STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **10**, September 2022, pp 1187–1203.
Published online in International Academic Press (www.IAPress.org)

The k-NN Classification of Histogram- and Trapezoid-Valued Data

Fathimah Al-Ma'shumah[1], Mostafa Razmkhah[1,*] and Sohrab Effati[2]

[1] *Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran*
[2] *Department of Applied Mathematics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran*

Abstract A histogram-valued observation is a specific type of symbolic objects that represents its value by a list of bins (intervals) along with their corresponding relative frequencies or probabilities. In the literature, the raw data in bins of all histogram-valued data have been assumed to be uniformly distributed. A new representation of such observations is proposed in this paper by assuming that the raw data in each bin are linearly distributed, which are called *trapezoid-valued data*. Moreover, new definitions of union and intersection between trapezoid-valued observations are made. This study proposes the k-nearest neighbour technique for classifying histogram-valued data using various dissimilarity measures. Further, the limiting behaviour of the computational complexities based on the performed dissimilarity measures are compared. To study the effect of using a distance instead of a dissimilarity measure, the Wasserstein distance is also used and the accuracy of the classification is compared. Some simulations are done to study the performance of the proposed procedures. Also, the results are applied to three various real data sets. Eventually, some conclusions are stated.

Keywords Dissimilarity measure; Histogram-valued data (HVD); Supervised learning; Trapezoid-valued data (TVD), Wasserstein distance.

AMS 2010 subject classifications 62H30, 68T10

DOI: 10.19139/soic-2310-5070-1451

1. Introduction

In classical data analysis, one usually describes a unit of data as a single (numerical, ordinal, or nominal) value for each variable. On the other side, symbolic data analysis (SDA) introduced many other alternatives of representing a unit of data with some more detailed descriptions than some single-valued variables. Such structured descriptions are commonly called symbolic objects (SOs). One of the well-known types of SO introduced at the early era of SDA is the interval-valued data, in which each variable contains a pair of upper and lower bounds. The interested reader may refer to the key books by [10, 9]; see, also [8]. Global computerization of data gathering causes data sets growing much more massive. Some data sets can naturally consist of some SOs. Meanwhile, many SOs can result from some pre-defined aggregations from extra-large classical data sets into more manageable smaller data sets. Nowadays, it is common to do some pre-defined aggregations or pre-clusterings of raw data to explore as much information as possible from such massive data. Usually, the specific scientific questions of interests play a significant role in determining the rules in pre-defined aggregations. A histogram-valued variable is a specific type of SOs, which represents its values by a list of intervals along with their corresponding relative frequencies or probabilities. Therefore, it is obvious that an interval-valued variable is a particular case of a histogram-valued variable, where the list contains just one interval with the consequence probability of one. A more general form of SOs is modal-valued variable, which consists of a list of qualitative or quantitative categories along with their corresponding relative frequencies or probabilities.

*Correspondence to: Mostafa Razmkhah (Email: razmkhah_m@um.ac.ir).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

In practice, we can naturally find data in the form of HVD. For example, in collecting financial data, usually the survey forms consists of ranges such as income, expenditure, etc. When the data are grouped by multiple measurements or institutions, then we will have a list of ranges along with the corresponding relative frequencies. In many experiments, accurate measurements may be difficult, but one can specify an interval for each subject, and the ratio of the number of times the subject is placed in that interval is also known. For various technical, political, or economic reasons, we might not have the access to obtain the raw data and only have summary data in the form of histograms.

We can also view the act of building HVD as a data reduction technique. Aside from reducing the size of the data, there may be some other analytical reasons behind the act of building HVD. For example, if the subject of interest is in the form of groups of observations such as countries, institutions, or other forms of grouped measurements, the HVD may be built. Kejzar et al. [24] brought an interesting example which consists of modal-valued variables. Specifically, one of them is in the form of histogram-valued variable. They aimed to cluster the European household patterns using three variables: "gender", "category of household members" and "age-groups of household members". For every household, the variable "age-groups of household members" was represented by histograms consisting nine 10-year groups along with their corresponding relative frequency. Obviously, presenting the variable in the form of HVD provides much more information than presenting it with just a real value such as mean. On the other hand, presenting all the raw data will make the data size much larger. Therefore, building HVD can also be viewed as a dimensionality reduction.

There are some interesting researches on HVD learning in both supervised and unsupervised learning. In supervised learning, there are some studies about the classification of HVD using learning random-forest [19] and learning decision trees [18]. There are also some researches studying linear regression of histogram-valued variables such as [23] and [13]. Some related works have been proposed in the fields of k-Nearest Neighbor (k-NN) method for forecasting histogram time series [5, 16], smoothing method for histogram time series [3], and k-NN regression on a classic dependent variable but using histograms for representing the observations [4]. However, there are some studies in unsupervised learning methods such as principal component analysis [29] and clustering [25, 27] of HVD. More recently, there are some works in discriminant analysis, such as discriminant analysis for interval-valued data [15, 32], distance based discriminant analysis for interval-valued data [31], and discriminant analysis of distributional data via fractional programming [14]. Verde et al. [33] have also studied the dimensionality reduction for distributional symbolic data. Different pattern classifiers for interval data based on the logistic regression methodology have been introduced by [12]. A nonlinear regression model to interval-valued data has been studied by [28]. Beyaztas et al. [7] introduced the functional forms of some well-known regression models for interval-valued data.

This work focuses on the classification method for HVD using the k-NN method. This method is a specific type of non-parametric classification method that is simple but effective in many cases [20]. The k-NN is a well-known algorithm used not only in classification but also regression [34, 2]. In classification, each training observation consists of a vector of features and its associated class label as its target value. Given a new observation, the k-NN finds its k-most similar training observations, called k-nearest neighbours, according to any chosen distance metric such as the Wasserstein distance or any dissimilarity measures such as those that will be presented in the next section. After that, the k-NN predicts its value or class label as an aggregation of the target values associated with its nearest neighbours. Usually, in classification, the class label will be the plurality vote of its neighbors. Hence, the k-NN classification assigns the new observation to the most common class label among its k-nearest neighbours. Despite its high computational query time, the k-NN can outperform the other classifiers and can take a role in a variety of applications such as text categorization [17], economic forecasting [21], stock market prediction [1], and genetics [35].

As a new point of view, a new representation of successive bins with associated frequencies is defined in this paper. Note that in HVD, it is assumed that the observations in each bin are uniformly distributed. This assumption may be extended in different ways to arise analogous but different representations of HVD. Precisely, we assume that the observations in each bin are linearly distributed, and call such data as *trapezoid-valued data* (TVD). The classification accuracy of both HVD and TVD will then be compared via simulation and real data sets. To compute dissimilarity measures between trapezoid-valued observations, new definitions for the union and intersection are

proposed. Further, the limiting behaviour of the computational complexities of the k-NN method is compared using different dissimilarity measures. To study the effect of using a distance instead of a dissimilarity measure, the Wasserstein distance is also used and the accuracy of the classification is compared for both of HVD and TVD.

The rest of this paper is organized as follows. Some preliminaries are presented in Section 2. The TVD is defined in Section 3, and some of its properties are studied. Section 4 proposes the k-NN classification for HVD and TVD based on different dissimilarity measures; the computational complexities of these measures are also discussed. Section 5 contains a simulation study to compare the accuracy of the classification using given dissimilarity and distance measures. The proposed method is applied to some real data sets in Section 6. Eventually, some conclusions are stated in Section 7.

2. Preliminaries

2.1. Histogram-valued variables

Let $\mathbf{X} = (X_1, \ldots, X_m)$ be an m-dimensional random variable, and let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be n copies of $\mathbf{X}$. For $i = 1, \ldots, n$, an outcome or observation of the $\mathbf{X}_i = (X_{i1}, \ldots, X_{im})$ is represented as

$$\mathbf{x}_i = \{x_{ij}; j = 1, \ldots, m\} = \{[b_{ijk}, b_{ij(k+1)}), p_{ijk}; j = 1, \ldots, m, k = 1, \ldots, t_{ij}\}, \tag{1}$$

where t_{ij} is the number of bins in x_{ij} and $[b_{ijk}, b_{ij(k+1)})$ stands for the kth bin; note that $\sum_{k=1}^{t_{ij}} p_{ijk} = 1$.

A HVD can also be represented by its quantile function. The quantile function of a histogram-valued random variable is basically the inverse of its cumulative distribution function (cdf). By assuming that the observations in each bin of the histogram-valued random variable are uniformly distributed, the X_{ij} in (4) has the following probability density function (pdf)

$$f_{ij}(x) = \begin{cases} p_{ij1}, & b_{ij1} \le x < b_{ij2}, \\ p_{ij2}, & b_{ij2} \le x < b_{ij3}, \\ \vdots & \\ p_{ijt_j}, & b_{ijt_j} \le x < b_{ij(t_j+1)}. \end{cases}$$

Therefore, the cdf of X_{ij} is given by

$$F_{ij}(x) = \begin{cases} 0, & x < b_{ij1}, \\ z_{ij0} + p_{ij1}\left(\frac{x - b_{ij2}}{b_{ij2} - b_{ij1}}\right), & b_{ij1} \le x < b_{ij2}, \\ z_{ij1} + p_{ij2}\left(\frac{x - b_{ij2}}{b_{ij3} - b_{ij2}}\right), & b_{ij2} \le x < b_{ij3}, \\ \vdots & \\ z_{ij(t_j-1)} + p_{ijt_j}\left(\frac{x - b_{ijt_j}}{b_{ij(t_j+1)} - b_{ijt_j}}\right), & b_{ijt_j} \le x < b_{ij(t_j+1)}, \\ 1 & x \ge b_{ij(t_j+1)}, \end{cases}$$

where

$$z_{ij\ell} = \begin{cases} 0, & \text{if } \ell = 0, \\ \sum_{h=1}^{\ell} p_{ijh}, & \text{if } \ell = 1, \ldots, t_{ij}. \end{cases} \tag{2}$$

It is not hard to show that the quantile function of X_{ij} is

$$Q_{ij}(q) = \begin{cases} b_{j1} + \frac{q - z_{ij0}}{p_{ij1}}(b_{ij2} - b_{ij1}), & 0 \le q < z_{ij1}, \\ b_{j2} + \frac{q - z_{ij1}}{p_{ij2}}(b_{ij3} - b_{ij2}), & z_{ij1} \le q < z_{ij2}, \\ \vdots & \\ b_{j(l+1)} + \frac{q - z_{ijl}}{p_{ij(l+1)}}(b_{ij(l+1)} - b_{ijl}), & z_{ijl} \le q < z_{ij(l+1)}, \\ \vdots & \\ b_{j(t_j+1)} + \frac{q - z_{ijt_j}}{p_{ij(t_j+1)}}(b_{ij(t_j+1)} - b_{ijt_j}), & z_{ijt_j} \le q < z_{ij(t_j+1)}. \end{cases} \tag{3}$$

The representation of a HVD based on its quantile function may be used to calculate the Wasserstein distance.

2.2. Dissimilarity and distance measures for HVD

Usually, the dissimilarity measures for continuous classical type of data are interpreted as the distance between two observations based on their location in p-dimensional space. Contrasting with this fact, even when the center of two histogram-valued variables are located at the same position in p-dimensional space, their dispersion degrees might be different. Kim and Billard [26] introduced some extension of dissimilarity measures for HVD which reflect both their location and dispersion. Some of their proposed dissimilarity measures include the extended Gowda-Diday D_{GD}, the generalized Minkowski distance D_M^q, the extended DeCarvalho's measure D_{DC}, and the normalized cumulative distribution function measure D_{NCDF}.

To compute the dissimilarity measures between two HVDs, they must have the same length and number of bins. From (1), it is observed that each HVD is assumed to have different lengths and different numbers of bins across i and j. Histogram-valued observations can be aggregated from classical raw data if they are given. By first pre-specifying the same bins and then assigning the corresponding relative frequencies, one can obtain the aggregated histogram data having common bins with the same lengths and the same number of bins for each variable. However, there are some situations that raw data are not available. For example, suppose that we want to compare some districts by the distribution of companies considering the features such as asset, total wage, and the number of employees. These data might originate from statistical tables published by each district, where these tables might be of the histogram form. Since the data might be from different sources, the lengths and numbers of bins might be different across districts. It is not easy to computationally handle histogram data obtained in this case. To solve this, we can consider a transformation of such data to obtain common bins across observations. More details can be found in the Appendix of [26]. Based on their idea, histogram-valued observations in (1) can be transformed such that all HVDs have common bin lengths and the same number of bins for each observation. That is

$$\mathbf{x}_i = \{x_{ij}; j = 1, \ldots, m\} = \{[b_{jk}, b_{j(k+1)}), p_{ijk}; j = 1, \ldots, m, k = 1, \ldots, t_j\}, \tag{4}$$

where the bins are non-overlapped for given j across k; in this case $\sum_{k=1}^{t_j} p_{ijk} = 1$.

In the sequel, to compute the dissimilarity measures, we use the transformed HVD as presented in (4). Denote by M_{ij} and S_{ij} the empirical mean and standard deviation of the HVD $\mathbf{x}_{ij}$, as defined by Billard and Diday [9], respectively. Then, the dissimilarity measures D_{NCDF}, D_{GD}, D_M^q and D_{DC} between two transformed HVDs $\mathbf{x}_{i_1}$ and $\mathbf{x}_{i_2}$ are briefly described as

$$D_{NCDF}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) = \sum_{j=1}^{m} \left\{ \frac{1}{b_{j(t_j+1)} - b_{j1}} \sum_{k=1}^{t_j} \left((b_{j(k+1)} - b_{jk}) \left| \sum_{\ell=1}^{k} p_{i_1 j\ell} - \sum_{\ell=1}^{k} p_{i_2 j\ell} \right| \right) \right\},$$

$$D_{GD}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) = \sum_{j=1}^{m} \left\{ \frac{|S_{i_1 j} - S_{i_2 j}|}{S_{i_1 j} + S_{i_2 j}} + \frac{S_{i_1 j} + S_{i_2 j} - 2S_{(i_1 \cap i_2) j}}{S_{i_1 j} + S_{i_2 j}} + \frac{|M_{i_1 j} - M_{i_2 j}|}{b_{j(t_j+1)} - b_{j1}} \right\},$$

$$D_M^q(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) = \left\{ \sum_{j=1}^{m} \left(S_{(i_1 \cup i_2)j} - S_{(i_1 \cap i_2)j} + \gamma\big(2S_{(i_1 \cap i_2)j} - S_{i_1 j} - S_{i_2 j}\big) \right)^q \right\}^{1/q}$$

and

$$D_{DC}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) = \pi(\mathbf{x}_{i_1 \cup i_2}) - \pi(\mathbf{x}_{i_1 \cap i_2}) + \gamma\Big(2\pi(\mathbf{x}_{i_1 \cap i_2}) - \pi(\mathbf{x}_{i_1}) - \pi(\mathbf{x}_{i_2})\Big),$$

respectively, where γ is a prespecified constant such that $0 \leq \gamma \leq 0.5$, and $\pi(\mathbf{x}_i) = \Pi_{j=1}^{m} S_{ij}$; also, $S_{(i_1 \cup i_2)j}$ and $S_{(i_1 \cap i_2)j}$ stand for the standard deviation of the union and intersection of the $x_{i_1 j}$ and $x_{i_2 j}$, respectively. The dissimilarity measure D_{NCDF} has a value between 0 and m, where m is the number of variables. For more details see [26].

To study the effect of using a distance instead of a dissimilarity measure, the Wasserstein distance is also used in the paper. This distance for two histogram-valued random variables X_{ij} and X_{lj} is defined by

$$D_W(X_{ij}, X_{lj}) = \sqrt{\int_0^1 \big(Q_{ij}(t) - Q_{lj}(t)\big)^2 dt}, \tag{5}$$

where $Q_{ij}(\cdot)$ is as defined in (3). For more details see [22] and [23].

3. Trapezoid-valued data

In this section, we define a trapezoid-valued observation, which is in fact an analogous representation of a histogram-valued observation. Note that a HVD contains more information which will produce more accurate results than those obtained by interval data. The only information one may obtain from a HVD is a list of bins or intervals and their corresponding relative frequencies. The common assumption on the bins of the HVD is that the raw data are distributed uniformly in each bin. Here, we change this assumption such that the raw data in each bin are linearly distributed. In this case, an observation looks like a set of trapezoidal stems and hence it is named as trapezoid-valued data (TVD). The constant and slope of a linear function in each bin are computed based on the information of the adjacent bins. Of course, it is also possible to assume other kinds of distributions for the raw data in different bins rather than the linear one like normal or skew-normal distributions. In any way, any assumption about a possible distribution rather than the uniform needs other additional, external or expert information about the data distribution inside the bins.

The definition as well as some characteristics and properties of the TVD are presented in the sequel. Generally we proceed with the case that there are n observations each containing m features such that the jth feature has the number of t_j bins. Also, we assume that the data are transformed to have common bins across each observation.

Definition 1
Let $\mathbf{x}_i$ be a transformed observation as defined in (1). Then, it is said to be a TVD if the raw data in each bin are distributed linearly with the pdf

$$f_{ijk}(x) = \alpha_{ijk} + \beta_{ijk} x, \quad b_{ijk} \leq x < b_{ij(k+1)}, i = 1, \ldots, n; j = 1, \ldots, m; k = 1, \ldots, t_{ij}, \tag{6}$$

where α_{ijk} and β_{ijk} are real constants, such that for given i, j and k,

$$\int_{b_{ijk}}^{b_{ij(k+1)}} f_{ijk}(x)\, dx = 1. \tag{7}$$

Note that the distribution of the data on each bin depend on the frequencies of available data in that bin and it's adjacent bins. In other words, if the relative frequency of data at the $(k+1)$th bin is more (less) than the kth

bin, then it is reasonable to assume that the data are concentrated at the end of the jth bin with larger (smaller) probability. This information must be considered to determine the slope of each bin. Hence, for fixed values of i and j ($i = 1, \ldots, n; j = 1, \ldots, m$), it is logical to consider the average of slopes of both the $(k-1)$th and the $(k+1)$th bins as the slope of the pdf assumed on the kth bin ($k = 1, \ldots, t_{ij}$). Hence, for $i = 1, \ldots, n, j = 1, \ldots, m$ and $k = 1, \ldots, t_{ij}$, we get

$$\beta_{ijk} = \frac{1}{2}\left(\frac{p_{ij(k+1)} - p_{ijk}}{m_{ij(k+1)} - m_{ijk}} + \frac{p_{ijk} - p_{ij(k-1)}}{m_{ijk} - m_{ij(k-1)}}\right), \tag{8}$$

where p_{ijk} and m_{ijk} stand for the relative frequency and midpoint of data included in the kth bin of the jth feature of the ith observation, respectively, such that $p_{ij0} = p_{ij(t_{ij}+1)} = 0$. Since, all observations are assumed to have same bins for each feature, the midpoints are also the same for all observations, that is, for $i = 1, \ldots, n$, we get $m_{ijk} = m_{jk}$, such that

$$\begin{aligned} m_{j0} &= b_{ij1} - \frac{b_{ij2} - b_{ij1}}{2}, \\ m_{jk} &= \frac{b_{ijk} + b_{ij(k+1)}}{2}, \quad k = 1, \ldots, t_{ij} - 1, \\ m_{jt_{ij}} &= b_{ij(t_{ij}+1)} + \frac{b_{ij(t_{ij}+1)} - b_{ijt_{ij}}}{2}. \end{aligned}$$

Therefore, (8) can be rewritten as

$$\beta_{ijk} = \frac{1}{2}\left(\frac{p_{ij(k+1)} - p_{ijk}}{b_{ij(k+2)} - b_{ijk}} + \frac{p_{ijk} - p_{ij(k-1)}}{b_{ij(k+1)} - b_{ij(k-1)}}\right),$$

where

$$\begin{aligned} b_{ij0} &= b_{ij1} - (b_{ij2} - b_{ij1}) = 2b_{ij1} - b_{ij2}, \\ b_{ij(t_{ij}+2)} &= b_{ij(t_{ij}+1)} + (b_{ij(t_{ij}+1)} - b_{ijt_{ij}}) = 2b_{ij(t_{ij}+1)} - b_{ijt_{ij}}. \end{aligned}$$

Now, the constants in (6) are determined such that they satisfy the condition (7). Therefore, for $i = 1, \ldots, n, j = 1, \ldots, m$ and $k = 1, \ldots, t_{ij}$, by doing some algebraic calculation, we obtain

$$\alpha_{ijk} = \frac{p_{ijk}}{b_{ij(k+1)} - b_{ijk}} - \frac{\beta_{ijk}}{2}(b_{ij(k+1)} + b_{ijk}).$$

To more clarification about the proposed data, let us compare the representations of both HVD and TVD for the special case of one observation containing one feature with four bins, that is $i = j = 1$ and $t_1 = 4$. For example, consider the symbolic observation below

$$\{[0,1), 0.1, [1,2), 0.4, [2,3), 0.3, [3,4), 0.2\}.$$

Note that assuming uniform or linear distribution for the raw data in each bin leads to different representations of the either HVD or TVD, respectively, which are plotted in Figure 1.

Lemma 1
The mean and standard deviation of the jth feature ($j = 1, \ldots, m$) of the ith ($i = 1, \ldots, n$) TVD are given by

$$M_{ij} = \sum_{k=1}^{t_{ij}} p_{ijk}\left(\alpha_{ijk}\frac{(b^2_{ij(k+1)} - b^2_{ijk})}{2} + \beta_{ijk}\frac{(b^3_{ij(k+1)} - b^3_{ijk})}{3}\right) \tag{9}$$

and

$$S_{ij} = \left\{\sum_{k=1}^{t_{ij}} p_{ijk}\left(\alpha_{ijk}\frac{(b^3_{ij(k+1)} - b^3_{ijk})}{3} + \beta_{ijk}\frac{(b^4_{ij(k+1)} - b^4_{ijk})}{4}\right) - (M_{ij})^2\right\}^{\frac{1}{2}}, \tag{10}$$

respectively.

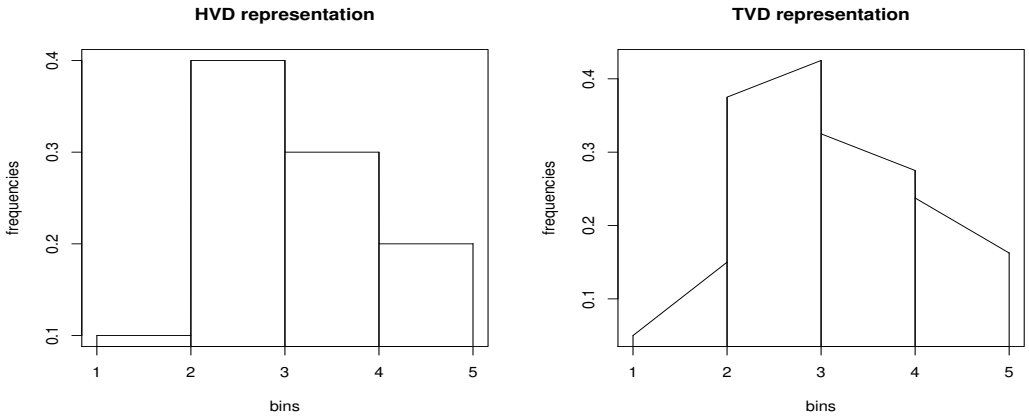

Figure 1. Examples of HVD and TVD representations.

Proof
Let us denote the pdf of the jth feature of the ith sample observation by $f_{ij}(\cdot)$, and denote the corresponding random variable by X_{ij}. Moreover, let us define

$$f_{ijk}(x) = P(X_{ij} = x | b_{ijk} \leq X_{ij} < b_{ij(k+1)}).$$

Since all of consecutive bins constitute a partition for the domain of X_{ij}, we get

$$\begin{aligned} f_{ij}(x) &= \sum_{k=1}^{t_{ij}} P(X_{ij} = x | b_{ijk} \leq X_{ij} < b_{ij(k+1)}) P(b_{ijk} \leq X_{ij} < b_{ij(k+1)}) \\ &= \sum_{k=1}^{t_{ij}} p_{ijk} f_{ijk}(x) I\big(x \in [b_{ijk}, b_{ij(k+1)})\big), \end{aligned} \tag{11}$$

where $I(\cdot)$ is an indicator function. From (7), we easily get $\int_{-\infty}^{\infty} f_{ij}(x)dx = 1$. Therefore, the expectation of X_{ij} may be obtained as follows:

$$\begin{aligned} M_{ij} = E(X_{ij}) &= \int_{-\infty}^{\infty} x f_{ij}(x)\, \mathrm{d}x \\ &= \sum_{k=1}^{t_{ij}} p_{ijk} \int_{b_{ijk}}^{b_{ij(k+1)}} x f_{ijk}(x)\, \mathrm{d}x \\ &= \sum_{k=1}^{t_{ij}} p_{ijk} \int_{b_{ijk}}^{b_{ij(k+1)}} x(\alpha_{ijk} + \beta_{ijk} x)\, \mathrm{d}x \\ &= \sum_{k=1}^{t_{ij}} p_{ijk} \left(\alpha_{ijk} \frac{(b_{ij(k+1)}^2 - b_{ijk}^2)}{2} + \beta_{ijk} \frac{(b_{ij(k+1)}^3 - b_{ijk}^3)}{3} \right). \end{aligned}$$

The variance is also derived as

$$\begin{aligned} S_{ij}^2 = \mathrm{Var}(X_{ij}) &= E(X_{ij}^2) - (E(X_{ij}))^2 \\ &= \sum_{k=1}^{t_{ij}} p_{ijk} \int_{b_{ijk}}^{b_{ij(k+1)}} x^2 f_{ijk}(x)\, \mathrm{d}x - (M_{ij})^2 \\ &= \sum_{k=1}^{t_{ij}} p_{ijk} \int_{b_{ijk}}^{b_{ij(k+1)}} x^2 (\alpha_{ijk} + \beta_{ijk} x)\, \mathrm{d}x - (M_{ij})^2 \\ &= \sum_{k=1}^{t_{ij}} p_{ijk} \left(\alpha_{ijk} \frac{(b_{ij(k+1)}^3 - b_{ijk}^3)}{3} + \beta_{ijk} \frac{(b_{ij(k+1)}^4 - b_{ijk}^4)}{4} \right) - (M_{ij})^2. \end{aligned}$$

So, the proof is complete. □

Remark 1
Note that the HVD is a special case of the TVD. Indeed, if $\beta_{ijk} = 0$ in equation (6), then by (7), we get $\alpha_{ijk} = 1$, hence, a TVD turns to a HVD. Therefore, all the results obtained above for the TVD are satisfied for the HVD by considering the mentioned changes. Of course in this case, the quantile function (3) is used. Though, if $\beta_{ijk} \neq 0$, the quantile function is a bit different which is presented in the following Lemma.

Lemma 2
Let the jth feature of the ith TVD by X_{ij}. Then, the cdf and quantile function of X_{ij} are given by

$$F_{ij}(x) = \sum_{k=1}^{t_{ij}} \left\{ z_{ij(k-1)} + p_{ijk}\alpha_{ijk}(x - b_{ijk}) + \frac{\beta_{ijk}}{2}(x^2 - b_{ijk}^2) \right\} I(b_{ijk} \leq x < b_{ij(k+1)}), \tag{12}$$

where $I(\cdot)$ stands for the indicator function, and z_{ijk} is as defined in (2). Also, the quantile function of X_{ij} is given by

$$Q_{ij}(q) = \sum_{k=1}^{t_{ij}} \frac{1}{\beta_{ijk}} \Big(-\alpha_{ijk} + \mathrm{sgn}(\beta_{ijk}) w_{ijk}(q) \Big) I(z_{ij(k-1)} \leq q < z_{ijk}), \tag{13}$$

provided that $\beta_{ijk} \neq 0$ $(1 \leq k \leq t_{ij})$, where

$$\mathrm{sgn}(\beta_{ijk}) = \begin{cases} -1, & \beta_{ijk} < 0, \\ 1, & \beta_{ijk} > 0, \end{cases}$$

and

$$w_{ijk}(q) = \left(\alpha_{ijk}^2 + 2\beta_{ijk} \left(\frac{q - z_{ij(k-1)}}{p_{ijk}} + \alpha_{ijk} b_{ijk} + \frac{\beta_{ijk}}{2} b_{ijk}^2 \right) \right)^{\frac{1}{2}}.$$

Proof
Using (11) and doing some algebraic calculations, it can be shown that the cdf of X_{ij} is as follows

$$F_{ij}(x) = \begin{cases} 0, & x < b_{ij1}, \\ p_{ij1} \int_{b_{ij1}}^{x} f_{ij1}(w)\, \mathrm{d}w, & b_{ij1} \leq x < b_{ij2}, \\ z_{ij1} + p_{ij2} \int_{b_{ij2}}^{x} f_{ij2}(w)\, \mathrm{d}w, & b_{ij2} \leq x < b_{ij3}, \\ \vdots & \\ z_{ij(t_{ij}-1)} + p_{ijt_{ij}} \int_{b_{ijt_{ij}}}^{x} f_{ijt_{ij}}(w)\, \mathrm{d}w & b_{ijt_{ij}} \leq x < b_{ij(t_{ij}+1)}, \\ 1 & x \geq b_{ij(t_j+1)}. \end{cases}$$

It is easy to rewrite the cdf as

$$F_{ij}(x) = \sum_{k=1}^{t_{ij}} \left[z_{ij(k-1)} + p_{ijk} \int_{b_{ijk}}^{x} f_{ijk}(w)\, \mathrm{d}w \right] I(b_{ijk} \le x < b_{ij(k+1)})$$
$$= \sum_{k=1}^{t_{ij}} \left[z_{ij(k-1)} + p_{ijk} \left\{ \alpha_{ijk}(x - b_{ijk}) + \frac{\beta_{ijk}}{2}(x^2 - b_{ijk}^2) \right\} \right] I(b_{ijk} \le x < b_{ij(k+1)}).$$

By solving the equation $q = F_{ij}(x)$ for $b_{ijk} \le x < b_{ij(k+1)}$ and $z_{ij(k-1)} \le q < z_{ijk}$, we get the quadratic equation

$$\frac{b_{ijk}}{2}\, x^2 + \alpha_{ijk}\, x = \frac{q - z_{ij(k-1)}}{p_{ijk}} + \alpha_{ijk} b_{ijk} + \frac{\beta_{ijk}}{2} b_{ijk}^2.$$

Hence, for $z_{ij(k-1)} \le q < z_{ijk}$, the inverse of the cdf is obtained as

$$F_{ij}^{-1}(q) = x = \frac{-\alpha_{ijk} + \mathrm{sgn}(\beta_{ijk}) \sqrt{\alpha_{ijk}^2 + 2\beta_{ijk}(\frac{q - z_{ij(k-1)}}{p_{ijk}} + \alpha_{ijk} b_{ijk} + \frac{\beta_{ijk}}{2} b_{ijk}^2)}}{\beta_{ijk}},$$

where $\mathrm{sgn}(\beta_{ijk})$ guarantees the quantile function to be increasing. Considering this function for different values of $0 < q < 1$, the the quantile function (13) is obtained. Hence, the proof is complete. □

Note that in the first glance, it seems that the same dissimilarity measures as those used for the HVD may be also used for the TVD by the difference that the means and standard deviations in the associated formulas are replaced by (9) and (10), respectively. But, this has the unfortunate consequence of leading to the possibility of negative values of variances of the trapezoid-valued observations, as computed by the extant formulas for these variances. To cope with this problem, we propose new definitions for the union and intersection between two trapezoid-valued observations. The proposed definitions formed on the basis of the principles of probability theory regarding the intersection and union between two independent events A and B, such that the probabilities of intersection and union are given by $P(A \cap B) = P(A)P(B)$ and $P(A \cup B) = P(A) + P(B) - P(A)P(B)$, respectively.

Definition 2
Let $\mathbf{x}_{i_1}$ and $\mathbf{x}_{i_2}$ be two independent transformed TVDs of the form (4). Then, the intersection of these observations is defined as

$$\mathbf{x}_{i_1 \cap i_2} = \{[b_{jk}, b_{j(k+1)}), p_{(i_1 \cap i_2)jk}; j = 1, \ldots, m, k = 1, \ldots, t_j\},$$

where

$$p_{(i_1 \cap i_2)jk} = p_{i_1 jk} p_{i_2 jk}, k = 1, \ldots, t_j. \tag{14}$$

Further, the union of $\mathbf{x}_{i_1}$ and $\mathbf{x}_{i_2}$ is given by:

$$\mathbf{x}_{i_1 \cup i_2} = \{[b_{jk}, b_{j(k+1)}), p_{(i_1 \cup i_2)jk}; j = 1, \ldots, m, k = 1, \ldots, t_j\},$$

where

$$p_{(i_1 \cup i_2)jk} = p_{i_1 jk} + p_{i_2 jk} - p_{i_1 jk} p_{i_2 jk}, k = 1, \ldots, t_j. \tag{15}$$

Note that for given i_1, i_2 and j, we get $\sum_{k=1}^{t_j} p_{(i_1 \cap i_2)jk} \le 1$ and $\sum_{k=1}^{t_j} p_{(i_1 \cup i_2)jk} \ge 1$. So, the frequencies $p_{(i_1 \cap i_2)jk}$ and $p_{(i_1 \cup i_2)jk}$ could not be used to measure the mean of $\mathbf{x}_{i_1 \cap i_2}$ and $\mathbf{x}_{i_1 \cup i_2}$, respectively. Thus, $p_{(i_1 \cap i_2)jk}$ and $p_{(i_1 \cup i_2)jk}$ need to be standardized by dividing them by $\sum_{k=1}^{t_j} p_{(i_1 \cap i_2)jk}$ and $\sum_{k=1}^{t_j} p_{(i_1 \cup i_2)jk}$, respectively. This problem was also occurred in the definition of the union and intersection of two HVDs introduced by [26]. It is also worthwhile to note that the standard deviations of both union and intersection satisfy the conditions $S_{(i_1 \cup i_2)j} \ge \max\{S_{i_1 j}, S_{i_2 j}\}$ and $S_{(i_1 \cap i_2)j} \le \min\{S_{i_1 j}, S_{i_2 j}\}$, respectively. So, they may be used to compute the dissimilarity measures and guarantee the non-negative values for them.

Remark 2
The frequencies (14) and (15) proposed in Definition 2 may be used to define the intersection and union of two HVDs. Accordingly, the dissimilarity measures between two HVDs may be computed by using the new definitions instead of the traditional ones used by [26].

4. The k-NN method

Let $(y_i, \mathbf{x}_i)$, for $i = 1, \ldots, n$, be a learning sample, where $\mathbf{x}_i$ is either a histogram- or trapezoid-valued observation with the corresponding label y_i. Moreover, let $\mathbf{x}^*$ be a given observation with unknown class membership y^*, and let $D(\cdot\,,\cdot\,)$ denote a metric. Then, the observations are ordered such that

$$D(\mathbf{x}^*, \mathbf{x}_{(1)}) \leq \cdots \leq D(\mathbf{x}^*, \mathbf{x}_{(k)}) \leq \cdots \leq D(\mathbf{x}^*, \mathbf{x}_{(n)}).$$

Now, define the neighborhood $\mathcal{N}(\mathbf{x}^*)$ of the k nearest neighbors of $\mathbf{x}^*$ by

$$\mathcal{N}(\mathbf{x}^*) = \{\mathbf{x}_j : D(\mathbf{x}^*, \mathbf{x}_j) \leq D(\mathbf{x}^*, \mathbf{x}_{(k)})\}. \tag{16}$$

Assuming there are c different classes, the estimated probability $\hat{\pi}_g$ that the covariate $\mathbf{x}^*$ belongs to class g is given by

$$\hat{\pi}_g = \frac{1}{k} \sum_{\mathbf{x}_j \in \mathcal{N}(\mathbf{x}^*)} I(y_j = g), \quad g = 1, \ldots, c, \tag{17}$$

where $I(\cdot\,)$ stands for the indicator function. The unknown y^* is assigned to the class that is most frequent within the neighborhood $\mathcal{N}(\mathbf{x}^*)$, i.e., to the class of highest predicted probability, that is, $y^* = \arg\max_g(\hat{\pi}_g)$.

Since different dissimilarity measures are used to construct the neighborhood (16), comparing the computational complexity of them may be of great interest, which is discussed in the following proposition.

Proposition 3
Consider the transformed HVDs or TVDs as presented in (4). Let $O_D(\cdot)$ be the limiting behaviour of the computational complexity of the k-NN algorithm using the dissimilarity measure D, when the number of observations n, the number of features m and the number of bins for the jth feature t_j, tends towards a particular value or infinity. Then, for both of HVD and TVD, we have

(i) The limiting computational complexity is of order $O(n(k + (\sum_{j=1}^{m} t_j)^2))$ when one of the dissimilarity measures D_{GD}, D_M^1, D_M^2 or D_{DC} is used.

(ii) The limiting computational complexity is of order $O(n(k + \sum_{j=1}^{m} t_j))$ when the algorithm uses the dissimilarity measure D_{NCDF}.

Proof
First consider the k-NN algorithm for some m-dimensional HVD with the number of t_j bins for the jth variable, $j = 1, \ldots, m$. For each single observation $\mathbf{x}^*$ having unknown label y, with n training samples, a k-NN function returning k selected indices of k nearest neighbor should have the following algorithm:

1. Initialize $selected_i = 0$ for each observation $\mathbf{x}_i$ in the training set;
2. For each observation $\mathbf{x}_i$ in the training set, compute $D(\mathbf{x}^*, \mathbf{x}_i)$, the distance between a new given observation $\mathbf{x}^*$ and $\mathbf{x}_i$;
3. For 1 to k, loop through all observations in the training set to select the index i with the smallest value of $D(\mathbf{x}^*, \mathbf{x}_i)$. Select this observation to be included in the k-nearest neighbour by setting $selected_i = 1$;
4. Return the k selected indices.

Note that, in the second step of the algorithm mentioned above, each dissimilarity measure has its own computational complexity, which differs with the others. From Definition 3.1-3.7 in [26], clearly, D_{GD}, D_M^q, and D_{DC} are computationally more complex than D_{NCDF}. Namely, each dissimilarity measure including the union and intersection between histograms adds more complexity, both in computation time and space. Meanwhile, the dissimilarity measure D_{NCDF} doesn't need any descriptive statistics to compute. The complexity of including union and intersection measures is $O(t_j)$ for the jth variable, $j = 1, \ldots, m$. Hence, the computational complexity of D_{GD}, D_M^q, and D_{DC} for the jth variable in the second step is $O((\sum_{j=1}^{m} t_j)^2)$, while the complexity of D_{NCDF} is $O(\sum_{j=1}^{m} t_j)$. Moreover, for each iteration in the third step, we perform $O(n)$ operations by looping through the training set observations, so the step overall requires $O(nk)$ operations. Further, the first and fourth steps only require $O(n)$ operations, so we get $O(n(k + \sum_{j=1}^{m} t_j))$ runtime if we use D_{NCDF} and $O(n(k + (\sum_{j=1}^{m} t_j)^2))$ if we use the other dissimilarity measures.

For the TVD, additional computation of the slopes and the constants of the pdf are needed before computing all dissimilarity measures. Hence, the limiting computational complexity is of order $O(n \sum_{j=1}^{m} t_j)$. Therefore, if one of D_{GD}, D_M^q $(q = 1, 2)$ or D_{DC} is used, then the limiting behavior of the computational complexity is of order $O(n(k + (\sum_{j=1}^{m} t_j)^2) + n \sum_{j=1}^{m} t_j) = O(n(k + (\sum_{j=1}^{m} t_j)^2))$. Also, the computational complexity of the distance D_{NCDF} in classification algorithm is $O(n(k + \sum_{j=1}^{m} t_j) + n \sum_{j=1}^{m} t_j) = O(n(k + \sum_{j=1}^{m} t_j))$. So, the proof is complete. □

Corollary 4

From Proposition 3, it is easy to see that if the numbers of bins of all features are the same, i.e., $t_j = t$, for $j = 1, \ldots, m$, then, for both of HVD and TVD, we have

(i) If one of the dissimilarity measures D_{GD}, D_M^1, D_M^2 or D_{DC} is used, then, the limiting computational complexity is of order $O(n(k + m^2 t^2)$;

(ii) If the algorithm uses the distance D_{NCDF}, then, the computational complexity is of order $O(n(k + mt))$.

5. Simulation study

To investigate the accuracy of the classification algorithm in both of HVDs and TVDs, we do a simulation study in this section. In a binary classification problem, we first generate the numbers of N_1 and N_2 observations from two classes, labeled by C_1 and C_2, respectively, each having a specific distribution. The classes C_1 and C_2 are studied for different levels of dissimilarity to analyze the behaviour of our method. Then, the distinctions between classes C_1 and C_2 are made by taking into account of both parameters of the distributions. The data from C_1 are generated from various parametric distributions, each of them having two parameters. Then, the data from the other class, C_2 are generated from the same distribution as C_1 with a $\theta_1\%$ increase on the first parameter and a $\theta_2\%$ increase on the second parameter. For example, if the class C_1 consists of normally distributed samples with mean μ and variance σ^2, denoted by $N(\mu, \sigma^2)$, then the distribution of class C_2 will be $N(\mu + \theta_1, \sigma^2)$, where θ_1 is the percentage of the change on the first parameter. In the same manner, the class C_2 can also be taken from $N(\mu, \sigma^2 + \theta_2)$, where θ_2 is the percentage of the change on the second parameter.

In the second step, we build the number of N_{SO}^k symbolic observations of the form (4) from some groups of observations in class k $(k = 1, 2)$, such that $N_{SO} = N_{SO}^1 + N_{SO}^2$ stands for the total number of SOs. Here, we assume that all SOs have the number of ten bins with the same bin lengths. After obtaining these data with binary class labels, we assume that the raw data are not available and we only have access to the SOs. Then, we consider the uniform and linear distributions for the raw data in each bin to get two kinds of HVD and TVD representations of SOs, respectively. To compute the accuracies of the k-NN classification, let us choose randomly 80% of N_{SO} as the training set, and the remaining data as the test set. To study the variety of distributions, four different cases are considered as below. In all cases, the observations of class C_1 are denoted by X_i $(i = 1, \ldots, N_1)$, and the observations of class C_2 are denoted by Y_j $(j = 1, \ldots, N_2)$.

- Case 1 (Normal distribution): Let $X_i \sim N(\mu, \sigma^2)$, where $N(\mu, \sigma^2)$ stands for the normal distribution with the probability density function (pdf)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

  where μ is the mean of the distribution and σ is the standard deviation. To study the effect of change in the mean and standard deviation, we consider $Y_j \sim N((1+\theta_1)\mu, \sigma^2)$ and $Y_j \sim N(\mu, \sigma^2 + \theta_2\sigma^2)$, respectively. So, the mean is shifted by $100\theta_1\%$ and the variance is enlarged by $100\theta_2\%$.
- Case 2 (Gamma distribution): Assume $X_i \sim \Gamma(\alpha, \beta)$, where $\Gamma(\alpha, \beta)$ denotes the Gamma distribution with shape parameter α and scale parameter β, which has the pdf

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, \quad x > 0,\ \alpha > 0, \beta > 0,$$

  where $\Gamma(\cdot)$ is the complete gamma function. The class C_2 is drawn from either $\Gamma((1+\theta_1)\alpha, \beta)$ or $\Gamma(\alpha, (1+\theta_2)\beta)$ distributions, where θ_1 and θ_2 show the amount of increase in the shape and scale parameters, respectively.
- Case 3 (Beta distribution): Suppose that $X_i \sim Beta(\alpha, \beta)$, where $Beta(\alpha, \beta)$ stands for the beta distribution with shape parameters α and β, and the pdf

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1, \alpha > 0, \beta > 0.$$

  To investigate the effect of changes in parameters, we consider either $Y_j \sim Beta((1+\theta_1)\alpha, \beta)$ or $Y_j \sim Beta(\alpha, (1+\theta_2)\beta)$, where θ_1 and θ_2 show the amount of changes on the shape parameters.
- Case 4 (LN distribution): Let $X_i \sim LN(\mu, \sigma)$, where $LN(\mu, \sigma)$ shows the Log-normal (LN) distribution with the pdf

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\Big\{ - \frac{(\log x - \mu)^2}{2\pi\sigma^2} \Big\}.$$

  The class C_2 is drawn from either $LN((1+\theta_1)\mu, \sigma)$ or $LN(\mu, (1+\theta_2)\sigma)$ distributions to study the effect of changes in the location or scale parameters, respectively.

Some one-dimensional (one feature) cases with various distributions and some selected values of θ_1 and θ_2 are simulated for both kinds of HVD and TVD for SOs. Here, the values of θ_1 and θ_2 are considered to be 20%, 50% and 100%. The accuracies of the k-NN classification using six different dissimilarity measures have been calculated based on 1000 iterations of the simulation algorithm, and the results are presented in Table 1 for $N_{SO}^1 = N_{SO}^2 = 500$.

From Table 1, it is deduced that:

1. In most of the cases, the bigger θ_1 and θ_2 result on the higher accuracies. This means that the proposed methods can recognize the different classes well.
2. The accuracy of the classification based on the TVD is always more than or equal to that based on the HVD for all dissimilarity and distance measures.
3. The accuracies obtained by using the distance D_{NCDF} for both HVD and TVD are always the same. This is trivial, since this measure uses neither the union nor intersection. Also, the distribution of the raw data in each bin does not affect on this distance.
4. The accuracies based on D_M^1, D_M^2 and D_{DC} are almost the same for all cases.
5. Usually, one of the either Gowda-Diday dissimilarity measure D_{GD} or Wasserstein distance D_W outperforms the other measures.

Class			Dissimilarity measure					
C_1	C_2		D_{GD}	D_M^1	D_M^2	D_{DC}	D_{NCDF}	D_W
N(1,1)	N(1.2,1)	HVD	0.8115	0.7479	0.7446	0.7351	0.8299	0.8266
		TVD	0.8470	0.7834	0.7801	0.7706	0.8299	0.8621
	N(1.5,1)	HVD	0.9884	0.7929	0.7929	0.7929	0.9940	0.9940
		TVD	0.9907	0.9886	0.9885	0.9849	0.9940	0.9922
	N(2,1)	HVD	1.0000	0.9971	0.9971	0.9971	1.0000	1.0000
		TVD	1.0000	0.9971	0.9971	0.9971	1.0000	1.0000
	N(1,1.2)	HVD	0.8870	0.5037	0.5053	0.4990	0.5214	0.8799
		TVD	0.9012	1.0000	1.0000	1.0000	0.5214	0.8369
	N(1,1.5)	HVD	0.9963	0.4995	0.5046	0.4950	0.5317	0.9961
		TVD	0.9984	1.0000	1.0000	1.0000	0.5317	0.9982
	N(1,2)	HVD	0.9990	0.5019	0.5017	0.4970	0.6454	1.0000
		TVD	1.0000	1.0000	1.0000	1.0000	0.6454	1.0000
Gamma(1,1)	Gamma(1.2,1)	HVD	0.5811	0.6205	0.6467	0.6439	0.8169	0.7366
		TVD	0.8320	0.8879	0.8879	0.8879	0.8169	0.8321
	Gamma(1.5,1)	HVD	0.9158	0.8432	0.8377	0.8510	0.9842	0.9871
		TVD	0.9884	0.9988	0.9988	0.9988	0.9842	0.9896
	Gamma(2,1)	HVD	0.9974	0.9962	0.9959	0.9960	1.0000	0.9999
		TVD	0.9996	1.0000	1.0000	1.0000	1.0000	1.0000
	Gamma(1,1.2)	HVD	0.7689	0.5110	0.5034	0.5033	0.6782	0.6630
		TVD	0.8523	0.5179	0.5184	0.5128	0.6782	0.7802
	Gamma(1,1.5)	HVD	0.9475	0.5139	0.5046	0.5031	0.9497	0.9477
		TVD	0.9844	0.5638	0.5816	0.5724	0.9497	0.9683
	Gamma(1,2)	HVD	0.9960	0.5806	0.5765	0.5801	0.9968	0.9962
		TVD	0.9988	0.6005	0.5866	0.5893	0.9968	0.9992
Beta(2,5)	Beta(2.4,5)	HVD	0.8040	0.7590	0.7590	0.7590	0.8057	0.8229
		TVD	0.8745	0.8238	0.8198	0.8123	0.8057	0.8838
	Beta(3,5)	HVD	0.9921	0.9810	0.9810	0.9810	0.9935	0.9961
		TVD	0.9961	0.9922	0.9926	0.9930	0.9935	0.9971
	Beta(4,5)	HVD	1.0000	0.9955	0.9955	0.9955	1.0000	1.0000
		TVD	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Beta(2,6)	HVD	0.8525	0.5276	0.5469	0.5410	0.8669	0.7963
		TVD	0.9841	0.5619	0.5460	0.5470	0.8669	0.8702
	Beta(2,7.5)	HVD	0.9841	0.6926	0.6990	0.6961	0.9943	0.9914
		TVD	0.9998	0.7944	0.7912	0.7880	0.9943	0.9953
	Beta(2,10)	HVD	0.9995	0.9238	0.9287	0.9198	1.0000	1.0000
		TVD	1.0000	0.9473	0.9448	0.9485	1.0000	1.0000
LN(0.25,0.44)	LN(0.3,0.44)	HVD	0.6597	0.5553	0.5664	0.5571	0.6846	0.4768
		TVD	0.6969	0.8133	0.8133	0.8133	0.6846	0.6790
	LN(0.375,0.44)	HVD	0.7077	0.6241	0.6270	0.6204	0.9052	0.8513
		TVD	0.8898	0.9978	0.9978	0.9978	0.9052	0.9025
	LN(0.5,0.44)	HVD	0.9768	0.8265	0.8363	0.8244	0.9961	0.9925
		TVD	0.9891	1.0000	1.0000	1.0000	0.9961	0.9965
	LN(0.25,0.308)	HVD	0.8166	0.4988	0.4915	0.4828	0.5596	0.7552
		TVD	0.9691	1.0000	1.0000	1.0000	0.5596	0.8309
	LN(0.25,0.66)	HVD	0.9732	0.5027	0.4978	0.4995	0.7878	0.9785
		TVD	0.9996	1.0000	1.0000	1.0000	0.7878	0.9876
	LN(0.25,0.88)	HVD	0.9986	0.5003	0.5039	0.5148	0.9912	0.9982
		TVD	1.0000	1.0000	1.0000	1.0000	0.9912	0.9996

Table 1. Simulation accuracies of k-NN classification in one-dimensional HVD and TVD.

Class			Dissimilarity measure					
C_1	C_2		D_{GD}	D_M^1	D_M^2	D_{DC}	D_{NCDF}	D_W
Feature 1: N(1,1)	N(1.2,1.2)	HVD	0.9309	0.498	0.5069	0.5111	0.8087	0.462
Feature 2: Gamma(1,1)	Gamma(1.2,1.2)	TVD	0.992	0.918	0.884	0.999	0.8087	0.661
Feature 1: N(1,1)	N(1.5,1.5)	HVD	0.9993	0.9753	0.9747	0.9997	0.965	0.9185
Feature 2: Gamma(1,1)	Gamma(1.5,1.5)	TVD	0.999	1	1	1	0.965	0.9997
Feature 1: Gamma(1,1)	Gamma(1,1)	HVD	0.8082	0.5138	0.4962	0.5085	0.7738	0.8321
Feature 2: LN(0.25,0.44)	LN(0.3,0.528)	TVD	0.894	1	1	0.9994	0.7738	0.8709
Feature 1: Gamma(1,1)	Gamma(1.5,1.5)	HVD	0.993	0.5047	0.5038	0.5808	0.9362	0.9874
Feature 2: LN(0.25,0.44)	LN(0.375,0.66)	TVD	0.9938	1	1	1	0.9362	0.9909
Feature 1: N(1,1)	N(1.2,1.2)	HVD	0.7053	0.5051	0.4929	0.5043	0.8718	0.9535
Feature 2: Gamma(1,1)	Gamma(1.2,1.2)	TVD	0.9649	1	1	1	0.8718	0.9604
Feature 3: Beta(2,5)	Beta(2.4,6)							
Feature 4: LN(0.25,0.44)	LN(0.3, 0.528)							
Feature 1: N(1,1)	N(1.5,1.5)	HVD	0.9876	0.5029	0.4986	0.492	0.99	0.9998
Feature 2: Gamma(1,1)	Gamma(1.5,1.5)	TVD	1	1	1	1	0.99	1
Feature 3: Beta(2,5)	Beta(3,7.5)							
Feature 4: LN(0.25,0.44)	LN(0.375, 0.66)							

Table 2. Simulation accuracies of k-NN classification in the case of multidimensional HVD and TVD.

	Dissimilarity measure					
	D_{GD}	D_M^1	D_M^2	D_{DC}	D_{NCDF}	D_W
HVD	0.42	0.33	0.34	0.35	0.11	2.12
TVD	1.09	0.91	0.89	1.16	0.11	2.33

Table 3. CPU time for k-NN classification using various dissimilarity measures for HVD and TVD.

Remark 3
As previously mentioned in Remark 2, both of new and traditional definitions of union and intersection have been used to classify the HVDs. Our simulated results show that the new definitions have no effect on the accuracies of HVDs.

To study the case of multidimensional (more than one feature) binary classification problem, a bit simulation is also done with 1000 iterations when there are two or four features with various distributions and different changes on the first and second parameters θ_1 and θ_2. The accuracies of k-NN classification method are presented in Table 2. It is observed that all the results deduced from Table 1 are also true for the multidimensional case.

Table 3 contains the CPU time for running the classification algorithm for the HVD and TVD one time in one-dimensional case for given dissimilarity and distance measures. The running time is recorded in minutes. From this table, it is deuced that:

1. The D_{NCDF} always performs faster than the other measures. This is clear because of its lower computational complexity, which was proved in Proposition 3.
2. For all measures, running the classification algorithm for the TVD needs more time than HVD.
3. The running time for the distance D_W is more than all given dissimilarity measures.

From the results above, it can be concluded that the distances and dissimilarity measures have their own characteristics. If one wants to obtain the most precise solution, one can use them altogether by using ensemble technique or by comparing these measures. However, sometimes there exists a condition that the computational time is limited. In this case, one can choose the suitable measure based on the available time and desired accuracy. It is also worth to note that even though in some cases the distance D_W works better than all given dissimilarity measures, it's computational time is more than twice longer than the others. Moreover, even though TVD can boost the performance of the given measures significantly, it does not change significantly the computational time for the distance D_W, but it changes computational time for other dissimilarity measures.

Data set	N_{data}	N_{SO}	N_{member}	N_{bins}	N_{train}	N_{test}	Class labels
Computers	360000	500	720	20	450	50	$Desktop\backslash Laptop$
Worms	232200	258	900	11	181	77	$Mutant\backslash Non-mutant$
Wafer	76000	500	152	10	400	100	$Normal\backslash Abnormal$

Table 4. Summary of the real data sets.

6. Applications to real data sets

Here, we use three data sets to perform binomial classification of the proposed SO representations. The data are taken from [6]. To study the performance of the proposed procedure in the paper, these data are transformed into both of HVD and TVD. Let us denote the number of raw data by N_{data}, in which every N_{member} of observations are grouped to build number of N_{SO} symbolic observations of any kind of HVD or TVD, each containing N_{bins} bins, such that they are split into training and test sets containing N_{train} and N_{test} symbolic observations, respectively. All three data sets are categorized into two classes. The summarized description of the data sets are reported in Table 4.

In the sequel some brief descriptions of the data sets are presented.

1. Computers data set
   These data are recorded as a part of a government-sponsored study called Powering the Nation. The aim was to collect behavioral data about how consumers use electricity in their houses to help reduce the country's carbon footprint. The data contains electricity readings from 251 households, sampled in two-minute intervals over a month. Hence, each observation contains a group of 720 measurements (24 hours of readings taken every 2 minutes). The two classes are either Desktop or Laptop. This way, without surveying one-by-one, hopefully, we can know how many consumers uses the desktop or laptop only from the records of their electricity.

2. Worms data set
   Caenorhabditis elegans is a special type of worm commonly used as a model organism in the study of genetics. The movement of these worms is known to be a useful indicator for understanding behavioral genetics. [11] described a system for recording the traces of the worms' motions. They captured these traces by four scalars representing the amplitudes along each dimension based on four "eigenworms". The data report 258 traces of worms converted into four eigenworm series. The eigenworm data are of the lengths from 17984 to 100674 (sampled at 30 Hz, so from 10 minutes to 1 hour) and in four dimensions (eigenworm 1 to 4). There are five classes: N2, goa-1, unc-1, unc-38, and un63, such that N2 is wildtype (i.e., normal) and the others are mutant strains. This data set is a two-class version: mutant vs non-mutant.

3. Wafer data set
   Wafer data set relates to the fabrication of semiconductor microelectronics. This data set is a record of the inline process control measurements through various sensors during the processing of silicon wafers for the fabrication of semiconductors. [30] formatted this data set as a part of his thesis. The formatted version constitutes the wafer database, in which each data set contains the measurements during the processing of one wafer by one tool recorded by one sensor. The two-class labels are normal and abnormal, having a large imbalance between normal and abnormal (10.7% of the training set and 12.1% of the test set are abnormal).

The k-NN classification method proposed in the paper has been done for the mentioned data sets above. The accuracies of different schemes have been presented in Table 5 for given dissimilarity and distance measures. The results show that:

1. Classifying the TVD is more accurate than HVD for all data sets.

Data set		Dissimilarity measure					
		D_{GD}	D_M^1	D_M^2	D_{DC}	D_{NCDF}	D_W
Computers	HVD	0.692	0.86	0.86	0.86	0.819	0.68
	TVD	0.78	0.94	0.89	0.94	0.819	0.694
Worms	HVD	0.64	0.59	0.59	0.6	0.644	0.58
	TVD	0.67	0.63	0.63	0.63	0.644	0.68
Wafer	HVD	0.92	0.852	0.852	0.852	0.91	0.94
	TVD	0.93	0.903	0.903	0.903	0.91	0.959

Table 5. The accuracies of k-NN classification for the real data sets.

2. For the Computers data set, D_M^1 and D_{DC} perform better than the other measures, and the accuracy for the TVD improves significantly.
3. For the Worms data set, it is better to use either D_{GD} or D_{NCDF} for HVD, and it is suggested to use either D_{GD} or D_W for TVD.
4. For the Wafer data set, using the distance D_W leads to more accurate classification than the other dissimilarity measures for both HVD and TVD.

7. Conclusion

In this paper a TVD representation was proposed for a symbolic observation consisting of different bins with the associated frequencies by assuming the linear distribution for the raw data in each bin. However, the HVD representation for the same observation had been previously used. The problem of classifying both types of HVD and TVD was studied using the k-NN technique based on some dissimilarity and distance measures. To classify the HVDs and TVDs, the suitable measures were investigated for different distributions using a simulation study. It was shown that the dissimilarity measure D_{NCDF} has a significantly lower computational complexity compared to the others. Therefore, this distance can be a suitable choice for classifying the HVDs or TVDs if the resources are limited and a fast running-time is needed. Also, from the experimental results, it was deduced that the TVD improves the accuracy of the classification rather than HVD, however, the TVD has higher computational complexity than HVD. Hence, using TVD is preferred if improving the accuracies of the classification is of interest, but the HVD is suggested if the resources do not support high-cost computations.

Some future researches on the field of k-NN classification of HVD or TVD include the usage of other dissimilarity measures or distances. Moreover, other assumptions for distribution of the raw data in each bin rather than uniform or linear distributions may be of interest.

Acknowledgment

The authors express their sincere thanks to anonymous referees and the Associate Editor for their useful comments and constructive criticisms on the original version of this manuscript, which led to this considerably improved version.

REFERENCES

1. Alkhatib, K., Najadat, H., Hmeidi, I., and Shatnawi, M. K. A. (2013). Stock price prediction using k-nearest neighbor (knn) algorithm. *International Journal of Business, Humanities and Technology*, 3(3):32–44.
2. Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.
3. Arroyo, J., Gonzalez-Rivera, G., Mate, C., and Roque, A. M. S. (2011). Smoothing methods for histogram-valued time series: an application to value-at-risk. *Statistical Analysis and Data Mining*, 4(2):216–228.

4. Arroyo, J., Guijarro, M., and Pajares, G. (2016). An instance-based learning approach for thresholding in crop images under different outdoor conditions. *Computers and Electronics in Agriculture*, 127:669–679.
5. Arroyo, J. and Mate, C. (2009). Forecasting histogram time series with k-nearest neighbours methods. *International Journal of Forecasting*, 25(1):192–207.
6. Bagnall, A., Lines, J., Bostrom, A., Large, J., and Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3):606–660.
7. Beyaztas, U., Shang, H. L., and Abdel-Salam, A.-S. G. (2020). Functional linear models for interval-valued data. *Communications in Statistics - Simulation and Computation*, pages 1–20.
8. Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal of American Statistical Association*, 98:470–487.
9. Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley and Sons, West Sussex, UK.
10. Bock, H. and Diday, E. (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Berlin.
11. Brown, A. E. X., Yemini, E. I., Grundy, L. J., Jucikas, T., and Schafer, W. R. (2013). A dictionary of behavioral motifs reveals clusters of genes affecting caenorhabditis elegans locomotion. 110(2):791–796.
12. de Souza, R. M. C. R., Queiroz, D. C. F., and Cysneiros, F. J. A. (2011). Logistic regression-based pattern classifiers for symbolic interval data. *Pattern Analysis and Applications*, 14:273–282.
13. Dias, S. and Brito, P. (2015). Linear regression model with histogram-valued variables. *Statistical Analysis and Data Mining*, 8(2):75–113.
14. Dias, S., Brito, P., and Amaral, P. (2021). Discriminant analysis of distributional data via fractional programming. *European Journal of Operational Research*, 294(1):206–218.
15. Duarte Silva, A. and Brito, P. (2006). Linear discriminant analysis for interval data. *Computational Statistics*, 21(2):289–308.
16. Gonzalez-Rivera, G. and Arroyo, J. (2012). Time series modeling of histogram-valued data: The daily histogram time series of s&p500 intradaily returns. *International Journal of Forecasting*, 28(1):20–33. Special Section 1: The Predictability of Financial Markets Special Section 2: Credit Risk Modelling and Forecasting.
17. Guo, G., Wang, H., Bell, D., Bi, Y., and Greer, K. (2004). An knn model-based approach and its application in text categorization. *In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2004. Lecture Notes in Computer Science*, 2945:559–570.
18. Gurung, R. B., Lindgren, T., and Bostrom, H. (2015). Learning decision trees from histogram data. In *Proceedings of the 2015 International Conference on Data Mining : DMIN 2015*, pages 139–145.
19. Gurung, R. B., Lindgren, T., and Bostrom, H. (2018). Learning random forest from histogram data using split specific axis rotation. *International Journal of Machine Learning and Computing*, 8(1):74–79.
20. Hand, D. J., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. The MIT Press.
21. Imandoust, S. and Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events theoretical background. *International Journal of Engineering Research and Applications*, 3:605–610.
22. Irpino, A. and Verde, R. (2006). A new wasserstein based distance for the hierarchical clustering of histogram symbolic data. In Batagelj, V., Bock, H.-H., Ferligoj, A., and Žiberna, A., editors, *Data Science and Classification*, pages 185–192, Berlin, Heidelberg. Springer Berlin Heidelberg.
23. Irpino, A. and Verde, R. (2015). Linear regression for numeric symbolic variables: a least squares approach based on wasserstein distance. *Advances in Data Analysis and Classification*, 9:81–106.
24. Kejzar, N., Korenjak-Cerne, S., and V, V. B. (2021). Clustering of modal valued symbolic data. *Advances in Data Analysis and Classification*, 15:513–541.
25. Kim, J. and Billard, L. (2011). A polythetic clustering process and cluster validity indexes for histogram-valued objects. *Computational Statistics and Data Analysis*, 55(7):2250 – 2262.
26. Kim, J. and Billard, L. (2013). Dissimilarity measures for histogram-valued observations. *Communications in Statistics - Theory and Methods*, 42(2):283–303.
27. Kim, J. and Billard, L. (2018). Double monothetic clustering for histogram-valued data. *Communications for Statistical Applications and Methods*, 25:263–274.
28. Lima Neto, E. d. A. and de Carvalho, F. d. A. T. (2017). Nonlinear regression applied to interval-valued data. *Pattern Analysis and Applications*, 20:809–824.
29. Nagabhushan, P. and Pradeep Kumar, R. (2007). Histogram pca. *Advances in Neural Networks–ISNN 2007. Lecture Notes in Computer Science*, 4492:1012–1021.
30. Olszewski, R. T. (2001). *Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, 2001.
31. Ramos-Guajardo, A. B. and Grzegorzewski, P. (2016). Distance-based linear discriminant analysis for interval-valued data. *Information Sciences*, 372:591–607.
32. Silva, A. P. D. and Brito, P. (2015). Discriminant analysis of interval data: An assessment of parametric and distance-based approaches. *Journal of Classification*, 32(3):516–541.
33. Verde, R., Irpino, A., and Balzanella, A. (2016). Dimension reduction techniques for distributional symbolic data. *IEEE Trans Cybern*, 46:344–355.
34. Wu, X., Kumar, V., Quinlan, R., Ghosh, J., Yang, Q., Motoda, H., Mclachlan, G., Ng, S. K. A., Liu, B., Yu, P., Zhou, Z.-H., Steinbach, M., Hand, D., and Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14:1–37.
35. Yao, Z. and Ruzzo, W. L. (2006). A regression-based k nearest neighbor algorithm for gene function prediction from heterogeneous data. *BMC Bioinformatics*, 7:S11.