DEVELOPMENT OF ADVANCED STATISTICAL METHODS FOR MULTIVARIATE CLASSIFICATION

MARIO CARDENAS JR

Master's Program in Physics

APPROVED:

Marian Manciu, Ph.D., Chair

Felicia Manciu, Ph.D.

Giulio Francia, Ph.D.

Stephen L. Crites, Jr., Ph.D.
Dean of the Graduate School

DEVELOPMENT OF ADVANCED STATISTICAL METHODS FOR MULTIVARIATE CLASSIFICATION

by

MARIO CARDENAS JR

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Physics

THE UNIVERSITY OF TEXAS AT EL PASO

May 2020

Table of Contents

Table of Contents iv

List of Tables v

List of Figures vi

Chapter 1: Introduction 1

Chapter 2: Multivariate Methods 4

2.1 Principal Component Analysis 4

2.2 Linear Discriminant Analysis 12

2.3 Linear Support Vector Machines 16

2.4 Random Forest 19

Chapter 3: Proposed Classification Method 22

3.1 Overview of Method 22

3.2 Experimental Design 23

3.3 Experimental Results 25

Chapter 4: Accurate classification of normal/disease sample using Raman Confocal Microscopy 29

4.1 Motivation 29

4.2 Preliminary data analysis 30

4.3. Results and Discussion 31

Chapter 5: Conclusions 42

References 43

Vita 45

List of Tables

Table 3.1: GDS4336 dataset 25
Table 4.1. Confusion matrix for single spectrum LDA classification (4 variables). 38
Table 4. 2. Confusion matrix for single spectrum LSVM classification (~300 variables). 38
Table 4.3. Confusion matrix for 11 spectra classification. 41

List of Figures

Figure 2.1: Plot of words found in a dictionary. Here the number of words has been graphed vs. the number of lines in the definition of said words. 7
Figure 2.2: Example of a Scree graph found in Jolliffe, 2002. 9
Figure 2.3: Plot of 50 observation vs two component 10
Figure 2.4: Plot of 50 observations with respect to principal components 11
Figure 2.5: Is an example of two classes exhibiting an overlap along the axes X_1 and X_2 axis, but full separation along the discriminant function, [18]. 15
Figure 2.7: Example of a decision tree [19]. 20
Figure 3.1: Plot containing the first two principal components for $i = 1$ 26
Figure 3.1: Plot containing the first two principal components for $i = 4$ 27
Figure 3.1: Plot containing the first two principal components for $i = 10$ 28
Figure 4.1: Integral Raman spectra (each averaged over 22500 spectra with the background individually subtracted) of 7 samples (three "Normal" and four "ROD") 32
Figure 4.2a: Ratio 2 of the individual Raman spectra vs. Ratio1. 33
Figure 4.2b: Ratio 4 of the individual Raman spectra vs. Ratio3. 34
Figure 4.3a: Values of Ratio 2 vs. Ratio 1 for individual spectra. 35
Figure 4.3b: Values of Ratio 4 vs. Ratio 3 for individual spectra. 36
Figure 4.4: Distribution of scores for "Normal" and "ROD" individual spectra; classification assumes a score of less than one for "Normal" and larger than one for "ROD" samples. The complete details of the confusion matrix and related parameters for individual spectra are provided in Table 4.1. 37
Figure 4.5: Probability of type I and type II errors vs. the number of randomly chosen spectra employed in classification. 40

Chapter 1: Introduction

Research into the diagnosis and treatment of a variety of diseases has been a longstanding area of interest. The use of biomarkers is one way in which this area is explored. Biomarkers have several potential clinical applications some of which include treatment response predictions, risk assessment, and class identification [3]. A biomarker is defined as "any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome of disease" [5]. Biomarkers are also used to track disease progression, serve as surrogate clinical endpoints, and measure and detect the effects of a drug [1,2]. Thus, biomarkers have the potential to improve the early detection of a disease present in a subject and lead to an improved life expectancy. As an example, the early detection of the presence of a disease may lead to shorter treatment response and may ultimately result in a lower mortality rate [2]. Another example of where lower mortality can occur is with risk assessment. If a risk assessment can be made based on biomarkers, a patient can take certain actions to reduce their risk of developing a particular disease by taking preventive measures, like that of a lifestyle changes [3].

In biospectroscopy one can make use of multivariate and univariate methods for biomarker identification [2]. The field of biospectroscopy provides a wide range of spectra data via techniques like IR Spectroscopy, Fourier-transform IR, and Raman Spectroscopy. Raman spectroscopy uses Raman scattering where one measures the vibrational energy of chemical bonds present in cell or tissue samples. Here, a beam of monochromatic light is directed at a sample of interest typically within the mid-Infra-red range (λ=5-25 μm). During this process, the incident photons are polarizing the present molecules' electron cloud and promote them into excited states. These states lie above the molecule's ground state and are considered unstable or short-lived. Because of the instability of this virtual state, the incident photons are quickly scattered (re-emitted) and then

captured using a detector. Whereas most of the scattering is elastic (and therefore carries no information about the scattering material), Raman spectroscopy is concerned with the inelastic scattering, which provides information about the chemical structure of the scatterer [1]. Raman Confocal Microscopy is an experimental imaging technique, which provides individual Raman spectra on a (usually) 250x250 map, where this spectra is employed for the image reconstruction. In this process and other similar techniques typically result in large data sets providing ample information regarding the molecular composition of the sample measured. For example, Raman measurement discussed in the last part of the thesis contains 250x250x1024 data points. To analyze such large data sets, one can turn to principal component analysis, linear discriminant analysis, and other multivariate analysis methods to aid in the extraction of information [3]. The combination of biospectroscopy and multivariate methods can result in cell type identification, biomarker identification, and other information regarding the measured sample(s). Methods like principal component analysis are of particular interest because of their dimensionality reduction capabilities which can reduce the computation power required used to analyze them.

With the help of microarray technology, gene expression profiles, like spectra data sets, can be used for disease classification. In fact, an important aspect of microarray analysis is cancer classification. This is because cancer may be a genetic disease, and so the analysis of gene mutations may lead to the identification of the gene(s) responsible for cancer [4]. Gene expressions data sets typically consist of low sample (observations) size in the order of tens and a high number of genes (variables), in the order of $\sim 10^4$. The dimensionality of the data sets presents researchers with a problem commonly referred to as the "curse of dimensionality". This can lead to data overfitting in microarray cancer classification [4].

Class prediction (classification) and feature selection are two important types of analysis employed when analyzing gene expressions. In feature or gene selection the analysis focuses on finding the most informative genes. To do this, three types of approaches can be used: the filter, wrapper, and hybrid approach. The filer approach uses techniques like that of Random Forest Ranking, where features are ranked based on decision trees. The wrapper approach typically involves bio-inspired algorithms like the Genetic Algorithm, Ant Colony Optimization, and others. Finally, the hybrid method combines the two mentioned approaches by reducing the features present in a data set followed by feature optimization of the reduced data set or subset. In classification, supervised learning techniques can be used by creating classifiers based on learning data sets. A classifier can then be used for class prediction when applied to other data sets not used in the training phase. Some examples of algorithms currently used include Support Vector Machines, Neural Networks, K Nearest Neighbor, and others [4].

In the following chapter, I will give a brief description of the commonly used supervised and unsupervised learning techniques which will serve as an overview of the novel method proposed in Chapter 3 for the classification of data sets. Chapter 4 details another novel alternative classification method, which was shown to provide an unprecedented accuracy for the classification of a particular disease, using a reduced set of data points [paper].

Chapter 2: Multivariate Methods

2.1 Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised statistical data analysis tool commonly used to process genomic datasets [2]. This is due to the high dimensionality of the matrices obtained when, for example, you measure the gene expression levels for a given cell or tissue sample. This dimensionality reduction technique used in the field of multivariate analysis where one can benefit from reducing the computational cost or time of analysis. PCA analysis a dataset via orthogonal transformations using Singular Value Decomposition (SVD).

Singular Value Decomposition provides us with a manner to calculate the Principal Components of a matrix without having to compute the covariance matrix [14]. Using Singular value decomposition to find the Principal Components of a matrix has been regarded as the best computational approach to finding Principal Components [12]. Finding Principal Components of a given matrix, as well as the covariance matrix, serve a very important role in multivariate classification which I will briefly describe in the following pages.

Here matrices will be denoted in bold upper-case letters and the transpose of said matrices with an upper case T superscript. Elements of a matrix will be denoted by the lowercase letter of the corresponding matrix uppercase letter with subscripts defining the element index. The matrix that is to be processed, matrix **A**, will be separated into three main matrices when subjected to Singular Value Decomposition:

$$\mathbf{A} = \mathbf{SVD}^T$$

Where **A** represents an (n x p) matrix composed of n observations and p variables. **S** is an (n x r) matrix referred to as the left singular matrix, **V** is an (r x r) matrix referred to as the diagonal matrix, and **D** is a (p x r) matrix referred to as the right singular matrix. The diagonal matrix **D** is