Pattern Recognition 153 (2024) 110547

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A general framework for implementing distances for categorical variables

Michel van de Velden [a], Alfonso Iodice D'Enza [b], Angelos Markos [c], Carlo Cavicchia [a,*]

[a] Econometric Institute, Erasmus University Rotterdam, Netherlands
[b] Department of Political Sciences, University of Naples Federico II, Italy
[c] Department of Primary Education, Democritus University of Thrace, Greece

ARTICLE INFO

Dataset link: https://alfonsoiodicede.github.io/blogposts_archive/distances_experiment_superv_unsuperv.html, https://github.com/alfonsoIodiceDE/catdist_package

Keywords:
Categorical data
Distance
Cluster analysis
Classification
K-NN

ABSTRACT

The degree to which objects differ from each other with respect to observations on a set of variables, plays an important role in many statistical methods. Many data analysis methods require a quantification of differences in the observed values which we can call distances. An appropriate definition of a distance depends on the nature of the data and the problem at hand. For distances between numerical variables, there exist many definitions that depend on the size of the observed differences. For categorical data, the definition of a distance is more complex as there is no straightforward quantification of the size of the observed differences. In this paper, we introduce a flexible framework for efficiently computing distances between categorical variables, supporting existing and new formulations tailored to specific contexts. In supervised classification, it enhances performance by integrating relationships between response and predictor variables. This framework allows measuring differences among objects across diverse data types and domains.

1. Introduction

In many statistical methods, the quantification of dissimilarity, that is, the degree to which objects differ from each other, plays an important role [1]. We can refer to such dissimilarity quantification as a distance. Classification methods such as K-Nearest Neighbors (KNN, [2]), but also clustering methods such as K-means, Partitioning Around Medoids (PAM) and hierarchical linkage methods [3], and data visualization methods such as MultiDimensional Scaling (MDS, [4]) and biplots [5], require a definition of distance between objects. The way to select a definition of distance depends on the nature of the data and problem at hand.

The distance measures for numerical data are usually based on the magnitude of the observed differences in values. For categorical data, however, the situation is more complex as we do not directly observe, and hence cannot directly quantify, sizes of differences. We can only establish whether there is a difference or not. For distance calculations in multivariate contexts, two cases can be distinguished. First, the distances are calculated for each variable independently and then added. Second, the association between the variables is taken into account when calculating the distances. For numerical variables, several well-known distances, e.g., Euclidean or Manhattan distances, implicitly assume independence between the variables and require that the measurement scales must be commensurable. For categorical variables, there are also several measures that take the sum of distances per variable when considering a multivariate distance. For example, in simple matching, distance between two observations is defined as the number of times that the categories of corresponding variables do not match.

For numerical variables, the association between variables can be accounted for using the Mahalanobis distance, where (sample) covariances are used to weigh observed differences to account for correlation between the variables. For categorical data, so-called association-based distances exist. In such distances, the association between categorical variables is used to quantify the differences between observations. The question of how to account for associations in a categorical setting is not trivial. Several recent proposals for distances between categorical variables are indeed association-based distances (see, e.g., [6–11]).

In this paper, we propose a general framework for implementing categorical distances. By incorporating existing distances in our framework, it becomes possible to assess the differences and similarities between them. Currently, such comparison is not trivial due to the wide variety of notation and research fields (and hence objectives) in which methods have been proposed. In addition, our framework makes it possible to construct and define new and highly customizable distances. For example, in a supervised classification context, the framework can be used to define a distance that takes into account association with the classes of the response variable. As our framework is not method- or application-specific, it can be used to calculate distance matrices for any method or application that requires distance calculations. For example, MDS, cluster analysis, or, in a supervised context,

* Corresponding author.
E-mail address: cavicchia@ese.eur.nl (C. Cavicchia).

https://doi.org/10.1016/j.patcog.2024.110547
Received 11 January 2023; Received in revised form 17 April 2024; Accepted 28 April 2024
Available online 6 May 2024

KNN. The framework offers flexibility and transparency by defining the categorical variable distances as a function of category dissimilarities. The category dissimilarities can either be data-driven or theory-based. Furthermore, our framework allows for fast and efficient distance calculations, potentially representing a substantial enhancement compared to current implementations. In particular, we show that for the distance for categorical variables proposed in [7], our implementation is much faster than existing implementations.

An important issue with regard to the definition of distance measures is the validation of the measures. That is, how does one know that the chosen measure is appropriate? Although the new framework does not provide an answer to this question, having a general formulation simplifies both theoretical and empirical comparisons between different choices. We illustrate our method by applying distance-based data analysis methods to several well-known categorical data sets using a selection of known and new, association-based, categorical distances. We implemented functions to perform all categorical distance calculations using our general framework in the R package `catdist`, which is available on GitHub.[1]

This paper is organized as follows. After introducing some notation in Section 2, we describe our general framework in Section 3. In Section 4, we introduce several common distance measures for categorical variables and show how they can be incorporated. Categorical distances based on co-occurrences are introduced in Section 5, with particular attention to the distance measure proposed by [7]. In Section 6, we show how supervised distances can be constructed and implemented using our framework. Tuning of distance definitions is described in Section 7, after which we illustrate our methodology using several data sets in Section 8. Section 9 concludes the paper.

2. Notation

Suppose that we have n observations on Q categorical variables and let the number of categories for the $j \in 1 \dots Q$th variable be q_j. We can then code the categorical data by using indicator matrices. That is, for each categorical variable $j \in 1 \dots Q$, we create an $n \times q_j$ binary matrix $\mathbf{Z}_j$, where the n rows correspond to observations and the q_j columns to categories. The observed category is indicated by a one, and all other categories are assigned zeros. Furthermore, for each observation of a categorical variable, exactly one category is observed, and we only include categories that have been observed at least once in the data set. Therefore, each column of $\mathbf{Z}_j$ contains at least one element equal to one and $\mathbf{Z}_j \mathbf{1}_{q_j} = \mathbf{1}_n$, where, generically, $\mathbf{1}_i$ denotes an i-dimensional vector of ones. That is, the sum over the columns is 1. Using these indicator matrices, we can code data on Q categorical variables into a so-called super-indicator matrix by collecting all indicator matrices next to each other. That is, $\mathbf{Z} = (\mathbf{Z}_1 \dots \mathbf{Z}_Q)$. Furthermore, define

$$\mathbf{P} = \frac{1}{n}\mathbf{Z}'\mathbf{Z}, \tag{1}$$

and $\mathbf{P}_d = \frac{1}{n}(\mathbf{Z}'\mathbf{Z}) \odot \mathbf{I}_{Q^*}$, where $\odot$ indicates the Hadamard product, that is, element-wise multiplication, and $Q^* = \sum_{j=1}^{Q} q_j$. Note that $\mathbf{P}_d$ is a diagonal matrix with as its diagonal elements the observed relative frequencies (within each variable) for the categories. Moreover, let $\mathbf{p} = \mathbf{P}_d \mathbf{1}_{Q^*}$ denote the vector of observed relative frequencies, and $\mathbf{p}^- = \mathbf{P}_d^{-1}\mathbf{1}_{Q^*}$ is the vector of inverse observed relative frequencies.

Note that the ijth off-diagonal block of $\mathbf{P}$ gives the relative frequencies of co-occurrences for the categories of variables i and j. They can be seen as (empirical) joint probability distributions for variables i and j. For the calculation of association-based distances in Section 5, we also define

$$\mathbf{R} = \mathbf{P}_d^{-1}(\mathbf{P} - \mathbf{P}_d). \tag{2}$$

The rows of the ijth off-diagonal block of $\mathbf{R}$ give, for the categories of the ith variable, the distributions over the categories of the jth variable. These can be interpreted as (empirical) conditional distributions.

[1] https://github.com/alfonsoIodiceDE/catdist_package.

3. Categorical distance calculations based on category dissimilarities

For a categorical variable, it is not obvious how to quantify differences between different categories. For example, suppose that we observe three individuals, one from the Netherlands, one from Italy, and one from Greece. Geographically, and perhaps also culturally, Italy and Greece are more similar than the Netherlands. How to take such differences into account is, however, not trivial. In our framework, we do so by defining *category dissimilarities*.

Definition 1. A matrix $\boldsymbol{\Delta}_j$ is the category dissimilarity matrix for variable j. The elements of this matrix, δ_{ab}, where a and b indicate two categories of variable j, quantify the dissimilarities between the categories a and b of the jth variable.

We can impose conditions on the dissimilarity matrix that are consistent with typical distance definitions. That is, (1) the dissimilarity of a category from itself is zero ($\delta_{aa} = 0$, for all categories). (2) Dissimilarities are symmetric ($\delta_{ab} = \delta_{ba}$, for all pairs of categories). (3) Dissimilarities satisfy the triangle inequality. That is, if a, b and c denote different categories for a variable j, then, for all categories a, b and c, $\delta_{ac} \leq \delta_{ab} + \delta_{bc}$. If all three of these conditions are satisfied, the dissimilarities can be considered as metric distances between categories. If they are non-negative and satisfy only the first two conditions, they can be interpreted as non-metric distances between categories. However, we refer to them as category dissimilarities and reserve the term "distance" for the distances between observations.

If we have Q categorical variables, each with a category dissimilarity matrix $\boldsymbol{\Delta}_j$, we can construct a $Q^* \times Q^*$, block diagonal matrix $\boldsymbol{\Delta}$, with separate category dissimilarity matrices as diagonal blocks.

The category dissimilarity matrices can be used to calculate a between observations distance matrix as follows. First, consider the n by q_j indicator matrix $\mathbf{Z}_j$ corresponding to the jth categorical variable. Furthermore, we have the corresponding category dissimilarity matrix $\boldsymbol{\Delta}_j$. We can formulate the following theorems:

Theorem 1. *The distances between the observations for the categorical variable j are $\mathbf{D}_j = \mathbf{Z}_j \boldsymbol{\Delta}_j \mathbf{Z}_j'$.*

Proof. The matrix multiplication of the row i of $\mathbf{Z}_j$ with $\boldsymbol{\Delta}_j$ selects the row of $\boldsymbol{\Delta}_j$ corresponding to the category chosen by the individual i. Similarly, matrix multiplication of this row by the i'th column of $\mathbf{Z}_j'$ (i.e., the i'th observation) selects the element corresponding to the category chosen by the individual i'. Hence, the (i, i')th element of $\mathbf{D}_j$ is the dissimilarity between the categories chosen by individuals i and i'. ■

Theorem 2. *If we define the distance between observations on Q categorical variables as the sum of Q distances for each categorical variable, the $n \times n$ distance matrix can be calculated as*

$$\mathbf{D} = \mathbf{Z}\boldsymbol{\Delta}\mathbf{Z}'. \tag{3}$$

Proof.

$$\mathbf{D} = (\mathbf{Z}_1 \dots \mathbf{Z}_Q) \begin{pmatrix} \boldsymbol{\Delta}_1 & & \\ & \ddots & \\ & & \boldsymbol{\Delta}_Q \end{pmatrix} \begin{pmatrix} \mathbf{Z}_1' \\ \vdots \\ \mathbf{Z}_Q' \end{pmatrix} = \sum_{j=1}^{Q} \mathbf{Z}_j \boldsymbol{\Delta}_j \mathbf{Z}_j' = \sum_{j=1}^{Q} \mathbf{D}_j \quad ■$$

From (3), it follows that distances between observations of categorical variables depend on the choices of the category dissimilarity matrices $\boldsymbol{\Delta}_j$. This allows for great flexibility in defining a suitable distance measure for a set of categorical variables. In the next section, we briefly review some choices for $\boldsymbol{\Delta}$, and we show how they relate to existing distances.

Note that associations between categorical variables are not explicitly incorporated in this formulation. That is, the differences between

the categories observed for one variable are not related to the differences in the categories observed for other variables. There are, however, ways to account for such observations. For example, rather than creating an indicator matrix for each categorical variable, one could construct an indicator matrix for all possible combinations (or subsets thereof) of observations. That is, one can create, for each (or a subset of) combination of categories one indicator matrix. The number of columns of such a matrix is therefore $\prod_{j=1}^{Q} q_j$ and only one category dissimilarity matrix is needed where each category is a combination of the categories for all Q variables. However, with several categorical variables, the total number of combinations and hence the number of categories of the final indicator matrix quickly becomes large. Furthermore, finding an appropriate category dissimilarity matrix for the combinations is not a trivial task. An alternative way to account for associations between the categorical variables is to use them in the construction of the category dissimilarity matrices. That is, by defining the dissimilarities between the categories of a variable in $\boldsymbol{\Delta}_j$, based on the associations with other variables. In Section 5, we give examples of such category dissimilarity measures.

3.1. Distances between sets

Suppose that we have two separate sets of observations on the same Q categorical variables. Data for these two sets can be collected in the $n_1 \times Q^*$ and $n_2 \times Q^*$ super indicator matrices $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$. Then, for a known category dissimilarity matrix $\boldsymbol{\Delta}$, it is easily verified that the distances between the observations for the two sets can be calculated as $\mathbf{D}^{(12)} = \mathbf{Z}^{(1)}\boldsymbol{\Delta}\mathbf{Z}^{(2)\prime}$.

Note that the matrix $\mathbf{D}^{(12)}$ is of order $n_1 \times n_2$. Calculating distances between sets can be useful when considering distance-based classification problems. In KNN, for example, distances between "new" (unlabeled) observations and observations in a labeled data set are required. The KNN predictions are based on (usually by majority vote) the labels of the K nearest neighbors in the training set. Similarly, in PAM – a popular clustering algorithm similar to K-means, where instead of considering within-cluster variation around the mean, variation around an actual observation, the medoid, is considered – computing the distance between sets can be of interest. If the medoids are collected in $\mathbf{Z}^{(1)}$ and "new" data points in $\mathbf{Z}^{(2)}$, we can assign the new points to existing clusters considering the distances $\mathbf{D}^{(12)}$ and selecting the smallest distances.

4. Independent category dissimilarity matrices

We first consider several definitions of dissimilarity for categorical data that do not take into account the association between variables. Hence, in a multivariate context, distances are calculated as the sum of distances per variable, and for each variable, the category dissimilarities are independent of the observations on other variables. However, category dissimilarities may depend on the observed frequencies for a variable. We do not aim to be complete with respect to the different definitions of dissimilarity. Instead, we select definitions from [12] (which contain several definitions also reviewed in [13]), and transform them into dissimilarities. We show how these dissimilarities can be incorporated into our framework by defining the appropriate category dissimilarity matrices $\boldsymbol{\Delta}$. For a more detailed description, as well as study on the relative performances of these dissimilarity definitions in a cluster analysis setting, see [12].

Simple Matching (SM). The idea of this measure is that the distance between observations is 1 if the categories do not match and 0 if they do. Consequently, all different between category dissimilarities are 1. For the jth categorical variable with q_j categories, we define $\boldsymbol{\Delta}_{M_j} = \mathbf{1}_{q_j}\mathbf{1}'_{q_j} - \mathbf{I}_{q_j}$.

That is, the distance between each category is exactly 1. If we have Q categorical variables and want to calculate a simple matching for all variables, we simply collect all $\boldsymbol{\Delta}_{M_j}$ in a block diagonal matrix $\boldsymbol{\Delta}_M = \mathbf{K}_b - \mathbf{I}_{Q^*}$, where $\mathbf{K}_b$ is a $Q^* \times Q^*$ block diagonal matrix with, for $j = 1 \dots Q$, $q_j \times q_j$ matrices $(\mathbf{K}_{b_j})$ of ones as its diagonal blocks.

Eskin. For Eskin distance [14], category dissimilarities depend on the number of categories. Dissimilarities for variables with more categories are smaller than dissimilarities for variables with fewer categories. In particular, the dissimilarity between different categories for a variable with q_j categories is $\frac{2}{q_j^2}$. Therefore, for the jth categorical variable with q_j categories, the category dissimilarity matrix is defined as $\boldsymbol{\Delta}_{E_j} = \frac{2}{q_j^2}\left(\mathbf{1}_{q_j}\mathbf{1}'_{q_j} - \mathbf{I}_{q_j}\right)$. Collecting all Q dissimilarity matrices in a block diagonal matrix produces the Eskin category dissimilarity matrix $\boldsymbol{\Delta}_E$. If all variables have the same number of categories, Eskin merely re-scales the simple matching dissimilarity.

Lin. [15] proposed an information-theoretic measure that gives more weight to matches on frequent values and lower weight to mismatches on infrequent values. We implement Lin's proposal as follows: define $\mathbf{P}_r = \mathbf{p}\mathbf{1}'_{Q^*}$ and $\mathbf{P}_c = \mathbf{1}_{Q^*}\mathbf{p}'$. Furthermore, let $\widetilde{\mathbf{P}} = \mathbf{P}_r + \mathbf{P}_c - \mathbf{P}_d$. Then, the category dissimilarity matrix can be defined as $\boldsymbol{\Delta}_{Lin} = \left[\log(\mathbf{P}_r) + \log(\mathbf{P}_c) - 2\log(\widetilde{\mathbf{P}})\right] \oslash 2\log(\mathbf{P}_r + \mathbf{P}_c)$, where $\oslash$ indicates the Hadamard division (i.e., element-wise) and $\log(\cdot)$ takes the logarithms of the elements of the parenthesized object and collects them in an object of the same size. Note that Lin's dissimilarity for a category with itself is zero. Furthermore, in our implementation, for each variable, the dissimilarity between different categories, say categories a and b, is $\left[\log(p_a) + \log(p_b) - 2\log(p_a + p_b)\right]/2\log(p_a + p_b)$, where p_a and p_b are, respectively, the relative frequencies of categories a and b.

Inverse Occurrence Frequency (IOF) and *Occurrence Frequency* (OF). For inverse occurrence frequency (IOF, [13]), a higher dissimilarity is assigned when categories are more frequently observed. In particular, the category dissimilarity matrix is defined as $\boldsymbol{\Delta}_{IOF} = \left[\log(n\mathbf{p})\right]\left[\log(n\mathbf{p})\right]' - \left[\log(n\mathbf{p})\right]\left[\log(n\mathbf{p})\right]' \odot \mathbf{I}_{Q^*}$. It is worth observing that, for each variable, IOF dissimilarity for a category with itself is zero, and the dissimilarity between two different categories, say a and b, corresponds to $\log(np_a)\log(np_b)$. The IOF measure is related to the concept of inverse document frequency (TF-IDF) from information retrieval, where it is used to account for document relevance for a given term [16]. In other words, since a rare term contributes more information than a more frequent term, the IOF measure accounts for how rare the term is, and a lower IOF dissimilarity corresponds to a rarer term. Log frequency is used to reduce the impact of terms of very high frequencies. For occurrence frequency (OF) dissimilarity, dissimilarities are higher if the categories are observed less frequently. The category dissimilarity matrix is defined as $\boldsymbol{\Delta}_{OF} = \left[\log(\mathbf{p}^-)\right]\left[\log(\mathbf{p}^-)\right]' - \left[\log(\mathbf{p}^-)\right]\left[\log(\mathbf{p}^-)\right]' \odot \mathbf{I}_{Q^*}$. Therefore, OF dissimilarity for a category with itself is zero, and the dissimilarity between two different categories a and b is $\log(p_a)\log(p_b)$.

Goodall dissimilarities. In [13], four variations of Goodall's similarity are considered. These are based on Goodall's original proposal [17]. After transforming similarities into dissimilarities, where dissimilarity is 1− similarity, the four measures have in common that dissimilarities between different categories are, as is the case with simple matching, always equal to one. However, the dissimilarity of category with respect to the same category depends on the observed proportions of the categories. Below we provide the category dissimilarity matrices for Goodall 3 and Goodall 4. For Goodall 1 and 2, we can also construct such matrices. However, these definitions require conditional sums of proportions. In particular, for Goodall 1, the dissimilarity for category a with itself, is defined as the sum of squared observed proportions that are smaller or equal to the observed proportion of category a. For Goodall 2, it is the sum of squared observed proportions that are larger or equal to the observed proportion of category a. The Goodall 3 and 4 measures do not require the calculation of a (conditional) sum and have the squared proportion and one minus the squared proportion of a category, respectively, on the diagonal blocks of $\boldsymbol{\Delta}$. That is, $\boldsymbol{\Delta}_{G_3} = \mathbf{K}_b - \mathbf{I}_{Q^*} + \mathbf{P}_d^2$, and $\boldsymbol{\Delta}_{G_4} = \mathbf{K}_b - \mathbf{P}_d^2$. In these definitions, dissimilarity of a category with the same category is not zero. Consequently, the resulting "distances" do not satisfy the typical requirements of a distance. Note

that for the Goodall 1 and 3 measures, a higher dissimilarity is assigned when the matching categories are frequent, whereas for the Goodall 2 and 4 measures a higher dissimilarity is assigned when the matching categories are infrequent.

Variable Entropy (VE) and *Variable Mutability dissimilarities* (VM). [12] proposed two variability-based dissimilarity measures that are related to Goodall 1 and 2, respectively. These dissimilarities are equal to one if the categories do not match, while, if they do match, VE uses the entropy and VM uses the Gini coefficient to quantify "dissimilarity". In particular, for the jth categorical variable with q_j categories, the category dissimilarity matrices are defined as $\boldsymbol{\Delta}_{VE_j} = \mathbf{K}_{b_j} + \left(\frac{1}{\log q_j}\sum_{l=1}^{q_j} p_l \log p_l\right)\mathbf{I}_{q_j}$ and $\boldsymbol{\Delta}_{VM_j} = \mathbf{K}_{b_j} - \left[\frac{q_j}{q_j-1}\left(1-\sum_{l=1}^{q_j} p_l^2\right)\right]\mathbf{I}_{q_j}$, respectively. Collecting all Q dissimilarity matrices in a block diagonal matrix returns the VE and VM category dissimilarity matrices $\boldsymbol{\Delta}_{VE}$ and $\boldsymbol{\Delta}_{VM}$.

4.1. Ordered categories

If categories are ordered, the order can be reflected in the dissimilarities. A simple choice would be to define the dissimilarities as the difference in category numbers. That is, the dissimilarity between categories a and b is simply $b - a$. If the data are rank order data or rating (e.g., Likert) scale data, this definition would imply treating the data as interval data. However, implementation of alternative, custom, definitions of ordered between-category distances is also straightforward. For example, if the categories correspond to bins on a numerical scale (e.g., age or income groups), differences between the midpoints of the bins can be used to define dissimilarities that better reflect the underlying values. More generally, let $\boldsymbol{\Delta}_o^i$ denote the ith diagonal block of $\boldsymbol{\Delta}_o^*$, and δ_{ab}^i its abth element. Then, dissimilarities between ordered categories can be imposed by letting $\delta_{ab}^i \leq \delta_{ab^*}^i$, for $a, b \in 1 \dots i_{q_i}$, $b^* \neq b$ and $b^* > a$. Note that this definition does not guarantee that the triangle inequality holds. That is, it may be the case that the direct distances between two categories are larger than the indirect distances between those categories. To enforce the triangle inequality to hold, one could impose appropriate additional constraints on the elements of the category dissimilarity matrix.

5. Association-based category dissimilarity matrices

The general idea of association-based measures, similar to the case of the Mahalanobis distance for numerical variables, differences that are in line with the association between variables are less informative (i.e., should correspond to smaller dissimilarity values) than differences that are not in line with the general association. How to exactly implement this idea depends on the calculation of the association between categorical variables, and how to incorporate this association in the category dissimilarities. Here, we present a general form to calculate and collect association-based dissimilarities that can be directly implemented in our general framework in Section 3. We then present some specific variants and link them to recent proposals.

Fundamental in the calculation of association-based distances is the matrix of proportions of co-occurrences $\mathbf{P}$ and the corresponding profile matrix $\mathbf{R}$ as defined in Section 2. In particular, recall that the off-diagonal blocks of $\mathbf{P}$ and $\mathbf{R}$ can be interpreted as (empirical) joint and conditional probability distributions, respectively. By considering different ways to quantify the dissimilarities between the conditional distributions (i.e., the rows of the off-diagonal blocks of $\mathbf{R}$) we can construct different category dissimilarity matrices $\boldsymbol{\Delta}$ that, by applying (3) can be used to obtain the between-observation distances.

Let $\mathbf{R}^{ij}$ denote the ijth off-diagonal block of $\mathbf{R}$, and let $\mathbf{r}_a^{ij}$ denote its ath row. Note that the elements of each row of $\mathbf{R}^{ij}$ add up to 1. Hence, these elements can be seen as (empirical) conditional probabilities. We define the dissimilarities between categories for all pairs of categories of variable i (for $i = 1 \dots Q$) based on the association with variable j (with $j \neq i$), as

$$\delta^{ij}(a,b) = \boldsymbol{\Phi}^{ij}\left(\mathbf{r}_a^{ij}, \mathbf{r}_b^{ij}\right), \tag{4}$$

where, generically, a and b indicate categories of variable i and $\boldsymbol{\Phi}^{ij}$ is the dissimilarity function that quantifies the differences between profiles based on the association between variables i and j. The overall between-category dissimilarities for all pairs of categories of variable i can be defined as

$$\delta^{i}(a,b) = \sum_{j\neq i}^{Q} w_{ij}\delta^{ij}(a,b) = \sum_{j\neq i}^{Q} w_{ij}\boldsymbol{\Phi}^{ij}\left(\mathbf{r}_a^{ij}, \mathbf{r}_b^{ij}\right). \tag{5}$$

The weights w_{ij} in (5) allow flexibility with respect to the importance of different variables in the calculation of association-based category dissimilarities, as defined by $\boldsymbol{\Phi}^{ij}$. By collecting, for each variable i, the elements $\delta^i(a,b)$ in a category dissimilarity matrix $\boldsymbol{\Delta}_{\boldsymbol{\Phi}}^i$, and organizing them on the diagonal of a block diagonal matrix, we obtain a dissimilarity matrix $\boldsymbol{\Delta}_{\boldsymbol{\Phi}}$, which can be used to calculate the distances between the observations using (3). Thus, (5) provides a very general way to define category dissimilarities using pair-specific weights and dissimilarity functions.

For the association-based dissimilarity functions $\boldsymbol{\Phi}^{ij}$, any function that quantifies the difference between two distributions can be used. A brief overview of 46 different functions and their implementation in the R package `philentropy` is described in [18]. Concerning the choice of weights w_{ij}, we distinguish two options. In the first, all weights are equal. Usually, 1 or $1/(Q-1)$, so that sums or averages are obtained. Alternatively, different weights can be used for different pairs. These weights can either be selected using expert knowledge (e.g., based on the experience and preferences of the researcher) or by using a data-driven approach. For example, one could set certain weights to zero and others to some constant based on some predetermined data dependent criterion (e.g., a measure of association like Cramér's V). Pairs with non-zero weights can then be referred to as "context" variables. Approaches using such context-based dissimilarities are described in [8,9]. If an objective measure of overall fit of a solution is available, one could consider w_{ij} and $\boldsymbol{\Phi}^{ij}$ as tuning parameters, and search combinations of these parameters to make a choice. In Section 7 we shall further explore the tuning of w_{ij} and $\boldsymbol{\Phi}^{ij}$. In the next subsections, we describe some specific choices of dissimilarity functions. In particular, we provide definitions for category dissimilarities between categories a and b of variable i, based on the association between variables i and j. That is, we present specific choices of $\boldsymbol{\Phi}^{ij}$ in (4). Inserting these definitions into (5) results in a dissimilarity matrix $\boldsymbol{\Delta}_{\boldsymbol{\Phi}}$ that can be used to calculate the between-observation distances. For ease of notation, we now drop the superscripts ij.

Total variation distance between profiles (TVD). The total variation distance between two discrete probability distributions can be defined as $1/2$ times the L_1 norm between the distributions. We can implement this in our framework by defining the category dissimilarity function $\boldsymbol{\Phi}$ as

$$\boldsymbol{\Phi}\left(\mathbf{r}_a, \mathbf{r}_b\right) = \frac{1}{2}\sum_{l=1}^{q_j}|r_{al} - r_{bl}|, \tag{6}$$

where r_{al} and r_{bl} denote the lth element of $\mathbf{r}_a$ and $\mathbf{r}_b$, respectively.

Ahmad and Dey's categorical variable distance. [7] argue that the dissimilarity between categories should be computed as a function of their distribution in the overall data set and in co-occurrence with other categories, rather than in isolation. The idea is to take into account co-occurrences of categories when constructing distances. The way they do this, is by considering all combinations of categories of one variable, and selecting the partitioning (that is, a combination of categories) for which the sum of proportions in the two complementary partitions for the two categories is maximal. Following [7], we can define the dissimilarity between categories a and b of variable i, with

respect to the distribution over the categories of variable j, as $\delta(a,b) = \max_{\omega_j}\left(P(\omega_j|a) + P(\bar{\omega}_j|b) - 1\right)$, where ω_j and its complement $\bar{\omega}_j$ define a binary partition with respect to the categories of variable j, and $P(\omega_j|a)$ denotes the proportion of observations with the category a of variable i, corresponding to the set of categories of variable j as defined by ω_j. Note that the term -1 is only introduced to fix the upper limit of the dissimilarities at 1. The number of binary partitions for variable j, excluding the partitions containing all or no categories, equals $2^{q_j} - 2$, where q_j gives number of categories of variable j, and hence this number grows exponentially when the number of categories for a variable increases. [7] propose an algorithm to calculate their distances. The order of their algorithm is $O(Q^{*2}n + Q^{*2}\bar{q}^3)$, where Q^* gives the total number of categories, n is the number of observations and $\bar{q}$ denotes the average number of categories per variable.

The Ahmad and Dey's categorical variable distance [7] is equivalent to TVD, and computing the latter is much more efficient. However, as this relationship is not trivial and appears to be unknown, we present this below by formulating the following theorem.

Theorem 3. *Ahmad and Dey's distance for categorical variables [7] is equivalent to the total variation distance between profiles.*

Proof. When going from $\delta(a,b)$ to $\delta(b,a)$, the optimal partition ω_j, that is, the combination of categories that maximizes the sum, is simply flipped (i.e., the complement is taken), hence Ahmad and Dey's distance is symmetric: $\delta(a,b) = \max_{\omega}(P(\omega|a) + P(\bar{\omega}|b) - 1) = \max_{\omega}(P(\omega|b) + P(\bar{\omega}|a) - 1) = \delta(b,a)$, where, for convenience, we dropped the subscripts j for the ω's. As $P(\bar{\omega}|\cdot) = 1 - P(\omega|\cdot)$, we have $\delta(a,b) = \max_{\omega}(P(\omega|a) - P(\omega|b)) = \max_{\omega}(P(\omega|b) - P(\omega|a))$. This shows that Ahmad and Dey's distance is equal to finding the maximum difference between all combinations of observed proportions. This implies that we can express this distance as the supremum norm of a vector of differences between probabilities. The total variation distance can also be defined as the largest difference between probabilities from two probability distributions that can be assigned to the same event. Let $\mathbf{K}_{q_j}$ be a design matrix that defines all, except the empty and complete, binary partitions for the q_j categories of variable j. Hence, $\mathbf{K}_{q_j}$ is a $q_j \times q_j^\star$ matrix of zeros and ones, where $q_j^\star = \sum_{l=1}^{q_j-1}\binom{q_j}{l} = 2^{q_j} - 2$. Then $\delta(a,b) = \|\mathbf{K}'_{q_j}\left(\mathbf{r}_a - \mathbf{r}_b\right)\|_\infty = \|\mathbf{K}'_{q_j}\mathbf{d}_{ab}\|_\infty$, where $\mathbf{d}_{ab} = \left(\mathbf{r}_a - \mathbf{r}_b\right)$ and $\|\mathbf{x}\|_\infty$ denotes the supremum norm of vector $\mathbf{x}$, that is, the maximum element of $\mathbf{x}$ in absolute value. The number of columns of $\mathbf{K}_{q_j}$ and hence the size of the vector from which we need to take the norm, grows exponentially with the number of categories. A more efficient way to calculate the distances between categories a and b can be obtained using

$$\|\mathbf{K}'_{q_j}\mathbf{d}_{ab}\|_\infty = \frac{1}{2}\|\mathbf{d}_{ab}\|_1, \tag{7}$$

where $\|\mathbf{x}\|_1$ denotes the L_1 norm of vector $\mathbf{x}$, that is $\|\mathbf{x}\|_1 = \sum_i \|x_i\|$. To see that (7) holds, note that the maximum, in absolute value, for the combinations of elements of $\mathbf{d}_{ab}$ is obtained by selecting the combination consisting of elements that have the same sign. As $\mathbf{d}_{ab}\mathbf{1}_{q_j} = \left(\mathbf{r}_a - \mathbf{r}_b\right)\mathbf{1}_{q_j} = 0$, the sum of all positive elements equals the sum of all negative values. Therefore, $\|\mathbf{K}'_{q_j}\mathbf{d}_{ab}\|_\infty = \sum_{l:d_l>0}\|d_l\| = \sum_{l:d_l<0}\|d_l\|$, where d_l denotes the lth element of $\mathbf{d}_{ab}$. Finally, as $\|\mathbf{d}_{ab}\|_1 = \sum_l \|d_l\| = \sum_{l:d_l>0}\|d_l\| + \sum_{l:d_l<0}\|d_l\|$, the equivalence in (7) immediately follows and we can express Ahmad and Dey's distance between categories a and b with respect to the categories of variable j, as $\delta(a,b) = \frac{1}{2}\|\mathbf{d}_{ab}\|_1 = \frac{1}{2}\|\mathbf{d}_{ab}\|_1 = \frac{1}{2}\sum_{l=1}^{q_j}|r_{al} - r_{bl}|$. ■

Kullback–Leibler divergence between profiles (KL). Kullback–Leibler divergence [19] is an entropy-based measure of dissimilarity between probability distributions. [6] define category dissimilarities for the categories of variable i by taking the sum of KL-divergences between the (empirical) conditional probability distributions for all other variables. Using similar notation as before, we can implement this divergence by setting all weights w_{ij} equal to one and by defining $\Phi\left(\mathbf{r}_a, \mathbf{r}_b\right) = \sum_{l=1}^{q_j}\left[r_{al}\log\left(\frac{r_{al}}{r_{bl}}\right) + r_{bl}\log\left(\frac{r_{bl}}{r_{al}}\right)\right]$, where $\log()$ is the binary logarithm and r_{al} and r_{bl} denote, as before, lth element of $\mathbf{r}_a$ and $\mathbf{r}_b$, respectively. It is important to note that KL is not symmetric. Hence, distance calculations using $\boldsymbol{\Delta}_{KL}$ may result in non-symmetric distances.

Chi-squared distance between profiles (Chi2). A distance for categorical data, that has a strong link to the data visualization technique correspondence analysis, is the Chi-squared distance. There exist several forms and implementations of the chi-square distance that differ with respect to the chosen standardization. For an $n \times p$ contingency matrix $\mathbf{F}_j = \mathbf{Z}'_i\mathbf{Z}_j$, the Chi2 between rows a and b can be defined as $s\sum_{l=1}^{p}\frac{1}{f_{\bullet l}}\left(\frac{f_{al}}{f_{a\bullet}} - \frac{f_{bl}}{f_{b\bullet}}\right)^2$, where s is the sum of all elements of $\mathbf{F}$ and $\bullet$ denotes the summation in the appropriate dimension (rows or columns) of the matrix (see, e.g., [20], p.266). In our notation, we can implement the Chi-squared distances as category dissimilarities by defining $\Phi\left(\mathbf{r}_a, \mathbf{r}_b\right) = \sum_{l=1}^{q_j}\frac{1}{p_l}\left(r_{al} - r_{bl}\right)^2$, where we dropped the constant s, p_l corresponds to the lth element of the jth block of $\mathbf{P}_d$ and, as before, r_{al} and r_{bl} denote the lth element of $\mathbf{r}_a$ and $\mathbf{r}_b$, respectively.

6. Supervised association-based distances

In a supervised setting, where we want to assign observations to classes (i.e., categories) for one variable, say y, based on observations on categorical variables x_j where $j = 1, \ldots, Q$, we can define a supervised variant of association-based categorical variable distances. That is, we can define category dissimilarities that take into account the association between variables y and x. Next, we can make predictions using either the K-nearest neighbors or a distance-based clustering method, where we fix the number of clusters to the number of classes and do a post-hoc comparison of clusters and classes. That is, we match the clusters to the true classes and assign labels accordingly. To define supervised association-based distances we create an $n \times c$ indicator matrix $\mathbf{Z}_y$, where c corresponds to the number of classes of y. If we add, to the right, this indicator matrix to $\mathbf{Z}$ and insert this supplemented $\mathbf{Z}$ into (1) through (2), we can calculate category dissimilarities using (4) and (5). Note that in this new setting, we have $Q+1$ variables and consequently $Q+1$ association-based category dissimilarity matrices $\boldsymbol{\Delta}^i$. However, the $(Q+1)$th diagonal block gives the category dissimilarities between the categories of the y variable. In a supervised setting, a category (class) of y is to be predicted based on data from the other Q variables. The category dissimilarities for y should therefore not be used. This is easily achieved by simply ignoring these in the overall block diagonal category dissimilarity matrix $\boldsymbol{\Delta}$. That is, we construct $\boldsymbol{\Delta}$ by collecting only the first Q category dissimilarity matrices on its diagonal. As before, our framework allows for great flexibility in how to incorporate the information of variable y. In particular, the dissimilarity functions $\boldsymbol{\Phi}^{ij}$ and the weights w_{ij} are pair-specific. One could, as suggested in Section 5, set all weights w_{ij} equal to 1, so that the category dissimilarities take into account all associations. We refer to this choice as "full supervised" dissimilarity.

Alternatively, in a supervised setting, one may choose to have the category dissimilarities depend only on the association with the variable y. This corresponds to the choice $w_{ij} = 1$ for $j = Q+1$ and 0 for all other pairs. In this case, category dissimilarities may better discriminate with respect to the classes of y. We refer to this choice as "supervised" dissimilarity. Note that both supervised variants lead to new categorical variable distances. Moreover, as each of these require a choice of association-based dissimilarity functions $\boldsymbol{\Phi}^{ij}$, for all pairs of variables, one could consider these as a family of new, supervised categorical variable distances.

7. Aggregation and dissimilarity tuning

Our general framework introduced in Section 3 allows for great flexibility in the implementation of distances between categorical variables. Indeed, as is clear from Definition (5), all separate category

dissimilarity measures can be combined and aggregated according to the researcher's preferences. How to exactly determine which category dissimilarity and aggregation strategy is the most appropriate is non-trivial, and this choice may depend on the properties of the data and the research objectives.

In exploratory distance-based methods, e.g., cluster analysis, a clear measure of fit is not available, since the goal is to find and interpret patterns in the data. However, if a measure of fit can be calculated, it can be used to select an aggregation and category-dissimilarity definition. That is, we can apply several aggregation and category dissimilarity definitions, and compare the fit for each of them by considering the selected measure. In a supervised classification setting, a choice for aggregation and category dissimilarity definitions can then be made based on the discrepancy between true classes with "predicted" classes. Therefore, the aggregation state (i.e., w_{ij}) and the category dissimilarity function $\boldsymbol{\Phi}^{ij}$ can be treated as tuning parameters. Note, however, that (5) allows many combinations and some choices need to be made to restrict the total search space.

8. Applications

To illustrate how our general framework can be used in practice, we consider the following category dissimilarity measures: SM, Eskin, Lin, OF, Goodall 3, Goodall 4, IOF, VE and VM (independent, 4), and, TVD, KL, Chi2, Supervised TVD and Supervised TVD full (association-based, 5). In a supervised setting, we employ K-nearest neighbors in which each new observation is labeled according to a set of K *close* training points (neighbors). In an unsupervised setting, we consider PAM as clustering method. For cluster analysis to find clusters that match classes, one needs to assume that homogeneous groups in the data are related to the class-variable. The supervised association-based distances do require classes to be known as these are used for defining the distances. Consequently, a cluster analysis using a supervised distance can no longer be considered to be unsupervised. Furthermore, when considering matching of clusters to classes, it is expected that supervised distances outperform the non-supervised distances.

We consider both synthetic and real-world datasets. The synthetic datasets consist of 1000 observations described by 12 categorical variables with an underlying cluster structure corresponding to 4 equal sized clusters. The datasets can be characterized by mild and strong *clusterability* through the association of 4 of these 12 variables with a cluster membership variable. Hence, for these four variables, the mild and strong *clusterability* levels correspond to variable-to-cluster Cramér's V values of 0.5 and 0.7, respectively. The other eight variables serve as noise variables. They are not associated with the underlying clusters. However, they are related to each other. That is, they have pairwise associations among themselves with a Cramér's V value of approximately 0.3. The two synthetic data sets are available in the `cat_dist` package,[2] and were generated using an evolutionary algorithm. The complete data generating process is described in a step-by-step fashion in [21].

Regarding the nine labeled real-world data sets in Table 1, all are available via the UCI Machine Learning repository.[3] These data sets vary with respect to the number of observations, variables and categories, and have been used in previous papers (see, e.g., [9,13]) on categorical distances. For our analysis, each data set is split into five folds for cross-validation. For the training subsets, we compute a block diagonal matrix $\boldsymbol{\Delta}$, representing pairwise category dissimilarities for every dissimilarity measure. The testing subset is used for performance assessment of the considered methods. The chosen performance metrics are (i) *Accuracy* for the nearest neighbors classifier, which represents the proportion of test observations that were correctly classified, and (ii) the *Adjusted Rand Index* (*ARI*, [22]), which measures the similarity between the assigned cluster of the test observations and their *true* cluster (i.e., the labels). These are common and well-documented choices for the specific supervised or unsupervised context, respectively. The procedure is iterated until each fold is used as test. The cross-validation procedure is repeated 10 times.

Table 1
Data sets information.

Data set	n	p	min q_j	max q_j	# clusters	Cramér's V
australian	690	8	2	14	2	0.26
balance	625	4	5	5	3	0.28
cars	1728	6	3	4	4	0.20
lymphography	148	18	2	8	4	0.38
soybean (large)	307	35	2	8	19	0.65
tae	151	5	2	46	3	0.30
tictactoe	958	9	3	3	2	0.14
vote	435	16	3	3	2	0.50
wbcd	699	9	9	11	2	0.77

8.1. K-Nearest neighbors classification of categorical data

The KNN classification of the test observations is based on the calculation of the distance between each test observation and the training observations, as described in Section 3.1. Let $\boldsymbol{\Delta}_{train}$ denote the category dissimilarity matrix, where the subscript $train$ indicates that if the category dissimilarities are data dependent, only observations of the training set were used. Furthermore, $\mathbf{Z}_{test}$ and $\mathbf{Z}_{train}$ are the indicator matrices of the test and training observations, respectively. The distances of interest are in the columns of $\mathbf{Z}_{train}\boldsymbol{\Delta}_{train}\mathbf{Z}'_{test}$ and the nearest neighbors for the jth test observation are the K smallest values in the jth column. The number of neighbors, K, is a hyper-parameter that can be tuned via cross-validation as described in Section 8.

To isolate the effect of the distance measure in the classification and clustering performance, we consider different values for K for each combination of distance metric and dataset. However, the choice of an optimal K is not crucial for our experiments and we limit ourselves to the following range of values $K \in \{1, 3, 7, 13, 21, 31, 43\}$.

Fig. 1 presents the accuracy assessment of the tuned KNN classifier across the two scenarios and the category dissimilarity definitions. In particular, the left panel of the figure refers to mild variable-to-cluster associations, the right panel to strong variable-to-cluster associations. The size of each point is proportional to the chosen tuning hyper-parameter (i.e., the number of neighbors). The position of each point represents the median cross-validation accuracy across the 20 repetitions, and the error bars signify the interquartile range of the accuracy values across these repetitions. Within each data set or scenario, performance for the different category dissimilarity measures are ranked. In this way, the most effective measures in each scenario are immediately apparent as well as performances across scenarios.

We see that a strong structure yields higher accuracy, irrespective of the selected dissimilarity measure. Moreover, the Supervised TVD category dissimilarity (described in Section 6) consistently ranks as best performing measure. This is in line with expectations as the structure in this study concerns the association strength of four of the variables with the clusters. The Supervised TVD category dissimilarity directly, and exclusively, incorporates these associations. The Supervised TVD full dissimilarity does not perform as well as Supervised TVD. This is because this measure also incorporates association among all the explanatory variables. As our data generating process also includes noise variables that are independent of the clusters, the pair-wise associations between them obscure the underlying cluster structure. This is also known as the *cluster masking problem* (see, for instance, [21]).

[2] The synthetic dataset names are `simcatdata1` for the mild structured dataset and `simcatdata2` for the strong structured dataset.

[3] https://archive.ics.uci.edu/ml/index.php.

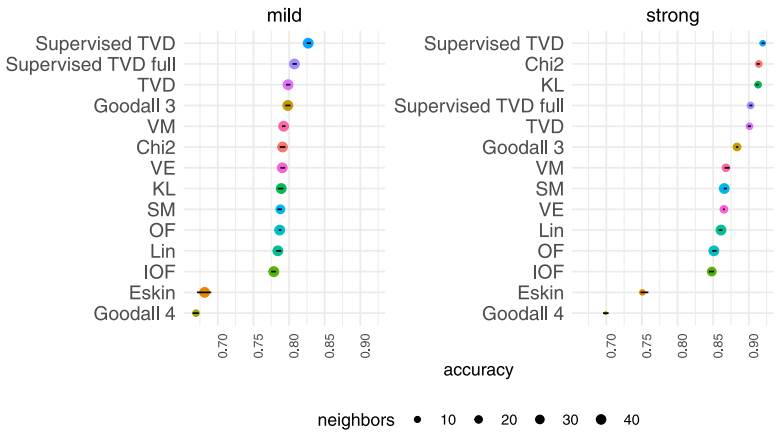

Fig. 1. KNN classification accuracy for each considered distance measure and scenario.

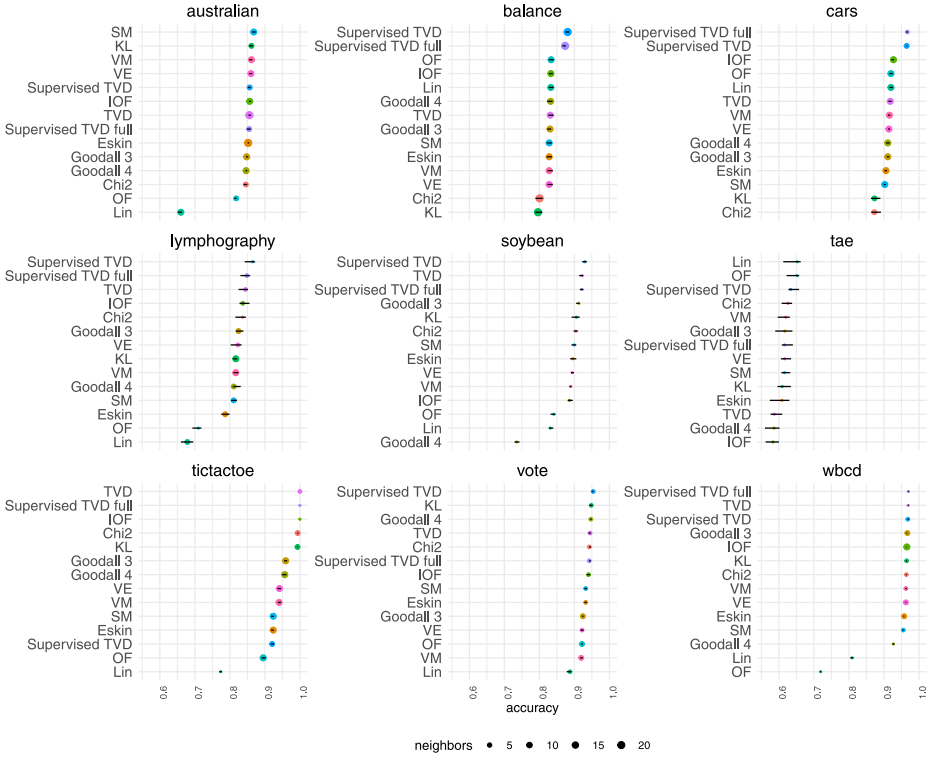

Fig. 2. KNN classification accuracy for each considered distance measure and data set.

For the application on real data, we use the same range of values for K as for the experiments on synthetic data; however, we exclude the two largest levels (31 and 43) from due to the smaller size of some real-world datasets. For each data set/category dissimilarity combination, the value of K that minimizes the cross-validation estimate of the classifier's test accuracy is selected.

Fig. 2 depicts the accuracy assessment of the tuned KNN classifiers for each data set and category dissimilarity definition under study. The presentation format and the variability assessment of the results mirror those of Fig. 1, with each panel in the figure corresponding to a specific real data set. For some data sets, e.g., *australian*, *vote*, and *wbcd*, the accuracy is high almost regardless of the chosen distance, with no variability over the 20 cross-validation repetitions. This indicates that in these three data sets, the considered categorical predictors have high discriminatory power for the underlying classes. Furthermore, since both the independent and the association-based measures perform well, taking into account interactions (associations) between predictors is neither beneficial nor detrimental.

8.2. Partitioning around medoids of categorical data

PAM is a distance-based iterative clustering procedure. Within a cluster, a medoid corresponds to the *median* observation, just like a centroid in K-means corresponds to the mean. The starting set of the K medoids is random, and each observation is assigned to the closest medoid; given the allocation of the obtained clusters, the medoids are updated accordingly. The procedure stops when there are no further changes in the set of medoids. Although we are working in an unsupervised context, the performance of PAM on each data set/category dissimilarity combination can be assessed via cross-validation by calculating distances based on the supervised association dissimilarities; this also allows for consistency with the KNN-based application. In particular, for each data set and each category dissimilarity definition, a medoids set is obtained by applying PAM to the training data. That is, $\mathbf{D} = \mathbf{Z}_{train}\boldsymbol{\Delta}_{train}\mathbf{Z}'_{train}$ where for $\boldsymbol{\Delta}_{train}$ we consider all category dissimilarity matrices described in Sections 4 and 5. Next, we compute

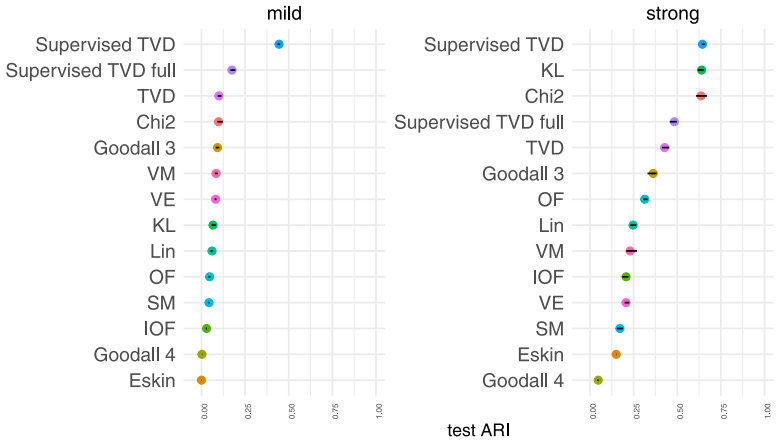

Fig. 3. ARI results for PAM for each considered dissimilarity measure and scenario.

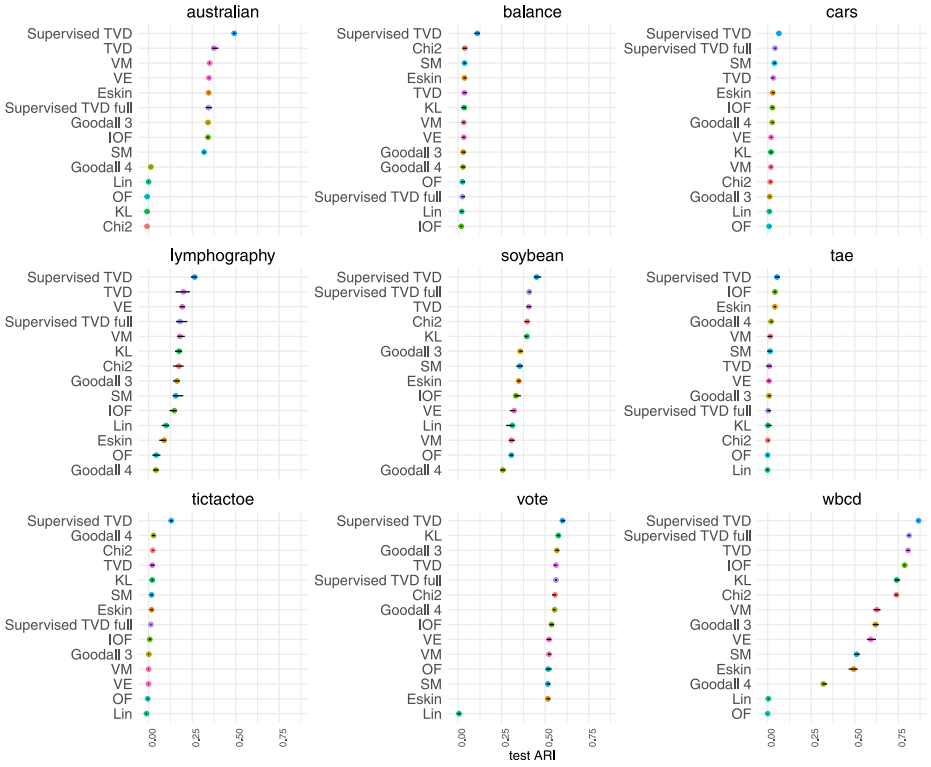

Fig. 4. ARI results for PAM for each considered dissimilarity measure and data set.

the test observation-to-medoid distance matrix as $\mathbf{Z}_{medoid}\boldsymbol{\Delta}_{train}\mathbf{Z}'_{test}$, and assign each test observation to the cluster corresponding to the nearest medoid.

The PAM results are reported in Figs. 3 (for the synthetic scenarios) and 4 (for real data). These figures have similar characteristics to Figs. 1 and 2, with a couple of distinctions: they present the ARI, which compares the clustering solution to the *true* cluster membership, and the size of the points is constant. As observed in the KNN-based application, a strong structure yields high ARI values, especially for the highest-ranking distance variants, which are all association-based distances. Note that, when there is strong association, the supervised total variance variant only performs slightly better than the unsupervised variants KL and Chi2. The strong association of the four variables to the cluster structure also leads to strong association among these variable that helps in picking up the cluster structure. Furthermore, as the 8 noise variables are associated amongst each other, this association obscures the association with the cluster structure so that the supervised TVD variant only gives slightly better result than TVD, and considerably worse results than the supervised variant that only considers the association with the class variable. When there is only mild structure, cluster retrieval is much worse. All unsupervised variants fail to adequately retrieve the clusters. Regarding the supervised variants, we observe that the supervised TVD variant that only uses the association of the variables with the class variable clearly stands out. As before, the association among the noise variables, obscures the association with the class variable.

The results on real data scenarios are reported in Fig. 4. We observe that the supervised association-based distance generally performs well. This is not a surprise as, unlike for the unsupervised variants, the class labels are explicitly used when determining the category dissimilarities and hence the distances. For the *tae* data set, the PAM performance variability is very low, due to PAM failing to find clusters that are related to the class variable. The cross-validated ARI values obtained for each of the 20 random splits are zero, or close to zero, hence variation

is low. For slightly better results, e.g. using the Supervised TVD, higher ARI values have an increased variability. In the *lymphography* data set case, the PAM procedure does find some cluster structure (most ARI's are above .15), therefore the 20 random splits of a small number of observations into five folds do affect the cross-validated ARI values and cause the observed variability.

8.3. Summary of results

Our application of KNN and PAM to various data sets shows that supervised and unsupervised performance is usually positively correlated. However, this was not the case for some datasets like *balance*, *lymphography*, and *tictactoe*, where clusters did not match classes well. Using supervised distance slightly improved results. On synthetic data, association-based measures are preferred for clustering related to classes. Real datasets with high ARI values also benefited from association-based distances, such as in the *wbcd* dataset where four out of five measures with ARI value exceeding 0.7 were association-based.

9. Conclusion

In this paper, we introduced a comprehensive framework for implementing distances between categorical variables that is flexible, efficient, and transparent. We showed that both independent and association-based distances can be integrated within our framework. Consequently, our approach can be used to apply existing measures, thus enhancing implementation transparency, or to introduce new, highly customizable ones.

Our framework is not limited to specific methods or applications, allowing for domain-specific dissimilarity measures. Specifically in supervised contexts, our approach accommodates measures considering associations with response variables. The dissimilarity definitions for independent categories outlined in Section 4 are integrated into the `nomclust` R package [12], excluding ordered category dissimilarities. The package does not support association-based measures. Conversely, the `catdist` package[4] implements both the independent and the association-based measures discussed in this paper.

To underscore the significance of selecting the "optimal" distance for a given problem, we employed our framework in both supervised (KNN) and unsupervised (PAM) settings. For real-world datasets, the choice of measure did not significantly affect results unless variables lacked discriminatory power or strong clustering, or when KNN/PAM were not suitable methods for the problem. In some scenarios, all measures delivered comparable results. However, setting aside extreme cases, opting for the "most appropriate" measure often improved outcomes, with association-based measures generally performing better than independent category measures. While further validation with more datasets would be beneficial, it falls outside the scope of this paper.

Our framework allows for implementing new or customized distances easily, as demonstrated by the supervised total variation distances introduced in Section 6. Instead of introducing and evaluating new measures, a more valuable approach would be to systematically compare existing dissimilarity measures for categorical variables in the literature, highlighting their strengths and weaknesses.

Finally, while our focus is on categorical variables, our framework can be adapted for mixed-variable settings with numerical variable distances. This integration is not a trivial task and requires addressing scale differences within and between variable types, yet our framework offers a solid starting point for this task.

[4] Available on GitHub at https://github.com/alfonsoIodiceDE/catdist_package.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Michel van de Velden: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Alfonso Iodice D'Enza:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Angelos Markos:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Carlo Cavicchia:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Extended results and the code to reproduce them are available online at https://alfonsoiodicede.github.io/blogposts_archive/distances_experiment_superv_unsuperv.html. The data used in this study are available in the `catdist` package[5].

References

[1] E. Blanco-Mallo, L. Morán-Fernández, B. Remeseiro, V. Bolón-Canedo, Do all roads lead to rome? Studying distance measures in the context of machine learning, Pattern Recognit. 141 (2023) 109646.
[2] G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor, An Introduction to Statistical Learning: With Applications in Python, Springer Nature, 2023.
[3] L. Kaufman, P. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons, New York, 1990.
[4] I. Borg, P. Groenen, Modern Multidimensional Scaling: Theory and Applications, Springer Science & Business Media, 2005.
[5] J.C. Gower, S.G. Lubbe, N.J. Le Roux, Understanding Biplots, John Wiley & Sons, 2011.
[6] S. Le, T. Ho, An association-based dissimilarity measure for categorical data, Pattern Recognit. Lett. 26 (16) (2005) 2549–2557.
[7] A. Ahmad, L. Dey, A k-mean clustering algorithm for mixed numeric and categorical data, Data Knowl. Eng. 63 (2) (2007) 503–527.
[8] H. Jia, Y. Cheung, J. Liu, A new distance metric for unsupervised learning of categorical data, IEEE Trans. Neural Netw. Learn. Syst. 27 (2014) 1065–1079.
[9] M. Ring, F. Otto, M. Becker, T. Niebler, D. Landes, A. Hotho, ConDist: A context-driven categorical distance measure, in: A. Appice, P. Rodrigues, V. Santos Costa, C. Soares, J. Gama, A. Jorge (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer International Publishing, Cham, 2015, pp. 251–266.
[10] E. Mousavi, M. Sehhati, A generalized multi-aspect distance metric for mixed-type data clustering, Pattern Recognit. 138 (2023) 109353.
[11] H. Rezaei, N. Daneshpour, Mixed data clustering based on a number of similar features, Pattern Recognit. 143 (2023) 109815.
[12] Z. Šulc, H. Řezanková, Comparison of similarity measures for categorical data in hierarchical clustering, J. Classification 36 (1) (2019) 58–72.
[13] S. Boriah, V. Chandola, V. Kumar, Similarity measures for categorical data: A comparative evaluation, in: Proceedings of the 2008 SIAM International Conference on Data Mining, SIAM, 2008, pp. 243–254.

[5] https://github.com/alfonsoIodiceDE/catdist_package

[14] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, S. Stolfo, A geometric framework for unsupervised anomaly detection, in: Applications of Data Mining in Computer Security, Springer, 2002, pp. 77–101.
[15] D. Lin, An information-theoretic definition of similarity, in: Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998, pp. 296–304.
[16] K. Spärck Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. 28 (1972) 11–21.
[17] D. Goodall, A new similarity index based on probability, Biometrics (1966) 882–907.
[18] H. Drost, Philentropy: information theory and distance quantification with R, J. Open Sour. Softw. 3 (26) (2018) 765.
[19] S. Kullback, R. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1) (1951) 79–86.
[20] A. Gifi, Nonlinear Multivariate Analysis, John Wiley & Sons Ltd., 1990.
[21] M. van de Velden, A. Iodice D'Enza, F. Palumbo, Cluster correspondence analysis, Psychometrika 82 (2017) 158–185.
[22] L. Hubert, P. Arabie, Comparing partitions, J. Classification 2 (1) (1985) 193–218.

Michel van de Velden is an associate Professor of Statistics at the Econometric Institute of the Erasmus University Rotterdam. His main research interests are exploratory data analysis. In particular, dimension reduction and cluster analysis methods with a strong focus on data visualization. In addition, he is in involved in several supervised machine learning projects involving tree-based machine learning methods.

Alfonso Iodice D'Enza is an associate Professor of Statistics at the University of Naples Federico II (Italy). His areas of interest include statistical learning, clustering, dimension reduction, computational statistics and visualization, with applications in behavioral sciences.

Angelos Markos is an associate Professor of Data Analysis in the Social Sciences at the Democritus University of Thrace, School of Education (Greece). His areas of interest include multivariate analysis (especially dimension reduction and clustering), psychological testing and measurement in the social sciences and statistics education.

Carlo Cavicchia is an assistant Professor of Statistics at the Econometric Institute of the Erasmus University Rotterdam. His research is focused on the methodological and computational aspects of data analysis. He is currently working on latent variable models, unsupervised classification and model-based composite indicators.