A comparison of two dissimilarity functions for mixed-type predictor variables in the δ-machine

Yuan, B.; Heiser, W.; Rooij, M. de

Advances in Data Analysis and Classification
https://doi.org/10.1007/s11634-021-00463-6

REGULAR ARTICLE

A comparison of two dissimilarity functions for mixed-type predictor variables in the δ-machine

Beibei Yuan[1] · **Willem Heiser**[1] · **Mark de Rooij**[1]

Received: 12 July 2019 / Revised: 2 August 2021 / Accepted: 6 September 2021

Abstract

The δ-machine is a statistical learning tool for classification based on dissimilarities or distances between profiles of the observations to profiles of a representation set, which was proposed by Yuan et al. (J Claasif 36(3): 442–470, 2019). So far, the δ-machine was restricted to continuous predictor variables only. In this article, we extend the δ-machine to handle continuous, ordinal, nominal, and binary predictor variables. We utilized a tailored dissimilarity function for mixed type variables which was defined by Gower. This measure has properties of a Manhattan distance. We develop, in a similar vein, a Euclidean dissimilarity function for mixed type variables. In simulation studies we compare the performance of the two dissimilarity functions and we compare the predictive performance of the δ-machine to logistic regression models. We generated data according to two population distributions where the type of predictor variables, the distribution of categorical variables, and the number of predictor variables was varied. The performance of the δ-machine using the two dissimilarity functions and different types of representation set was investigated. The simulation studies showed that the adjusted Euclidean dissimilarity function performed better than the adjusted Gower dissimilarity function; that the δ-machine outperformed logistic regression; and that for constructing the representation set, K-medoids clustering achieved fewer active exemplars than the one using K-means clustering while maintaining the accuracy. We also applied the δ-machine to an empirical example, discussed its interpretation in detail, and compared the classification performance with five other classification methods. The results showed that the δ-machine has a good balance between accuracy and interpretability.

Keywords Dissimilarity · Nonlinear classification · Mixed-type data · Monte Carlo

Mathematics Subject Classification 62H30

✉ Beibei Yuan
b.yuan@fsw.leidenuniv.nl

[1] Institute of Psychology, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands

Published online: 16 September 2021

1 Introduction

Classification is a process that assigns objects to categorical outcomes (James et al. 2013). Examples of classification problems include medical diagnosis, spam email detection, and credit card fraud detection. A typical scenario is that we have an outcome variable that we would like to predict based on a set of predictor variables. Instead of building a classification rule on the set of predictor variables, we can build a classification rule using the (dis)similarities between objects and a set of selected or stored *exemplars* or *prototypes*, where an exemplar is an actual object of a category, while a prototype is defined as an abstract average of the objects of a category (Nosofsky 1992; Ross and Makin 1999). The decision to classify a new object to a certain category is based on the dissimilarities of the new object to the selected exemplars or prototypes (Medin and Schaffer 1978).

Bergman and Magnusson (1997) called the predictor-based approach and the dissimilarity-based approach as the variable-oriented approach and the person-oriented approach. Specifically, in a variable-oriented approach, the analytical unit is the variable, and the obtained results from a variable-oriented method are interpreted in terms of the constructed relations among the variables. By contrast, in a person-oriented approach the analytical unit is the individual. Examples of person-oriented approaches are latent class and latent profile analysis, and their longitudinal extension, latent transition analysis (Hickendorff et al. 2018). These models can be used to model the heterogeneity between and within individuals, e.g., examining subgroups of individuals. By contrast, variable-oriented approaches present all individuals in a general model that is unable to deal with heterogeneity. Moreover, linear variable-oriented methods, such as linear regression, are unable to find complex, non-linear relations among the predictor variables. Although variable-oriented approaches exist that do allow for non-linear patterns, the choices of non-linear patterns are still limited (Hickendorff et al. 2018).

Yuan et al. (2019) proposed a statistical learning tool for classification based on dissimilarities or distances between profiles of the objects on the predictor variables, named the δ-machine. Objects is here a more general name than persons, i.e., a person-oriented approach could also be termed an object-oriented approach. We use the terms objects and persons interchangeable. By changing the basis of the classifier from predictor variables to dissimilarities, it is possible to achieve non-linear classification boundaries in the original predictor space. Because of its focus on profiles of objects, the δ-machine is a person-oriented approach as contrasted with the more usual variable-oriented approaches. Given a new object, the decision of assigning this object to a specific class is determined by the dissimilarities of this object towards the selected exemplars or prototypes. The δ-machine showed very promising predictive performance (Yuan et al. 2019), i.e., it is competitive and often superior to other classification methods including support vector machines (Cortes and Vapnik 1995), the Lasso logistic regression (Tibshirani 1996), and classification trees (Breiman et al. 1984). Meanwhile the usage of *the representation set* in the δ-machine, containing a small number of informative exemplars or prototypes, results in accurate person-oriented models.

The existing the δ-machine only considers continuous predictor variables. As many real world data consist of a mix of binary, nominal, ordinal and continuous variables, an extension of the δ-machine to mixed-type predictor variables is of great interest. Gower's (dis)similarity measure (Gower 1971) is commonly suggested in multidimensional scaling (Borg and Groenen 2005) when one needs to compute (dis)similarities among objects described by mixed-type predictor variables. Gower's dissimilarity measure shows the properties of the Manhattan distance (see below). In a similar vein, we propose an adjusted Euclidean dissimilarity function (AEDF) for mixed-type data, which is an extension of the ordinary Euclidean distance function. Because the Euclidean distance is one of the most popular and easily understandable distance metrics for continuous predictor variables, the Euclidean distance is suggested as the default measure in the δ-machine (Yuan et al. 2019). Various abbreviations frequently used in the manuscript are listed in the Appendix along with their full forms for quick reference.

We will distinguish between five types of variables: symmetric binary, asymmetric binary, nominal, ordinal, and continuous (Cox and Cox 2000). The first four types of variables are collectively called categorical variables and described as follows:

- *symmetric binary*, if the predictor variable X_p has only two possible levels, where each level is a label of a relatively homogeneous group; e.g. gender with categories 'male', 'female';
- *asymmetric binary*, if the predictor variable X_p has only two possible levels, and the two levels are not equally homogeneous; e.g. a variable represents the presence or absence of a specific disease. We can say that two patients with this disease have something in common, while it may not hold for two patients without the disease;
- *nominal*, if the predictor variable X_p has a finite and discrete set of levels, but the levels are not ordered; e.g. a variable represents the color of cars with categories 'black', 'blue', 'red' and 'white';
- *ordinal*, if the predictor variable X_p has a finite and discrete set of levels, and the levels are ordered, e.g. a variable with categories 'strongly dislike', 'dislike', 'neutral', 'like' and 'strongly like'; The distances between levels are unknown but the categories are ordered.

In this paper, we extend the δ-machine to handle mixed-type predictor variables. We define two dissimilarity functions, the adjusted Gower dissimilarity functions (AGDF) and the AEDF. The main goals of this paper are to compare the predictive performance of the δ-machine to logistic regression and to compare the predictive performance of the δ-machine with the two adjusted dissimilarity functions via simulation studies. This paper is organized as follows. Section 2 briefly reviews the δ-machine and extends it to handle mixed-type data. Section 3 presents simulation studies investigating the performance of the δ-machine on mixed-type data, the comparison of the two adjusted dissimilarity functions, and the comparison of different types of representation set. In Sect. 4, we apply the δ-machine on an empirical example and compare it to five other classification methods. In Sect. 5, we draw conclusions from the simulation studies and the empirical example and discuss some limitations and open issues.

2 The δ-machine

2.1 A brief review

We will first present the notation. The predictor matrix is represented as the $I \times P$ matrix $\mathbf{X}$, where P is the number of predictor variables and I is the number of objects. Each row of the predictor matrix is called a *row profile* for object i, $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iP}]^\mathsf{T}$, and the measurements are denoted by lower case letters, i.e. x_{i1} represents the measurement for object i on predictor variable X_1. The outcomes for the I objects are collected in $\mathbf{y}$, a vector of length I. Our focus will be on binary outcomes, i.e. $y \in \{0, 1\}$.

Besides the predictor matrix, there is a representation matrix $\mathbf{R}$ collecting the information of R highly informative exemplars or prototypes on the P predictor variables. There are several choices for the representation set. The simplest choice is to define the complete training set as the representation set. Two other choices are to select them through K-medoids clustering (Kaufman and Rousseeuw 1990) or K-means clustering (MacQueen 1967). No matter which clustering method is chosen, it will be applied separately on each outcome category. Subsequently, the representation set is the collection of exemplars or prototypes on both outcome categories. With K-medoids clustering, the resulting medoids are real existing objects; in other words, the representation set is a collection of exemplars. In contrast, when applying K-means clustering, the representation set is defined by prototypes. For example, Fig. 1 depicts three types of representation sets for the same example data with two continuous predictor variables. When the representation set is equal to the training set, the entire training set is collected in the representation set, i.e., all points in the figure. When the representation set is selected by K-medoids clustering the representation set has four exemplars (indicated by filled bullets), while the representation set selected by K-means clustering has four prototypes (indicated by filled squares).

Suppose we have a dissimilarity function $\delta(\cdot)$. The first step is to apply this dissimilarity function on the predictor matrix and the representation matrix to calculate the dissimilarity matrix. The pairwise dissimilarity between objects i from $\mathbf{X}$ and r from $\mathbf{R}$ is a scalar d_{ir}, i.e.

$$d_{ir} = \delta(\mathbf{x}_i, \mathbf{x}_r).$$

The pairwise dissimilarities d_{ir}, $(i = 1, \ldots, I, r = 1, \ldots, R)$, will be collected in the $I \times R$ dissimilarity matrix $\mathbf{D}$. Rows of the matrix $\mathbf{D}$ are given by

$$\mathbf{d}_i = [d_{i1}, d_{i2}, \ldots, d_{iR}]^\mathsf{T},$$

collecting the dissimilarities of an object i towards the R exemplars/prototypes. The dissimilarity matrix $\mathbf{D}$ defines an R-dimensional dissimilarity space (Pekalska and Duin 2005).

The second step is to build a linear classifier by logistic regression with the Least absolute shrinkage and selection operator (Lasso) (Tibshirani 1996; Friedman et al. 2010) on the R-dimensional dissimilarity space. The probability for a specific object

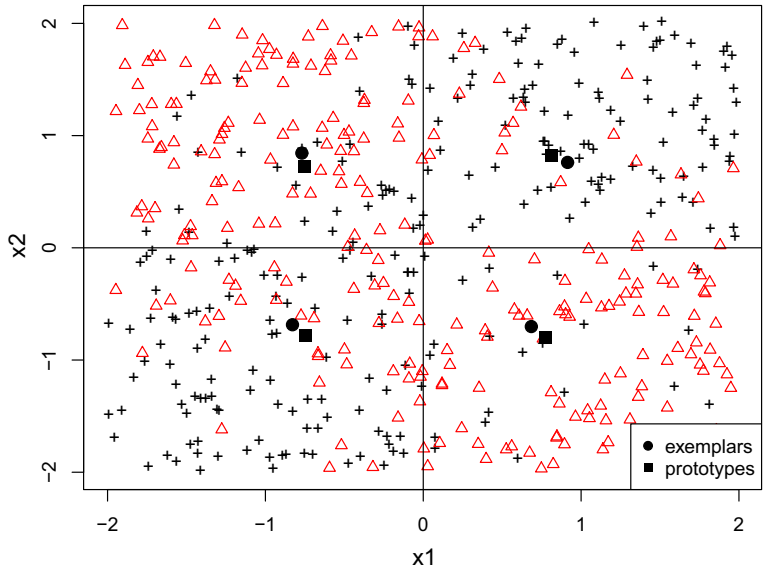


Fig. 1 Examples of the three types of representation set in two dimensions. The triangle and the cross denote objects in observed classes 0 and 1, respectively. The four filled bullets are the exemplars identified by K-medoids clustering. The four filled squares are the prototypes determined by K-means clustering

will depend on the R dissimilarities to members of the representation set,

$$\pi(\mathbf{d}_i) = Pr(Y = 1|\mathbf{d}_i) = \frac{\exp\left(\alpha + \mathbf{d}_i^{\mathsf{T}}\boldsymbol{\beta}\right)}{1 + \exp\left(\alpha + \mathbf{d}_i^{\mathsf{T}}\boldsymbol{\beta}\right)},$$

where α represents the intercept, and $\boldsymbol{\beta}$ represents an R vector of regression coefficients, and $\pi(\mathbf{d}_i)$ is the conditional probability that the outcome Y belongs to class 1 (i.e., $\pi(\mathbf{d}_i) = Pr(Y = 1|\mathbf{d}_i)$) and $1 - \pi(\mathbf{d}_i)$ represents the probability of being class 0. The model can be fitted by minimizing the penalized binomial deviance (Friedman et al. 2010). The resulting exemplars/prototypes for which $\beta_r \neq 0$ are called the *active* exemplars/prototypes.

2.2 Extending the δ-machine to mixed-type data

2.2.1 The transformed matrix

Given a mixed-type predictor matrix $\mathbf{X}$ of size $I \times P$, we construct the transformed matrix $\mathbf{X}^{\mathrm{M}}$ of size $I \times P^{\star}$, where $P^{\star} \geq P$. The matrix $\mathbf{X}^{\mathrm{M}}$ consists of two parts, the first part is a group of indicator variables obtained from different types of categorical variables, and the second part is a group of standardized continuous variables. Consider a row profile of object i from a predictor matrix of five variables

$$\mathbf{x}_i = (x_{ia}, x_{is}, x_{io}, x_{in}, x_{ic})^{\mathsf{T}},$$

Table 1 The indicator matrix $\mathbf{G}_p$ of the nominal predictor variable X_p with 3 levels

X_p	$\mathbf{G}_p$		
	g_{p1}	g_{p2}	g_{p3}
Level 1	1	0	0
Level 2	0	1	0
Level 3	0	0	1

Table 2 The cumulative indicator matrix $\mathbf{F}_p$ of the ordinal variable X_p with 3 levels

X_p	$\mathbf{F}_p$	
	f_{p_1}	f_{p_2}
Level 1	0	0
Level 2	1	0
Level 3	1	1

where x_{ia}, x_{is}, x_{io}, x_{in}, x_{ic} are values from asymmetric binary, symmetric binary, ordinal, nominal, and continuous variables. The corresponding transformed row profile of this object is

$$\mathbf{x}_i^{\mathrm{M}} = (b_{ia}, b_{is}, f_{i1}, \ldots, f_{i,(k-1)}, g_{i1}, \ldots, g_{ik}, z_{ic})^{\mathsf{T}},$$

where b_{ia} and b_{is} are the values of the indicator variables from asymmetric binary, symmetric binary variables respectively; $\{f_{i1}, \ldots, f_{i,(k-1)}\}$ and $\{g_{i1}, \ldots, g_{ik}\}$ are the values of the (cumulative) indicator variables from ordinal and nominal variables; z_{ic} is a standardized continuous variable. The way of computing z_{ic} is dissimilarity function-dependent.

Here we present how to convert different types of categorical variables into indicator variables.

- If the variable X_p is binary (no matter symmetric or asymmetric), X_p is represented by an indicator variable B_p with values of 0 and 1.
- If the variable X_p is nominal with K_p levels, X_p is represented by an *indicator matrix* $\mathbf{G}_p$ of column size K_p. An example of constructing an indicator matrix is given in Table 1, where g_{p_1}, g_{p_2} and g_{p_3} are the indicator variables of the indicator matrix $\mathbf{G}_p$; This indicator matrix has a disjoint structure, that is, each row profile from this matrix has a single occurrence of 1.
- If the variable X_p is ordinal with K_p levels, X_p is represented by a *cumulative indicator matrix* $\mathbf{F}_p$ of column size $(K_p - 1)$ (see Table 2). This was called conjoint coding in Heiser (1981, pp. 123–124).

2.2.2 Two adjusted dissimilarity functions

We will define two adjusted dissimilarity functions based on the transformed matrix $\mathbf{X}^{\mathrm{M}}$. The definitions of the pairwise adjusted Gower dissimilarity and the pairwise

Table 3 Dissimilarities and weights of the four possible combinations of an asymmetric binary variable

Object i	Object r	d_{irp}^{G}	w_{irp}
Presence	Presence	0	1
Presence	Absence	1	1
Absence	Presence	1	1
Absence	Absence	0	0

adjusted Euclidean dissimilarity between objects i and r are as follows

$$d_{ir}^{\mathrm{G}} = \frac{\sum_{p=1}^{P} w_{irp} d_{irp}^{\mathrm{G}}}{\sum_{p=1}^{P} w_{irp}}, \tag{1a}$$

$$d_{ir}^{\mathrm{E}} = \sqrt{\sum_{p=1}^{P} (d_{irp}^{\mathrm{E}})^2}, \tag{1b}$$

where d_{irp}^{G} and $(d_{irp}^{\mathrm{E}})^2$ are the adjusted Gower dissimilarity and the squared adjusted Euclidean dissimilarity between objects i and r on variable X_p respectively. In Eq. (1a) $w_{irp} = 1$ if objects i and r can be compared on variable X_p and $w_{irp} = 0$ otherwise. All types of variable have $w_{irp} = 1$ except for asymmetric binary. Below we define d_{irp}^{G} and $(d_{irp}^{\mathrm{E}})^2$ for the five types of predictor variables according to the order in Sect. 1.

– If the variable X_p is binary,

$$d_{irp}^{\mathrm{G}} = \begin{cases} 0 & x_{ip} = x_{rp} \\ 1 & x_{ip} \neq x_{rp} \end{cases}, \tag{2a}$$

$$(d_{irp}^{\mathrm{E}})^2 = (b_{ip} - b_{rp})^2 = \begin{cases} 0 & x_{ip} = x_{rp} \\ 1 & x_{ip} \neq x_{rp} \end{cases}, \tag{2b}$$

where b_{ip} and b_{rp} are the values of the indicator variable collected in the transformed matrix $\mathbf{X}^{\mathrm{M}}$ for objects i and r. For the AEDF, symmetric and asymmetric binary variables are treated in the same way. For the AGDF, the only difference lies in w_{irp} with the absence-absence matches.
Table 3 shows the four possible combinations that may occur for the two objects on X_p, and gives the values of w_{irp} for these combinations accordingly. The last row of Table 3 shows the irrelevance of the negative matches ($w_{irp} = 0$). The idea of the irrelevance of negative matches was proposed by Jaccard (1912) in the well-known Jaccard similarity function.

– If the variable X_p is nominal with K_p levels,

$$d_{irp}^{\mathrm{G}} = \begin{cases} 0 & x_{ip} = x_{rp} \\ 1 & x_{ip} \neq x_{rp} \end{cases}, \tag{3a}$$

$$(d_{irp}^{\mathrm{E}})^2 = \sum_{k=1}^{K_p} (g_{ipk} - g_{rpk})^2 = \begin{cases} 0 & x_{ip} = x_{rp} \\ 2 & x_{ip} \neq x_{rp} \end{cases}, \tag{3b}$$

where g_{ip_k} and g_{rp_k} are the values of the Kth indicator variable of the indicator matrix $\mathbf{G}_p$ (see Table 1) collected in the transformed matrix $\mathbf{X}^{\mathrm{M}}$ for objects i and r.

– If the variable X_p is ordinal with K_p levels,

$$d_{irp}^{\mathrm{G}} = \frac{\sum_{k=1}^{K_p-1} |f_{ip_k} - f_{rp_k}|}{K_p - 1}, \tag{4a}$$

$$(d_{irp}^{\mathrm{E}})^2 = \sum_{k=1}^{K_p-1} (f_{ip_k} - f_{rp_k})^2, \tag{4b}$$

where f_{ip_k} and f_{rp_k} are the values of the Kth indicator variable of the cumulative indicator matrix $\mathbf{F}_p$ (see Table 2) collected in the transformed matrix $\mathbf{X}^{\mathrm{M}}$ for objects i and r.

– If the variable X_p is continuous,

$$d_{irp}^{\mathrm{G}} = |z_{ip}^{\mathrm{G}} - z_{rp}^{\mathrm{G}}|, \qquad \text{where} \quad Z_p^{\mathrm{G}} = \frac{X_p^{\mathrm{G}}}{r(X_p)}, \tag{5a}$$

$$(d_{irp}^{\mathrm{E}})^2 = (z_{ip}^{\mathrm{E}} - z_{rp}^{\mathrm{E}})^2, \qquad \text{where} \quad Z_p^{\mathrm{E}} = \frac{(X_p - \mu_{X_p})}{\phi_{X_p}}. \tag{5b}$$

where Z_p^{G} and Z_p^{E} are the dissimilarity function-dependent standardized continuous variable in the transformed matrix $\mathbf{X}^{\mathrm{M}}$, and $r(X_p)$, ϕ_{X_p} and μ_{X_p} represent the range, the standard deviation and the mean of the variable X_p respectively.

2.2.3 Dissimilarity functions for purely categorical data

When all predictors are categorical, the predictor matrix $\mathbf{X}$ has a limited set of potential row profiles. We store all possible row profiles in a new matrix, namely, *the exemplar matrix* $\mathbf{E}$. The dissimilarity matrix is calculated from the exemplar matrix. Therefore, for a large categorical data set, we can generate a smaller dissimilarity matrix of size that is equal to the number of all possible row profiles. Suppose that we have $I = 10,000$ and one ordinal predictor with three levels and one binary predictor. Then the data cannot have more than $3 \times 2 = 6$ different row profiles. The size of the dissimilarity matrix, therefore, is 6 by R instead of 10,000 by R. The corresponding response is not a vector $\mathbf{y}$ of length I but a 6 by 2 matrix $\mathbf{Y}$, where the first column is the number of "failures" and the second column is the number of "successes".

2.2.4 Selection of the representation set

Following Yuan et al. (2019), K-means clustering and K-medoids clustering can be used to select the representation set, also for mixed-type. They applied the Partitioning around medoids (PAM) algorithm (Kaufman and Rousseeuw 1990) for K-medoids clustering. As PAM accepts any dissimilarity matrix, we can simply use the AEDF or the AGDF. For K-means clustering, however, we need a specific adaptation to derive the prototypes. Because the results of K-means clustering are prototypes rather than exemplars, the indicator variables of a selected prototype do not have two values (i.e., 0 and 1) but the average values from the data points from this cluster. Given a binary or nominal variable, the AGDF checks if two objects have the same value on this variable. For the average of the indicator variables, the values are unlikely to be 0 or 1. This leads to the higher chance of obtaining $d_{irp}^{\mathrm{G}} = 1$, because the prototype and an object are unlikely to have the same value on this indicator variable. To solve this problem, two approaches can be considered, (a) we round average indicator variables, or (b) we treat the indicator variable as continuous variable. Specially, in approach (a), if the value of an average indicator variable is 0.8, the new value for this indicator variable is rounded to be 1 (as $0.8 > 0.5$). However, for a nominal variable, we assign 1 to the highest value of each row of the indicator matrix and set the rest to be 0s. We do not directly round the values because it could have all zeros for a row of the indicator matrix. It is impossible to happen to a nominal variable, because each object belongs to a certain category. Whereas, for an ordinal variable, a row of all zeros means that it belongs to the first category. After rounding (or) assigning, the obtained variables have values of 0 and 1, and the AGDF can be applied. By contrast, in approach (b), we directly apply the AGDF with the formula of the continuous variable, i.e. Eq. (5a). Note that by taking all indicator variables as continuous, the property of asymmetric binary variables for the AGDF is lost. Although for the AEDF we can simply apply the average indicator variables in the defined equations, which falls in the proposed approach (b), we still implement approach (a) to the AEDF to make a fair comparison.

To determine the number of clusters, we follow the idea of Yuan et al. Specifically, for K-means clustering, they used an automatic stopping rule, similar to the decision rule in Mirkin (1999) and Steinley and Brusco (2011). That is, the algorithm stops searching for more clusters if the percentage of explained variance exceeds a user-specific threshold. For PAM, Yuan et al. tried K from 2 to 10 and chose the best number of clusters by the optimum average silhouette width.

2.3 Detailed theoretical comparison of the two adjusted dissimilarity functions

In this section we discuss the relationship between the two dissimilarity functions in three possible situations: the data consist of only continuous predictor variables, only categorical predictor variables, and mixed-types.

2.3.1 Data consisting of only continuous predictors

In this situation, the pairwise adjusted Gower dissimilarity and the pairwise adjusted Euclidean dissimilarity between objects i and r, i.e., Eqs. (1a) and (1b) become

$$d_{ir}^{\mathrm{G}} = \frac{1}{P}\sum_{p=1}^{P}\frac{|x_{ip} - x_{rp}|}{r(X_p)}, \tag{6a}$$

$$d_{ir}^{\mathrm{E}} = \sqrt{\sum_{p=1}^{P}\frac{(x_{ip} - x_{rp})^2}{var(X_p)}}. \tag{6b}$$

In Eq. (6a) the variables are feature scaled by the range ($r(X_p)$) and then the Manhattan distance function is used to calculate the dissimilarities. Feature scaling is used to bring all values into the range [0,1]. In Eq. (6b) the variables are first replaced by z-scores and then the ordinary Euclidean distance function is used to calculate the dissimilarities.

2.3.2 Data consisting of only categorical predictors

In this situation, the predictors can be asymmetric binary, symmetric binary, nominal, ordinal, or a mix of the four. A categorical predictor variable will be replaced by the indicator matrix $\mathbf{G}_p$ or the cumulative indicator matrix $\mathbf{F}_p$, or the binary indicator variable B_p, according to its type. The square of the difference gives the same value as the absolute value of the difference between indicator variables. For example, in Eqs. (2a) and (2b), the results of $(b_{ip} - b_{rp})^2$ and $|b_{ip} - b_{rp}|$ are equal. To make the comparison simpler, we replace the squared difference with the absolute difference in the AEDF.

– If the data consist of purely symmetric binary predictor variables,

$$d_{ir}^{\mathrm{G}} = \frac{1}{P}\sum_{p=1}^{P}|b_{ip} - b_{rp}|, \tag{7a}$$

$$d_{ir}^{\mathrm{E}} = \sqrt{\sum_{p=1}^{P}|b_{ip} - b_{rp}|}. \tag{7b}$$

  In this situation, the result of the AGDF is proportional to the squared AEDF.
– If the data consist of purely asymmetric binary predictor variables, the pairwise adjusted Gower dissimilarity is

$$d_{ir}^{\mathrm{G}} = \frac{\sum_{p=1}^{P} w_{irp}|b_{ip} - b_{rp}|}{\sum_{p=1}^{P} w_{irp}}, \tag{8}$$

where $w_{irp} = 0$ for negative matches. The pairwise adjusted Euclidean dissimilarity is the same as Eq. (7b). In this situation, the results of the AGDF are not proportional to the squared AEDF because of the negative match-related w_{irp}.

– If the data consist of purely nominal predictor variables,

$$d_{ir}^{\mathrm{G}} = \frac{1}{P}\sum_{p=1}^{P} d_{irp}^{\mathrm{G}}, \quad d_{irp}^{\mathrm{G}} = \begin{cases} 0 & x_{ip} = x_{rp} \\ 1 & x_{ip} \neq x_{rp}, \end{cases} \tag{9a}$$

$$d_{ir}^{\mathrm{E}} = \sqrt{\sum_{p=1}^{P} (d_{irp}^{\mathrm{E}})^2}, \quad (d_{irp}^{\mathrm{E}})^2 = \begin{cases} 0 & x_{ip} = x_{rp} \\ 2 & x_{ip} \neq x_{rp} \end{cases}, \tag{9b}$$

Eq. (9b) is the square root of the sum of 0s and 2s, whereas Eq. (9a) is that the sum of 0s and 1s divided by the number of variables. Therefore in this situation, the results of the AGDF are also proportional to the squared AEDF.

– If the data consist of purely ordinal predictor variables,

$$d_{ir}^{\mathrm{G}} = \frac{1}{P}\sum_{p=1}^{P} \frac{\sum_{k=1}^{K_p-1} |f_{ip_k} - f_{rp_k}|}{(K_p - 1)}, \tag{10a}$$

$$d_{ir}^{\mathrm{E}} = \sqrt{\sum_{p=1}^{P}\sum_{k=1}^{K_p-1} |f_{ip_k} - f_{rp_k}|}. \tag{10b}$$

For the special case that all ordinal predictors have K levels, the results of the AGDF are proportional to the squared AEDF.

In summary, if the data consist of purely symmetric binary, or purely nominal, or purely ordinal variables with same levels, the result of the AGDF is proportional to the squared AEDF.

2.3.3 Data consisting of continuous and categorical predictors

Suppose that the data consist of P_1 continuous predictors and P_2 categorical predictors, so that $P = P_1 + P_2$. The AGDF and the AEDF are the sum of the continuous part and the categorical part as shown in the previous two situations. More specifically, the AGDF between two objects is the sum of the Manhattan distances of the P_1 continuous predictors and the adjusted Gower dissimilarities of the P_2 categorical predictors. The AEDF between two objects is the squared root of the sum of the squared ordinary Euclidean distances of the P_1 continuous predictors and the squared adjusted Euclidean dissimilarities of the P_2 categorical predictors.

3 Simulation studies

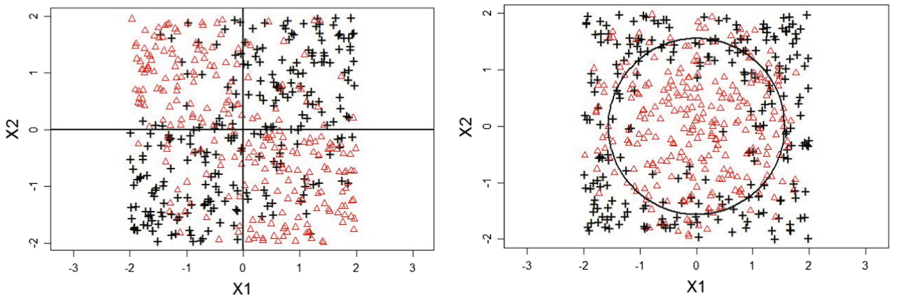
The main goals of the simulation studies are to show the predictive performance of the δ-machine in comparison to logistic regression models and to compare the two adjusted dissimilarity functions under two main situations: data with mixed-type and data with purely categorical variables. Two logistic regression models are used as reference methods: logistic regression with (LR_+) and without two-way interactions (LR). The comparison of logistic regression and the δ-machine is to show the difference between building a classifier in the predictor space (i.e., variable-oriented approach) and building it in the dissimilarity space (i.e., person-oriented approach).

3.1 Data generation

We first generate data with binary outcomes and continuous predictors, and then convert the continuous predictors into categorical ones according to different conditions. In the conversion process, the underlying relationships between predictors and the outcomes are maintained. Therefore, despite the different conditions for predictor variables, the relationships are similar on the same problem so that we can make comparisons across different types of predictor variables on the same problem.

We consider two of the three artificial problems studied in Yuan et al. (2019): the four blocks problem and the Gaussian ordination problem (see Fig. 2). Because Yuan et al. (2019) found the δ-machine had very good predictive performance on the third problem, we do not include it here. For the four blocks problem, the data have a pure interaction between the predictor variables, that is, no main effects. The predictor variables are independent and identically generated from the uniform distribution in the range $[-2, 2]$. An observation with a higher product of predictor variables (i.e. $\prod_{p=1}^{P} x_{ip}$) has a higher chance to be assigned to class 1. For the Gaussian ordination problem, the relationship between the outcome and the predictor variables is single peaked in the multivariate space. The predictor variables are independent and identically generated from the uniform distribution in the range $[-2, 2]$. An observation with a higher sum of squares of predictor variables (i.e. $\sum_{p=1}^{P} x_{ip}^2$) has a higher chance to be assigned to class 1 (see Yuan et al. 2019 for details).

We convert continuous predictors into categorical ones. We consider three types of categorical variables, ordinal, nominal and binary. We distinguish between balanced and unbalanced categorical variables, where in the balanced condition the objects are evenly distributed among the levels whereas in the unbalanced condition they are not. Suppose that the original data have two continuous predictor variables. To obtain mixed-type data, the first continuous variable is converted to ordinal with balanced levels, and the second variable keeps unchanged. We use the set of quantiles $\{1/3, 2/3\}$ to categorize the first variable X_1, and the objects are equally assigned to the three levels (see Fig. 3a). Similarly, to categorize X_1 as ordinal with unbalanced levels, the set of quantiles $\{1/2, 5/6\}$ is applied. In this situation X_1 still has a clear ordering but the objects are unequally distributed across levels (see Fig. 3b). To convert the first variable X_1 into nominal with balanced levels, we use the same sets of quantiles as the ordinal ones, but we reverse the levels of predictors. For instance, the objects of X_1



(a) Four blocks. **(b)** Gaussian ordination.

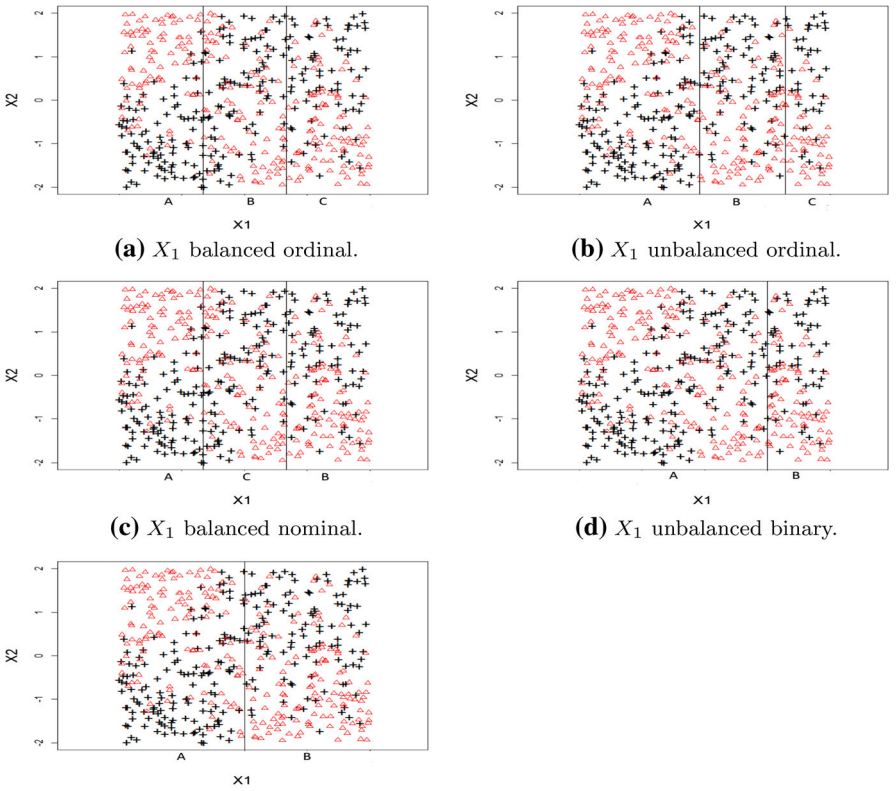
Fig. 2 Examples of two artificial problems in two dimensions. The triangle and the cross denote the object classified as class 0 and class 1 respectively. The lines are the decision boundaries in each problem

smaller than the first quantile are still assigned to the first level, but the second level and the third level are reversed, i.e., $\{A, C, B\}$ (see Fig. 3c). Likewise, a continuous variable can be categorized to nominal with unbalanced levels using the set quantiles $\{1/2, 5/6\}$. We recode X_1 to binary data in two ways: unbalanced and balanced. In the unbalanced case we dichotomize on the 0.2 quantile (see Fig. 3d); In the balanced case we dichotomize on the 0.5 quantile (see Fig 3e). The converted predictor variable has two possible values, where the levels of A and B stand for the values of 1 and 0 respectively.

3.2 Five simulation studies

Simulation studies 1 to 3 evaluate the performance on data with mixed-type predictor variables. Simulation studies 4 and 5 evaluate the performance on data with purely categorical predictor variables. Specifically, in Study 1 we consider data with ordinal/nominal and continuous predictor variables. The generated data vary in the following factors: the number of predictor variables, the type of categorical variables, and the distribution of categorical variables (see Table 4). In Study 2 we generate data with binary and continuous predictor variables. We investigate the performance of the asymmetric and symmetric measures in the adjusted Gower dissimilarity function. The factors in Table 4 are used except the last row. In Study 3, we generate data with 15 predictor variables which are a mix of ordinal, nominal, binary, and continuous variables. Moreover, we investigate different ways of selecting the representation set. Three selection methods are considered: (a) use the training set; (b) use PAM; (c) use K-means clustering (using both thresholds $v_e = 0.5$ and 0.9). As discussed in Sect. 2.2.4, we proposed two approaches for K-means clustering for the data with mixed-type predictor variables. We will apply both approaches. In Study 4 and 5 we generate purely categorical predictor variables, where in Study 4 the data have two ordinal/nominal predictors and in Study 5 two binary predictor variables.

All generated data have size of 500. We split the data randomly into two parts (2/3 and 1/3) referred to as a training set and a test set. The performance results are reported in terms of misclassification rate (MR) on the test set. We do not use other criteria to



(a) X_1 balanced ordinal. **(b)** X_1 unbalanced ordinal.

(c) X_1 balanced nominal. **(d)** X_1 unbalanced binary.

(e) X_1 balanced binary.

Fig. 3 Five examples of partition for the four blocks problem. The triangle and the cross denote the object classified as class 0 and class 1. The vertical lines are the quantiles of X_1. The labels from A to C refer to the levels of the categorical predictor. The variable X_2 is always continuous. **a** and **b** The first continuous predictor X_1 is converted into an ordinal predictor of (un)balanced levels. **c** The predictor X_1 is converted into a nominal predictor of balanced levels. **d** and **e** The predictor X_1 is converted into unbalanced and balanced binary predictors respectively

Table 4 Summary of factors and levels for simulation Study 1–2

Factor	#	Levels
Artificial problem	2	Four blocks, Gaussian ordination
Number of predictor variables	2	2, 5
Distribution of categorical predictor	2	Balanced, unbalanced
Type of categorical predictor (only for Study 1)	2	Ordinal (3 levels), Nominal (3 levels)

The last row is the extra factor for simulation study 1

evaluate the performance, because the generated data have balanced outcomes. For each condition we use 100 replications. The performance is the average MR over 100 replications.

Table 5 The results obtained from Study 1 for the two problems

Problem	m	t	b	v	$m:t$	$m:b$	$m:v$
(a) Estimated effect size η^2 for each model component.							
Four blocks	0.32	0.00	0.00	0.26	0.00	0.00	0.24
Gaussian ordination	0.62	0.00	0.02	0.01	0.00	0.01	0.02

Problem	LR	LR_+	δ_Gower	δ_Euc
(b) The average misclassification rate of the four methods.				
Four blocks	0.50 (0.05)	0.38 (0.12)	0.49 (0.04)	**0.37** (0.10)
Gaussian ordination	0.40 (0.05)	0.42 (0.05)	0.31 (0.03)	**0.30** (0.04)

Note: (a) The method related effect sizes which are larger than or equal to 0.06 are underlined. m, t, b, v, $m:b$, $m:t$, $m:v$ are abbreviations of the method, the type of categorical predictor, the distribution of categorical predictor, the number of predictor variables and their interaction terms with the method factor respectively

(b) Best results are bold. The standard deviations are shown in brackets. LR, LR_+, δ_Gower, δ_Euc are abbreviations of logistic regression, logistic regression with two-way interactions, the δ-machine using the AGDF, and the AEDF, respectively

The results of the simulations will be analyzed using analysis of variance (ANOVA), and statistics for factors and their interactions will be tested. The two artificial problems will be analyzed separately. To assess the effect size of the factors and their interactions, η squared (η^2) (Cohen 1973) is used, which ranges from 0 to 1. A common rule of thumb is that η^2 values of 0.01, 0.06, 0.14 represent a small, a medium and a large effect size, respectively (Cohen 1988, pp. 280–287).

To perform the studies, we use the open source statistical analysis software R (R Core Team 2015). All source code is available at (https://osf.io/9gz3j/?view_only=d04da7c14c2e46999c32720f65a7a054).

3.3 Results

3.3.1 Study 1: Nominal/ordinal and continuous predictor variables

The results obtained from ANOVA are shown in Table 5a. For the four block problems, the method effect was large ($\eta^2 = 0.32$), showing that applying different methods had large effect on the MR. Table 5b shows that logistic regression with two-way interactions (LR_+) and the δ-machine with the AEDF had the lowest misclassification rates (0.38 and 0.37). Using these two methods instead of LR and the δ-machine with the AGDF the MR deceases sharply. The interaction between the method factor and the number of predictor variables ($m:v$) had a large effect size. Figure 4 illustrates the effect size of the ($m:v$) interaction. As the number of predictor variables increases, the difference between the four methods disappears.

For the Gaussian ordination problem, the method factor had a large effect on the MRs. The δ-machine, regardless of the dissimilarity function, had significantly lower MRs than logistic regression. The interaction between the method factor and the dis-

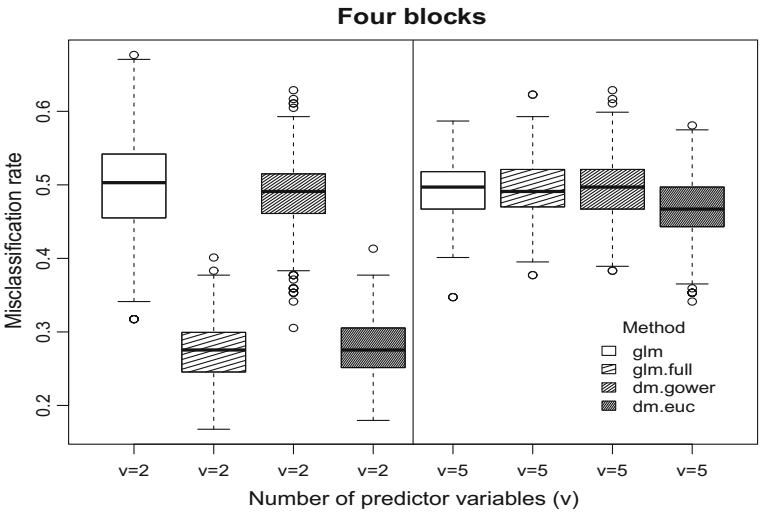


Fig. 4 Box plots of the misclassification rate for logistic regression with (LR_+) and without two-way interactions (LR) and the δ-machine with the AGDF (Gower) and the AEDF (Euclidean) on the data of two predictors (left panel) and the data of five predictors (right panel)

tribution of categorical predictor ($m : b$) and the interaction between the method factor and the number of predictor variables ($m : v$) had only small effect sizes ($\eta^2 > 0.01$).

In this study, for the four blocks problem, the δ-machine using the AEDF had lower MRs than the AGDF. The δ-machine had competitive MRs to logistic regression. As the number of predictor variables increases, all methods failed to make accurate predictions. For the Gaussian ordination problem, the two adjusted dissimilarity functions had similar MRs. The δ-machine had lower MRs than logistic regression regardless the chosen dissimilarity function.

3.3.2 Study 2: Binary and continuous predictor variables

As for the AGDF, there are two ways of treating binary variables, in this study we have five methods instead of four: logistic regression with (LR_+) and without interactions (LR), the δ-machine using the AGDF asymmetric binary (δ_Gower_asy), using the AGDF symmetric binary (δ_Gower_sy), and using the AEDF (δ_Euc). For the four block problem, the δ-machine using the AEDF, the AGDF asymmetric binary and logistic regression with two-way interactions (LR_+) had the lowest MR (see Table 6a). The method factor (m) had a large effect size, also reflecting that using different methods had large influences on MRs. The interaction between the method and the number of predictor variables ($m : v$) had a large effect size. Figure 5 illustrates the interaction ($m : v$) that as the number of predictors increases, the difference in terms of MR between these methods has vanished. When the data have five predictor variables, all methods failed to make accurate predictions. For the Gaussian ordination problem, the δ-machine had significant lower MR than logistic regression regardless of the dissimilarity function chosen.

Table 6 The results obtained from Study 2 for the two problems

Problem	m	b	v	$m:b$	$m:v$
(a) Estimated effect size η^2 for each model component.					
Four blocks	<u>0.18</u>	0.03	0.35	0.02	<u>0.18</u>
Gaussian ordination	<u>0.66</u>	0.07	0.00	0.04	0.00

Problem	LR	LR_+	δ_Gower_asy	δ_Gower_sy	δ_Euc
(b) The average misclassification rate of the five methods.					
Four blocks	0.50 (0.05)	**0.40** (0.11)	0.41 (0.11)	0.49 (0.04)	0.41 (0.11)
Gaussian ordination	0.46 (0.06)	0.46 (0.06)	**0.32** (0.04)	**0.32** (0.04)	**0.32** (0.04)

(a) The method related effect sizes which are larger than or equal to 0.06 are underlined. m, b, v, $m:b$, $m:v$ are abbreviations of the method, the distribution of categorical predictor, the number of predictor variables and their interaction terms with the method factor respectively.
(b) Best results are bold. The standard deviations are shown in brackets. LR, LR_+, δ_Gower_asy, δ_Gower_sy, δ_Euc are abbreviations of logistic regression, logistic regression with two-way interactions, the δ-machine using the AGDF with asymmetric measure, the AGDF with symmetric measure, the AEDF respectively

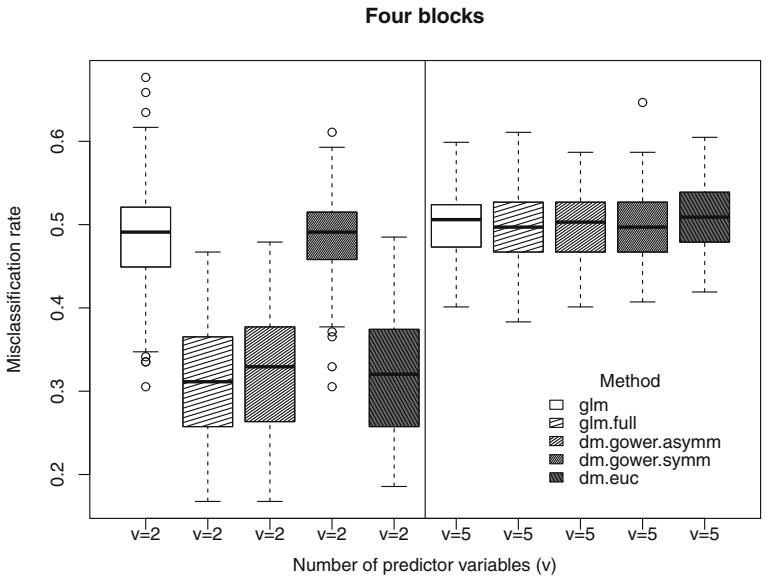


Fig. 5 Box plots of the misclassification rate for logistic regression with (LR_+) and without two-way interactions (LR) the δ-machine with the AGDF (Gower_asy and Gower_sy) and the AEDF (Euc) on the data of two predictors (left panel) and the data of five predictors (right panel)

Table 7 Description of the predictor variables in Study 3

Variable	Type	Level	Quantiles
X_1	Binary	2	0.5/0.5
X_2	Binary	2	0.5/0.5
X_3	Binary	2	0.2/0.8
X_4	Binary	2	0.2/0.8
X_5	Ordinal	4	0.25/0.25/0.25/0.25
X_6	Ordinal	4	0.25/0.25/0.25/0.25
X_7	Nominal	3	1/3/1/3/1/3
X_8	Nominal	3	1/3/1/3/1/3
X_9–X_{15}	Numerical	–	

3.3.3 Study 3: Fifteen mixed-type predictor variables

The results obtained from Studies 1 and 2 showed that for the four blocks problem when the number of predictor variables was 5, all methods failed to make accurate predictions. Therefore, we do not further consider the four block problem in the current study, but only the Gaussian ordination problem with 15 mixed-type predictor variables. A description of the predictor variables is given in Table 7.

Table 8 summarizes the results obtained from this study. The δ-machine using the AEDF had the lowest MR (0.33), but the difference between the two adjusted dissim-

ilarity functions was small (0.01). The δ-machine had lower misclassification rates than logistic regression (MR $>$ 0.4) regardless of the dissimilarity function chosen.

The comparison of the different representation sets is shown in Parts I, II and III. Overall, the δ-machine using the training set as representation set had slightly lower MR than using the other types. However, the obtained models had many more active exemplars or prototypes, which makes the obtained models difficult to interpret. The models obtained from a smaller representation set, especially the representation set selected by PAM, were sparser than the ones using the training set. The δ-machine using the AEDF and PAM had a comparable misclassification rate (0.35) and the smallest number of active exemplars (7.27) compared to the δ-machine using the other types. The δ-machine using K-means clustering did not decrease the number of active prototypes. For the two approaches of K-means clustering, we found that the average misclassification rates were very similar, whereas using approach (a) had fewer active prototypes than using approach (b).

In conclusion, the δ-machine outperformed logistic regression. The two adjusted dissimilarity functions had similar performance in this study. Using a smaller representation set, e.g., a representation set selected by PAM had a comparable misclassification rate but far fewer active exemplars.

3.3.4 The case of purely categorical data

We also performed two simulation studies with purely categorical data, to compare the two dissimilarity functions. One was on data with two nominal or two ordinal predictors. The other was on data with two binary predictors. It turned out that in both studies the δ-machine using the AEDF had a better performance than the AGDF on the four blocks problem. For the Gaussian ordination problem, these two dissimilarity functions showed similar results. Compared to logistic regression, the δ-machine performed equally well. In conclusion, for purely categorical data, it is recommended to use the AEDF in the δ-machine. The δ-machine had the same performance as logistic regression, which is therefore the preferred method.

4 Application

The δ-machine is illustrated using the Statlog heart data. The Statlog heart data contains 270 objects, who were patients referred for coronary arteriography at the Cleveland Clinic (Detrano et al. 1989). Each patient was described by 13 predictor variables. Six predictors are continuous; one predictor is ordinal; three predictors are binary, and three predictors are nominal (see Table 9). The outcome variable refers to the presence or absence of heart disease. For more details about this data set see Detrano et al. (1984). This data set is available from the UCI repository of machine learning database (Dheeru and Karra Taniskidou 2017).

Quite some papers compared classification methods using the Statlog heart data (Şahan et al. 2005; Brown 2004; Kotsiantis and Pintelas 2004). In these studies, Classification trees and Naive Bayes were commonly considered good candidate classifiers. Naive Bayes, particularly, often achieved the best predictive performance. Kotsiantis

Table 8 The average results of the compared methods in Study 3

Part	Representation set	Method	MR	Act.Exp
I	Training	δ_Euc	**0.33** (0.04)	66.08 (56.13)
		δ_Gower_asym	0.34 (0.04)	27.74 (11.27)
		δ_Gower_sym	0.34 (0.04)	26.12 (7.58)
II	PAM	δ_Euc	0.35 (0.04)	**7.27** (4.77)
		δ_Gower_asym	0.36 (0.04)	10.45 (4.62)
		δ_Gower_sym	0.38 (0.04)	7.97 (4.19)
III	K-means_$v_e = 0.5$	δ_Euc_approach (a)	0.35 (0.04)	35.45 (8.82)
		δ_Gower_asym_approach (a)	0.35 (0.04)	21.42 (9.21)
		δ_Gower_sym_approach (a)	0.35 (0.04)	23.70 (8.89)
		δ_Euc_approach (b)	0.36 (0.04)	40.33 (3.24)
		δ_Gower_approach (b)	0.35 (0.04)	24.92 (10.47)
	K-means_$v_e = 0.9$	δ_Euc_approach (a)	0.35 (0.04)	122.13 (71.22)
		δ_Gower_asym_approach (a)	0.34 (0.04)	31.45 (15.22)
		δ_Gower_sym_approach (a)	0.34 (0.04)	30.55 (15.42)
		δ_Euc_approach (b)	0.37 (0.04)	163.05 (46.18)
		δ_Gower_approach (b)	0.35 (0.04)	40.22 (23.80)
IV		LR	0.43 (0.04)	
		LR_+	0.49 (0.04)	

Best results are bold. The standard deviations over the 100 replications are shown in brackets. The δ_Euc, δ_Gower_asym, δ_Gower_sym, LR, LR_+ are abbreviations of the δ-machine using the AEDF, using the AGDF (asymmetric and symmetric measure), logistic regression without and with two-way interactions. Training, PAM, K-means_$v_e = 0.5$, and K-means_$v_e = 0.9$ are abbreviations of the representation set using the complete training set, the representation set selected by PAM and by K-means clustering with the threshold as 0.5 and 0.9 respectively. The approaches (a) and (b) are the two approaches that proposed for K-means clustering for mixed-type data (see Sect. 2.2.4 for details). MR and Act.Exp are abbreviations of misclassification rate and the number of active exemplars or prototypes. The cells of Act.Exp of logistic regression are left blank, because they used all the available predictors for classification

and Pintelas (2004) showed that the difference between single Naive Bayes and ensemble methods like bagging (Breiman 1996), Adaboost (Freund and Schapire 1996), Multiboost (Webb 2000), and DECORATE (Melville and Mooney 2003) were not substantial, although generally these sophisticated methods were slightly more accurate than a single classifier (Opitz and Maclin 1999).

The aims of this application are to compare the two adjusted dissimilarity functions in the δ-machine, to compare the δ-machine to five other classification methods, to present how the choice of the representation set influences the performance of the δ-machine, and to interpret the results of the δ-machine from a person-oriented perspective.

The five classification methods we compare the δ-machine to are logistic regression, Lasso logistic regression, SVM with radial basis kernel (SVM(RBF)), Classification trees, and Naive Bayes. To show the comparison of a linear classifier in the dissimilarity space versus the predictor space, we compare logistic regression and the Lasso with the δ-machine. For logistic regression, we only consider the main effects model. SVM

Table 9 Descriptions of variables in Statlog heart data

	Variable	Type	Level	Possible values
1	Age	Numerical	–	Min. = 29; Max. = 77
2	Gender	Binary	2	Male; Female
3	CPT: Chest pain type	Nominal	4	Typical angina; Atypical angina; Non-anginal pain; Asymptomatic
4	RBP: Resting blood pressure	Numerical	–	Min. = 94; Max. = 200
5	SC: Serum cholesterol in mg/dl	Numerical	–	Min. = 126; Max. = 564
6	FBS: Fasting blood sugar > 120 mg/dl	Binary	2	Yes; No
7	RECG: Resting electrocardiographic results	Nominal	3	Normal; Having ST-T wave abnormality; Showing probable or definite left ventricular hypertrophy by Estes' criteria
8	MHT: Maximum heart rate achieved	Numerical	–	Min. = 71; Max. = 202
9	EIA: Exercise induced angina	Binary	2	Yes; No
10	ST: ST depression induced by exercise relative to rest	Numerical	–	Min.= 0; Max. = 6.2
11	Slope: The slope of the peak exercise ST segment	Ordinal	3	Upsloping; Flat; Downsloping
12	MVC: Number of major vessels that appeared to contain calcium	Numerical	–	Min. = 0; Max. = 3
13	Thal: Exercise thallium defects	Nominal	3	Normal; Fixed; Reversible
14	y: Heart disease	Binary	2	Absence (150); Presence (120)

is considered a good candidate because of its high generalization performance (James et al. 2013, p. 337). Parameter tuning is performed by a grid search over the two tuning parameters using 10-fold cross validation. Classification trees and Naive Bayes are considered because they were commonly applied on this dataset.

Because the data set has mixed continuous and categorical variables, here we explain how the different classification methods treat the categorical variables. We have shown how the δ-machine handles mixed-type data for the two dissimilarity functions. For logistic regression, Lasso logistic regression, and support vector machines, we do

not further distinguish categorical predictors but simply treated them as nominal and represent them with indicator matrices (Hsu et al. 2003). For Classification trees, a nominal predictor variable with K levels is ordered with the proportion falling in outcome class 1. Then this nominal variable is split as if it was an ordinal variable. In this way, the search for the best split is reduced from $2^{K-1} - 1$ to $K - 1$ splits (Ripley 1996; Breiman et al. 1984). For an ordinal predictor variable with K levels, there are $K - 1$ different possible splits (Breiman et al. 1984). Naive Bayes accepts all types of predictors (Friedman et al. 2009, pp. 210–211). Among all the classification methods, only the δ-machine using the AGDF distinguishes between symmetric and asymmetric binary predictors. For this data set, the three binary predictor variables are gender, fasting blood sugar > 120 mg/dl (FBS), and exercise induced angina (EIA). For the adjusted Gower dissimilarity function, the gender variable is treated as symmetric and the remaining two are treated as asymmetric, because the two levels of gender are equally homogeneous while the two levels of FBS and EIA are not. For example, considering the two levels of the variable EIA, we can say that the patients with angina have something in common, but that does not hold for the patients without angina.

In order to compare the predictive performance, we randomly split the data into a training ($n = 170$) and a test set ($n = 100$) and repeat this procedure ten times. We consider the area under the receiver operating characteristic (ROC) curve (AUC) (Fawcett 2006), MR and the sparsity of the models as the criteria. The average MR, AUC and sparsity over ten runs are computed on the test sets. To illustrate the interpretation we use one of these ten replications and develop the interpretational rules in detail.

The Lasso is implemented in the `glmnet` package (Friedman et al. 2010a). Support vector machines and Naive Bayes are implemented in the `e1071` package (Meyer et al. 2014), and Classification trees are implemented in the `rpart` package (Therneau et al. 2015).

4.1 Predictive performance

Table 10 displays the average results of the ten replications, where in Part I we compare the two dissimilarity functions. The results show that using the adjusted Euclidean dissimilarity function had a slightly lower MR than the adjusted Gower dissimilarity function. The AGDF had less active exemplars than the AEDF, however the difference was marginal. Therefore, the δ-machine using these two dissimilarity functions performed similarly on this data set.

Parts I and IV of Table 10 show the comparison of the δ-machine to the five other classification methods on the three criteria. Among all the methods, Naive Bayes achieved the lowest MR, followed by the δ-machine, the Lasso and SVM (RBF), but the difference was minor (0.01). In addition, these four methods had the same AUC value. Logistic regression had moderate results (MR = 0.21), but showed large standard deviations (0.09). Classification trees had the worst results (MR = 0.26). The δ-machine had sparser models than SVM. The average number of support vectors was 103, while for the δ-machine the number of active exemplars was around 11 for both dissimilarity functions. Lasso selected on average 12.3 predictors. Classification

Table 10 The average results of the all compared methods on Statlog heart data set

Part	Representation set	Method	MR	AUC	sparsity
I	Training	δ_Euc	0.17 (0.03)	**0.90** (0.02)	11.40 (4.62)
		δ_Gower	0.18 (0.03)	**0.90** (0.03)	10.90 (5.17)
II	PAM	δ_Euc	0.18 (0.03)	0.89 (0.02)	5.00 (2.26)
		δ_Gower	0.17 (0.03)	**0.90** (0.02)	8.30 (4.37)
III	K-means_$v_e = 0.5$	δ_Euc_approach (a)	0.18 (0.02)	0.89 (0.03)	12.20 (4.98)
		δ_Gower_approach (a)	0.18 (0.02)	0.89 (0.01)	12.30 (5.38)
		δ_Euc_approach (b)	0.19 (0.03)	0.88 (0.03)	25.60 (21.70)
		δ_Gower_approach (b)	0.37 (0.12)	0.87 (0.06)	13.30 (4.37)
	K-means_$v_e = 0.9$	δ_Euc_approach (a)	0.22 (0.04)	0.89 (0.03)	6.80 (1.32)
		δ_Gower_approach (a)	0.18 (0.02)	0.89 (0.02)	15.20 (8.12)
		δ_Euc_approach (b)	0.21 (0.03)	0.88 (0.03)	12.20 (4.87)
		δ_Gower_approach (b)	0.51 (0.10)	0.88 (0.02)	16.50 (8.38)
IV		Logistic regression	0.21 (0.09)	0.84 (0.12)	
		the Lasso	0.17 (0.02)	**0.90** (0.02)	12.30 (1.49)
		SVM (RBF)	0.17 (0.03)	**0.90** (0.02)	103.00 (11.12)
		Classification trees	0.26 (0.05)	0.77 (0.04)	**4.00** (0.82)
		Naive Bayes	**0.16** (0.03)	**0.90** (0.03)	

Best results are bold. The standard deviations over the ten replications are shown in brackets. The $v_e = 0.5$ and $v_e = 0.9$ are abbreviations of the representation set selected by K-means clustering with the threshold as 0.5 and 0.9, respectively. The approaches a and (b) are the two approaches that proposed for K-means clustering for mixed-type data. The definition of sparsity of the methods in this table is described as follows: the sparsity of the δ-machine is the number of active exemplars or prototypes. The sparsity of the Lasso is the number of selected predictors. The sparsity of support vector machines is the number of support vectors. The sparsity of classification trees is the number of splits. The sparsity of logistic regression and Naive Bayes are left blank, as they used all the available predictors for classification

trees had on average four splits, showing that the models can be well interpreted. The sparsity of logistic regression and Naive Bayes are left blank in the table, because they used all the available predictors for classification.

Parts II and III of Table 10 lists the results obtained from the δ-machine using smaller representation sets selected by either PAM or K-means. The δ-machine using PAM had remarkable results. That is, the average MR and AUC were equal to the δ-machine using the entire training set (see Part I), but the models had on average only 5.0 and 8.30 active exemplars. Compared with the sparsest method, Classification trees, the δ-machine using PAM had smaller MR and higher AUC. The representation set selected by K-means clustering performed less well. In Part III we show the comparison of the two proposed approaches for K-means. For the AEDF, using approach (a) had similar MRs and AUCs to approach (b), but with a smaller number of active prototypes. For the AGDF, using approach (a) resulted in lower MRs and higher AUCs.

Our findings from this empirical example can be summarized as follows. The δ-machine using two dissimilarity functions performed similarly on this data set. The δ-machine provided a good balance between sparsity and prediction accuracy compared to the other methods we considered. Using a smaller representation set (selected by PAM) can result in a sparser model but still comparable prediction performance. We recommend to use approach (a) for K-means clustering.

4.2 Interpretation of the obtained the δ-machine model

As given in Table 10, the δ-machine using a smaller representation set (selected by PAM) resulted in sparser models, and meanwhile had the same predictive performances as the one used the entire training set. Here we present a detailed interpretation of the model with K-medoids clustering using the Euclidean distance on the first replication. In this case, there are two active exemplars: Patient 155 who has no disease, and Patient 164 who has the disease. Table 11 gives an overview of the values of the predictor variables for these two active exemplars. The estimated model is

$$\text{logit}[p(\mathbf{x}_i)] = 1.95 + 0.93 \times d_{i,155} - 1.34 \times d_{i,164}, \tag{11}$$

where $d_{i,155}$ and $d_{i,164}$ are pairwise dissimilarities of object i towards to the two objects with row profiles $\mathbf{x}_{155}$ and $\mathbf{x}_{164}$, receptively. With every unit increase in the dissimilarity towards patient 155 the log odds of having the heart disease go up by 0.93. With every unit increase in the dissimilarity towards observation 164 the log odds go down by 1.34. In other words, if a patient is more dissimilar to the patient 155, he or she has a higher probability to get the heart disease. If a patient is more dissimilar to the patient 164, he or she has a lower probability to get the heart disease.

The prediction of the label of participants in the test set was based on the dissimilarities towards these two active exemplars. Figure 6 gives the 2D dissimilarity space built on these two active exemplars, and the the straight line is the estimated decision boundary. The coordinates of the patients from the test set are the pairwise dissimilarities of the patients towards the two active exemplars. The triangle and the cross denote the patients classified as “absence" and “presence" of the disease respectively,

Table 11 Descriptions of variables of patients $\mathbf{x}_{155}$ and $\mathbf{x}_{164}$

	Patient 155	Patient 164
Age	51	58
Gender	Female	Male
CPT	Non-anginal pain	Asymptomatic
RBP	130	128
SC	256	259
FBS	No	No
RECG	Left ventricular hypertrophy	Left ventricular hypertrophy
MHT	149	130
EIA	No	Yes
ST	0.5	3.0
Slope	Upsloping	Flat
MVC	0	2
Thal	Normal	Reversable defect
y	Absence	Presence

which are the observed class labels. If a case falls above the decision line it is classified as healthy (disease is absent), whereas if the case falls below the decision line it will be classified as diseased (disease is present). Despite some overlap, we can see a clear separation of the patients with or without the disease. The model with only two exemplars had good accuracy (MR = 0.15), which suggests that the efficiency of using the dissimilarities as predictors can be high. Furthermore, because the dissimilarity space could separate the objects relatively well, these two exemplars are well chosen.

Using dissimilarities as predictor variables makes it difficult to see the value of the original variables. Therefore, Yuan et al. used variable importance measures by using a permutation test approach, similar to the importance measure for random forests (Breiman 2001). Besides investigating the importance of a particular predictor variable, Yuan et al. used partial dependence plots (Friedman 2001; Berk 2008) to interpret the marginal relationship between predictor variables and the response (Yuan et al. 2019). For this data set, the variable importance plot suggests that the important predictors are MVC, ST, CPT and Gender (see Fig. 7). The other predictors are of minor importance. Subsequently, we made the partial dependence plots for the predictor variables considered to be important. The partial dependence plot of MVC illustrates the positive relationship between the number of major vessels that appeared to contain calcium and the presence of heart disease. The more major vessels contained calcium the higher probability to have coronary heart disease. The levels of CPT from one to four stand for typical angina, atypical angina, non-anginal pain, asymptomatic, respectively. The results showed that asymptomatic patients were more likely to have coronary heart disease than symptomatic patients. The partial dependence plot of ST illustrates the positive relationship between the number of ST depression included by exercise relative to rest and the presence of heart disease (Fig. 8).

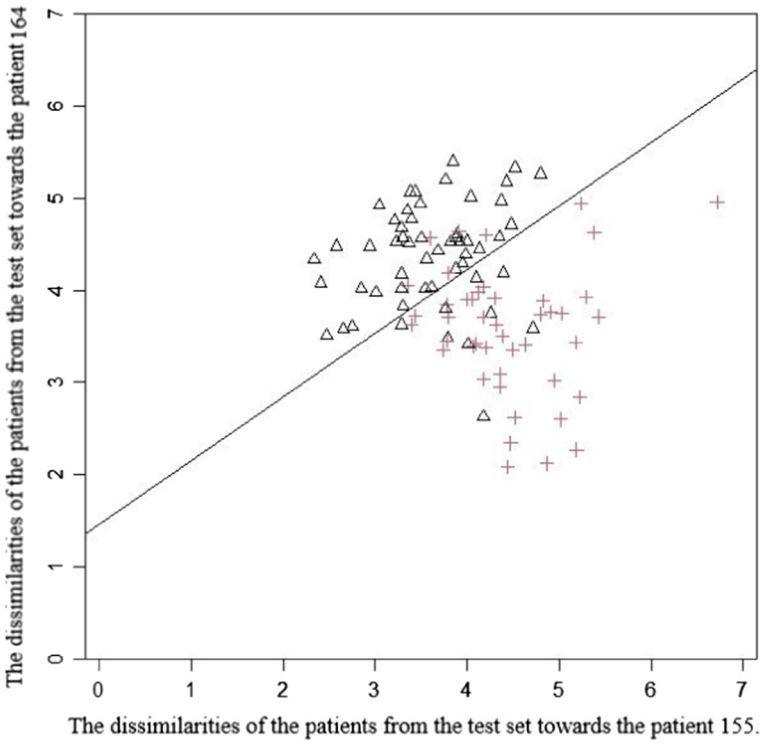


Fig. 6 The patients from the test set are represented in a 2D dissimilarity space, where the coordinates are the pairwise dissimilarities of the patients from the test set towards the two active exemplars $\mathbf{x}_{155}$ and $\mathbf{x}_{164}$. The triangle and the cross denote the patients labeled as "absence" and "presence" of the disease respectively. The straight line is the decision line of making predictions

5 Conclusions and discussions

The δ-machine is a classification technique based on dissimilarities towards prototypes. There are essentially three steps of the δ-machine: (1) clustering is performed to each outcome class in order to select highly informative exemplars or prototypes; (2) a dissimilarity matrix is calculated between profiles of the objects and the profiles of the selected exemplars or prototypes; (3) penalized logistic regression is fitted on the dissimilarity space. Step (1) is not compulsory in terms of the purpose of analysis; clustering results in a smaller representation set, which could result in a lower number of active exemplars in the final step and therefore may lead to a sparser model.

In this paper, we extended the δ-machine to handle mixed-type predictor variables. We defined two dissimilarity functions, the adjusted Gower and the adjusted Euclidean dissimilarity functions (the AGDF and the AEDF). Five simulation studies were conducted to compare the performance of the δ machine with the two adjusted dissimilarity functions. Furthermore, we compared the performance of the δ-machine to logistic regression with and without interactions. We also studied how the selection of the representation set influences the performance of the δ-machine. The general

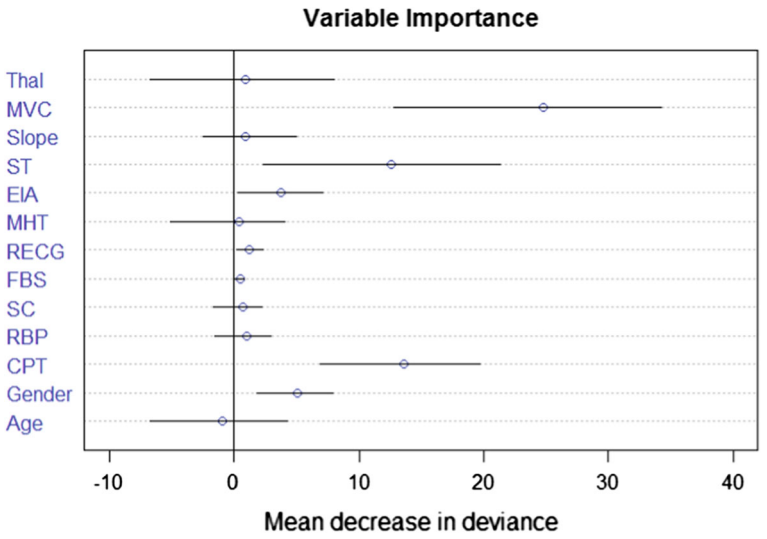


Fig. 7 Variable importance plot for the δ-machine using the adjusted Euclidean dissimilarity. The representation set was selected by PAM. The horizontal lines on the variable importance plot are the 95% confidence intervals. The variable names are written on the y-axis

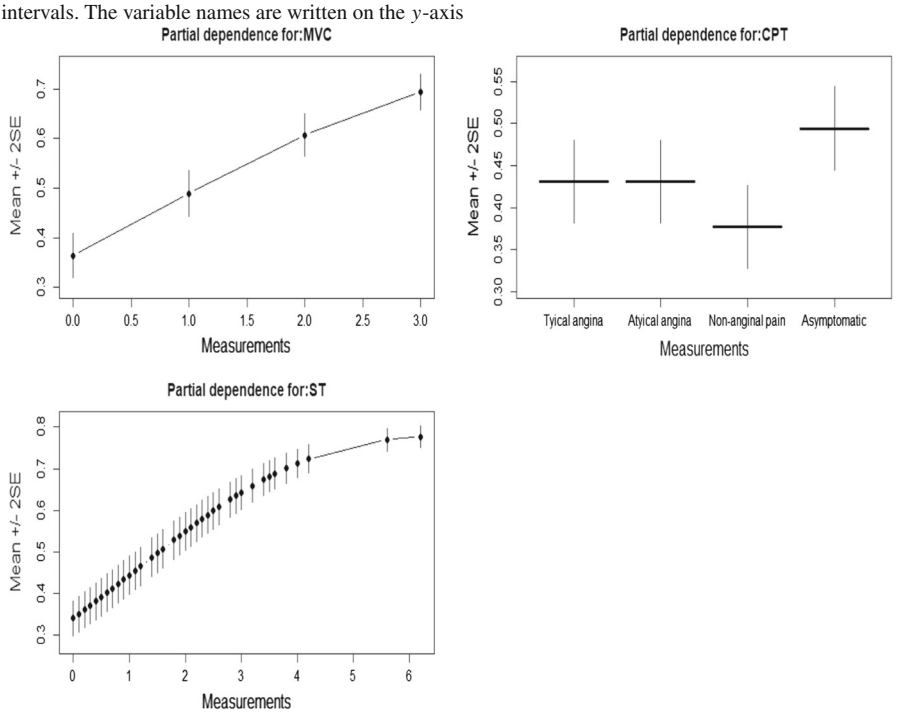


Fig. 8 Partial dependence plots of probabilities on the variables MVC, CPT, and ST for the δ-machine. The plots correspond to the solution of the δ-machine using the adjusted Euclidean dissimilarity, where the representation set was selected by PAM. The vertical lines on the partial dependence plots present the 95% confidence intervals

conclusions drawn from these studies are: (1) the δ-machine using the AEDF had better performance than using the AGDF; (2) the δ-machine using PAM to construct a representation set results in sparser models than using K-means clustering or using the complete training set; (3) the predictive performance of the δ-machine in comparison to logistic regression models was superior for mixed type data but inferior for purely categorical predictors.

The δ-machine is a person-oriented approach and offers a different perspective of interpretation (i.e. via the relations between objects/persons) than variable-oriented approaches such as logistic regression (i.e. via the relations between original predictor variables). In the empirical example, we have compared the predictive performance of the Lasso on the original predictor space (variable-oriented) with the δ-machine which is the Lasso on the dissimilarity space (person-oriented). The Lasso built on the original predictor space had on average 12.3 regression coefficients, which is not easy to interpret. On the contrary, the δ-machine using a representation set selected by PAM had on average five exemplars showing sparser models. Moreover, the selected exemplars (patients) may be potentially useful in a further study.

The dissimilarity function is important for the predictive performance of the δ-machine. When a good dissimilarity function is determined, the discriminatory power of the dissimilarities might be large (Pekalska et al. 2001). The two adjusted dissimilarity measures are the modified functions of the Manhattan distance and the Euclidean distance for mixed-type predictor variables. The Manhattan distance and the Euclidean distance are two special cases of the Minkowski distance,

$$d(\mathbf{x}_i, \mathbf{x}_r)^{\text{Min}} = \left(\sum_{p=1}^{P} d_{irp}^{\text{Min}} \right)^{\frac{1}{\omega}} = \left(\sum_{p=1}^{P} |x_{ip} - x_{rp}|^{\omega} \right)^{\frac{1}{\omega}},$$

where $\omega (\geq 0)$ is called the Minkowski exponent, which equals 1 for the Manhattan distance and 2 for the Euclidean distance.

The Minkowski distance with other ω values could also be implemented in the δ-machine. However, some modifications of their equations are required to adopt for mixed-type predictor variables. More specifically, the definitions of the adjusted functions for different types of variables should be defined accordingly. Meanwhile, the way of standardizing continuous variables also needs to be decided such as to use z-scores (e.g., the AEDF) or to use each continuous variable divided by the range (e.g., the AGDF). For instance, suppose that $\omega = 3$, for different types of categorical variables, the definitions between objects i and r are given below
If the variable X_p is binary,

$$d_{irp}^{\text{Min}} = |b_{ip} - b_{rp}|^3 = \begin{cases} 0 & x_{ip} = x_{rp} \\ 1 & x_{ip} \neq x_{rp} \end{cases}. \tag{12a}$$

If the variable X_p is nominal with K_p levels,

$$d_{irp}^{\text{Min}} = \sum_{k=1}^{K_p} |g_{ipk} - g_{rpk}|^3 = \begin{cases} 0 & x_{ip} = x_{rp} \\ 2 & x_{ip} \neq x_{rp} \end{cases}. \tag{12b}$$

If the variable X_p is ordinal with K_p levels,

$$d_{irp}^{\text{Min}} = \sum_{k=1}^{K_p - 1} |f_{ip_k} - f_{rp_k}|^3. \tag{12c}$$

In multidimensional scaling, a common strategy is to try out many Minkowski exponent values and then to choose the one with the lowest stress (Okada and Shigemasu 2010). Okada and Shigemasu proposed a new Bayesian method for the estimation of the Minkowski exponent. We believe that the idea of considering the Minkowski exponent as a parameter can be applied in the δ-machine. The performance of the δ-machine may improve, but the computational cost will increase dramatically.

A particular data set may have its own data-specific (dis)similarity function, i.e., the optimal dissimilarity function is application dependent. We used two types of artificial problems to evaluate the performance of the δ-machine with the two dissimilarity functions. The four blocks problem creates data containing high-order interaction terms, and the Gaussian ordination problem creates data containing quadratic terms. For the four blocks problem, the AGDF in the δ-machine failed to make accurate predictions while the AEDF not. The AGDF treats symmetric binary and nominal in the same way, therefore the AGDF had the same performance in these two cases. By contrast, using the asymmetric binary measure had different performances. Because the asymmetric binary measure discards the absent-absent matches, it breaks the link with the nominal measure. Therefore, it may bring extra information to achieve lower MRs. This could be the reason that the AGDF with asymmetric binary had satisfactory results on the four blocks problem. For the Gaussian ordination problem, both dissimilarity functions worked well. The AEDF showed good predictive performances in both problems, therefore, we consider it as the default dissimilarity function.

We would like to point out two issues for these two tailored dissimilarity functions. First, we choose a straightforward way to code categorical variables into indicator matrices. The pairwise dissimilarities are computed from the (cumulative) indicator matrices. However, for ordinal variables, the underlying assumption is that the numerical distance between each level is equal. In other words, ordinal variables are treated as continuous by taking the orderings of levels as the integers. This assumption may not hold for some real world data. Some form of optimal scaling could be used to replace categorical variables by optimally quantified variables (Meulman et al. 2019). In optimal scaling, each categorical predictor variable is replaced by a set of quantifications. Instead of creating dummy variables, optimal quantifications are assigned directly to the categories of the predictor. Second, for binary variables, for the ease of computation of the AEDF, we did not include asymmetric binary condition as the AGDF.

The δ-machine takes the dissimilarities as the predictor variables in the Lasso rather than the original predictor variables. By doing so, the original predictor matrix is dis-

carded. Another option is to concatenate the dissimilarity matrix and the original predictor matrix and build a classifier on the resulting matrix. The resulting classifier might have better predictive performance in some applications. However, the resulting model cannot be interpreted from a person-oriented nor a variable-oriented perspective; but it becomes a hybrid perspective.

The approach of Huang (1997, 1998) is very similar to our way to extend K-means clustering, i.e., first construct the transformed matrix $\mathbf{X}^{\mathrm{M}}$ and later apply the clustering methods on it. Huang proposed a K-prototypes algorithm which is based on K-means clustering. The K-prototypes algorithm removes the limitation of K-means clustering of accepting only numerical variables. The essence of this algorithm lies in the way of treating a mix of continuous variables and categorical variables. Huang proposed a distance measure defined by a sum of squared Euclidean distances of continuous variables and sample mismatch measures of categorical variables, i.e.,

$$d_{ir}^{\mathrm{H}} = d(\mathbf{x}_i, \mathbf{x}_r)^{\mathrm{H}} = \sum_{m=1}^{Q} (x_{im} - x_{rm})^2 + \lambda \sum_{m=Q+1}^{P} dis(x_{im}, x_{rm}),$$

where the first Q variables are continuous and the other variables are categorical. The function $dis()$ is a simple mismatch measure where if the two objects $\mathbf{x}_i$ and $\mathbf{x}_r$ have the same value on the categorical variable X_m, $dis(x_{im}, x_{rm}) = 0$, otherwise $dis(x_{im}, x_{rm}) = 1$. The parameter λ controls the weights between the groups of continuous and categorical variables. For example, if $\lambda = 0$ the distance measure only takes the group of numerical variables into account. The differences between our extension and Huang's are

- the sum of the squared dissimilarities (Huang's) versus the square root of the sum of the squared dissimilarities (ours);
- treating all categorical variables as nominal (Huang's) versus distinguishing between ordinal, nominal, asymmetric and symmetric binary (ours);
- the weight parameter λ between the two groups of categorical and continuous variables (Huang's) versus the same weight on the two groups (ours).

Of course, it is possible to assign the relative weights to the predictor variables, but it will increase the computational cost of the method, and an exhaustive search need to be performed to find the optimal predictor weights. Huang (1997) suggested to use the average standard deviation of continuous predictors as a guidance in specifying the weight, but then Huang (1998) concluded it is too early to consider it as the general rule. The prior information is of importance to specify the weight, but in practical applications such information may be rarely available.

Appendix: Overview of abbreviations used in the paper

Abbreviations	
AGDF	Adjusted Gower dissimilarity function
AEDF	Adjusted Euclidean dissimilarity function
Lasso	Least absolute shrinkage and selection operator
PAM	Partitioning around medoids
MR	Misclassification rate
ANOVA	Analysis of variance
LR	Logistic regression
LR_+	Logistic regression with two-way interactions
Act.Exp	the number of active exemplars or prototypes
SVM(RBF)	Support Vector Machines with Radial Basis Kernel
δ_Gower	The δ-machine using the Adjust Gower Dissimilarity Function
δ_Gower_asy	The δ-machine using the Adjust Gower Dissimilarity Function with asymmetric measure
δ_Gower_sy	The δ-machine using the Adjust Gower Dissimilarity Function with symmetric measure
δ_Gower_asy_approach (a)	The δ-machine using the Adjust Gower Dissimilarity Function with asymmetric measure with the proposed approach (a) for K-means clustering for mixed type of predictor variables
δ_Gower_sy_approach (a)	The δ-machine using the Adjust Gower Dissimilarity Function with symmetric measure with the proposed approach (a) for K-means clustering for mixed type of predictor variables
δ_Gower_approach (b)	The δ-machine using the Adjust Gower Dissimilarity Function with the proposed approach (b) for K-means clustering for mixed type of predictor variables
δ_Euc	The δ-machine using the Adjust Euclidean Dissimilarity Function
δ_Euc_approach (a)	The δ-machine using the Adjust Euclidean Dissimilarity Function with the proposed approach (a) for K-means clustering for mixed type of predictor variables
δ_Euc_approach (b)	The δ-machine using the Adjust Euclidean Dissimilarity Function with the proposed approach (b) for K-means clustering for mixed type of predictor variables

References

Bergman LR, Magnusson D (1997) A person-oriented approach in research on developmental psychopathology. Dev Psychopathol 9(02):291–319

Berk RA (2008) Statistical learning from a regression perspective, 1st edn. Springer, New York

Borg I, Groenen PJ (2005) Modern multidimensional scaling: theory and applications. Springer Science & Business Media, Berlin

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Chapman and Hall/CRC

Brown G (2004) Diversity in neural network ensembles. PhD thesis, University of Birmingham

Cohen J (1973) Eta-squared and partial eta-squared in fixed factor anova designs. Educ Psychol Measure 33(1):107–112

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Academic press, New York

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
Cox TF, Cox M (2000) Multidimensional scaling. CRC Press
Detrano R, Yiannikas J, Salcedo EE, Rincon G, Go RT, Williams G, Leatherman J (1984) Bayesian probability analysis: a prospective demonstration of its clinical utility in diagnosing coronary disease. Circulation 69(3):541–547
Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid JJ, Sandhu S, Guppy KH, Lee S, Froelicher V (1989) International application of a new probability algorithm for the diagnosis of coronary artery disease. Am J Cardiol 64(5):304–310
Dheeru D, Karra Taniskidou E (2017) UCI machine learning repository. http://archive.ics.uci.edu/ml. University of California, Irvine, School of Information and Computer Sciences
Fawcett T (2006) An introduction to roc analysis. Patt Recogn Lett 27(8):861–874
Freund Y, Schapire RE et al (1996) Experiments with a new boosting algorithm. In: Saitta L (ed) Machine learning: proceedings of the thirteenth international conference, vol 96. Morgan Kauffman Inc., San Francisco, pp 148–156
Friedman J (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232
Friedman J, Hastie T, Tibshirani R (2009) The elements of statistical learning, 2nd edn. Springer series, New York (**in Statistics**)
Friedman J, Hastie T, Tibshirani R (2010a) glmnet: regularization paths for generalized linear models via coordinate descent. R package version 1.6-4
Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22
Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27(4):857–871
Heiser WJ (1981) Unfolding analysis of proximity data. Ph.D. dissertation, Department of Data Theory, Leiden University
Hickendorff M, Edelsbrunner PA, McMullen J, Schneider M, Trezise K (2018) Informative tools for characterizing individual differences in learning: latent class, latent profile, and latent transition analysis. Learn Indiv Diff 66:4–15
Hsu CW, Chang CC, Lin CJ et al (2003) A practical guide to support vector classification. Technical Report, Department of Computer Science, National Taiwan University
Huang Z (1997) Clustering large data sets with mixed numeric and categorical values. In: Motoda H (ed) Proceedings of the 1st Pacific-Asia conference on knowledge discovery and data mining,(PAKDD), Singapore. World Scientific Publishing Co., Inc., pp 21–34
Huang Z (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining Knowl Discov 2(3):283–304
Jaccard P (1912) The distribution of the flora in the alpine zone. 1. New Phytol 11(2):37–50
James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer, New York
Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, New York
Kotsiantis S, Pintelas P (2004) Combining bagging and boosting. Int J Comput Intell 1(4):324–333
MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: LeCam LM, Neyman J (eds) Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol 1. Oakland, CA, USA, pp 281–297
Medin DL, Schaffer MM (1978) Context theory of classification learning. Psychol Rev 85(3):207–238
Melville P, Mooney RJ (2003) Constructing diverse classifier ensembles using artificial training examples. In: Gottlob G, Walsh T (eds) Proceedings of the eighteenth international joint conference on artificial intelligence, vol 3, pp 505–510
Meulman JJ, van der Kooij AJ, Duisters KL et al (2019) Ros regression: integrating regularization with optimal scaling regression. Stat Sci 34(3):361–390
Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2014) e1071: misc functions of the department of statistics (e1071), TU Wien. R package version 1.6-4. http://CRAN.R-project.org/package=e1071
Mirkin B (1999) Concept learning and feature selection based on square-error clustering. Mach Learn 35(1):25–39
Nosofsky R (1992) Exemplars, prototypes, and similarity rules. In: Healy AF, Estes WK, Kosslyn SM, Shiffrin RM (eds) From learning theory to connectionist theory, vol 1. Lawrence Erlbaum Associates Inc, pp 49–167

Okada K, Shigemasu K (2010) Bayesian multidimensional scaling for the estimation of a minkowski exponent. Behav Res Methods 42(4):899–905
Opitz D, Maclin R (1999) Popular ensemble methods: an empirical study. J Artif Intell Res 11:169–198
Pekalska E, Duin RP (2005) The dissimilarity representation for pattern recognition: foundations and applications. World Scientific, Singapore
Pekalska E, Paclik P, Duin RP (2001) A generalized kernel approach to dissimilarity-based classification. J Mach Learn Res 2(Dec):175–211
R Core Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/
Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, New York
Ross BH, Makin VS (1999) Prototype versus exemplar models in cognition. In: Sternberg RJ (ed) The nature of cognition. MIT Press Cambridge, MA, pp 205–241
Şahan S, Polat K, Kodaz H, Güneş S (2005) The medical applications of attribute weighted artificial immune system (awais): diagnosis of heart and diabetes diseases. In: International conference on artificial immune systems, Springer, pp 456–468
Steinley D, Brusco MJ (2011) Choosing the number of clusters in k-means clustering. Psychol Methods 16(3):285
Therneau T, Atkinson B, Ripley B (2015) rpart: recursive partitioning and regression trees. https://CRAN.R-project.org/package=rpart, r package version 4.1-10
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc Ser B (Methodological) 58(1):267–288
Webb GI (2000) Multiboosting: A technique for combining boosting and wagging. Mach Learn 40(2):159–196
Yuan B, Heiser W, de Rooij M (2019) The δ-machine: classification based on distances towards prototypes. J Classif 36(3):442–470