>
<td>57.1</td>
</tr>
<tr>
<td>10</td>
<td>61.3</td>
<td>59.8</td>
<td>71.2</td>
<td>61.7</td>
</tr>
<tr>
<td>12</td>
<td>69.2</td>
<td>66.5</td>
<td>75.6</td>
<td>68.8</td>
</tr>
<tr>
<td>14</td>
<td>68.1</td>
<td>69.9</td>
<td>75.9</td>
<td>69.2</td>
</tr>
<tr>
<td>16</td>
<td>64.5</td>
<td>68.4</td>
<td>75.6</td>
<td>66.9</td>
</tr>
</tbody>
</table>

with RBF kernel function is able to accurately predict the relation between inputs (the PCs of tobacco NIR spectrum data) and outputs (predicted tobacco growing areas). A 5-fold cross-validation procedure is employed for the selection of parameter pair $(C,\sigma)$, as well as for avoiding the overfitting problem.

Specifically, a grid-search approach is employed to identify the most proper parameter pair with the best cross-validation prediction accuracy. The grid-search is a pretty straightforward method. It simply uses each point on the grid of parameter values to build models. Each of the value on the girds is applied to calculate the performance of the model. The point with the best performance is chosen to set up the model at the final stage. In order to reduce the computational cost for this exhaustive search method, the combination of coast grid-search and fine grid-search is adopted. The coarse grid-search is able to find out the best possible field for the optimal parameter pair, and shrink the next step search region to this small grid field. And then the fine grid-search is conducted around this region, and identifies the true optimal parameter pair with best cross-validation prediction accuracy. Furthermore, because of the independence of parameter $C$ and $\sigma$, it is easy to conduct grid-search.
Table 4.6: The parameters selected by grid-search and the performance of SVM model for different number of inputs. \( n \): the number of inputs, that is, the first \( n \) PCs; \( C \): the penalty factor, the important parameter for SVM model with Gaussian RBF kernel function; \( P_{cva} \): the best 5 fold cross-validation prediction accuracy; \( P_a \): the prediction accuracy; \( t \): the time consumption for selecting \( C \) and \( \sigma \) using grid-search.

<table>
<thead>
<tr>
<th>( n )</th>
<th>( C )</th>
<th>( \sigma )</th>
<th>( P_{cva} ) (%)</th>
<th>( P_a ) (%)</th>
<th>( t ) (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>2.000</td>
<td>2.000</td>
<td>59.0</td>
<td>56.1</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>1.414</td>
<td>2.000</td>
<td>66.2</td>
<td>69.7</td>
<td>21</td>
</tr>
<tr>
<td>10</td>
<td>2.000</td>
<td>1.414</td>
<td>72.2</td>
<td>71.20</td>
<td>23</td>
</tr>
<tr>
<td>12</td>
<td>2.000</td>
<td>1.414</td>
<td>75.6</td>
<td>74.2</td>
<td>24</td>
</tr>
<tr>
<td>14</td>
<td>2.000</td>
<td>2.000</td>
<td>78.6</td>
<td>80.3</td>
<td>24</td>
</tr>
<tr>
<td>16</td>
<td>2.000</td>
<td>0.707</td>
<td>75.6</td>
<td>75.8</td>
<td>28</td>
</tr>
</tbody>
</table>

in a parallel way. The time for grid-search would be greatly reduced.

In this section, the data that are preprocessed by SG smoothing and PCA, are used to demonstrate the selection of parameter pair \((C, \sigma)\). The best cross-validation accuracies \( P_{cva} \) and the prediction accuracies \( P_a \) using the corresponding parameter pair are presented in Table 4.6. Additionally, the time consumed by grid-search is also recorded for different numbers of inputs. As Table 4.6 shows, the optimal parameter pair is obtained with the best cross-validation along with different inputs. When the first 14 PCs are input into this model, the best 5-fold cross-validation prediction accuracy is 78.6\% with the parameter pair \((C = 2, \sigma = 2)\). Therefore, this parameter pair for SVM model with RBF kernel is adopted for the number of inputs of 14 in the final stage. Based on this parameter pair, SVM tobacco growing area classification model is trained and tested. The prediction accuracy for independent test dataset is 80.3\%, which is also the best performance among different inputs. The time of grid-search for parameter pair on Matlab platform running in Window Vista operating system depends on the number of inputs. The searching time for 14 inputs is 24s and
for others are all around 24s. With the increase of the number of inputs, the time for parameter pair searching is slightly longer but no more than 30s. More inputs must need longer time to calculate the performance for each pair of parameters.

4.4 Summary

The SVM algorithm aims to finding OSH so that the samples from different classes can be separated as accurately as possible. The introduction of kernel function and slack variable are the highlights of SVM. The kernel function is adopted to map data, which are nonlinearly separable in the input space, into the high-dimensional feature space, in which the data can be linearly separated. The OSH guarantees the margins between different classes are maximized, so that the structural error is minimized. In order to enlarge the generalization of SVM, a slack variable is introduced to deal with the minor nonlinear separable instances. The slack variable reduces the margin between different classes to an acceptable scale. Consequently, the minor samples, which are still nonlinear separable in feature space, are treated appropriately. The proposed SVM classifier is validated on the adopted tobacco NIR database. The best classification accuracy on test data set is up to 80%, which fully demonstrates the effectiveness of the proposed SVM approach. The selection of penalty factor, kernel function, and parameters of kernel function are also discussed for better discrimination capability of SVM.
Chapter 5

The Proposed Genetic Algorithm Optimized SVM Classifier

In this chapter, GA is introduced to selecting model-sensitive input subset from the results of PCA process. The sensitive input subset is selected and used to investigate the relation between growing areas and tobacco NIR spectra. With the aid of GA, the GA-SVM model can take both the importance of individual input and the interaction of inputs into account, which can provide more knowledge for the investigated problem.

5.1 Introduction

Considering the high dimensionality of tobacco NIR spectrum data, PCA technique is applied to extract limited features for tobacco growing area discrimination. Most of information for the tobacco NIR spectrum data is concentrated on the first several PCs, whose importance are indicated by the corresponding eigenvalues. The higher the eigenvalue is, the more significant the PC is, and the more information is conveyed by this PC. The PCA outputs PCs in the order of their significance. For instance, the amount of information for each PC with the SG smoothed tobacco NIR spectra is listed in Table 5.1. The first PC occupies 73.3491% of total information. The second PC occupies 23.6801%, which is less information than the first one. However,
the PC carrying more information does not mean that it contributes more to the classification problem, especially when considering the cooperation with other PCs. In the previous work of this study, only the first few fixed number of PCs are selected as the inputs of developed models. According to the observation of experimental results, the prediction accuracy did not monotonically increase along with feeding more PCs, as shown in Table 4.4. Therefore, a conclusion can be drawn that not each of PC contributes positively for the performance of classifiers.

Table 5.1: The amount of information for each PC ranging from 1 to 25 using the SG smoothing processed tobacco NIR spectra. *n*-th: the *n*-th PC of tobacco NIR spectra; and *I* (%): the amount of information of the corresponding the *n*-th PC.

<table>
<thead>
<tr>
<th><em>n</em>-th</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>I</em> (%)</td>
<td>73.3491</td>
<td>23.6801</td>
<td>1.3320</td>
<td>1.0265</td>
<td>0.4154</td>
<td>0.0813</td>
<td>0.0379</td>
</tr>
<tr>
<td><em>n</em>-th</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td><em>I</em> (%)</td>
<td>0.0211</td>
<td>0.0166</td>
<td>0.0107</td>
<td>0.0083</td>
<td>0.0054</td>
<td>0.0034</td>
<td>0.0026</td>
</tr>
<tr>
<td><em>n</em>-th</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td><em>I</em> (%)</td>
<td>0.0019</td>
<td>0.0010</td>
<td>0.0009</td>
<td>0.0008</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0005</td>
</tr>
<tr>
<td><em>n</em>-th</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>I</em> (%)</td>
<td>0.0005</td>
<td>0.0004</td>
<td>0.0003</td>
<td>0.0003</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In order to obtain the optimum performance of the developed models in terms of prediction accuracy, the redundant and irrelevant information, which may obscures the decision boundary between classes, should be eliminated from the inputs. Moreover, the large number of inputs will not only increases the complexity of models, but also make the model prone to overfitting problem. Thus, the selection of sensitive features is very important for classifiers to achieve better performance, as well as to reduce the model complexity.

Typically, for an *n*-input recognition problem with *m* features patterns, there are
\( m \times (m - 1) \times (m - 2) \ldots (m - n + 1) \) feasible solutions. The search space for the optimal input subset is huge and infeasible to conducting exhaustive search. GA, as a heuristic search method, is a promising solution to solve this challenging problem, which has been widely accepted. It is capable to find out the optimal or at least suboptimal solution efficiently in huge search space, with proper settings.

Based on the description above, GA is adopted to select the most model-sensitive input subset for tobacco growing area classification. In this part, a GA based SVM approach (GA-SVM) is developed. With the aid of GA, the most sensitive input subset for SVM model is selected. The better performance of classifiers is achieved, and the complexity of models is reduced.

5.2 Genetic Algorithm Optimized Support Vector Machine Approaches

In order to feed the more sensitive features into the tobacco growing area classifier, a GA optimized SVM scheme is proposed. The schematic digraph of the proposed GA-SVM model is illustrated in Figure 5.1. In this model, the tobacco NIR spectra are preprocessed by the SG de-noising approach, which has been proved to be the best de-noising method for tobacco NIR data in previous chapters. Based on the results of PCA method, the PCs are encoded into chromosomes and fed into the SVM classifier to conduct fitness evaluation of genes and selection of excellent genes in terms of prediction accuracy. After several generations of evolution, the chromosome with the best fitness value will survived as the final result for optimal input subset. The terminal conditions for evolution usually including preset maximum iteration number is achieved or target fitness value is obtained.

Because of its attractive merits, GA is popular to conduct feature selection. It mimics the evolutionary processes of nature and is based on the Darwinian principle of survival of the fittest.

In the search space of input subsets, each individual is a potential candidate of
Figure 5.1: The framework of the proposed GA-SVM classifier for tobacco growing areas.
## Genetic Algorithm

<table>
<thead>
<tr>
<th>Genetic Algorithm</th>
<th>Principle of Survival of Fittest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoding Feasible Solutions</td>
<td>Genes, Chromosomes</td>
</tr>
<tr>
<td>Initialization the Population</td>
<td>Creation</td>
</tr>
<tr>
<td>Evaluation/Fitness Function</td>
<td>Environment</td>
</tr>
<tr>
<td>Selection of Parents</td>
<td>Reproduction</td>
</tr>
<tr>
<td>Genetic Operators</td>
<td>Crossover and Mutation</td>
</tr>
</tbody>
</table>

Figure 5.2: The corresponding elements for genetic algorithm and the Darwinian principle of survival of the fittest.

optimal input subset and, therefore, is called feasible solution. It corresponds to a chromosome in the principle of survival of fittest. GA can encode feasible solutions using any data structure according to the need of real applications. Generally, binary vectors and real number vectors are commonly used to represent feasible solutions and chromosomes. Corresponding to the principle of survival of fittest, GA mainly includes five elements: feasible solution, initial population, goodness or fitness evaluation, parent selection and genetic operators. The corresponding relation between GA and principle of survival of fittest is shown in Figure 5.2. The left parts are the elements of GA, and the right parts are the ones of the principle of survival of fittest. In order to obtain the same power with the survival of fittest, it is import to adjust the parameters to suit for the real application.

The whole procedure of a standard GA is illustrated in Figure 5.3. In the first step, a set of population is initialized by randomly selected from first 25 PCs. The number of feasible solutions in the initialized population is pre-configured. The value of 25 is chosen because the first 25 PCs extracted by PCA occupies almost 100% information of tobacco NIR spectra. Each individual input subset in the population is encoded into a string of positive integers ranged from $[1, 25]$. The length of chromosome exactly equals to the number of inputs. For instance, if one 6-digital feasible solution
Figure 5.3: The flow diagram of GA to implement the procedure of evolution.
is \{1, 3, 5, 7, 12, 15\} , it means this chromosome consists of the 1-th, 3-th, 5-th, 7-th, 12-th and 15-th PC. Secondly, the fitness evaluation is described by the steps within the dashed rectangle. Prediction accuracy of SVM model is used to measure fitness value for each individual. Thirdly, fitness value biased roulette method is adopted to conduct parent selection. The larger fitness value the input subset has in population, the higher probability the input subset is selected as parents. All parents keep better or more sensitive genetic information for the SVM classification model. At the fourth step, parents are mated randomly, and reproduction is conducted. Genetic operations: crossover and mutation, are functioned to generate the offspring and develop new population, respectively. In this study, one point crossover is adopted, and uniform mutation is employed. Generation after generation, the genes with better fitness value are passed on and inherited. The evolution is processed continuously until termination criteria are satisfied. Eventually, the optimal input subset are selected, which has the best fitness for the SVM classification model. Eventually, the algorithm identifies the individuals with the optimizing fitness values, and those with lower fitness will naturally get discarded from the population.

5.3 Experimental Results

In the experiment, the raw tobacco NIR spectra are preprocessed by SG smoothing method and PCA, as the same as the previous two models. The division of training and testing data sets is also identical. The Gaussian RBF kernel is selected. Grid search and five-fold cross validation are applied to find out the best parameter pair \((C\text{ and }\sigma)\). All the algorithms and methods are implemented by MATLAB. LIBSVM toolbox (Chang and Lin, 2011) and Genetic Algorithm toolbox are employed. The different number combinations of the first 25 PCs are investigated. The relation between the information of a PC and its sensitivity for classification task is examined, as well as the interaction with other PCs in input subset. In this study, the GA-SVM model is settled as followings: the population size (Pop): 200; maximum generation: 100; probability of crossover: 0.6; probability of mutation: 0.01; the strategy of
Table 5.2: The best individual PCs are selected using GA-SVM model based on SG smoothed tobacco NIR spectra. $N_f$ is the number of inputs

<table>
<thead>
<tr>
<th>$N_f$</th>
<th>Input number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>0</td>
</tr>
</tbody>
</table>

68
Table 5.3: The prediction accuracy of GA-SVM and SVM for different number of inputs, and the corresponding information amount based on testing set with SG smoothed tobacco NIR spectra. \( N_f \) is the number of inputs; \( P_a \) is prediction accuracy; and \( I_m \) is the amount of information.

<table>
<thead>
<tr>
<th>( N_f )</th>
<th>GA-SVM</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_a ) (%)</td>
<td>( I_m ) (%)</td>
<td>( P_a ) (%)</td>
</tr>
<tr>
<td>6</td>
<td>72.7</td>
<td>72.42</td>
</tr>
<tr>
<td>8</td>
<td>75.8</td>
<td>74.12</td>
</tr>
<tr>
<td>10</td>
<td>75.8</td>
<td>74.66</td>
</tr>
<tr>
<td>12</td>
<td>74.2</td>
<td>74.18</td>
</tr>
<tr>
<td>14</td>
<td>83.3</td>
<td>98.78</td>
</tr>
<tr>
<td>16</td>
<td>78.8</td>
<td>99.84</td>
</tr>
</tbody>
</table>

parents selection is biased roulette method; one point crossover and uniform mutation are adopted.

The selected PCs are shown in Table 5.2 when the number of input is 6, 8, 10, 12, 14 and 16, separately. The columns depict the selected PCs for different number of inputs. The selected PCs are indicated using the value of 1s, and the unselected ones are marked using 0s. This experiment for each fixed number of inputs is repeated 3 times and the results are rather stable. It can be observed that the first, sixth, seventh, eighth, tenth and eleventh PCs are significant factors, which are selected regardless how many inputs are, as indicated in Table 5.2. But along with the increase of input number, the selection of other PCs is based on their interaction rather than how much information the PCs possessed.

The prediction accuracy (\( P_a \)) of GA-SVM using input subset selected by GA are shown in Table 5.3, as well as that of SVM model using the first corresponding number
of PCs. The corresponding information amount \( I_m \) of inputs is also listed. The GA-SVM has the better performance for almost all the selected number of inputs in terms of prediction accuracy on testing set. When 14 inputs are selected, the prediction accuracy is up to 83.3\%, which is superior to the best performance 80.3\% of SVM model. Meanwhile, by conducting the feature selection, GA-SVM studies the relationship between growing areas and tobacco NIR spectra based on sensitive PCs, rather than only based on the ones having more information. For the same number of inputs, the input subset of GA-SVM has less information than that of SVM, but makes the model achieve much better performance. Figure 5.4 depicts the average and the best fitness of GA-SVM model along with the generations for training.

In order to give a comprehensive evaluation for the two growing area classifiers, the assessment criteria of \( \gamma_{TP}, \gamma_{TN}, \gamma_{PP} \) and \( \gamma \) each class are also examined when the number of input is 14. The results are shown in Table 5.4. GA-SVM outperforms
Table 5.4: The identification capability of GA-SVM for each class are analyzed in terms of $\gamma_{TP}$, $\gamma_{TN}$, $\gamma_{PP}$, and $\gamma$ based on testing set with tobacco NIR spectra using SG smoothing method. The input number is 14; Class 1: the north of Guizhou Province in China; Class 2: the middle of Guizhou Province in China; Class 3: the northwest of Guizhou Province in China; Class 4: the southwest of Guizhou Province in China; $\gamma_{TP}$: true positive rate; $\gamma_{TN}$: specificity or true negative rate; $\gamma_{PP}$: positive predictive rate; and $\gamma$: F1 score.

<table>
<thead>
<tr>
<th></th>
<th>Class 1: North</th>
<th></th>
<th>Class 2: Middle</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\gamma_{TP}$</td>
<td>$\gamma_{TN}$</td>
<td>$\gamma_{PP}$</td>
<td>$\gamma$</td>
</tr>
<tr>
<td>GA-SVM</td>
<td>0.92</td>
<td>1.00</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td>SVM</td>
<td>0.67</td>
<td>0.98</td>
<td>1.00</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Class 4: Southwest</td>
</tr>
<tr>
<td></td>
<td>$\gamma_{TP}$</td>
<td>$\gamma_{TN}$</td>
<td>$\gamma_{PP}$</td>
<td>$\gamma$</td>
</tr>
<tr>
<td>GA-SVM</td>
<td>0.67</td>
<td>0.93</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>SVM</td>
<td>0.75</td>
<td>0.94</td>
<td>0.71</td>
<td>0.75</td>
</tr>
</tbody>
</table>
the SVM model in identifying the tobacco leaves from the north and the southwest of Guizhou Province. It can be seen that for class 1, the $\gamma_{TP}$ and $\gamma$ of GA-SVM are 0.25 and 0.2 higher than that of SVM model, respectively. For tobacco leaves from the middle region, these two discriminators have almost the same performance. SVM can more correctly discriminate the tobacco leaves from the northwest region. Considering the overall evaluation, the GA-SVM model is preferred for application to tobacco growing area discrimination.

5.4 Discussion

In order to improve the performance of GA-SVM classification model for tobacco growing areas, it is important to employ proper parameters for the model setting. The parameters, including population size, crossover rate and mutation rate, are discussed in this section.

5.4.1 GA-SVM Model Setting

The solution representation is an important step for GA-SVM model. In our experiments, the feasible solution (feasible input subset) is encoded into positive integer ranged from 1 to 25. The length of the individual chromosome equals to the number of inputs. For instance, when the number of input is 6, the individual feasible solution is represented as six non-repeated integer value ranged from 1 to 25. Those integers stand for the corresponding PCs. The length of chromosome is six. In the GA-SVM classification model, the prediction accuracy by SVM classifier is adopted to measure the goodness of candidates. The initial population is randomly selected and the biased Roulette selection is employed for parents selection.

5.4.2 Population Size

The population size in GA is an important factor because it can affect the search speed and convergence rate. If the population size is too small, the GA can not
explore enough candidates in solution space, which may lead to the missing of good solutions. But too large population size would certainly lead to the increase of search complexity. There is no arithmetical method to decide the size of population in the literature, because the problems solved by GA are particularly different from the complexity and types. Generally, the trial method is applied to find the appropriate population size for the specific application. In this study, different population sizes are tested. Undergoing parents selection and genetic operations, the population size that makes the GA-SVM model reach best performance is selected at the final stage. The average and best fitness value with different population sizes is shown in Figure 5.5. It can be easily observed that the large population size makes the GA-SVM model achieve the optimal solution quicker. However, too large population size leads to the exponential increase of complexity. So the proper population size to be used in this study is 200.
5.4.3 Genetic Operators

In GA, the genetic operators (crossover and mutation) make the diversity of gene, which cause heuristically searching in the search space. The probabilities of crossover and mutation greatly influence the performance, as well as the convergence speed. Mutation is necessary to ensure that potential solutions are not lost. Generally, the probability of crossover is good to select within the range of \([0.5, 1]\), and the probability of mutation is usually less than 0.1. In this section, different crossover rate and mutation rate are tried in order to find out the proper values to keep the variety of searching space. The biased Roulette wheel is applied for parent selection. The one point crossover is adopted to reproduce the off-springs with better performance for GA-SVM model. Uniform mutation is employed to the off-springs to restore lost information to the population so that more fit individual may be created.

The relationship between the crossover rate and the best fitness of the population
Figure 5.7: The best fitness along with different mutation rate for GA-SVM classifier on training.

is explored in Figure 5.6. The best fitness value is increasing when the crossover rate increases, but decreases when it is over 0.6. This phenomenon reflects that the high crossover rate could produce more new populations to get more chance of better solution. However, performance degradation is observed when the crossover rate increases too big. Therefore, the crossover rate equals to 0.6 is selected.

The best fitness value along with the increase of generation for different mutation rate within [0.005, 0.01, 0.02, 0.03] are depicted in Figure 5.7. With the increase of mutation rate from 0.005, the best fitness value rises up. However, when the mutation rate is larger than 0.01, the best fitness decreases. That means when the mutation rate is too high, the GA-SVM model loses the good solution. Therefore mutation rate equals to 0.01 is finally selected.
5.5 Summary

Considering the influence of interactions between the inputs (PCs) on discrimination performance, genetic algorithm is applied to systematically investigate the optimal or sensitive input subset for discrimination problem. From the potentially 25 inputs (the first 25th PCs), a series of inputs are tried to evaluate the classification model. The prediction results demonstrate that the GA-SVM model possess superior classification performance and outperforms the previous methods, which just simply choose the first several PCs as the inputs. The parameters for GA-SVM model, including population size, crossover and mutation probability, are also discussed in order to improve the performance of classification model. From the reported results and discussions, it can be concluded that the tobacco growing area discrimination performance using soft computing classifiers can be enhanced by GA feature selection.
Chapter 6

Conclusions and Future Work

In this thesis, the problem of discriminating tobacco growing areas is investigated using three soft computing approaches based on NIR spectra. An ANN with a three-layer structure is firstly employed to explore the relationship between growing areas and corresponding tobacco NIR spectra. Then, an SVM model with Gaussian kernel is applied to make use of its ability of nonlinear classification. Finally, in conjunction with SVM, GA is used for optimal feature selection, so that the performance of classification model would be enhanced. The conventional statistical model, the MD, is developed as comparison. These models are evaluated mainly in terms of prediction accuracy. Furthermore, the classification performance of those three models for individual class is also assessed on the basis of three criteria: true positive rate, true negative rate and positive predictive value. F1 score is introduced as a more reliable evaluation for each model, so that the influence of training and testing dataset selection can be eliminated. The experimental results strongly demonstrate the effectiveness of the developed soft computing methods. It can be observed that the proposed classification models outperform the statistical model for all five evaluations.

Among the three investigated discrimination models, the experimental results show that GA-SVM model possesses the best discrimination performance, followed by SVM. Because the estimation principle of ANN is based on empirical risk minimization, it is less prone to achieving global minima, especially for complex classification
problems. Meanwhile, SVM model uses the structural risk minimization principle, which provides a unique hyperplane, so that the margin between two classes is maximized. As a result, SVM classifier effectively avoids the curse of dimensionality and alleviates overfitting problem. However, when applying ANN or SVM to tobacco growing areas discrimination problem, the input subsets are selected based on maximum information principle. This may not provide the optimal classification results, since the interactions among individual inputs are ignored. On the other hand, there is no evidence in the literature demonstrating the equivalence between the inputs’ information and their contribution to classification performance. Therefore, as an effective heuristic search approach, GA is employed to combine with SVM for classification performance enhancement. The GA-SVM can eliminate redundant and irrelevant information from input subsets by exploring the interactions among them with maximum classification accuracy. This leads to the best performance of GA-SVM than other two models. In addition, staying at the same prediction accuracy level, the reduced number of inputs decreases the complexity of GA-SVM model.

In NIR spectra de-noising, the SG smoothing method is employed to remove the spectral difference of original tobacco spectrum from the baseline drifts. The discussion is conducted among original SG smoothing, first-order, and second-order derivative methods, and original version is selected at last. PCA is applied to extract the features and to eliminate redundant information from NIR spectra, thereby improving the SNR. The effectiveness of the proposed approaches are demonstrated by testing on a tobacco dataset from the real world collection.

The future work would include following three aspects.

1. The database employed in this thesis is from a real-world environment, and is used for model development without representative sample screening. However, a representative dataset is critical for model development, especially in training process, one potential future work is to add a data pre-selection process to eliminate the samples that have negative effect on model construction. Suitable statistical method can be used for representative sample selection so that the prediction ability of the developed models can be improved.
2. For each sample in the tobacco dataset, noise exists in the NIR spectrum due to measurement error. Advanced de-noising methods can be adopted to facilitate the de-noising process. For example, the de-noising operators in spectrum domain, like wavelet, have been reported to eliminate noise effectively in many applications (Zhu et al., 2007; Lin et al., 2011).

3. The investigated classification problem in this thesis shares the similar essential with many other applications. The classification models are of great importance in those scenario. Because of the great generalization ability and robustness of the developed models shown in this thesis, they can be easily extended to other applications through training and testing process. For example, employing GA-SVM model into the cultivation area/quality/brand discrimination of agricultural products.
References


and recovery of the (Bio)chemical interesting variables in data analysis with support vector machine classification. *Analytical Chemistry* **82**(16), 7000–7007.


Maleki, M.R., L. Van Holm, H. Ramon, R. Merckx, J. De Baerdemaeker and A.M.


Upendar, J., C.P. Gupta, G.K. Singh and G. Ramakrishna (2010). PSO and ANN-


White, F.H., R.S. Pandeya and V.A. Dirks (1979). Correlation studies among and between agronomic, chemical, physical and smoke characteristics in flue-cured tobacco (nicotiana tabacum l.). *Canadian Journal of Plant Science* 59(1), 111–120.


Zhu, D., B. Ji, C. Meng, B. Shi, Z. Tu and Z. Qing (2007). Study of wavelet de-


Appendix A

The First 25 PCs of the Preprocessed Data and the Raw Data

In this thesis, the raw tobacco NIR spectrum data are preprocessed by SG smoothing method and PCA technique for de-noising and feature extraction, respectively. The results of PCA, which account much information of NIR spectra, are applied to to be the inputs of the developed classifiers.

In order to give an example of the preprocessed data, Table A.1 shows the first 25 PCs of 16 NIR spectra samples, which are evenly selected from 226 training set and from 66 testing set, respectively, i.e., there are 8 NIR spectra samples are from training set and other 8 are from testing set. From each class, two samples are randomly selected. The selected samples are processed by SG smoothing methods and PCA technique, and are normalized in the range of [-1, 1]. The values, 1, 2, 3 and 4, represent the four different areas. As comparison, the corresponding PCs of the raw data are shown in Table A.2. Due to the space limit, all numbers in Table A.1 and Table A.2 are rounded into the three digits in decimal.
Table A.1: The first 25 PCs of the tobacco NIR spectra, which are preprocessed by SG smoothing and PCA method.

<table>
<thead>
<tr>
<th>Training Label</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-0.009</td>
<td>-0.325</td>
<td>0.212</td>
<td>0.169</td>
<td>0.072</td>
<td>0.538</td>
<td>0.337</td>
<td>0.191</td>
<td>-0.268</td>
<td>0.046</td>
<td>0.086</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-0.451</td>
<td>0.084</td>
<td>-0.340</td>
<td>-0.221</td>
<td>-0.401</td>
<td>-0.820</td>
<td>-0.806</td>
<td>-0.221</td>
<td>0.641</td>
<td>0.518</td>
<td>-0.261</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>-0.324</td>
<td>0.015</td>
<td>0.206</td>
<td>0.525</td>
<td>-0.052</td>
<td>0.167</td>
<td>0.279</td>
<td>-0.093</td>
<td>-0.08</td>
<td>0.001</td>
<td>-0.082</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>-0.427</td>
<td>0.117</td>
<td>0.167</td>
<td>-0.092</td>
<td>-0.299</td>
<td>0.001</td>
<td>0.134</td>
<td>-0.113</td>
<td>-0.111</td>
<td>0.236</td>
<td>0.334</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.087</td>
<td>0.356</td>
<td>0.291</td>
<td>-0.265</td>
<td>-0.441</td>
<td>-0.643</td>
<td>0.141</td>
<td>-0.062</td>
<td>0.029</td>
<td>-0.487</td>
<td>0.062</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0.079</td>
<td>-0.399</td>
<td>0.124</td>
<td>0.266</td>
<td>-0.174</td>
<td>-0.009</td>
<td>-0.012</td>
<td>0.015</td>
<td>0.577</td>
<td>-0.364</td>
<td>-0.101</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>0.226</td>
<td>0.101</td>
<td>0.090</td>
<td>-0.153</td>
<td>-0.305</td>
<td>-0.890</td>
<td>0.240</td>
<td>0.161</td>
<td>-1.000</td>
<td>-0.286</td>
<td>-0.220</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>0.178</td>
<td>-0.340</td>
<td>0.113</td>
<td>-0.318</td>
<td>-0.058</td>
<td>0.034</td>
<td>-0.112</td>
<td>-0.322</td>
<td>-0.267</td>
<td>-0.704</td>
<td>0.023</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Testing Label</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-0.118</td>
<td>-0.118</td>
<td>0.166</td>
<td>-0.050</td>
<td>-0.055</td>
<td>0.202</td>
<td>0.078</td>
<td>0.054</td>
<td>-0.282</td>
<td>-0.164</td>
<td>-0.040</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-0.079</td>
<td>-0.278</td>
<td>0.393</td>
<td>0.007</td>
<td>0.032</td>
<td>0.593</td>
<td>0.276</td>
<td>0.159</td>
<td>-0.473</td>
<td>-0.117</td>
<td>-0.041</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.731</td>
<td>0.287</td>
<td>-0.233</td>
<td>0.258</td>
<td>0.591</td>
<td>0.189</td>
<td>-0.041</td>
<td>0.308</td>
<td>0.002</td>
<td>-0.087</td>
<td>-0.034</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>-0.207</td>
<td>-0.138</td>
<td>0.324</td>
<td>0.460</td>
<td>0.007</td>
<td>0.167</td>
<td>0.052</td>
<td>-0.363</td>
<td>-0.089</td>
<td>0.181</td>
<td>-0.039</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>-0.162</td>
<td>-0.141</td>
<td>-0.048</td>
<td>-0.081</td>
<td>0.097</td>
<td>0.018</td>
<td>0.285</td>
<td>-0.139</td>
<td>0.347</td>
<td>-0.301</td>
<td>-0.056</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0.092</td>
<td>0.189</td>
<td>-0.195</td>
<td>0.397</td>
<td>0.519</td>
<td>-0.118</td>
<td>-0.395</td>
<td>-0.769</td>
<td>-0.058</td>
<td>-0.196</td>
<td>0.036</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>0.227</td>
<td>-0.067</td>
<td>0.213</td>
<td>-0.090</td>
<td>-0.273</td>
<td>-0.607</td>
<td>0.145</td>
<td>0.780</td>
<td>-0.147</td>
<td>-0.272</td>
<td>-0.259</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>0.121</td>
<td>-0.197</td>
<td>-0.322</td>
<td>0.348</td>
<td>0.140</td>
<td>0.249</td>
<td>0.344</td>
<td>0.233</td>
<td>-0.079</td>
<td>0.421</td>
<td>-0.082</td>
</tr>
</tbody>
</table>
Table A.1: The first 25 PCs of the tobacco NIR spectra, which are preprocessed by SG smoothing and PCA method.
Table A.2: The first 25 PCs of the raw tobacco NIR spectra, which are only preprocessed by PCA method.

<table>
<thead>
<tr>
<th>Training Label</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-0.009</td>
<td>-0.337</td>
<td>0.229</td>
<td>0.209</td>
<td>0.054</td>
<td>-0.384</td>
<td>-0.360</td>
<td>0.289</td>
<td>-0.403</td>
<td>-0.073</td>
<td>-0.216</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-0.459</td>
<td>0.097</td>
<td>-0.317</td>
<td>-0.318</td>
<td>-0.482</td>
<td>0.059</td>
<td>0.765</td>
<td>-0.951</td>
<td>0.233</td>
<td>0.602</td>
<td>-0.465</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>-0.324</td>
<td>0.006</td>
<td>0.180</td>
<td>0.540</td>
<td>-0.095</td>
<td>-0.081</td>
<td>-0.129</td>
<td>0.219</td>
<td>-0.019</td>
<td>0.008</td>
<td>-0.253</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>-0.432</td>
<td>0.109</td>
<td>0.229</td>
<td>-0.084</td>
<td>-0.306</td>
<td>-0.211</td>
<td>0.102</td>
<td>0.142</td>
<td>-0.090</td>
<td>-0.159</td>
<td>-0.216</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.080</td>
<td>0.344</td>
<td>0.418</td>
<td>-0.254</td>
<td>-0.411</td>
<td>-0.005</td>
<td>0.683</td>
<td>0.338</td>
<td>0.143</td>
<td>-0.046</td>
<td>0.327</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0.081</td>
<td>-0.403</td>
<td>0.097</td>
<td>0.273</td>
<td>-0.188</td>
<td>-0.151</td>
<td>0.093</td>
<td>-0.031</td>
<td>-0.167</td>
<td>-0.011</td>
<td>-0.378</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>0.224</td>
<td>0.102</td>
<td>0.126</td>
<td>-0.172</td>
<td>-0.248</td>
<td>0.586</td>
<td>0.637</td>
<td>0.428</td>
<td>0.059</td>
<td>-0.489</td>
<td>0.906</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>0.178</td>
<td>-0.348</td>
<td>0.183</td>
<td>-0.195</td>
<td>-0.029</td>
<td>-0.311</td>
<td>0.141</td>
<td>0.103</td>
<td>0.351</td>
<td>0.051</td>
<td>0.499</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Testing</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-0.118</td>
<td>-0.119</td>
<td>0.157</td>
<td>-0.034</td>
<td>-0.028</td>
<td>0.184</td>
<td>-0.260</td>
<td>0.072</td>
<td>-0.014</td>
<td>-0.494</td>
<td>0.146</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-0.079</td>
<td>-0.289</td>
<td>0.401</td>
<td>0.069</td>
<td>0.069</td>
<td>-0.047</td>
<td>-0.550</td>
<td>0.292</td>
<td>-0.277</td>
<td>-0.620</td>
<td>-0.041</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>-0.725</td>
<td>0.292</td>
<td>-0.319</td>
<td>0.292</td>
<td>0.601</td>
<td>-0.081</td>
<td>-0.196</td>
<td>-0.119</td>
<td>-0.372</td>
<td>0.165</td>
<td>0.089</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>-0.207</td>
<td>-0.152</td>
<td>0.323</td>
<td>0.495</td>
<td>-0.047</td>
<td>-0.231</td>
<td>-0.033</td>
<td>0.030</td>
<td>0.378</td>
<td>0.638</td>
<td>0.019</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>-0.160</td>
<td>-0.139</td>
<td>-0.059</td>
<td>-0.074</td>
<td>0.128</td>
<td>0.132</td>
<td>-0.069</td>
<td>0.326</td>
<td>-0.032</td>
<td>0.064</td>
<td>-0.219</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>0.099</td>
<td>0.193</td>
<td>-0.273</td>
<td>0.412</td>
<td>0.521</td>
<td>-0.049</td>
<td>0.151</td>
<td>-0.279</td>
<td>0.766</td>
<td>-0.026</td>
<td>-0.109</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>0.224</td>
<td>-0.072</td>
<td>0.254</td>
<td>-0.094</td>
<td>-0.269</td>
<td>0.206</td>
<td>0.484</td>
<td>0.053</td>
<td>-0.706</td>
<td>0.186</td>
<td>0.454</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>0.128</td>
<td>-0.184</td>
<td>-0.425</td>
<td>0.306</td>
<td>0.132</td>
<td>0.361</td>
<td>-0.427</td>
<td>0.291</td>
<td>-0.203</td>
<td>-0.005</td>
<td>0.032</td>
</tr>
</tbody>
</table>
Table A.2: The first 25 PCs of the raw tobacco NIR spectra, which are only preprocessed by PCA method.

<table>
<thead>
<tr>
<th></th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>-0.830</td>
<td>0.010</td>
<td>-0.009</td>
<td>-0.157</td>
<td>0.468</td>
<td>0.033</td>
<td>0.200</td>
<td>0.312</td>
<td>0.379</td>
<td>-0.194</td>
<td>0.422</td>
<td>0.095</td>
<td>-0.099</td>
</tr>
<tr>
<td>14</td>
<td>-0.069</td>
<td>-0.159</td>
<td>-0.017</td>
<td>-0.329</td>
<td>0.218</td>
<td>0.089</td>
<td>0.199</td>
<td>0.209</td>
<td>0.029</td>
<td>-0.157</td>
<td>0.053</td>
<td>0.043</td>
<td>0.386</td>
</tr>
<tr>
<td>15</td>
<td>-0.282</td>
<td>0.136</td>
<td>-0.059</td>
<td>-0.329</td>
<td>0.296</td>
<td>-0.207</td>
<td>0.166</td>
<td>-0.044</td>
<td>0.237</td>
<td>-0.506</td>
<td>-0.306</td>
<td>0.244</td>
<td>-0.645</td>
</tr>
<tr>
<td>16</td>
<td>-0.203</td>
<td>-0.069</td>
<td>-0.236</td>
<td>0.125</td>
<td>-0.235</td>
<td>0.585</td>
<td>-0.061</td>
<td>0.299</td>
<td>-0.351</td>
<td>0.068</td>
<td>0.026</td>
<td>-0.164</td>
<td>0.084</td>
</tr>
<tr>
<td>17</td>
<td>0.389</td>
<td>0.077</td>
<td>0.0713</td>
<td>0.170</td>
<td>0.559</td>
<td>0.027</td>
<td>0.095</td>
<td>0.199</td>
<td>-0.559</td>
<td>-0.213</td>
<td>0.266</td>
<td>-0.3241</td>
<td>-0.038</td>
</tr>
<tr>
<td>18</td>
<td>1.000</td>
<td>-0.168</td>
<td>0.624</td>
<td>0.301</td>
<td>-0.221</td>
<td>0.253</td>
<td>-0.200</td>
<td>-0.075</td>
<td>-0.335</td>
<td>-0.362</td>
<td>-0.188</td>
<td>-0.716</td>
<td>-0.058</td>
</tr>
<tr>
<td>19</td>
<td>-0.436</td>
<td>0.047</td>
<td>0.148</td>
<td>-0.149</td>
<td>-0.064</td>
<td>0.012</td>
<td>0.200</td>
<td>-0.327</td>
<td>-0.017</td>
<td>0.718</td>
<td>0.403</td>
<td>-0.073</td>
<td>0.062</td>
</tr>
<tr>
<td>20</td>
<td>0.016</td>
<td>-0.129</td>
<td>0.251</td>
<td>-0.224</td>
<td>0.149</td>
<td>-0.164</td>
<td>0.024</td>
<td>0.189</td>
<td>0.434</td>
<td>0.654</td>
<td>-0.409</td>
<td>-0.315</td>
<td>-0.312</td>
</tr>
<tr>
<td>21</td>
<td>0.161</td>
<td>-0.398</td>
<td>-0.217</td>
<td>0.262</td>
<td>-0.545</td>
<td>0.345</td>
<td>0.178</td>
<td>-0.235</td>
<td>0.369</td>
<td>0.067</td>
<td>0.507</td>
<td>0.537</td>
<td>0.217</td>
</tr>
<tr>
<td>22</td>
<td>-0.323</td>
<td>0.310</td>
<td>0.163</td>
<td>-0.165</td>
<td>0.034</td>
<td>0.168</td>
<td>-0.045</td>
<td>0.304</td>
<td>0.055</td>
<td>0.006</td>
<td>0.290</td>
<td>-0.808</td>
<td>0.190</td>
</tr>
<tr>
<td>23</td>
<td>-0.213</td>
<td>-0.211</td>
<td>0.146</td>
<td>-0.073</td>
<td>-0.339</td>
<td>0.044</td>
<td>-0.002</td>
<td>0.323</td>
<td>-0.004</td>
<td>-0.139</td>
<td>-0.308</td>
<td>0.143</td>
<td>0.256</td>
</tr>
<tr>
<td>24</td>
<td>-0.589</td>
<td>-0.666</td>
<td>0.360</td>
<td>-0.208</td>
<td>-0.340</td>
<td>0.075</td>
<td>-0.061</td>
<td>-0.431</td>
<td>-0.461</td>
<td>-0.135</td>
<td>0.032</td>
<td>0.207</td>
<td>-0.609</td>
</tr>
<tr>
<td>25</td>
<td>0.494</td>
<td>-0.036</td>
<td>-0.224</td>
<td>-0.005</td>
<td>-0.109</td>
<td>0.176</td>
<td>-0.002</td>
<td>0.613</td>
<td>-0.456</td>
<td>-0.258</td>
<td>-0.009</td>
<td>-0.098</td>
<td>-0.272</td>
</tr>
<tr>
<td></td>
<td>-0.083</td>
<td>0.316</td>
<td>-0.438</td>
<td>-0.165</td>
<td>0.462</td>
<td>-0.418</td>
<td>0.079</td>
<td>-0.169</td>
<td>0.319</td>
<td>-0.059</td>
<td>-0.204</td>
<td>0.352</td>
<td>-0.376</td>
</tr>
<tr>
<td></td>
<td>-0.281</td>
<td>-0.133</td>
<td>0.107</td>
<td>-0.267</td>
<td>0.209</td>
<td>-0.364</td>
<td>-0.236</td>
<td>0.133</td>
<td>-0.873</td>
<td>0.328</td>
<td>0.118</td>
<td>0.138</td>
<td>-0.006</td>
</tr>
<tr>
<td></td>
<td>-0.234</td>
<td>-0.119</td>
<td>0.241</td>
<td>0.219</td>
<td>-0.197</td>
<td>-0.061</td>
<td>-0.039</td>
<td>0.012</td>
<td>-0.380</td>
<td>0.022</td>
<td>-0.198</td>
<td>0.281</td>
<td>0.156</td>
</tr>
</tbody>
</table>