On machine learning algorithms and compositional data

Raimon Tolosana-Delgado
Helmholtz-Zentrum Dresden-Rossendorf
208 PUBLICATIONS **3,092** CITATIONS
SEE PROFILE

Hassan Talebi
Rio Tinto
19 PUBLICATIONS **119** CITATIONS
SEE PROFILE

Mahdi Khodadadzadeh
Faculty of Geo-information Science and Earth Observation (ITC), University of Twent…
55 PUBLICATIONS **737** CITATIONS
SEE PROFILE

On machine learning algorithms and compositional data

R. Tolosana-Delgado[1], **H. Talebi**[2] ,**M. Khodadadzadeh**[1], **and K.G. van den Boogaart**[1]
[1]Helmholtz Zentrum Dresden-Rossendorf,
Helmholtz Institute Freiberg for Resource Technology,
Freiberg, Germany; r.tolosana@hzdr.de
[2]CSIRO, Perth Australia

Abstract

Predictive methods such as Lasso regression, partition trees and random forests (RF), artificial neural networks (ANN) and deep learning, or support-vector machines (SVM) and other kernel methods have become in the last years increasingly popular, also in the compositional data community. However, most of the contributions using machine learning algorithms on compositional data just applied the relevant method to an additive, centered or isometric log-ratio (alr, clr, ilr) transformed version of the training data, without caring about the properties of the construct. In this contribution we briefly review the fundamental construction of these methods, and check in which way can they be tweaked or adapted to account for the compositional scale of the data.

As an example, a binary partition tree aims at constructing a hierarchy of classification, where each branch splits the data in two subgroups according to the one single covariable that provides highest purity of the two resulting subgroups; at the end of the hierarchy, all branches contain only data from one pure group. Random Forests Breiman (2001) were introduced to deal with the obvious over-fitting of partition trees, with a double randomisation strategy: first bootstrapping the number of observations, creating B different trees that form the forest; second, each branching of each tree is based not on the whole set of variables, but on a different random subset of them. The fact that at each branching only one variable is actively used makes the method non-invariant under the choice of possible log-ratio transformations. A way to allow for this one feature selection while keeping the relative nature of compositional information would be to build the trees on the set of pairwise log-ratios (pwlr). This applies to all kinds of tree-based methods with compositional covariables.

Key words: affine equivariance, subcompositional coherence, variable selection.

1 Introduction

Predictive methods such partition trees and random forests (RF), artificial neural networks (ANN) and deep learning, or support-vector machines (SVM) and other kernel methods have become in the last years increasingly popular, also in the compositional data community. However, most of the contributions using machine learning algorithms on compositional data just applied the relevant method to an additive, centered or isometric log-ratio (alr, clr, ilr) transformed version of the training data, without caring about the properties of the construct. In this contribution we briefly review the fundamental construction of these methods, and check in which way can they be tweaked or adapted to account for the compositional scale of the data.

After summarizing the most relevant ways of representing compositional data, the paper devotes a section to each family or group of machine learning algorithms. Each of these sections very briefly report what the method does, and several considerations on how to adapt them to compositional data.

2 Compositional data: ratios, logratios and isometric representations

A compositional data set is said to contain only relative information, which can be captured in the form of ratios or log-ratios (Aitchison, 1986). Aitchison (1997) highlighted as well the importance of subcompositions, as the counterpart of marginals for compositional data. For a D-part composition $\mathbf{x} = [x_1, x_2, \ldots, x_D]$, the subcomposition of the set of K parts $\{s_1, s_2, \ldots s_K\} \equiv S$ will be denoted as $\mathbf{x}_S = [x_{s_1}, x_{s_2}, \ldots, x_{s_K}]$.

Several ratios and log-ratios have been used in the literature to extract that relative information. Perhaps the first and simplest one was the closure, i.e.

$$\mathcal{C}[\mathbf{x}] = \left[\frac{x_1}{t(\mathbf{x})}, \frac{x_2}{t(\mathbf{x})}, \cdots, \frac{x_D}{t(\mathbf{x})}\right], \qquad \text{with} \qquad t(\mathbf{x}) = x_1 + x_2 + \cdots + x_D. \tag{1}$$

Before the work of Aitchison, another common way of treating compositions was through simple ratios (or logratios), e.g. via expressions such as

$$x^*_{ij} = \frac{x_i}{x_j}, \qquad \text{or} \qquad \xi_{ij} = \ln x^*_{ij}. \tag{2}$$

The pairwise logratio transformation $\mathrm{pwlr}(\mathbf{x}) = [\xi_{ij}; i < j = 1, 2, \ldots, D]$ was defined by Aitchison (1986), as well as the centered logratio transformation

$$\mathrm{clr}(\mathbf{x}) = \ln\left[\frac{x_1}{g(\mathbf{x})}, \frac{x_2}{g(\mathbf{x})}, \cdots, \frac{x_D}{g(\mathbf{x})}\right], \qquad \text{with} \qquad g(\mathbf{x}) = \sqrt[D]{x_1 x_2 \cdots x_D}, \tag{3}$$

with the logarithm applied component-wise. Another transformation found in this work is the additive logratio transformation,

$$\mathrm{alr}(\mathbf{x}) = \ln\left[\frac{x_1}{x_D}, \frac{x_2}{x_D}, \cdots, \frac{x_{D-1}}{x_D}\right]. \tag{4}$$

Aitchison (1986) introduced as well a notion of compositional distance between two vectors $\mathbf{x}$ and $\mathbf{y}$, as

$$d_A^2(\mathbf{x}, \mathbf{y}) = \frac{1}{D^2}\sum_{i<j}^{D}\left(\ln\frac{x_i}{x_j} - \ln\frac{y_i}{y_j}\right)^2,$$

also based on ratios. Later, with the establishment of the Aitchison geometry (Billheimer et. al. , 2001; Pawlowsky-Glahn and Egozcue, 2001), it was realized that the clr transformation represents an isometry, $d_A^2(\mathbf{x}, \mathbf{y}) \equiv d^2(\mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y}))$, being $d(\cdot, \cdot)$ the conventional sum-of-squares, Euclidean distance. Finally, Egozcue et al. (200X) introduced the isometric logratio transformation (ilr) as a Gramm-Schmidt orthogonalisation of the clr coefficients,

$$\mathrm{ilr}(\mathbf{x}) = \mathrm{clr}(\mathbf{x}) \cdot \mathbf{V}, \quad \text{with} \quad \mathbf{V} \cdot \mathbf{V}^t = \mathbf{1}_D - \frac{1}{D}\mathbf{1}_{D\times D} \quad \text{and} \quad \mathbf{V}^t \cdot \mathbf{V} = \mathbf{I}_{D-1}, \tag{5}$$

being $\mathbf{1}_D$ and $\mathbf{1}_{D\times D}$ resp. the $(D \times D)$-identity matrix and $(D \times D)$-ones matrix.

Closed compositional data pose the problem that the scores obtained for each variable depend on all other variables available at that observation. So, if different subcompositions are available, closed values would be not comparable anymore. The same happens with centered logratios. On the other hand, simple ratios, pairwise logratios, additive logratios and isometric logratios do not depend on which subcomposition was used, and they simply produce missing values in some of their coefficients if one or more components are lost. This is a very important issue in machine learning, because the training data fed to the algorithms is not necessarily complete, and ensuring consistency becomes a need.

3 Partition trees and random forests

Decision trees involve hierarchical segmentation of the predictor space into several simple regions. At each segmentation one single predictor that provides the highest purity of the two resulting regions is selected. To make a prediction for a given observation, we use the mode of the training observations in the region to which that observation belongs. Ensemble tree-based learners, such as bagging, boosting, and random forests, were introduced to deal with the high variance (overfitting) of decision trees and to improve prediction accuracy by generating multiple trees which are then combined to yield a single consensus prediction (Hastie et al, 2009). In the case of compositional predictors, the fact that at each segmentation only one predictor is actively used makes the tree-based methods non-invariant under the choice of possible log-ratio transformations. Indeed, using different log-ratio transformations leads to different predictor spaces and consequently different tree-based predictive models (potentially with different prediction accuracy).

To illustrate the effects of log-ratio transformations on tree-based learners (Random Forests, Breiman (2001), in this implementation) the multi-element near-surface geochemical compositions (compositional predictors) from the National Geochemical Survey of Australia (de Caritat and Cooper, 2011) are used to predict the exposed to deeply buried major crustal blocks (categorical response) of the Australian continent (Talebi et al, 2018). Samples of different sizes (10, 20, 30, 40, and 50) were taken randomly (without replacement) from geochemical components (hundred samples for each size). Each sample was closed (Eq. 1) or transformed via different log-ratio transformations (Eqs. 2-5). For each sample five random Forests classifiers were trained (using raw components, clrs, ilrs, pwlrs, and a combination of raw components plus all the log-ratios as input predictors). Figure 1 shows the distribution of Out-of-Bag (OOB) error estimation for each sample size and log-ratio transformation. As the number of component increases classifiers show more accuracy; however, pairwise log-ratios outperformed other options. The superiority of pwlr is clearer when more components are used to build the classifiers. Combining pwlrs with the other log-ratios does not improve the performance of the classifier and makes the interpretation of the predictor space more complicated. High-dimensionality of pwlrs is well addressed by tree-based predictive methods since they are working with subsets of predictors. In the case of compositional predictors, pairwise log-ratio transformation is recommended as a first choice to train a tree-based predictive model due to their ease of interpretation and superior performance. A recursive feature elimination with resampling technique may further improve the accuracy of the tree-based predicative models trained from pwlrs (Talebi et al, 2018).

4 Some remarks

This superior performance of pwlr on random forests (and in general, partition trees) does not apply to all machine learning methods. Regression methods (linear regression, logistic regression) can be proven to be affine equivariant, namely to produce the same predictions for every logratio transformation (alr, ilr, and even clr or pwlr if the appropriate inversion is used). Ridge regression penalizes the regression goodness of fit (least squares or deviance) by means of the square norm of the regression coefficients, in which case ilr appears to be necessary. Support vector machines, establishing a classifier based on distances, may be required to be applied on isometric representations of the data (ilr, clr but also pwlr). Finally, neural networks require further research, but preliminary results suggest that some implementations of the estimation algorithms may not be affine equivariant, in which case practitioners should use them carefully. These considerations apply to methods using compositional data on the role of predictors.

References

Aitchison J (1986) *The Statistical Analysis of Compositional Data.* Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (Reprinted in 2003 with additional material by The Blackburn

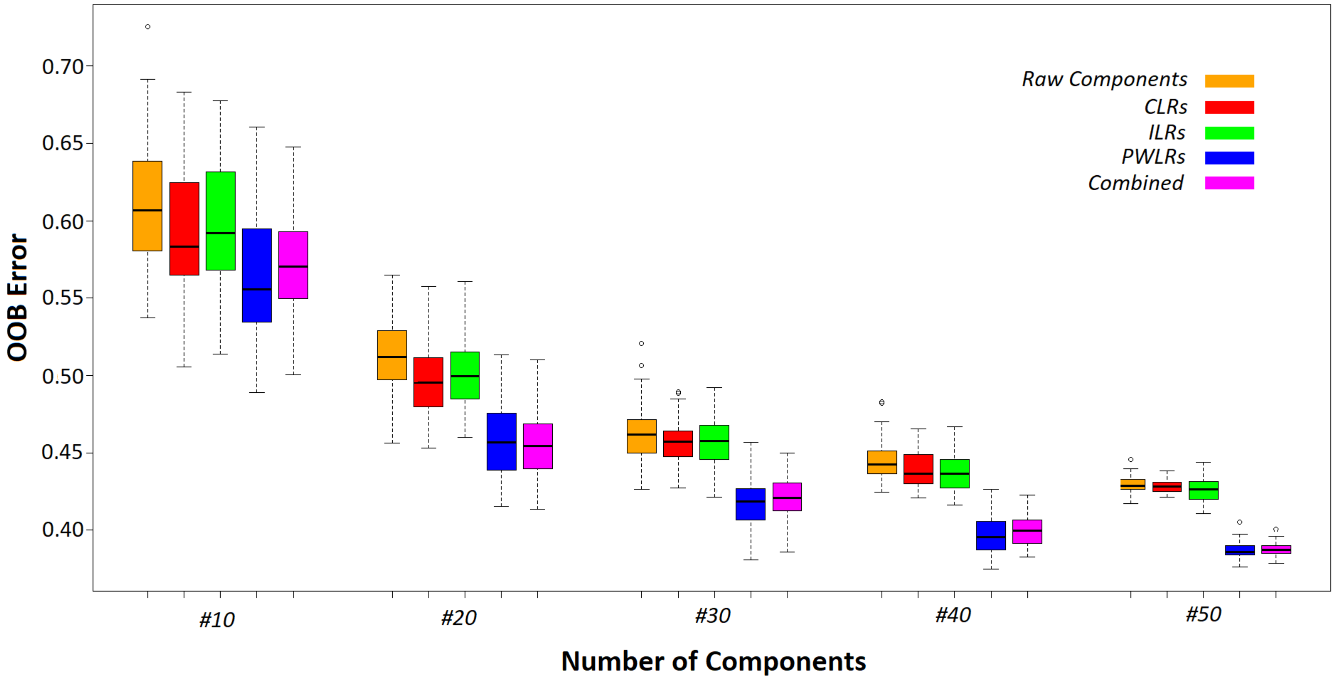

Figure 1: Out-of-bag error estimation for different logratio representations and sample sizes

Press)

Aitchison J (1997) The one-hour course in compositional data analysis or compositional data analysis is simple. In: Pawlowsky-Glahn V (Ed.) *Proceedings of IAMG'97 – The III annual conference of the International Association for Mathematical Geology*, pp 3-35

Billheimer, D. and Guttorp, P. and Fagan, W.F (2001) Statistical interpretation of species composition. *Journal of the American Statistical Association* 456(96):1205–1214

Breiman, L (2001) Random Forests. *Machine Learning* 45:5–32

Caritat P de, Cooper M (2011) *National geochemical survey of Australia: The geochemical atlas of Australia* Geoscience Australia, Record 2011/20

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. *Mathematical Geology* 35:279–300

Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning*. Springer, New York.

Talebi H, Mueller U, Tolosana-Delgado R, et al (2018). Surficial and Deep Earth Material Prediction from Geochemical Compositions. *Natural Resources Research* (online first) doi: 10.1007/s11053-018-9423-2

Pawlowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment* 15(5):384–398

Pawlowsky-Glahn V, Egozcue JJ (2002). BLU estimators and compositional data. *Mathematical Geology* 34(3):259–274