DOCTORAL THESIS

Computational methods for bioinformatics and image restoration

Liao, Haiyong

Date of Award:
2010

Computational Methods for Bioinformatics and Image Restoration

LIAO Haiyong

A thesis submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Principal Supervisor: Prof. Michael K. Ng

Hong Kong Baptist University

January 2010

Abstract

The main goal of Part I in this thesis is to study and analysis categorical data and Single Nucleotide Polymorphism (SNP) data in Human genome. To better understand the intrinsic structure in human genome, we propose and develop a unidimensional nonnegative scaling model to construct linkage disequilibrium (LD) maps in Chapter 2. The proposed constrained scaling model can be efficiently solved by transforming it to an unconstrained model. The method is implemented in PC Clusters at Hong Kong Baptist University. The LD maps are constructed for four populations from Hapmap data sets with chromosomes of several ten thousand single nucleotide polymorphisms (SNPs). The similarities and dissimilarities of the LD maps are studied and analyzed. Computational results are also reported to show the effectiveness of the method using parallel computation.

It is known that there are many categorical data existing in practice, for example in human genome analysis, market and customer analysis. However, many well-developed numerical methods cannot be applied to this kind of data directly. In Chapter 3, we investigate the problem of determining the number of clusters in the k-modes based categorical data clustering. We propose a new categorical data clustering algorithm with automatic selection of k. The new algorithm extends the k-modes clustering algorithm by introducing a penalty term to the objective function to make more clusters compete for objects. In the new objective function, we employ a regularization parameter to control the number of clusters in a clustering result. Instead of finding k directly, we choose a suitable value of regularization parameter such that the optimal clustering result is the most stable one among all the generated clustering results. Experimental results on synthetic data sets and real data sets are used to demonstrated the effectiveness of the proposed algorithm.

In Part II, some algorithms for image restoration are investigated. In the last decade total variation (TV) based methods has been being popular for image restoration because of edge-preserving property. There are many excellent TV-based algorithms proposed for image restoration. However, there is little research work on the choice of regularization parameter in TV based model. In Chapter 4, we try

different strategies to make the algorithms find the appropriate regularization parameter automatically. Numerical results are given to demonstrate the performance of the proposed algorithms. We also extend the proposed algorithm to blind image deconvolution (BID) problem in Chapter 5. Blind image deconvolution is a more challenging problem since the point spread function (PSF) is unknown. With Laplacian regularization for PSF and suitable imposed constraints on PSF, the proposed algorithm can recover both image and PSF successfully.

Contents