

Supervised Linear Dimension Reduction

Wei Bian

Faculty of Engineering and Information Technology
University of Technology, Sydney

A thesis submitted for the degree of

Doctor of Philosophy

2012

Certificate of Authorship/Originality

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature

Acknowledgements

On having completed this thesis, I am especially thankful to my advisor Prof. Dacheng Tao, who had led me to an at one time unfamiliar area of academic research, and trusted me and gave me as much as possible freedom to pursue my own research interests. Dacheng has taught me how to think and study independently and how to solve a difficult scientific problem in flexible but rigorous ways. He has sacrificed much of his precious time for developing my academic research skills. He has also given me great help and support in life.

I am thankful to the group members I met in the Hong Kong Polytechnic University, Nanyang Technological University, and University of Technology, Sydney, including Tianyi Zhou, Bo Geng, Chao Zhang, Bo Xie, Yang Mu, and many others. I learned a lot from these smart people, and I was always inspired by the interesting and in-depth discussions with them. I enjoyed the wonderful atmosphere, being with them, of both academic research and daily life.

I am incredibly grateful to my mother for her generosity and encouragement. This thesis is definitely impossible to be completed without her constant support and understanding. I am also thankful to my friends who have accompanied me, though not always at my side, through the arduous journey of four and a half years.

Abstract

Supervised linear dimension reduction (SLDR) is one of the most effective methods for complexity reduction, which has been widely applied in pattern recognition, computer vision, information retrieval, and multimedia data processing. This thesis explores SLDR by enriching the theory of existing methods and by proposing new methods.

In the first part of this thesis, we present theoretical analysis of Fisher's linear discriminant analysis (LDA), one of the most representative methods for SLDR. 1) Classical asymptotic analysis of LDA is based on a fixed dimensionality, and thus does not apply in the case where the dimensionality and the training sample number are proportionally large. Besides, the classical result does not provide quantitative information on the performance of LDA. To address these limitations, we present an asymptotic generalization analysis of LDA, allowing both the dimensionality and the training sample number to be proportionally large, from which we principally obtain an asymptotic generalization bound that quantitatively describes the performance of LDA in terms of the dimensionality and the training sample number. 2) We study a new regularization method for LDA, termed the block-diagonal regularization. By partitioning variables into small groups and treating them independently, block-diagonal regularization effectively reduces the dimensionality to training sample number ratio and thus improves the generalization ability of LDA. We present a theoretical justification of the block-diagonally regularized LDA by investigating its approximation and sample errors. We show that the block-diagonally regularized LDA performs competitively compared to other types of regularized LDA, e.g., with the Tikhonov regularization and the banded regularization.

In the second part of this thesis, we propose two new methods for SLDR. 1) The first method is for parametric SLDR, termed max-min distance analysis (MMDA). MMDA optimizes the projection matrix by maximizing the minimum pairwise distance of all class pairs in the dimension reduced space. Thus, it duly considers the separation of all classes and overcomes the “class separation” problem of existing parametric SLDR methods that close class pairs tend to merge in the dimension reduced space. 2) The second method is for nonparametric SLDR, which uses minimizing the asymptotic nearest neighbor classification error (MNNE) as the criterion for optimizing the projection matrix. Theoretically, we compare MNNE with other criteria, e.g., maximizing mutual information (MMI) and minimizing Bhattacharyya bound. We show that MNNE is superior to these two criteria in terms of the closeness to the Bayes optimal criterion. Empirical studies show that the proposed methods, MMDA and MNNE, achieve state-of-the-art performance for parametric and nonparametric SLDR, respectively.

Contents

Contents	v
List of Figures	ix
List of Tables	xi
1 Introduction	1
1.1 Dimension Reduction: an Overview	1
1.2 Supervised Linear Dimension Reduction	3
1.2.1 Parametric Methods	4
1.2.2 Nonparametric Methods	5
1.2.3 Proxies of the Bayes Optimal Criterion	6
1.3 Contributions of This Thesis	7
1.3.1 Theoretical Analyses	7
1.3.1.1 Chapter 2	8
1.3.1.2 Chapter 3	8
1.3.2 Algorithmic Extensions	9
1.3.2.1 Chapter 4	9
1.3.2.2 Chapter 5	9
1.4 Notations	10
I Theoretical Analyses	11
2 Asymptotic Generalization Analysis of Linear Discriminant Analysis	12

2.1	Introduction	12
2.2	LDA and its Asymptotic Optimality with Fixed Dimensionality	14
2.3	Asymptotic Generalization Analysis	17
2.3.1	Generalization Discrimination Power	18
2.3.2	Properties of the Auxiliary Estimates	20
2.3.2.1	Asymptotic Properties of $\widehat{\Sigma}_0$	20
2.3.2.2	Asymptotic Properties of \widehat{S}_0	21
2.3.3	Asymptotic Generalization Bound	22
2.4	Empirical Evaluations	25
2.4.1	On Synthetic Datasets	26
2.4.2	On Real Datasets	31
2.5	Discussions	35
2.6	Appendixes	37
2.6.1	Proof of Lemma 2.1	37
2.6.2	Proof of Lemma 2.2	39
2.6.3	Proof of Lemma 2.3	41
2.6.4	Proof of Lemma 2.4	45
2.6.5	Proof of Lemma 2.5	46
2.6.6	Proof of Lemma 2.6	48
2.6.7	Proof of Corollary 2.1	50
2.6.8	Proof of Corollary 2.2	51
2.6.9	Proof of Corollary 2.3	51
3	Block-Diagonal Regularization for Linear Discriminant Analysis 52	
3.1	Introduction	52
3.2	Block-Diagonally Regularized LDA	54
3.2.1	On the Approximation Error	56
3.2.2	On the Sample Error	57
3.2.3	Intuitive Variable Partitioning	59
3.3	Empirical Evaluations	61
3.3.1	On Variable Partitioning	61
3.3.2	On Comparison with Other Regularization Methods	64
3.4	Appendixes	66

3.4.1	Proof of Theorem 3.1	66
3.4.2	Proof of Theorem 3.2	68
II	Algorithmic Extensions	70
4	Max-Min Distance Analysis for Parametric SLDR	71
4.1	Introduction	71
4.2	Max-Min Distance Analysis	73
4.2.1	MMDA Criterion	73
4.2.2	Relationships with Other Criteria	75
4.2.2.1	MMDA vs. LDA, GMSS and HMSS	76
4.2.2.2	MMDA vs. the Bayes Optimal Criterion	78
4.3	Sequential Convex Relaxation	79
4.3.1	Global SDP Relaxation	79
4.3.2	Local SDP Relaxation	81
4.3.3	Iterative Local SDP Relaxation	83
4.4	Empirical Evaluations	85
4.4.1	Experiments on Synthetic Datasets	85
4.4.1.1	Data Generation and Evaluation Methods	85
4.4.1.2	Results and Analyses	86
4.4.2	Experiments on Face Recognition	91
5	Minimizing Asymptotic Nearest Neighbor Classification Error for Nonparametric SLDR	93
5.1	Introduction	93
5.2	Minimizing NN Error as a Proxy of the Bayes Optimal Criterion	95
5.2.1	Compared to Mutual Information	96
5.2.2	Compared to Bhattacharyya Bound	99
5.3	Algorithm	101
5.3.1	Kernel Density Estimation	101
5.3.2	Bandwidth Selection	102
5.3.3	Optimization on the Grassmann Manifold	103
5.4	Empirical Evaluations	105

CONTENTS

5.4.1	Bandwidth Selection	105
5.4.2	Experiments on UCI Machine Learning Repository	107
5.4.2.1	On the ImageSeg Dataset	107
5.4.2.2	On the TicTacToe Dataset	110
5.4.2.3	On More Datasets	112
5.5	Appendix	114
6	Conclusions	115
6.1	Summary of This Thesis	115
6.2	Future Works	116
6.2.1	SLDR for Structured Data	116
6.2.2	SLDR for Compressed Data	117
	Bibliography	118

List of Figures

2.1	Asymptotic generalization bound of LDA.	24
2.2	Evaluation of the asymptotic generalization bound on Example 1.	27
2.3	Evaluation of the asymptotic generalization bound on Example 2.	28
2.4	Evaluation of the asymptotic generalization bound on Example 3.	29
2.5	Evaluation of the asymptotic generalization bound on the Image-Seg dataset	32
2.6	Evaluation of the asymptotic generalization bound on the Image-Seg dataset	33
2.7	Evaluation of the asymptotic generalization bound on the OptDigits dataset	34
3.1	The sample error of the block-diagonally regularized LDA.	60
3.2	Evaluation of the block-diagonally regularized LDA on the Feret dataset	63
3.3	Evaluation of the block-diagonally regularized LDA on the Orl dataset	63
3.4	Evaluation of the block-diagonally regularized LDA on the Pie dataset	63
3.5	Evaluation of the block-diagonally regularized LDA on the Yale dataset	64

LIST OF FIGURES

3.6	Comparison of different regularization methods on the Feret, Orl, Pie, and Yale dataset. Block-D, the block-diagonally regularized LDA with deterministic variable partitioning; Block-R, the block-diagonally regularized LDA with random variable partitioning; Tikhonov, LDA with the Tikhonov regularization; and Banded, LDA with the banded regularization.	66
4.1	MMDA for three Gaussian distributions on the 2-dimensional space.	75
4.2	LDA, GMSS, and HMSS for three Gaussian distributions on the 2-dimensional space.	77
4.3	BLDA for three Gaussian distributions on the 2-dimensional space.	79
4.4	Evaluation on synthetic datasets by minimum pairwise distance. .	87
4.5	Evaluation on synthetic datasets by the average classification error rate and the standard deviation.	88
4.6	2-Dimensional data representation on the uniformly distributed dataset.	89
4.7	2-Dimensional data representation on the uniformly distributed dataset.	90
4.8	Evaluation on face recognition experiments	92
5.1	(a) The conditional NN error $r_{nn}(z)$ and half the posterior entropy $\frac{1}{2}H(Y Z = z)$ in binary classification. (b) Upper and lower bound of Bayes error P^* by the NN error P_{nn} and half the conditional entropy $\frac{1}{2}H(Y Z)$ in binary classification.	97
5.2	Evaluation of MNNE with respect to bandwidth selection.	106
5.3	Performance evaluation on the ImageSeg dataset. Lines denote average classification error rates and bars denote standard deviations.	108
5.4	2 dimensional visualization of data from the ImageSeg dataset. . .	109
5.5	Performance evaluation on the TicTacToe Dataset. Lines denote average classification error rates and bars denote standard deviations.	110
5.6	2 dimensional visualization of data from the TicTacToe dataset. .	111

List of Tables

5.1	Performance evaluation on the BalanceScale, Car, Ecoli, and Wine datasets: average classification error rate (standard deviation, best dimensionality)	113
-----	--	-----