Title

Giulio Prevedello* Eric Cramer† Felice Alessio Bava‡

Institut Curie, PSL Research University,
CNRS UMR 3348, Orsay, France
July 30, 2019

Abstract

Abstract

Keywords: key1 · key2

Mathematics Subject Classification: If · Needed

1 Introduction

Problem: field of biology, surge of highly multiplexed experimental protocols capable of recovering thousands of cells from which several measurements are obtained. This poses a challenge to analyse this multidimensional data. To do so, people recur to Machine Learning (ML) techniques, in order to investigate the data for structures that have biological relevance.

An ML method that fits the data can detect data structures that can then be challenged for biological relevance, thus helping researchers in their investigation.

Examples (Kara Davis's paper, and Garry Nolan's in general)

Feature selection is one of the desirable features. You run an experiment for several markers to classify cell types, you want to know which are the most relevant that should be retained in a follow up experiment (eventually with lower multiplexing power) to optimise the cell classification.

As cells are often classified by gating (describe gating), RF classification is an automatic routine that mimics gating. Although at present no automatic routine outperformes expert gating, thanks to the similarity with gating RF findings can be integrated, interpreted, verified with more ease from researchers.

*Corresponding author. Electronic address: p.giulio@hotmail.it
†Electronic address: cramerericm@gmail.com
‡Electronic address: alessio.bava@gmail.com

The metrics in output must also be well-defined in unbalanced data, where ofter rare populations are very important in the investigation.

For these reasons we aim at provide further explainability to the RF classifier by redefining feature importance. Must be theoretically sounding -¿ proper definition Be explainable -¿ plots that provide better understanding of the impact of feature to the class prediction in the model, even for non machine learning experts. Useful statistic -¿ on which researchers can base their feature selection decision to select for the top predicting features, tailored for specific classes of interest.

[RECYCLE] In particular, the recursive thresholding in Random Forest (RF) mirrors cytometry gating to classify data, still underperfomed by automatic solution that have difficulties in dealing with biological variability while taking previous field knowledge into account.

Background, Context, State-of-Art

Why we did this

What we did-Paper structure

2 Formal Definition

We seek to define a new measure for feature importance that do this as opposed to the what we call Global Importance

3 Application to dermatology data

Explain model: heatmap and expression figures Feature importance profile across classes -¿ high average for globally important features; large standard deviation for features with different impact among classes Comparison with global importance

3.1 Detailed model explanation

3.2 Feature selection

Feature importance distribution within each class Comparison with features selected via global importance

4 Discussion

Recap problem and what we achieve

Acknowledgements

The research leading to these results has received funding from REF. On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[2] N. Cressie and T. R. C. Read. Multinomial goodness-of-fit tests. *J R Stat Soc Series B Stat Methodol*, 46(3):440–464, 1984.

[3] J. M. Marchingo, G. Prevedello, A. Kan, S. Heinzel, P. D. Hodgkin, and K. R. Duffy. T-cell stimuli independently sum to regulate an inherited clonal division fate. *Nat Commun*, 7:13540, 2016.

[4] J. M. Marchingo, A. Kan, R. M. Sutherland, K. R. Duffy, C. J. Wellard, G. T. Belz, A. M. Lew, M. R. Dowling, S. Heinzel, and P. D. Hodgkin. Antigen affinity, costimulation, and cytokine inputs sum linearly to amplify T cell expansion. *Science*, 346:1123–1127, 2014.

[5] J. D. Wolchok, H. Kluger, M. K. Callahan, M. A. Postow, N. A. Rizvi, A. M. Lesokhin, N. H. Segal, C. E. Ariyan, R. Gordon, K. Reed, M. M. Burke, A. Caldwell, S. A. Kronenberg, B. U. Agunwamba, X. Zhang, I. Lowy, H. D. Inzunza, W. Feely, C. E. Horak, Q. Hong, A. J. Korman, J. M. Wigginton, A. Gupta, and M. Sznol. Nivolumab plus Ipilimumab in advanced melanoma. *N Engl J Med*, 369(2):122–133, 2013.

[6] Editorial. Rationalizing combination therapies. *Nat Med*, 23:1113, 10 2017.

[7] D. S. Moore and M. C. Spruill. Unified large-sample theory of general chi-squared statistics for tests of fit. *Ann Stat*, 3(3):599–616, 1975.

[8] D. S. Moore. Generalized inverses, Wald's method, and the construction of chi-squared tests of fit. *J Am Stat Assoc*, 72(357):131–137, 1977.

[9] D. P. Mihalko and D. S. Moore. Chi-square tests of fit for type II censored data. *Ann Stat*, 8(3):625–644, 1980.

[10] D. S. Moore. The effect of dependence on chi squared tests of fit. *Ann Stat*, 10(4):1163–1171, 1982.

[11] A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans Am Math Soc*, 54(3):426–482, 1943.

[12] A. S. Hadi and M. T. Wells. A note on generalized Wald's method. *Metrika*, 37(1):309–315, 1990.

[13] Q. H. Vuong. Generalized inverses and asymptotic properties of Wald tests. *Econ Lett*, 24(4):343–347, 1987.

[14] D. W. K. Andrews. Asymptotic results for generalized Wald tests. *Econ Theory*, 3(3):348–358, 1987.

[15] D. W. K. Andrews. Chi-square diagnostic tests for econometric models: Introduction and applications. *J Econom*, 37(1):135–156, 1988.

[16] J. R. Wilson and K. J. Koehler. Hierarchical models for cross-classified overdispersed multinomial data. *J Bus Econ Stat*, 9(1):103–110, 1991.

[17] B. Zhang. A chi-squared goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 86(3):531–539, 1999.

[18] D. E. Tyler. Asymptotic inference for eigenvectors. *Ann Stat*, 9(4):725–736, 1981.

[19] F. C. Drost. Generalized chi-square goodness-of-fit tests for location-scale models when the number of classes tends to infinity. *Ann Stat*, 17(3):1285–1300, 1989.

[20] V. Voinov, A. Roza, and N. Pya. Recent achievements in modified chi-squared goodness-of-fit testing. In F. Vonta, M. Nikulin, N. Limnios, and C. Huber-Carol, editors, *Statistical Models and Methods for Biomedical and Technical Systems*. 2008.

[21] G. G. Hamedani. Sub-independence: An expository perspective. *Communications in Statistics – Theory and Methods*, 42(3):3615–3638, 2013.