Identifying Genes and Interactions through a Forest Approach

Heping Zhang

And Xiang Chen, Ching-Ti Liu, Meizhuo Zhang

Presented at BIRS workshop, Canada

Outline

Introduction to trees and forests
Simulation studies
Genetic variants associated with age-related macular degeneration

Complex Traits

Diseases that do not follow Mendelian Inheritance Pattern
Genetic factors, Environment factors, G-G and G-E interactions
Interactions: effects that deviate from the additive effects of single effects

Genetic variants have been identified for Age-related Macular Degeneration, Diabetes, Inflammatory Bowel Disorders, etc.

Data Structure

1  *
2  *
...
25  *
...
72  *

Recursive Partitioning

A technique to identify heterogeneity in the data and fit a simple model (such as constant or linear) locally, and this avoids pre-specifying a systematic component.

Leukemia Data

Source: http://www-genome.wi.mit.edu/cancer

Contents:
- 25 mRNA - acute myeloid leukemia (AML)
- 38 - B-cell acute lymphoblastic leukemia (B-ALL)
- 9 - T-cell acute lymphoblastic leukemia (T-ALL)
- 7,129 genes

Question: are the microarray data useful in classifying different types of leukemia?
A Classification Tree

Node Splitting

Click to see the diagram

Recursive Partitioning

3-D View

Genomics

Adequate sample size for parameters of interest. Often, we have hundreds or thousands of observations for the inference on a few parameters. We can try to settle an “optimal” model.

In this information age, we have more and more variables but the access to the number of study subjects remains the same. One model can no longer provide an adequate summary of the information.

Forests

To identify a constellation of models that collectively help us understand the data.

For example, we can select and rank the genes whose expressions show a great promise of classifying tumor cells.
Bagging (Bootstrap Aggregating)

- Cancer Normal
- High
- Low

A random tree

A Random Forest

Choose 20 best splits

For the highlighted daughter nodes, we choose three best splits

Choose 3 best splits for each daughter node

Frequencies of Genes in Forest

SNPs vs. Haplotypes

SNPs
- Directly observed
- No uncertainty
- Less informative

Haplotypes
- Inferred from SNPs
- Uncertain
- More informative
- Forest approaches

Forest Forming Scheme

Haplotype Frequency Estimation

Existing haplotype frequency estimation software that output a set of haplotype pairs with corresponding frequencies for each subject in each region.

We used SNPHAP (Clayton 2006)
Unphased to Phased Data

One unphased data expands to a large number of phased datasets.
In each region, an individual’s haplotype pair is randomly selected based on the estimated frequencies to account for the uncertainty of the haplotypes.
Haplotypes with low frequencies (~5-10%) should have some representations.

Trees Based on Phased Data

A tree is grown for each phased data set.
A random forest is formed for all phased data sets.

Inference from the Forest

Importance of haplotype \( h \) in tree \( T \)

\[
V_h = \sum_{s \in T: \text{split by } h} 2^{-L_s} \chi^2,
\]

where \( L_s \) is the depth of node \( t \) and \( \chi^2 \) is the value of the \( \chi^2 \) - test statistic of independence.

Significance Level

Distribution of the maximum haplotype importance under null hypothesis is determined by permutation.
First, disease status is permuted among study subjects while keeping the genome intact for all individuals.
Then, each of the permuted data set is treated in the same way as the original data.

Simulation Studies (2 loci)

- 300 cases and 300 controls
- Each region has 3 SNPs
- 12 interaction models from Knapp et. al. (1994) and Becker et. al. (2005)
- 2 additive models with background penetrance
- 3 scenarios
  - Neither region is in LD with the disease allele
  - One of the regions is in LD (\( D' = 0.5 \)) with the disease allele
  - Both regions are in LD (\( D' = 0.5 \)) with the disease allele

<table>
<thead>
<tr>
<th>Region 1</th>
<th>Region 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( f_{00} ) ( f_{01} ) ( f_{02} )</td>
</tr>
<tr>
<td>1</td>
<td>( f_{10} ) ( f_{11} ) ( f_{12} )</td>
</tr>
<tr>
<td>2</td>
<td>( f_{20} ) ( f_{21} ) ( f_{22} )</td>
</tr>
</tbody>
</table>
Simulation Studies (2 loci)

**Model:**

- Ep-1: f / f / f / f / 0 / 0 / 0 / 0 = 0.349
- Ep-2: f / f / f / f / 0 / 0 / 0 / 0 = 0.349
- Ep-3: f / f / f / f / 0 / 0 / 0 / 0 = 0.799
- Ep-4: f / f / f / f / 0 / 0 / 0 / 0 = 0.050
- Ep-5: f / f / f / f / 0 / 0 / 0 / 0 = 0.050
- Ep-6: f / f / f / f / 0 / 0 / 0 / 0 = 0.100
- Ad-1: f / f / f / f / 0 / 0 / 0 / 0 = 0.324
- Ad-2: f / f / f / f / 0 / 0 / 0 / 0 = 0.150

**Benchmark:**

FAMHAP software from Becker et al. (2005)

**Result for Scenario I**

- False positive rate:
  - Our method: < 1%
  - FAMHAP: > 5%

**Result for Scenario II**

- Identify the correct haplotype (Forest)
- Identify an incorrect haplotype (Forest)
- Identify SNPs in the correct region (FAMHAP)
- Identify SNPs in the neutral region (FAMHAP)

**Result for Scenario III**

- Identify at least one haplotype (Forest)
- Identify both haplotypes (Forest)
- Identify SNPs in at least one region (FAMHAP)
- Identify SNPs in both regions (FAMHAP)
**Real Case Study**

**Age-related macular degeneration (AMD)**

- Leading cause of vision loss in elderly
- Affects more than 1.75 million individuals in the United States
- Projected to about 3 million by 2020

**Klein et al. (2005)**

- Case-control (96 AMD cases, 50 controls)
- ~100,000 SNPs for each individual
- CFH gene identified

**Analysis Procedure**

- **RTREE program**
  - Each SNP is used as one covariate
  - Two SNPs identified as potentially associated with AMD (rs1329428 on chromosome 1 and rs10272438 on chromosome 7)

- **Hapview program**: LD block construction
  - 6-SNP block for rs1329428
  - 11-SNP block for rs10272438

**Result**

- **Two haplotypes are identified**
  - Most significant: ACTCCG in region 1 (p-value = 2e-6)
  - Identical to Klein et al. (2005)
  - Located in CFH gene

- Another significant haplotype: TCTGGACGACA, in region 2 (p-value = 0.0024)
  - Not reported before
  - Protective
  - Located in BBS9 gene
**Expected Frequencies**

- **Haplotype 1**
  - Case
  - Control
- **Haplotype 2**
  - Case
  - Control

**Remarks**

- A
- B

---

**Books**


**Trees in Genetic Studies**

- Zhang and Bonney (2000)
- Nelson et al. (2001)
- Bastone et al. (2004)
- Cook, Zee and Ridker (2004)

---

**References on Forests**