

Minimum Redundancy Feature Selection and Extraction

Time: 2 hours.

Presenters:

Chris Ding, University of Texas at Arlington, chqding@uta.edu

Hanchuan Peng, Howard Hughes Medical Institute, pengh@janelia.hhmi.org

Goal and Scope.

This tutorial provides a step-by-step introduction to feature selection and feature extraction in bioinformatics. The emphasis will be on practical knowledge and the materials will be hands-on. We cover the state-of-art algorithms such as mRMR and ReliefF. Our goal is that after attending this tutorial, the attendee will be able to implement/use these state-of-art methods in their practical research. Classification methods such as SVM, Naive Bayes, LDA, are not the focus of the tutorial, but will be explained briefly.

In last 20 mins, we provide a broad perspective at research level, for both beginners and people familiar with the subject area. It is our view that although there exist a huge number of feature selection methods, simple, practical, efficient, and effective methods exist. This tutorial focus on these methods.

Introduction and outline

Out of thousands of genes in a DNA microarray data, only a small subset of them are relevant for the disease or genotype. How to select this small subset of genes/features/biomarkers is one of the most fundamental problems in bioinformatics, pattern recognition and machine learning.

Traditionally, most feature selection methods focus on relevance, i.e., features selected are most relevant to the discrete class variable, i.g., disease, genotype, molecular function etc. The relevance can be measured using mutual information, t-statistic, F-statistic, etc. Relevance can also be measured by ReliefR algorithm and a number of other methods. Given a fixed method, the relevance of each gene is computed. Genes are ranked by their relevance and top ranked (the most relevant) genes are selected in the traditional ranking method.

Minimum Redundancy feature selection.

A recurring problem with this popular ranking approach that the features thus selected are often redundant. A recent development is the Minimum Redundancy Maximum Relevance (mRMR) feature selection. Here the redundancy among the features in the subset are minimized while the relevance are maximized. An efficient algorithm computes the solution of these dual objective

optimization problem. We will also outline other approaches to deal with feature redundancy problem.

Filtering Approach of feature selection

In above feature selection methods, genes are selected independent of the classification method and thus are intrinsic to the data. These methods are called filter methods, because they resemble data filtering. These methods are popularly used in practice.

Wrapper Approach of feature selection

Another approach is called feature wrapper methods in which genes are selected in connection with their performance with a classification method (wrapped by the classification method). Here for a fixed classification method, we select a small subset which gives the best classification performance measured by leave-one-out or k-fold cross validation. There are several heuristic algorithms, such as the sequential forward search, backward search, or floating search etc. Feature wrapper method typically has high computational complexity, but the selected feature subset is often very small and gives high performance. A problem with this approach is the stability and sensitivity of the selected features — the selected subset changes substantially with different classification methods, and/or when additional data points are added (the generalization problem).

Unsupervised Feature Selection

All above feature selection methods assumes that class label information for each data point is given. This is called supervised feature selection. In practice, class labels for each data may not be available. For this unsupervised case, several methods have been developed. We will present gene shaving, two-way ordering, interactive feature filtering, etc. Biclustering can be viewed as simultaneous data clustering and feature selection.

Advanced concepts.

(1) conditional dependence, Markov blanket, feature relevance, etc. (2) L1-norm regularization and LASSO

Brief survey of major classification methods:

Support Vector Machine, K-Nearest Neighbor, Linear Discriminant Analysis, Naive Bayes, Logistic Regression.

Advances in class prediction

Several major new progress in class prediction, including Semi-supervised Learning: transduction, consistent field, Green's function.

Positive Samples Only Learning.

In many bioinformatics prediction problems such as structure prediction, we have only positive data samples, but no true negative data samples. In addition to positive samples, we have a pool of large unlabeled data, which are mostly negative (different from the positive data) and mixed with a small quantity of positive data. The task is to predict the positive sample from the pool of large unlabeled data. A good example is functional RNA gene prediction. A PSoL method will be

described that can handle this situation.

Feature extraction.

classification methods deal with feature vectors with the entries being of numerical values. From a DNA sequence of 4 letters, A,C,T,G, how can we extract numerical feature vectors? We will present several methods, (1) composition, (2) K-mers such as dimers, trimers, (also called stringer kernels), etc. (3) physical-chemical properties such as hydrophobicity, polarization.

Prerequisites.

Basic knowledge of bioinformatics and machine learning, especially class prediction.

Biography.

Chris Ding

Dr. Ding recently joined University of Texas at Arlington as a professor in computer science. He earned a Ph.D. from Columbia, worked at Caltech, Jet Propulsion Lab and Lawrence Berkeley National Lab. His research focus on bioinformatics and machine learning. In bioinformatics, he works on protein 3D structure prediction, gene selection, protein interaction, microarray. In machine learning, he works on spectral clustering, PCA, and other matrix based methods. He has given invited seminars on topics related to the tutorial at Stanford, Berkeley, Carnegie Mellon, U. Alberta, U. Hong Kong, IBM Research, Google Research. This tutorial grows out of his research experiences and previous tutorials on the subject areas.

Related Tutorials given by Chris Ding:

1. Bioinformatics and Machine Learning Methods, ICDM'03.
2. Bioinformatics and Bio-image Analysis, ICDM'05.
3. Spectral Clustering, ICML04
4. Principal Component Analysis and Matrix Factorizations for Learning, ICML05.
5. Spectral Clustering, ECML/PKDD 05
6. Spectral Clustering and nonnegative matrix factorization. IEEE Symp. Comp. Intel. Data Mining, April 2007.

Hanchuan Peng

Dr. Peng is currently a research scientist with Janelia Farm Research Campus, Howard Hughes Medical Institute. His research interests include bioimage informatics, computational biology, pattern recognition and machine learning, and neuronal networks. His recent work include building a 3D digital cell atlas for *C. elegans*, image mining of in situ gene expression patterns of fly embryos, minimum-Redundancy-Maximum-Relevance (mRMR) feature selection, and Bayesian structure-function analysis for brain images. His on-going projects include building a 3D high-resolution digital atlas for a fly brain and super-resolution image computing for PhotoActivated Localization Microscopy (PALM). Dr. Peng is active in the bioimage informatics/mining field. He was a program chair of the 2005 and 2006 International workshops on bioimage informatics held at Stanford and Santa Barbara, respectively, and a guest editor of a BMC Cell Biology supplement on bioimage informatics published in July 2007.