MATH1332: Minor Thesis Presentation

Identifying Optimal Set of Pairwise Interaction Terms by SP-FSR Algorithm and Empirical Comparison with Other Methods

Yong Kai Wong

30 May 2018

Background

- A motivating example: do gene interactions help predict cancer type?
- How to determine interactions in high dimensions *optimally*?
- Focus on pairwise interaction terms (no quadratic terms)
- Given p features with 2-way pairwise interaction terms, the number of possible features is

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

- SP-FSR algorithm [Aksakalli and Malekipirbazari, 2016] shows excellent performance in feature selection. Can we utilise it to identify interaction terms?

Organisation of the presentation

1. Basic terminology
2. Earlier and current methods to identify pairwise interaction terms
3. SP-FSR algorithm for interaction identification
4. Experimental setup and results
5. Discussion
6. Conclusion

Notations and terminology

- Y: response feature
- X_j: explanatory feature j for $j = 1, 2, ...p$
- Model formulation:

$$g(Y) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \sum_{i<j} \beta_{i:j} X_j X_i$$

- A precise definition [Lim and Hastie, 2018, p.1]:

"When a function $f(x_1, x_2)$ cannot be expressed as $h_1(x_1) + h_2(x_2)$ for some functions h_1 and h_2, we say that there is an interaction in f between x_1 and x_2."

- Introduction of "hierarchy"

Terminology: hierarchy

Lim and Hastie [2018] define:

Hierarchy	Description
Strong	Interactions are only among pairs of nonzero main effects
Weak	Each interaction has only one of its main effects present
Anti-hierarchical	Interactions are only among pairs of main effects that are not present
Pure interaction	No main effects present; only interactions

Terminology: an example of hierarchy

Consider three explanatory features: $\mathbf{X} = \{X_1, X_2, X_3\}$

1. Strong hierarchy: $\{X_1, X_2, X_1X_2\}$
2. Weak hierarchy: $\{X_1, X_1X_2, X_1X_3\}$
3. Anti-hierarchical: $\{X_2, X_1X_3\}$
4. Pure interaction: $\{X_2X_3, X_1X_3, X_1X_2\}$

In practice, how do we know? How can we detect?

Main methods to identify pairwise interaction terms

1. Statistical hypothesis test [Cox, 1984]
2. Regularisation, e.g. LASSO
3. Wrapper Feature Selection [Kohavi and John, 1997]

Statistical hypothesis test: an example

Considers two models:

1. $g(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
2. $g(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1:2} X_1 X_2$

Run hypothesis on test on $\beta_{1:2} = 0$.

Regularisation

- Let $l(y_i; \beta)$ be the **negative** log-likelihood contribution
- Elastic-net [Friedman et al., 2010]

$$\arg\min_{\beta} \frac{1}{n} \sum_{i=i}^{n} l(y_i; \beta) + \lambda[(1-\alpha) \parallel \beta \parallel_2^2 /2 + \alpha \parallel \beta \parallel_1]$$

- `glinternet` [Lim and Hastie, 2018]: group-based LASSO ($\alpha = 1$) by imposing additional constraints on $\beta_{k:j}$ and β_j

Wrapper feature selection

Extending feature selection by including interaction terms

1. SFFS: Sequential (Floating) Forward Selection [Pudil et al., 1994]
2. ~~SFBS: Sequential (Floating) Backward Selection~~[1994]
3. GA: Genetic Algorithm [Siedlecki and Sklansky, 2011]
4. SP-FSR [2016]

SP-FSR algorithm

- Introduced by Aksakalli and Malekipirbazari [2016]
- Based on Simultaneous Perturbation Stochastic Approximation [Spall, 1992]
- Refined by Yenice et al. [2018]
- Pseudo gradient descent method on the loss function
- `spFSR` package is now available in R [Aksakalli et al., 2018]

SP-FSR algorithm: pseudo code

SP-FSR algorithm to identify interactions

- Assume a strong hierarchy
- Simplified version of **two-step SP-FSR** algorithm:

1. Identify the optimal set of k main effects using SP-FSR
2. Search k', number of interactions from k main effects with SP-FSR

- k and k' can be determined via grid search or automatically

Experimental setup

- Assume strong hierarchy
- Comparison methods: SP-FSR, SFFS, GA and `glinternet`
- Accuracy evaluation: Area under curve (AUC)
- Information criteria: AIC and BIC
- Model: logistic regression
- Datasets with binary targets:

1. Ionosphere
2. Sonar

Experimental result: Ionosphere I

- Source: UCI (Lichman [2013])
- $n = 351, p = 33$

Method	k	k'	AUC	BIC	AIC
GA	17	71	0.9430	1747.094	1403.484
glinternet	25	23	0.9899	249.0249	434.3426
SFFS	6	4	0.9402	197.1674	239.6360
SP-FSR (Full grid search)	14	36	0.9963	151.2322	348.1323
SP-FSR (Automode)	13	14	0.9825	145.0256	253.127
Baseline	33	0	0.9815	179.0528	310.3195

Note: glinterent yields λ of 0.0005

Experimental result: Ionosphere II

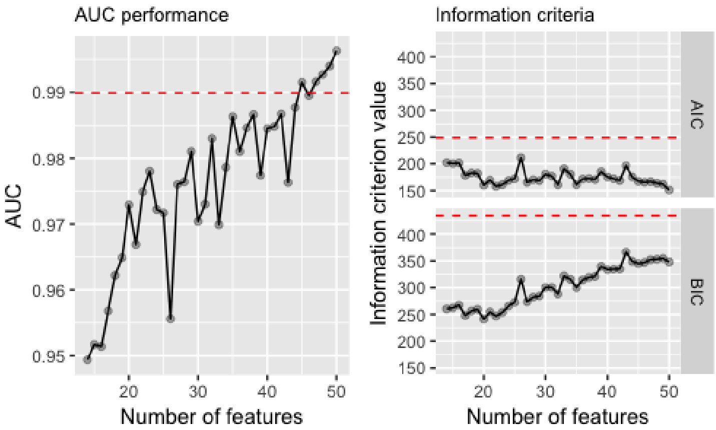

Figure 1: Comparison between glinternet and SP-FSR with full grid search

Experimental result: Ionosphere III

Key critiques

- No experiments on other continuous and multinomial Y
- No assessment of underlying assumptions
- No quadratic and higher-order terms

Conclusion

- SP-FSR
- The puzzle remains: do we really need to enforce (strong) hierarchy?

Acknowledgement

- Dr Vural Aksakalli (VA) as my supervisor and mentor
- VA, Dr Babak Abbasi, and Zeren D. Yenice for `spFSR` package
- Zeren D. Yenice, Niranjan Adhikari, Vural Aksakalli, Dr Alev Taskin Gumus, and Babak Abbasi for their previous works on SP-FSR algorithm

References I

Vural Aksakalli and Milad Malekipirbazari. Feature selection via binary simultaneous perturbation stochastic approximation. *Pattern Recognition Letters*, 75(Supplement C):41 – 47, 2016. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2016.03.002. URL http://www.sciencedirect.com/science/article/pii/S0167865516000702.

Vural Aksakalli, Babak Abbasi, and Yong Kai Wong. *spFSR: Feature Selection and Ranking by Simultaneous Perturbation Stochastic Approximation*, 2018. https://www.featureranking.com/, https://arxiv.org/abs/1804.05589.

D. R. Cox. Interaction. *International Statistical Review / Revue Internationale de Statistique*, 52(1):1–24, 1984. URL http://www.jstor.org/stable/1403235.

References II

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles*, 33(1):1–22, 2010. ISSN 1548-7660. doi: 10.18637/jss.v033.i01. URL https://www.jstatsoft.org/v033/i01.

R. Kohavi and G.H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 1(2):273–324, 1997.

M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Michael Lim and Trevor Hastie. *glinternet: Learning Interactions via Hierarchical Group-Lasso Regularization*, 2018. URL https://CRAN.R-project.org/package=glinternet. R package version 1.0.7.

References III

P. Pudil, F. J. Ferri, J. Novovicova, and J. Kittler. Floating search methods for feature selection with nonmonotonic criterion functions. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 - Conference C: Signal Processing (Cat. No.94CH3440-5)*, volume 2, pages 279–283 vol.2, Oct 1994. doi: 10.1109/ICPR.1994.576920.

W. Siedlecki and J. Sklansky. *A Note on Genetic Algorithm for Large-Scale Feature Selection*, pages 88–107. 2011. doi: 10.1142/9789814343138_0005. URL https://www.worldscientific.com/doi/abs/10.1142/9789814343138_0005.

James C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE*, 37(3):322–341, 3 1992.

Zeren D. Yenice, Niranjan Adhikari, Yong Kai Wong, Vural Aksakalli, Alev Taskin Gumus, and Babak Abbasi. Spsa-fsr: Simultaneous perturbation stochastic approximation in feature selection and ranking. 2018. URL https://arxiv.org/abs/1804.05589.