RF-ACE for uncovering nonlinear associations from heterogeneous cancer data

Timo Erkkilä
1st Annual TCGA symposium
National Harbor
Washington DC
Annotated Feature Matrix (AFM)

100-1000 samples
20000-50000 features:
- Categorical
- Numerical
- Binary
- String-literals
- Missing values
Problem: need algorithm for feature selection with heterogeneous data
Random Forest (RF)

Pros:
+ supports mixed-type data and missing values
+ predicted target can be of any type
+ no data transformations necessary
+ supports multivariate & nonlinear associations

Cons:
- importance score yields mere ranking of associations
- importance score is not normalized
- prediction performance could be better
- existing RF implementations often lack flexibility
Random Forest (RF)

Pros:
+ supports mixed-type data and missing values
+ predicted target can be of any type
+ no data transformations necessary
+ supports multivariate & nonlinear associations

Cons:
- importance score yields mere ranking of associations
- importance score is not normalized
- prediction performance could be better
- existing RF implementations often lack flexibility
RF-ACE
(Random Forests with Artificial Contrast Ensembles)

• RF implementation with added flexibility
  – support for string literals and various data formats
  – Easy interface with default parameter options
RF-ACE
(Random Forests with Artificial Contrast Ensembles)

• RF implementation with added flexibility
  – support for string literals and various data formats
  – Easy interface with default parameter options

• Normalized importance score
RF-ACE
(Random Forests with Artificial Contrast Ensembles)

• RF implementation with added flexibility
  – support for string literals and various data formats
  – Easy interface with default parameter options

• Normalized importance score

• Inclusion of statistical testing framework
  – p-values for associations
RF-ACE
(Random Forests with Artificial Contrast Ensembles)

• RF implementation with added flexibility
  – support for string literals and various data formats
  – Easy interface with default parameter options

• Normalized importance score

• Inclusion of statistical testing framework
  – p-values for associations

• Better predictive power with Gradient Boosting Trees
Pseudo-random example

- Find associations to PRAC in colorectal data
Pseudo-random example

• Find associations to PRAC in colorectal data

Reading file 'data.tsv', please wait... DONE

General configuration:
  nfeatures  = 39391
  nsamples   = 253 / 465 (45.5914 % missing)
  tree type  = Regression CART
  traindata  = data.tsv
  target     = N:GEXP:PRAC:chr17:44154161:44154753:- | index 17110
  associations = associations.PRAC.tsv
  testdata    = data.tsv
  predictions = predictions.PRAC.tsv
  optimized_split = NO

Random Forest configuration:
  --RF_ntrees = 1000
  --RF_mtry  = 198
  --RF_maxleaves = 100
  --RF_nodesize = 3

Significance analysis configuration:
  --RF_nperms = 20
  test type  = T-test
  --pthreshold = 0.1

Gradient boosting tree configuration for prediction:
  --GBT_ntrees = 1000
  --GBT_maxleaves = 6
  --GBT_shrinkage = 0.1
  --GBT_samplesize = 0.5

==> Uncovering associations... DONE
==> Filtering features... DONE, 19 / 39390 features (0.0402356 %) left
==> Predicting... DONE

190.49 seconds elapsed.

Association file 'associations.PRAC.tsv' created. Format:
TARGET PREDICTOR LCG10(P-VALUE) IMPORTANCE CORRELATION NSAMPLES

Prediction file 'predictions.PRAC.tsv' created. Format:
TARGET SAMPLE_ID DATA PREDICTION CONFIDENCE

RF-ACE completed successfully.
Top 3 associations for PRAC (out of 19 significant)

- HOXB13
- Anatomic organ subdivision
- Promoter methylation
Top 3 associations for PRAC (out of 19 significant)

HOXB13

Anatomic organ subdivision

Promoter methylation

"Core" features associated to PRAC

"AOS"  "Meth"  etc...

"HOXB13"

Gradient Boosting Trees
Builds a predictor for novel/missing data
Repeat the analysis for Tumor Stage

Lymphnode spread

Number of lymphnodes

PARADIGM ERCC4 act.

Etc.
Repeat the analysis for Tumor Stage

Lymphnode spread

Number of lymphnodes

PARADIGM ERCC4 act.

Low predictive power in low tumor stages?

Etc.
Repeat the analysis for Tumor Stage

Lymphnode spread  

Number of lymphnodes  

PARADIGM ERCC4 act.  

Etc.
Integrate data sources

Analyze with RF-ACE
- Inference of nonlinear associations was performed in a cluster of 1000 CPUs
- Each cancer type takes ~1-3 days of analysis

Store associations
- Each cancer type yields hundreds of megabytes of associations
- Associations are stored in a database for fast retrieval

Bring in other data
- MEDLINE literature mining
- Pathways from GO, KEGG, WikiPw, & PathwayCommons
- Reference genome

Explore cancer regulomes in a browser

http://explorer.cancerregulome.org
Summary

• RF-ACE combines good parts from various established algorithms
  – RF, GBT, ACE (Tuv et al., 2009)

• Generic & fast implementation
  – Suits well to TCGA data

• Novel aspects
  – P-values for associations (not available in RF)
  – GBT for prediction

http://code.google.com/p/rf-ace
Many Thanks!