Clustering algorithms for disease classification using mass spectrometry data

Vikram Chandramohan

A thesis submitted

in fulfillment of the requirements for the Degree of

Master of Science (Research)

School of Maths, Physics and Information Technology

James Cook University

May, 2008

Clustering algorithms for disease classification using mass spectrometry data

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Vikram Chandramohan
December 22, 2008

Clustering algorithms for disease classification using mass spectrometry data

Statement of Access

I, the undersigned, author of this work, understand that James Cook University will make this thesis available for use within the University library and, via the Australian Digital Thesis network, for use elsewhere.

I understand that, as an unpublished work, a thesis has significant protection under the Copyright Act and;

I do not wish to place any further restriction on access to this work.

Vikram Chandramohan
December 22, 2008

Acknowledgments

First, I wish to express my gratitude to my primary supervisor, A/Prof Tuan D. Pham for his supervison and his guidance during my research at James Cook University. I am also thankful for the advice, he gave me inspite his busy schedule and helping me in organizing the thesis.

I would like to take this oppurtunity to thank, A/Prof Bruce Litow, as my secondary supervisor. I would also like to thank Dr. Donggang Yu, Dominick Beck, Miriam Brandl and Peter Philips for their incredible support during my research at James Cook University. My gratitude is, to all lecturers and colleagues at the School of Maths, Physics and IT, James Cook University, Australia.

Many special thanks to my family members. I am indebted to my parents for the sacrifices they have made for me. I would like to thank my sister, who gave this encouragement to do my research in Australia.

Vita

Publications arising from this thesis include:

Chandramohan, V. and Tuan D. Pham (2008), Cancer classification using kernelized fuzzy c-means. In *9th WSEAS International Conference on FUZZY SYSTEMS.* Sofia, Bulgaria.

Chandramohan, V. and Tuan D. Pham (2007), Analysis of mass spectrometry data using kernel based fuzzy c-means. In *CMLS'07 International Symposium.*, Gold coast, Australia(Poster).

Clustering algorithms for disease classification using mass spectrometry data

Abstract

Besides the availability of genomic data, life-science researchers study proteomics in order to gain insight into the functions of cells by learning how proteins are expressed, processed, recycled, and their localization in cells. Proteomics are defined as the study of proteome which refers to the entire set of expressed proteins in a cell. In particular, functional proteomics involves the use of mass spectrometry (MS) to study the regulation, timing, and location of protein expression. It has been recently realized that the use of MS coupled with pattern recognition methodology can offer tremendous potential for the early detection of complex human diseases, and biomarker discovery. However, given the promising integration of several machine-learning methods and MS data in high-throughput proteomics, this biotechnology field still encounters several challenges in order to become a mature platform for clinical diagnostics and protein-based biomarker profiling. Some of the major challenges include noise filtering of MS data, feature extraction, feature reduction of MS datasets and selection of computational methods for MS-based classification. The main objective of this research is to classify diseases using MS data. First, we investigated feature extraction of MS data based on the fundamentals of signal processing such as the theory of linear predictive coding. Then we present an unsupervised kernel based fuzzy c-means (KFCM) approach, which is shown to be more robust to noise than fuzzy c-means (FCM) for mass spectrometry dataset. The KFCM is realized by modifying the original Euclidean distance in FCM by a kernel-induced distance. We evaluated the performance of our classification methods with some popular classification techniques such as support vector machine (SVM), principle component analysis (PCA), linear or quadratic discriminate analysis (LDA/QDA) and random forests.

Contents

List of Tables

List of Figures