Per-class feature importance

Giulio Prevedello* Eric Cramer† Felice Alessio Bava‡

Institut Curie, PSL Research University,
CNRS UMR 3348, Orsay, France
August 2, 2019

Abstract

Abstract

Keywords: key1 · key2

Mathematics Subject Classification: If · Needed

1 Introduction

In the past decades, the field of biology has witnessed a surge of new experimental techniques that measure hundreds of markers from thousands of cells, thus posing a challenge to analyse this multidimensional data. To analyse s a consequence, a call for new quantitative methods

MUST DEFINE HERE X AND Y FOR THE CLASSIFICATION PROBLEM

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) to identify markers 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.

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

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

If a ML model provides shows that an accurate prediction is achievable, it is of great interest to extrapolate which features contribute to the discrimination of one class form the others.

There are several of these metrics, but few are backed up by theoretical results and may suffer in the case of unbalanced data.

What we seek in particular is to define a measure of that is per-class. In particular, must be robust in the case of unbalanced dataset, considering that in biology ofter rare classes (i.e. rare cell types) are of great interest.

To guide feature selection and to improve explanability in the RF classifier, we defined feature importance on a per-class basis, as a modification of the MDI feature importance (REF). This method is then applied to a classical dataset with unbalanced to illustrate the information that can be extrapolated from the RF and comparing different strategies for feature selection (REF).

[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.

2 Formal definition

To define a measure how much each feature contributes to the prediction of one class, we follow the rationale behind the feature importance measured via mean importance decrease (MDI), which we now recall after introducing some notation as in REF. RF is represented by a collection of decision trees (Ti)i, where each decision tree is define by a rooted binary tree graph Ti = (Vi,Ei), whose nodes either have a left and right child, otherwise are leaves. Set of leaf node DEF. Each node is associated to a vector v. As the root node is associated to v = all data, at each non-leaf node a decision made based on a threshold t and one of the features j so that all data in v are passed to the left child if RULE, to the right child otherwise. For the RF, the such feature j and threshold t are determined so to maximize the impurity decrease, or gain, delta i DEF, where i is a measure for impurity such as the Gini or the Entropy. Then, the variables v reaching a leaf node determines the class that is predicted in the portion of features' space identified by all the nested decisions from the nodes in the path that join the root to such leaf. IMPROVE THIS

The decision tree's importance for a given feature j, is calculated as the sum the impurity gains over all nodes, that is EQ. Form hereafter we refer to the vector I = Ij as global feature importance.

We reason that, if Ij is the contribution the feature j to the overall prediction, the importance of j with respect to a class c must account for the contribution of only those nodes having at least one downstream leaf predicting class c. must be determined by

the sequence of splits that ultimately predict a portion of the space being labeled as c. Thus the feature importance relative to c accounts for the contribution of those, and only those, nodes having at least one downstream leaf assigned to the class c.

Moreover, from the perspective of a single class c, the node impurity should be affected only by the presence of variables that are not labeled to c, whichever their class. To this end, the impurity function i is replaced by the EQ of ic.

Finally, the per-class importance matrix, for the decision tree T, is defined as Icj for every class c in C and feature j in F. ALGORITHM

Normalisation TO DO Unbalanced data. TO DO

From equation REF, we remark that the feature importance vector relative to the class c, Ic, represents the global feature importance for the decision tree T with two classes, c and other-than-c. This property is evidenced in Figure 1 REF, which illustrates the algorithm for the per-class importance matrix, from one decision tree, as described.

In particular, being Ic is normalized in EQ REF so that the sum of its entries equals one, it is suited to represent feature importance for one class and to compare feature importance vectors between classes, in the case of imbalanced datasets.

DISEGNO

For ensemble learning methods that are defined by set of decision trees, such as RF, the importance matrix is calculated as the average of importance matrices across its trees, similarly as for the global feature importance REF.

3 Improved explainability

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

4 Feature selection strategies

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

5 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.