TOWARD BIG DATA ANALYSIS WORKSHOP

邁向巨量資料分析研討會摘要集

2015.06.05-06

巨量資料之矩陣視覺化

陳君厚

中央研究院統計科學研究所

摘要

視覺化 (Visualization) 與探索式資料分析 (Exploratory Data Analysis, EDA)在巨量資料的深層分析 (deep analytics) 將扮演重要的角色，但也有其待解決的問題與待開發的技術及環境。

本演講將先以幾個實例介紹矩陣視覺化 (Matrix Visualization: MV) 在象徵性資料分析 (Symbolic Data Analysis)的應用及方法論，與所開發之軟體模組 *i*GAP (Generalized Association Plots for Interval-Type Symbolic Data)。

接著我們試著以 *i*GAP 與 Hadoop 計算環境解決 MV 面對巨量資料時於關係矩陣計算、排序與呈現時所面臨的困難，提供一套可能可以看巨量資料之EDA工具。

APPLICATIONS OF STATISTICS TO CANCER GENOMICS BIG DATA

Yu-Chin Hsu[1], Jan-Gowth Chang[2] and Grace S. Shieh[1,*]

[1]Institute of Statistical Science, Academia Sinica, [2]Department of Laboratory Medicine, and Center of RNA Biology and Clinical Application, China Medical University Hospital and China Medical University, Taiwan.

Abstract

We first introduce two types of next generation-sequencing data, which are in the frontier of biotechnologies. Both DNA target-sequencing (illumine, 1000x) and exome-sequencing (illumine, 60x) are used to find mutations in genes of 10 local endometria cancer patients. There are 3.2×10^9 bps in human genome, and each aforementioned data set easily contains $10 \sim 50$ GBs. Here, to distinguish artifacts (due to PCR errors and sequencing) from true mutations in cancer is of interest. Although GATK (DePristo et al., 2011) has developed some statistics to filter artifacts from true variants (SNPs in particular), there is room for improvement for cancer genomics data.

We have developed two novel statistics, in addition to adopting four features of GATK, and applied it to aforementioned two NGS data sets of 10 endometria cancer patients. Performances of our methods and GATK are compared. We further applied our method, with trained thresholds for some statistics (by the endometria cancer NGS data), to exome-sequencing data of 12 thyroid cancer patients to predict true and false mutations; biological validations are on-going.

FACE RECOGNITION BY MPCA AND PIRE

Ting-Li Chen

Institute of Statistical Science, Academia Sinica

Abstract

Face recognition can be viewed as an image classification problem. Linear discriminant analysis (LDA) is a typical method for classification. One key step of LDA is to standardize the data. However, the standardization step which involves a matrix inversion can not be directly executed for this problem due to the dimension being larger than the sample size (large p small n). Sliced inverse regression (SIR) can be viewed as an extension to LDA, and partial inverse regression estimator (PIRE) adapts SIR approach to deal with the problem of large p small n. We apply PIRE on face recognition problem. Before the PIRE step, we process the images with multilinear principal component analysis (MPCA) to reduce the dimension first. From our simulation studies, MPCA+PIRE produces better results than the existing methods on several benchmark face recognition image datasets.

SOLVING FUSED GROUP LASSO PROBLEMS VIA BLOCK SPLITTING ALGORITHMS

Tso-Jung Yen
Institute of Statistical Science, Academia Sinica

Abstract

In this paper we propose a distributed optimization-based method for solving the fused group lasso problem, in which the penalty function is a sum of Euclidean distances between pairs of parameter vectors. As a result of that, the penalty function is not separable in terms of these parameter vectors. To make the penalty function separable, one common way is to introduce a set of auxiliary variables that represent the differences between pairs of parameter vectors. This representation can be seen as a linear operator on the joint vector of the parameter vectors, and the resulting augmented Lagrangian will have a coupling quadratic term involving the linear representation. Even though the linear representation is separable in terms of the parameter vectors, the coupling quadratic term is not. To make the coupling quadratic term separable, we further introduce a set of equality constraints that connect each parameter vector to a group of paired auxiliary variables. With these newly introduced equality constraints, we are able to derive a modified augmented Lagrangian that is separable either in terms of the parameter vectors or in terms of the paired auxiliary variables. This separable property further facilitates us to solve the fused group lasso problem by developing an iterative algorithm in that most tasks can be carried out independently in parallel. We evaluate performance of the parallel algorithm by carrying out fused group lasso estimation for regression models using simulated data sets. Our results show that the parallel algorithm has a massive advantage over its non-parallel counterpart in terms of computational time and memory usage. In addition, with additional steps in each iteration, the parallel algorithm can obtain parameter values almost identical to those obtained by the non-parallel algorithm.

Keywords: Fused lasso; Group lasso; Euclidean distance; Alternating direction method of multipliers; Block splitting algorithms.

APPLICATION OF SEQUENTIAL METHODS FOR ANALYZING BIG DATA FORM HEALTH CLOUD

Charlotte Wang[1,2], Yu-Tseng Chu[2], Yuan-Chin Ivan Chang[2], Mei-Shu Lai[1]

[1]Institute of Epidemiology and Preventive Medicine,
National Taiwan University

[2]Institute of Statistical Science, Academia Sinica

Abstract

With advance in information and biomedical technology, people have more chance to receive various medical examinations such as traditional biochemical tests and genetic tests, and a massive volume of health-related records will be recorded. In addition, governments, medical institutes and researchers also collect and set up datasets for diverse purposes such as the National Health Insurance database, all kinds of surveys, electronic health record, screening data, and epidemiological cohort databases. When researchers explore health-related issues based on combining a variety of databases, the volume, variety, and veracity of data become major challenges for analyzing data. Tackling these problems by reducing intractably huge among of data to more computationally manageable numbers will be a possible solution. Moreover, sequential methods, a widely used statistical hypothesis testing in clinical trials, do not assume a fixed sample size in advance and evaluate data as it is collected. Further sampling is stopped according to a pre-defined stopping rule as significant results are observed. In this study, we consider the characteristics and advantages of sequential methods and apply them to analyze big data from health cloud. Using National Health Insurance Database as an example, we demonstrated how to apply sequential methods for analyzing data and modified operating techniques for existing datasets rather than sequentially collected data in clinical trials.

Keywords: Big data, health cloud, sequential method.

DETECTION OF GENE-GENE INTERACTIONS USING MULTISTAGE SPARSE AND LOW-RANK REGRESSION

Hung Hung
Institute of Epidemiology & Preventive Medicine,
National Taiwan University

Abstract

Finding an efficient and computationally feasible approach to deal with the curse of high-dimensionality is a daunting challenge faced by modern biological science. The problem becomes even more severe when the interactions are the research focus. To improve the performance of statistical analyses, we propose a sparse and low-rank (SLR) screening based on the combination of a low-rank interaction model and the Lasso screening. SLR models the interaction effects using a low-rank matrix to achieve parsimonious parametrization. The low-rank model increases the efficiency of statistical inference and, hence, SLR screening is able to more accurately detect gene-gene interactions than conventional methods. Incorporation of SLR screening into the Screen-and-Clean approach (Wasserman and Roeder, 2009; Wu et al., 2010) is also discussed, which suffers less penalty from Boferroni correction, and is able to assign p-values for the identified variables in high dimensional model. We apply the proposed screening procedure to the Warfarin dosage study and the CoLaus study. The results suggest that the new procedure can identify main and interaction effects that would have been omitted by conventional screening methods.

BIG DATA CASE STUDY: MALWARES AND HONEYNET LOGS ANALYSIS

安興彥

國家高速網路與計算中心

Abstract

To resist the various threats from Internet, NCHC builds a honeynet on TANet. With the help of the honeynet we can observe the hackers who run the botnets to search and attack new victims. As our honeynet grows, the large mount of data cannot be easily handled by the relational databases. In this talk, I will share our experience of malwares and honeynet logs analysis.

SCORE-SCALE DECISION TREE FOR PAIRED COMPARISON DATA

Yu-Shan Shih

Department of Mathematics, National Chung Cheng University

Abstract

Paired comparison data are collected by comparing objects in couples. A new decision tree method for analyzing such data is proposed. It finds the preference patterns (ranks) of the subjects based on some covariates. A scoring system is implemented first and the total scores associated with each object for each subject are counted. The GUIDE regression tree for multi-responses is then applied to the score outcomes and the mean scores of the objects are used to give the preference scale of the subjects. This way of preference ranking is identical to that given by the Bradley-Terry model when the 2-1-0 scoring system is employed. The usefulness of our tree method is demonstrated through simulation and data analysis.

BAYESIAN VARIABLE SELECTION FOR MULTI-RESPONSE LINEAR REGRESSION

Ray-Bing Chen

Department of Statistics, National Cheng Kung University

Abstract

This paper studies the variable selection problem in high dimensional linear regression, where there are multiple response vectors, and they share the same or similar subsets of predictor variables to be selected from a large set of candidate variables. In the literature, this problem is called multi-task learning, support union recovery or simultaneous sparse coding in different contexts. In this paper, we propose a Bayesian method for solving this problem by introducing two nested sets of binary indicator variables. In the first set of indicator variables, each indicator is associated with a predictor variable or a regressor, indicating whether this variable is active for any of the response vectors. In the second set of indicator variables, each indicator is associated with both a predicator variable and a response vector, indicating whether this variable is active for the particular response vector. The problem of variable selection can then be solved by sampling from the posterior distributions of the two sets of indicator variables. We develop the Gibbs sampling algorithm for posterior sampling and demonstrate the performances of the proposed method for both simulated and real data sets.

ACTIVE LEARNING: SUBJECT AND VARIABLE SELECTION

Hsiang-Ling Hsu
Institute of Statistics, National University of Kaohsiung

Abstract

In this work, we are interested in active learning for the binary classification problems based on the logistic models. According to the concept of the active learning, we modify a sequential design procedure as a subject selection strategy to select the unlabeled points into training set. Furthermore, we implement a stepwise variable selection procedure to identify proper classification model based on the current training set. Thus the proposed active learning algorithm iteratively performs these two procedures to add more training points and then to update the classification model. Simulations are used to demonstrate the advantages of the proposed active learning algorithm. Comparing with the classification results obtained by the whole training set and the full model, the simulation results show that the proposed algorithm can obtain the almost the same classification rates with much fewer training points and a compact classification model. This is a joint work with Professor Yuan-chin Ivan Chang and Professor Ray-Bing Chen.

Keywords: active learning algorithm, backward elimination, D-efficiency criterion, forward selection, single feature optimization