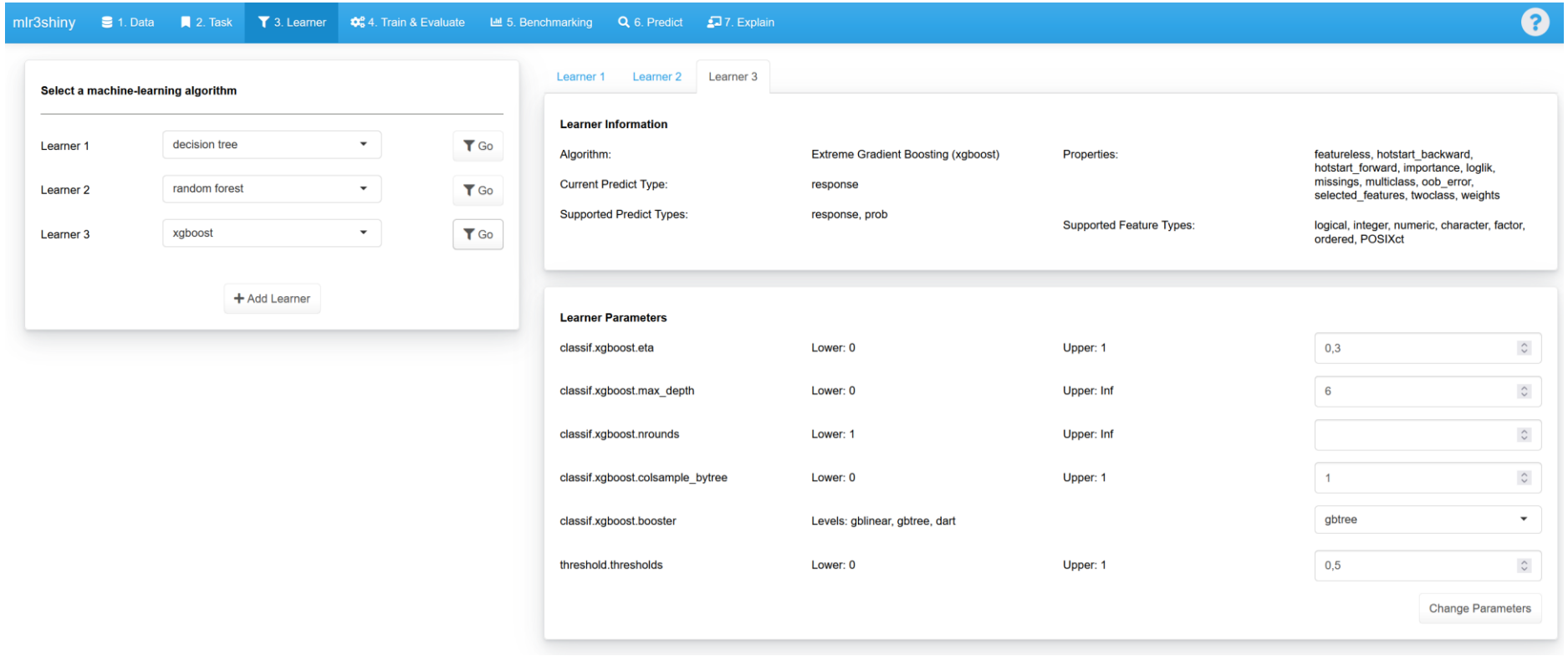
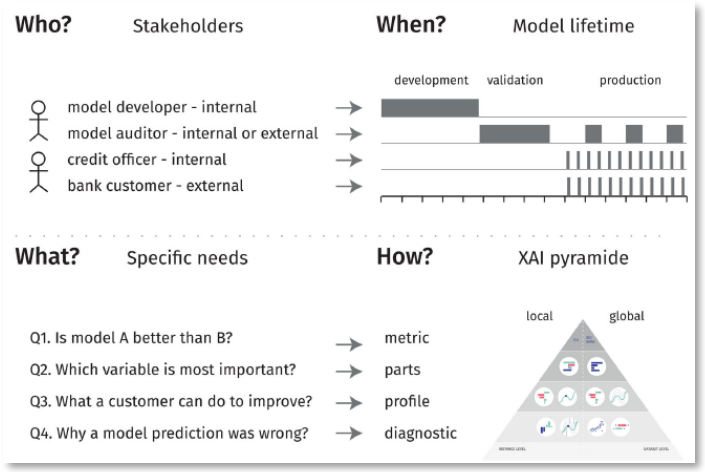
09.04.2025, FLI

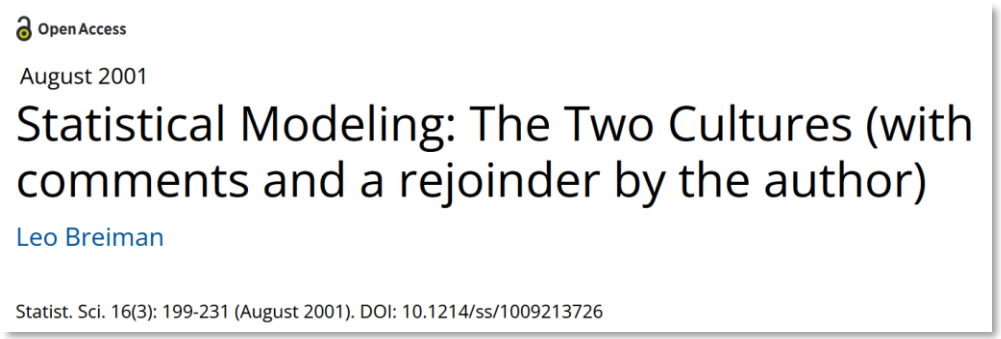
Gero Szepannek
Statistik, Wirtschaftsmathematik und Machine Learning
Hochschule Stralsund



https://github.com/LamaTe/mlr3shiny

(Tetzlaff and Szepannek, 2022)



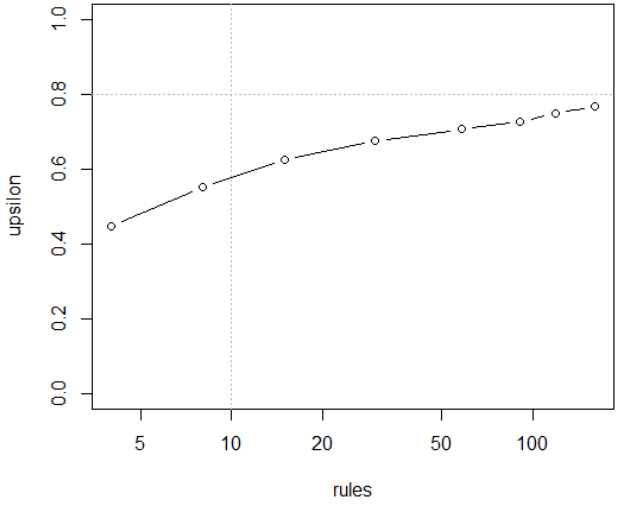

Open Access

August 2001

Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)

Leo Breiman

Statist. Sci. 16(3): 199-231 (August 2001). DOI: 10.1214/ss/1009213726

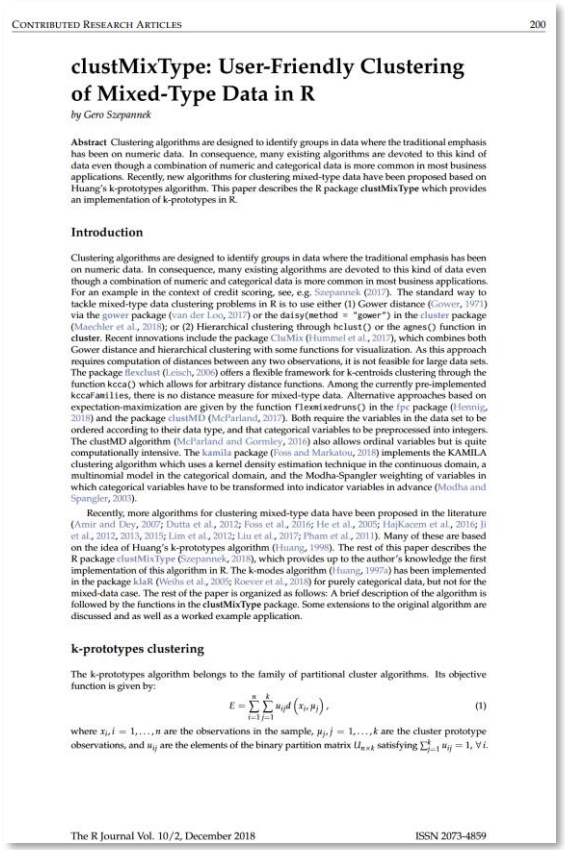


Variable	Split (<=)	Levels Left	pleft	pright
Intercept			0,7025	0,7025
duration	11.5		0,0880	-0,0188
duration	34.5		0,0315	-0,1477
history		delay in paying off in the past | critical account/other credits elsewhere	-0,1851	0,0180
history		delay in paying off in the past | no credits taken/all credits paid back duly | existing credits paid back duly till now	-0,0270	0,0518
purpose		others | repairs	-0,0771	0,0275
savings		unknown/no savings account | ... < 100 DM	-0,0281	0,0675
status		no checking account | ... < 0 DM	-0,1262	0,1499
status		no checking account | 0<= ... < 200 DM | ... >= 200 DM / salary for at least 1 year	-0,0267	0,0697


Can't see the forest for the trees

Analyzing grouves to explain random forests

(Bücker et al., 2021; Szepannek, 2024)



CONTRIBUTED RESEARCH ARTICLES 200

clustMixType: User-Friendly Clustering of Mixed-Type Data in R

by Gero Szepannek

Abstract Clustering algorithms are designed to identify groups in data where the traditional emphasis has been on numeric data. In consequence, many existing algorithms are devoted to this kind of data even though a combination of numeric and categorical data is more common in most business applications. Recently, new algorithms for clustering mixed-type data have been proposed based on Huang's k-prototypes algorithm. This paper describes the R package **clustMixType** which provides an implementation of k-prototypes in R.

Introduction

Clustering algorithms are designed to identify groups in data where the traditional emphasis has been on numeric data. In consequence, many existing algorithms are devoted to this kind of data even though a combination of numeric and categorical data is more common in most business applications. For an example in the context of credit scoring, see, e.g. Szepannek (2017). The standard way to tackle mixed-type data clustering problems in R is to use either (1) Gower distance (Gower, 1971) via the gower package (van der Loo, 2017) or the `daisy(method = "gower")` in the cluster package (Maechler et al., 2018); or (2) Hierarchical clustering through `hclust()` or the `agnes()` function in **cluster**. Recent innovations include the package CluMix (Hummel et al., 2017), which combines both Gower distance and hierarchical clustering with some functions for visualization. As this approach requires computation of distances between any two observations, it is not feasible for large data sets. The package flexclust (Leisch, 2006) offers a flexible framework for k-centroids clustering through the function `kcca()` which allows for arbitrary distance functions. Among the currently pre-implemented `kccaFamilies`, there is no distance measure for mixed-type data. Alternative approaches based on expectation-maximization are given by the function `flexmixedruns()` in the fpc package (Hennig, 2018) and the package clustMD (McParland, 2017). Both require the variables in the data set to be ordered according to their data type, and that categorical variables to be preprocessed into integers. The **clustMD algorithm** (McParland and Gormley, 2016) also allows ordinal variables but is quite computationally intensive. The kamila package (Foss and Markatou, 2018) implements the KAMILA clustering algorithm which uses a kernel density estimation technique in the continuous domain, a multinomial model in the categorical domain, and the Modha-Spangler weighting of variables in which categorical variables have to be transformed into indicator variables in advance (Modha and Spangler, 2003).

Recently, more algorithms for clustering mixed-type data have been proposed in the literature (Amir and Dey, 2007; Dutta et al., 2012; Foss et al., 2016; He et al., 2005; HajKacem et al., 2016; Ji et al., 2012, 2013, 2015; Lim et al., 2012; Liu et al., 2017; Pham et al., 2011). Many of these are based on the idea of Huang's k-prototypes algorithm (Huang, 1998). The rest of this paper describes the R package clustMixType (Szepannek, 2018), which provides up to the author's knowledge the first implementation of this algorithm in R. The k-modes algorithm (Huang, 1997a) has been implemented in the package klaR (Weihs et al., 2005; Roever et al., 2018) for purely categorical data, but not for the mixed-data case. The rest of the paper is organized as follows: A brief description of the algorithm is followed by the functions in the **clustMixType** package. Some extensions to the original algorithm are discussed and as well as a worked example application.

k-prototypes clustering

The k-prototypes algorithm belongs to the family of partitional cluster algorithms. Its objective function is given by:

$$E = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij} d\left(x_i, \mu_j\right), \tag{1}$$

where $x_i, i = 1, \ldots, n$ are the observations in the sample, $\mu_j, j = 1, \ldots, k$ are the cluster prototype observations, and u_{ij} are the elements of the binary partition matrix $U_{n \times k}$ satisfying $\sum_{j=1}^{k} u_{ij} = 1, \forall i$.

Advances in Data Analysis and Classification
https://doi.org/10.1007/s11634-024-00595-5

ORIGINAL ARTICLE

Clustering large mixed-type data with ordinal variables

Gero Szepannek[1] · **Rabea Aschenbruck**[1] · **Adalbert Wilhelm**[2]

Received: 18 August 2023 / Revised: 24 February 2024 / Accepted: 3 May 2024

Abstract
One of the most frequently used algorithms for clustering data with both numeric and categorical variables is the k-prototypes algorithm, an extension of the well-known k-means clustering. Gower's distance denotes another popular approach for dealing with mixed-type data and is suitable not only for numeric and categorical but also for ordinal variables. In the paper a modification of the k-prototypes algorithm to Gower's distance is proposed that ensures convergence. This provides a tool that allows to take into account ordinal information for clustering and can also be used for large data. A simulation study demonstrates convergence, good clustering results as well as small runtimes.

Keywords Cluster analysis · Mixed-type data · Ordinal data · K-prototypes · Gower's distance

Mathematics Subject Classification 62H30

1 Introduction

In cluster analysis, as an unsupervised learning approach, the purpose consists in identifying homogeneous groups of observations in a data. The problem can be addressed in different ways depending on the definition of homogeneity (for an overview cf. e.g., Jain 2010). In practice, data are often of mixed-type i.e., containing both numeric and categorical variables. Hennig and Liao (2013) give a practical step by step example of clustering mixed-type real world data and reviews on available algorithms are given in Hunt and Jorgensen (2011) and Ahmad and Khan (2018).

A common approach to deal with mixed-type data in Statistics consists in using Gower's distance (Gower 1971). Accordingly, an intuitive and widespread way of clustering mixed-type data is based on Gower's distance by combining it with the

✉ Gero Szepannek
gero.szepannek@hochschule-stralsund.de

[1] Stralsund University of Applied Sciences, Stralsund, Germany

[2] Constructor University Bremen, Bremen, Germany

Published online: 27 May 2024

Cluster Validation for Mixed-Type Data

Rabea Aschenbruck and Gero Szepannek

Abstract For cluster analysis based on mixed-type data (i.e. data consisting of numerical and categorical variables), comparatively few clustering methods are available. One popular approach to deal with this kind of problems is an extension of the *k-means* algorithm (Huang, 1998), the so-called *k-prototype* algorithm, which is implemented in the R package `clustMixType` (Szepannek and Aschenbruck, 2019).

It is further known that the selection of a suitable number of clusters k is particularly crucial in partitioning cluster procedures. Many implementations of cluster validation indices in R are not suitable for mixed-type data. This paper examines the transferability of validation indices, such as the Gamma index, Average Silhouette Width or Dunn index to mixed-type data. Furthermore, the R package `clustMixType` is extended by these indices and their application is demonstrated. Finally, the behaviour of the adapted indices is tested by a short simulation study using different data scenarios.

Rabea Aschenbruck · Gero Szepannek
University of Applied Sciences Stralsund, Zur Schwedenschanze 15, D-18435 Stralsund, Germany
✉ rabea.aschenbruck@hochschule-stralsund.de
✉ gero.szepannek@hochschule-stralsund.de

ARCHIVES OF DATA SCIENCE, SERIES A
(ONLINE FIRST) DOI: 10.5445/KSP/1000098011/02
KIT SCIENTIFIC PUBLISHING ISSN 2363-9881
Vol. 6, No. 1, 2020

+50K downloads

(Szepannek, 2018, 2024; Aschenbruck und Szepannek, 2020; …)

Startseite > HOST > Fakultäten > Wirtschaft > Studienangebot > Angewandte Data Science und Künstliche Intelligenz

Angewandte Data Science und Künstliche Intelligenz

Master-Studiengang

Wie funktioniert maschinelles Lernen, Deep Learning oder große Sprachmodelle und Anwendungen wie ChatGPT? Was sind die Potenziale und Grenzen solcher KI-Methoden und Werkzeuge und wie können sie kompetent und sicher in der Praxis angewandt werden? Wie werden KI-Anwendungen mit modernen Werkzeugen und Frameworks in Unternehmen effektiv umgesetzt? Wie wird die Einführung und Kommunikation von KI in Unternehmen erfolgreich, sicher und ethisch verantwortlich gemacht?

Mit dem Fokus auf Anwendungskompetenzen befähigen wir Sie auch ohne technisches Erststudium zum Konzipieren, Durchführen und Leiten von Projekten mit Data Science- und KI-Bezug. Durch unser innovatives didaktisches Konzept können auch QuereinsteigerInnen Schlüsselkompetenzen in Data Science und künstliche Intelligenz mit engagiertem Einsatz erfolgreich erwerben.

Der **Master-Studiengang Angewandte Data Science & Künstliche Intelligenz** vermittelt

Startseite > HOST > Fakultäten > Wirtschaft > Veranstaltungen > ECDA2026

European Conference on Data Analysis

9.-11.9.2026

About | Program | Practical Informations

ECDA is an international conference dedicated to analytical aspects of data science. It connects statistics, computer science, and practical application domains related to all aspects of data analysis.

The ECDA is organized under the auspices of the GfKl (Gesellschaft für Klassifikation).

It will be organzied in cooperation with the:

- European Association for Data Science (EuADS),
- Dutch/Flemish classification society (VOC),
- Classification and Data Analysis Group of the Italian Statistical Society (CLADAG),
- Classification and Data Analysis Section of the Polish Statistical Association (SKAD) and the
- British Data Science Society (BDSS)

- Tetzlaff, L., Szepannek, G. (2022): mlr3shiny—State-of-the-art machine learning made easy. SoftwareX, Volume 20, https://doi.org/10.1016/j.softx.2022.101246

- Bücker, M., Szepannek, G., Gosiewska, A., and Biecek, P. (2021). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, *73*(1), 70–90. https://doi.org/10.1080/01605682.2021.1922098
- Szepannek, G., Holt, BH.v. Can't see the forest for the trees. *Behaviormetrika* **51**, 411–423 (2024). https://doi.org/10.1007/s41237-023-00205-2
- Biecek, P., Kozak, A., Zawada, A., Szepannek, G. (2023): Per Anhalter durch die Galaxis des verantwortungsvollen maschinellen Lernens, SmarterPoland, ISBN: 978-8365291189.

- Szepannek, G.: clustMixType (2018): User-Friendly Clustering of Mixed-Type Data in R, R Journal, 10/2, 200 —208.
- Szepannek, G., Aschenbruck, R., Wilhelm, A. (2024): Clustering Large Mixed-Type Data with Ordinal Variables, Advances in Data Analysis and Classification, DOI: 10.1007/s11634-024-00595-5.
- Aschenbruck, R., Szepannek, G., Wilhelm, A. (2023): Random-Based Initialization Strategies for Clustering Mixed-Type Data with the k-Prototypes CLADAG 2023.
- Aschenbruck, R., Szepannek, G., Wilhelm, A. (2022): Imputation strategies for clustering mixed-type data with missing values, Journal of Classification, DOI: 10.1007/s00357-022-09422-y.
- Aschenbruck, R., Szepannek, G.; (2020): Cluster Validation for Mixed-Type Data. Archives of Data Science (Series A) 6(1), 1-12, DOI: 10.5445/KSP/1000098011/02.

Vielen Dank!