A SPARSE REPRESENTATION TECHNIQUE FOR CLASSIFICATION PROBLEMS

REINALDO SANCHEZ ARIAS

Program in Computational Science



APPROVED:

Miguel Argáez, Ph.D., Chair

Leticia Velázquez, Ph.D.

Rodrigo Romero, Ph.D.

Patricia Witherspoon, Ph.D.
Dean of the Graduate School

*A mi amada madre Alid,*

*mi padre Reinaldo, y mi hermano Juan Camilo*

*que son la luz de mi vida.*

# A SPARSE REPRESENTATION TECHNIQUE FOR CLASSIFICATION PROBLEMS

by

REINALDO SANCHEZ ARIAS



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

Program in Computational Science

THE UNIVERSITY OF TEXAS AT EL PASO

May 2011




# Acknowledgements

I would like to express my sincere gratitude to my mentors Dr. Argáez and Dr. Velázquez for their guidance and encouragement. I am thankful for the opportunity they have given to me to work with them and for their support and advice while living this experience away from home. I also want to acknowledge Dr. Rodrigo Romero for kindly accepting being part of my thesis committee, and for his valuable observations.

Thanks to my amazing friends Paula, Carlos, Anibal, Ron, Javier, Clemente, and all the other ones that have made me feel as part of their family. I am forever thankful for having the chance to share this adventure with them, for their patience, trust, help and encouragement. Thank you all.

Infinitely many thanks to my family for their endless support. To my dear mother Alid for her unlimited love and support in every moment; to my brother Juan Camilo for his motivation and kindness; and to my dad Reinaldo for always lighting my way and for taking care of us from Heaven. This is for you.

I also want to thank the Computational Science Program and Department of Mathematical Sciences professors and staff for all their hard work and dedication. This work was supported by the Department of the Army ARL Grant No. W911NF-07-2-0027.

# Abstract

In pattern recognition and machine learning, a classification problem refers to finding an algorithm for assigning a given input data into one of several categories. Many natural signals are sparse or compressible in the sense that they have short representations when expressed in a suitable basis. Motivated by the recent successful development of algorithms for sparse signal recovery, we apply the selective nature of sparse representation to perform classification. In order to find such sparse linear representation, we implement an $\ell_l$-minimization algorithm. This methodology overcomes the lack of robustness with respect to outliers. In contrast to other classification algorithms such as Support Vector Machines (SVM), no model selection dependence is involved. The minimization algorithm is a convex relaxation-like algorithm that has been proven to efficiently recover sparse signals. To study its performance, the proposed method is applied to six tumor gene expression datasets with a large number of features but few samples. Our numerical results compare favorably with various SVM methods. We also test the effectiveness of our classification algorithm in the Fisher's Iris dataset where a large number of samples but a small number of features are available.

Since the process and techniques for acquiring and analyzing data advance every day at high rates, we need to manage and analyze large amounts of data for several different scientific problems. Future work aims to study the performance of our classification method when dimensionality reduction techniques are applied, including feature selection and feature extraction strategies.

# Table of Contents

**Page**

Acknowledgements . . . . . v

Abstract . . . . . vi

Table of Contents . . . . . vii

**Chapter**

1 Introduction . . . . . 1

2 Sparse Solution of Linear Inverse Problems . . . . . 3

2.1 Problem Formulation . . . . . 3

2.2 Algorithmic Approaches . . . . . 4

2.3 $\ell_1$-minimization problem . . . . . 5

2.3.1 Restricted Isometry Property . . . . . 6

2.3.2 The Null Space Property . . . . . 6

2.3.3 Sparse Recovery Result . . . . . 6

2.4 Convex Relaxation Strategies . . . . . 7

2.4.1 Donoho, Saunders et al. - Basis Pursuit (BP) . . . . . 8

2.4.2 Boyd, Lustig et al. . . . . 9

2.4.3 Figueiredo, Wright et al. . . . . 10

2.4.4 Zhang et al. . . . . 11

2.4.5 M. Argáez et al. . . . . 12

3 Classification Problem . . . . . 13

3.1 Description . . . . . 13

3.2 Mathematical Formulation . . . . . 14

3.3 Discriminant Functions and Classifier . . . . . 16

3.4 Support Vector Machines (SVM) . . . . . 16

4 Solving the $\ell_1$ Optimization Problem . . . . . 18

4.1 Algorithmic Approach . . . . . 18
4.2 Algorithm Description and Methods . . . . . 19
5 Experiment Design and Numerical Experimentation . . . . . 22
5.1 $K$-fold cross validation . . . . . 23
5.2 Large number of features and few samples . . . . . 23
5.2.1 Dataset Description . . . . . 24
5.2.2 Numerical Results . . . . . 25
5.3 Large number of samples and few features . . . . . 28
5.3.1 Dataset Description . . . . . 28
5.3.2 Numerical Results . . . . . 30
6 Future Research and Conclusions . . . . . 32
6.1 Sparse Representation Capabilities . . . . . 32
6.2 Further Research . . . . . 32
6.2.1 Dimensionality Reduction . . . . . 33
6.2.2 Sparse Representation Technique Alternative . . . . . 35
References . . . . . 38
Curriculum Vitae . . . . . 42

# Chapter 1

# Introduction

Several engineering and science applications involve solving linear inverse problems that are usually ill-conditioned and for which the use of regularization techniques is required to be able to propose useful solutions. Recently, regularization via *sparsity* constraints has become very popular, where we look for an approximate solution to a linear system of equations, with the requirement that it has as few nonzero components as possible. This kind of problems can be found in several applications in machine learning, image and signal processing, and coding and information theory among others. Moreover, it has been proven that sparse signals can effectively approximate compressible signals [4].

In machine learning and pattern recognition, the term "classification" refers to an algorithm/technique for assigning a given set of input data into one of a given number of categories. An example would be assigning a given email into "spam" or "non-spam" classes, or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.). An algorithm that implements classification is referred to as a classifier.

Many natural signals are sparse or compressible in the sense that they have short representations when expressed in a suitable basis. Motivated by the recent successful development of algorithms for sparse signal recovery [11, 17, 21, 26], we apply the selective nature of sparse representation to perform classification. Any test sample is represented in an overcomplete dictionary with the training sample as base elements. In case we have sufficient training samples available for each class; test samples can be expressed as a linear combination of only those training samples belonging to the same class, therefore providing a naturally sparse representation. In order to find the sparsest linear representation we propose an algorithm