

JCU ePrints

This file is part of the following reference:

Chandramohan, Vikram (2008) *Clustering algorithms for disease classification using mass spectrometry data.*
Masters (Research) thesis,
James Cook University.

Access to this file is available from:

<http://eprints.jcu.edu.au/2122>



**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 Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.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I understand that, as an unpublished work, a thesis has significant protection under the Copyright Act and;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 supervision 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

Acknowledgments	iv
Vita	v
Abstract	vi
List of Tables	x
List of Figures	xi
1 Introduction	1
1.1 Proteomics for disease classification	2
1.2 Background and Motivation	7
1.3 Aims and Objectives	7
1.4 Problem description	8
1.5 Thesis outline	9
2 Literature review	11
2.1 Introduction	11
2.2 Preprocessing	12
2.3 Feature extraction	13
2.3.1 Wavelets/Principle Component Analysis (PCA)	13
2.3.2 Genetic algorithms (GA)	14
2.3.3 Peak detection techniques	15
2.4 Feature selection	16
2.4.1 Filter method	16
2.4.2 Wrapper or embedded methods	17
2.4.3 Nearest shrunken centroid	18
2.5 Classifiers	19
2.5.1 Support vector machine	19
2.5.2 Self-organizing maps (SOM)	21
2.5.3 Linear or quadratic discriminate analysis	22

2.5.4	Centroid classification methods	23
2.5.5	Boosting and Random forests (RF)	25
2.5.6	Principle component analysis	26
2.6	Conclusion	27
3	Reflection on the Research Method	28
3.1	Research method construction	28
3.2	Data collection procedure	29
3.3	Data analysis procedure	29
3.4	My contribution to the research community	29
3.5	Overview of the framework	30
4	Feature extraction from MS data	33
4.1	Introduction	33
4.2	Linear predictive coding (LPC)	34
4.2.1	LPC model	35
4.3	Variograms	36
4.3.1	Introduction	36
4.3.2	Semi-variogram	37
4.4	Conclusion	38
5	Clustering algorithms	40
5.1	Fuzzy clustering	40
5.1.1	Introduction	40
5.1.2	Fuzzy c -means algorithm (FCM)	41
5.1.3	Conditions for optimality	41
5.1.4	The algorithm	42
5.1.5	Strength and weakness	42
5.2	Kernel based fuzzy c -means algorithm(KFCM)	43
5.2.1	Kernel methods	43
5.2.2	Kernel fuzzy c -means	47
5.2.3	Strength and weakness	48
5.3	Cluster validation	48
5.3.1	Introduction	48
5.3.2	FCM-based model selection algorithm	48
5.3.3	Validity indices	49
5.4	Exponent value validation	52
5.4.1	Introduction	52
5.4.2	Estimation of m value	53
5.5	Conclusion	55

6	Classification measure	57
6.1	Clustering-based decision rule	57
6.1.1	Introduction	57
6.1.2	Cepstral distortion measure	58
6.1.3	Likelihood distortion measure	59
6.2	Accuracy estimation	61
6.2.1	Cross-validation	61
7	Experiments on PC-H4 Dataset	64
7.1	Overview of datasets	64
7.1.1	Prostate Cancer:	64
7.1.2	Dataset description	65
7.2	Experiment setup	66
7.2.1	Parameters of study	68
7.2.2	Results	69
7.2.3	Comparison	71
8	Conclusions and Future work	74
8.1	Conclusions	74
8.2	Directions of future work	75
	Bibliography	77

List of Tables

7.1	Classification accuracy with $m = 2$ using cepstrum distortion measure	70
7.2	Classification accuracy with $m = 2$ using likelihood distortion measure	70
7.3	Classification accuracy with $m = 1.12$ using cepstrum distortion measure	71
7.4	Classification accuracy with $m = 1.12$ using likelihood distortion measure	71
7.5	Comparative study of cepstrum and likelihood distortion measure . .	72
7.6	Comparative study with other techniques	73

List of Figures

1.1	<i>Overview diagram of mass spectrometer</i>	3
1.2	<i>MALDI Process</i>	5
2.1	<i>Diagramatic representation of SVM</i>	20
2.2	<i>Graphical representation of clustering algorithm</i>	24
3.1	<i>Overview of the analysis pipeline</i>	31
4.1	<i>Spherical representation of semi-variogram</i>	38
4.2	<i>Comparison of the exponential and spherical models</i>	39
5.1	<i>Two-dimensional classification example, using the second-order monomials x_1^2, $\sqrt{2}x_1x_2$ and x_2^2 as features a separation in feature space can be found using hyperplane [69]</i>	44
7.1	<i>Example of mass spectrum in which the relative intensity is plotted against mass-to-charge ratio(m/z). The data in this example are from the FDA-NCI Clinical Proteomics Program Databank. Every point of the mass-spectra is a candidate feature and usually the spectra of a cancer patient differs from that of a healthy person.</i>	66
7.2	<i>Experimental and spherical semi-variogram representation of SELDI-MS samples</i>	67
7.3	<i>Graphical representation of PC and CE</i>	68
7.4	<i>Graphical representation of XB</i>	68