DIMENSIONALITY REDUCTION FOR NON-VECTORIAL DATA

By

Yi Guo

B. Eng., M. Eng.

UNIVERSITY OF NEW ENGLAND

School of Science and Technology
University of New England

A thesis submitted for the degree of

PhD of Computer Science

of University of New England

January 2008

DECLARATION

I certify that the substance of this thesis has not already been submitted for any degree and is not currently being submitted for any other degree.

I certify that to the best of my knowledge, any help received in preparing this thesis, and all sources used, have been acknowledged in this thesis.

.... [redacted] ....

Acknowledgements

I would like to thank my principal supervisor Associate Professor Junbin Gao without whom the work of the thesis would have been impossible. Though he left University of New England in July 2005, his support and supervision stayed with me over the past three years. We had thousands of emails and telephone calls with ideas, discussions, encouragement and so on. I really appreciate his insights into the area of machine learning, especially in statistical machine learning, sharing with me his abundant knowledge of mathematics and computer science. I am grateful for his generosity and his passion for education. Most of our research arose from his brilliant ideas. He brought me into this wonderful field and, with his help, my understanding and perception of dimensionality reduction problems have been gradually formed.

I would also like to thank my co-supervisor Dr Paul Kwan. I have benefited greatly from his instruction in the methodology of research, from his assistance in writing papers and for his generous support of my research and everyday living. He always encouraged me to present our latest results and created great opportunities for me to introduce our algorithms, so that I can now stand before a platform with confidence. Sincere thanks to Dr Xiaodi Huang and the mathematics people in our school, Professor Yihong Du, Dr. Shusen Yan and Dr. Adam Harris. The discussions on mathematical issues with them helped me understand the existing methods more clearly. Many thanks to Professor A. S. M. Sajeev, Mr Norman Gaywood and Mr Saint Ashley. They provided me with extra computational support so that the large numbers of experiments could be performed in parallel. I would like to thank Dr Neil D. Lawrence for sharing his GPLVM code, which was partly used in our algorithms.

The research was funded by the International Postgraduate Research Scholarship (IPRS) and the University of New England Research Assistance (UNERA). I acknowledge the support from the research office. They made the research possible.

Finally, I want to thank sincerely my clear-sighted and strong-minded wife, Huiling Yang, who never took me away from my work and supported me all along. I am also very grateful to our parents and sisters. Their love always support me in completing my study.

Publications

Journal Papers

Yi Guo, Junbin Gao, and Paul W. Kwan. Visualization of protein structure relationships using constrained twin kernel embedding. *International Journal of Biomedical Science and Engineering*, submitted, 2008.

Yi Guo, Junbin Gao, and Paul W. Kwan. Twin kernel embedding. *IEEE Transaction of Pattern Analysis and Machine Intelligence*, to appear, 2008.

Referenced conference proceedings

Yi Guo, Junbin Gao, and Paul W. Kwan. Twin Kernel Embedding with back constraints. In *HPDM in ICDM*, 2007.

Yi Guo, Junbin Gao, and Paul W. Kwan. Learning optimal kernel from distance metric in twin kernel embedding for dimensionality reduction and visualization of fingerprints. In *International Conference on Advanced Data Mining and Applications*, 2007.

Yi Guo, Junbin Gao, Paul W. Kwan, and Kevin X. Hou. Visualization of protein structure relationships using twin kernel embedding. In *International Conference on Bioinformatics and Biomedical Engineering*, pages 1–4.

Yi Guo, Junbin Gao, and Paul W. Kwan. Learning out-of-sample mapping in non-vectorial data dimensionality reduction using constrained twin kernel embedding. In *International Conference on Machine Learning and Cybernetics*, 2007.

Yi Guo, Junbin Gao, and Paul W. Kwan. Visualization of non-vectorial data using twin kernel embedding. In *First International Workshop on Integrating AI and Data Mining*, volume 0, pages 11–17, 2006.

Yi Guo, Junbin Gao, and Paul W. Kwan. Kernel Laplacian eigenmaps for visualization of non-vectorial data. In *Lecture Notes on Artificial Intelligence*, volume 4304, pages 1179–1183, 2006.

Yi Guo and Junbin Gao. An integration of shape context and semigroup kernel in image classification. In *International Conference on Machine Learning and Cybernetics*, 2007.

Yi Guo and Junbin Gao. Manifolds of bag of pixels: A better representation for image recognition? In *2006 IEEE Conference on Systems, Man, and Cybernetics*, pages 3618–3622, Taipei, October 8-11 2006. IEEE, SMC society, IEEE.

Abstract

Dimensionality Reduction (DR) is an important step in many advanced applications such as exploratory data analysis and manifold learning. Its main goal is to discover the mappings of the input data in a much lower dimensional space or the so-called latent space without incurring unnecessary information loss.

In most existing DR algorithms, the main objective is to preserve relational structure among objects of the input space in the latent space by minimizing the inconsistency between two similarity/dissimilarity measures, one for the input data and the other for the embedded data, via a separate matching objective function. Based on this observation, a new dimensionality reduction method called Twin Kernel Embedding (TKE) is proposed. TKE addresses the problem of embedding non-vectorial data that is difficult for conventional methods in practice due to the lack of efficient vectorial representation. TKE solves this problem by minimizing the inconsistency between the similarity measures captured respectively by their kernel Gram matrices in the two spaces. This algorithm is proven to be effective on some real world data sets and has been successfully applied to protein visualization, kernel learning, fingerprint classification etc. TKE is further extended to novel samples by introducing the backward mapping which is incorporated into the objective function as either substitution of all embeddings or regularization terms which generate BCTKE and RCTKE algorithms respectively. Intuitively, the mapping function can be integrated into any other host DR algorithms as a solution to the so-called out-of-sample problem.

This thesis starts with the analysis of the existing DR methods. Based on the understanding of their common features, we will show the development of the TKE algorithms and the details

on their behaviors at length. We present not only a series of new algorithms, but also the aspects of the design including the origin of the ideas, observations and implementation. This research provides a stepping stone for new algorithmic design in Dimensionality Reduction.

Contents

List of Tables

List of Figures