2013 Presentations

Big Data Conference

4-16-2013

Predicting Survival Probability Based on Gene Expression Levels

Daniel Frobish
Grand Valley State University, frobishd@gvsu.edu

Dan Frobish (Department of Statistics)

Predicting Survival Probability Based on Gene Expression Levels

Predicting Survival from Microarray Gene Expression Data

Dan Frobish
Department of Statistics

April 26, 2013

What is the goal?

- We want to build a statistical model that can be used to predict a patient’s survival probability as a function of time, based on his/her gene expression profile
- Examples
  - Patient A’s predicted median survival time is 4 years
  - Patient B’s predicted survival rate (probability) at 5 years is 40%

Why is this “big data”?

- Many, many variables (columns), one for each gene
- Many more columns than rows presents a problem
  - Typical data set might have 50,000 or more columns and maybe only 100 rows
- Because of the high dimension issue, typical modeling strategies are useless
- Columns are also often correlated with each other, which can cause problems

What to do about this?

- Dimension reduction
- Reduce the number of columns down to a manageable size, with respect to the sample size
- Most dimension reduction methods have built-in ways of dealing with correlation between the columns
- Many different methods have been proposed, so it is necessary to compare them, in terms of predictive ability

Kinds of dimension reduction

- The three types I am studying are (there are others)
  - Principal components based methods (PC)
  - Partial least squares methods (PLS)
  - Random forest methods (RF)
- PC and PLS try to find “optimal” linear combinations (weighted averages) of the columns to form “principle predictors”
- RF goal is to partition the input variables (gene expressions) recursively to create survival trees, and then average over many trees to create a forest

Summarizing

- Goal is to predict an outcome of interest (e.g. survival), when the number of explanatory variables is much bigger than sample size
- Methods discussed here are applicable outside of predicting survival
- There is no reason why the inputs to the model have to be gene expression levels
- Goal of my research is to compare these dimension reduction methods to see which performs better in terms of prediction under various conditions