Identifying Interaction Terms Using SP-FSR Algorithm

Yong Kai Wong

Contents

Introduction

- Identifying interactions via feature selection with SPSA in the context of supervised learning
- Why interaction?
  1. Interactions can be scientifically interesting and predictive, especially in genome data
  2. Main effects might not be significant but their interactions are. For example, coffee and milk
- Why SPSA? To scale up identification process

  Given p features, 2^p set of features. With 2-way pairwise interaction terms (excluding quadratic terms), there are possible $2^{p(p+1)/2}$ set of features

$$p + \binom{p}{2} = p + \frac{p!}{2!(p-2)!} = p + \frac{p(p-1)}{2} = \frac{p(p+1)}{2}$$

  For three-way interactions, it would be $p + \binom{p}{2} + \binom{p}{3}$
- Ultimate objective is to end with an incisive interpretation @Cox1984 [, p.4]
- Goals of this study:
  1. Identifying interactions via feature selection and ranking methods.
  2. Focus on linear pairwise/three-way interactions by ignoring quadratic terms and cubic terms. The polynomial terms are ignored because it is not applicable on non-numeric (nominal and ordered) variables.
  3. Compare performance of various feature selection methods in identifying interactions
     - Time to identify interactions
     - Any consistency in subset of features
  4. Introduce new functions for mlr and refine spFSR packages
- Goals of the minor thesis:
  1. To develop an algorithm to identifying interaction terms based on SP-FSR algorithm by focusing on pairwise interactions by ignoring quadratic terms.
  2. Compare performance of various interaction-identifying algorithms using case studies (and simulated data).
  3. Use basic network analysis to visualise pairwise interactions and identify the "central" variable.

Literature Review

Related Works of identifying Interaction Terms

@Cox1984 [, p.1] Aspects of interaction terms

- the definition of interaction and the reasons for its importance;
- the detection of interaction;
- the interpretation and application of interaction

@Cox1984 [, p.3] Types of explanatory variables

- x = treatment variable
- z = intrinsic variable, e.g. gender, age
- u = nonspecific variable, e.g. replicates, block designs

We write

$$E(Y) = f(x, z, u)$$

For two-way interaction terms, we can consider:

- $x \times x$
- $x \times z$
- $x \times u$

A "simple" example:

- $f(x_1, x_2) = f_1(x_1) + f_2(x_2)$
- $f(x_1, x_2) = f_1(x_1) + e^{-\beta x_2} f_2(x_1)$ for $x_2 \geq 0$

@Cox1984 [, p.10 to p.11] Detection methods:

- F test: joint hypotheses
- Graphical method: high dimensionality

Other paramtric methods:

- @LASSO: glmnet
- @Yangli: forward/backward selection with information criterion.
- @Noah: Permutation approach
- @Rajen: Backtracking

Feature Selection

- Feature selection allows interpretability
- See @kohavi for overview on feature selection and comparison to feature extraction
- See @relief, @spfs, @rfimportance, @chisquared as feature selection methods.

SPSA for Feature Selection

- See @spall for SPSA
- See @vural which applies BSPSA in feature selection
- See @zeren for BB method to extend SPSA for feature selection

Key Issues of Including Interaction Terms

- Computational cost due to a larger search space and possibility $p > n$ @HDcurse
- Overfitting issue @Yangli @Noah
- Interpretability is beyond the scope of this study. See @Cox2007
- Significant interaction of insignificant main effect features
- Remove main effect features, but keep interaction $\implies$ better result?
- Keep main effect features and interaction $\implies$ better result?

SP-FSR Algorithm for Identifying Interactions

2-Step SP-FSR Algorithm

1. Identify top m main-effect features selected by SF-FSR algorithm where $m \leq p$

2. Consider k interaction terms among m main-effect features where $k \leq \binom{m}{2}$
3. Form a network with m nodes corresponding to main effects and k edges representing pairwise interaction
4. Obtain the most central node

Backward SP-FSR Algorithm

1. Identify top m features including selected by SF-FSR algorithm where $m \leq p$
2. Select top interaction termss from top m features and identify the main effects
3. Add the main effects to the feature set and compute evaluation measure

Case Studies

Benchmarking measures:

1. Information criteria: AIC, BIC, EBIC
2. Evaluation measure: accuracy rate for binary logistic regression and mean squared error for linear regression
3. 5-fold Stratified CV

Ionosphere

SFFS

SFBS

LASSO

2-Step/Backward SP-FSR Algorithm

- 5-fold Stratified CV
- num.cv.reps.grad.avg = 3L
- 3 cross-validation repetitions for gradient averaging
- 3 cross-validation repetitions for feature subset evaluation.
- With $p = 33$, run $m = \{3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 33\} + \{0.1, 0.2, 0.3, 0.4, 0.5\}$ % of main effects
- $n = 351$

Boston Housing

SFFS

SFBS

LASSO

2-Step/Backward SP-FSR Algorithm

- 5-fold CV
- num.cv.reps.grad.avg = 3L
- 3 cross-validation repetitions for gradient averaging
- 3 cross-validation repetitions for feature subset evaluation.
- With $p = 13$, run $m = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13\} + \{0.1, 0.2, 0.3, 0.4, 0.5\}$ % of main effects

Future works

Other considerations

- Other classifiers
- Incorporate quadratic terms and three-way interactions
- Mixed data
- Removing small columns say user has predefined quadratic terms.
- Non-bernoulli but symmetric distribution

Caveats:

- No mixed effects model
- GAM on interactions
- Linearity and non-linearity

Methods

- spsa methods
- network plot among terms

References