Revista Colombiana de Estadística
Junio 2014, volumen 37, no. 1, pp. 77 a 92

Three Similarity Measures between One-Dimensional Data Sets

Tres medidas de similitud entre conjuntos de datos unidimensionales

Luis Gonzalez-Abril[1,a], Jose M. Gavilan[1,b], Francisco Velasco Morente[1,c]

[1]Departamento de Economía Aplicada I, Facultad de Ciencias Económicas y Empresariales, Universidad de Sevilla, Sevilla, Spain

Resumen

Based on an interval distance, three functions are given in order to quantify similarities between one-dimensional data sets by using first-order statistics. The Glass Identification Database is used to illustrate how to analyse a data set prior to its classification and/or to exclude dimensions. Furthermore, a non-parametric hypothesis test is designed to show how these similarity measures, based on random samples from two populations, can be used to decide whether these populations are identical. Two comparative analyses are also carried out with a parametric test and a non-parametric test. This new non-parametric test performs reasonably well in comparison with classic tests.

***Palabras clave*:** Data mining, Interval distance, Kernel methods, Non-parametric tests.

Abstract

Basadas en una distancia intervalar, se dan tres funciones para cuantificar similaridades entre conjuntos de datos unidimensionales mediante el uso de estadísticos de primer orden. Se usa la base de datos Glass Identification para ilustrar cómo esas medidas de similaridad se pueden usar para analizar un conjunto de datos antes de su clasificación y/o para excluir dimensiones. Además, se diseña un test de hipótesis no parámetrico para mostrar cómo similaridad, basadas en muestras aleatorias de dos poblaciones, se pueden usar para decidir si esas poblaciones son idénticas. También se realizan dos análisis comparativos con un test paramétrico y un test no paramétrico. Este nuevo test se comporta razonablemente bien en comparación con test clásicos.

***Key words*:** distancia entre intervalos, métodos del núcleo, minería de datos, tests no paramétricos.

[a]Senior lecturer. E-mail: luisgon@us.es
[b]Senior lecturer. E-mail: gavi@us.es
[c]Senior lecturer. E-mail: velasco@us.es

1. Introduction

Today, in many tasks in which data sets are analysed, researchers strive to achieve some way of measuring the features of data sets, for instance, to distinguish between informative and non-informative dimensions. A first step could be to study whether several sets of data are similar. The similarity may be defined as a measure of correspondence between the data sets under study. That is, a function which, given two data sets X and Y, returns a real number that measures their similarity.

In data mining, there exist several similarity measures between data sets: for instance, in Parthasarathy & Ogihara (2000), a similarity is used which compares the data sets in terms of how they are correlated with the attributes in a database. A similar problem, studied in Burrell (2005), is the measurement of the relative inequality of productivity between two data sets using the Gini coefficient (González-Abril, Velasco, Gavilán & Sánchez-Reyes 2010). A similarity measure based on mutual information (Bach & Jordan 2003) is used to determine the similarity between images in Nielsen, Ghugre & Panigrahy (2004). Similarity between molecules is used in Sheridan, Feuston, Maiorov & Kearsley (2004) to predict the nearest molecule and/or the number of neighbours in the training set.

A common problem with the aforementioned similarity measures is that their underlying assumptions are often not explicitly stated. This study aims to use first-order statistics to explain the similarity between data sets. In this paper, the similarity is established in the sense that one-dimensional data sets are similar simply by comparing the statistics of the variables in each data set.

In statistics, other similarity measures between data sets are also available (González, Velasco & Gasca 2005), for instance, those which are used in hypothesis testing. In this way, a non-parametric hypothesis test based on the proposed similarity is presented in this paper and a comparative analysis is carried out with several well-known hypothesis tests.

The remainder of the paper is arranged as follows: In Section 2, we introduce some notation and definitions. Sections 3 and 4 are devoted to give two similarity measures between one-dimensional data sets. An example is presented in Section 5 to show their use. A non-parametric test is derived in Section 6 and experimental results are given to illustrate its behaviour and good features. Finally, some conclusions are drawn and future research is proposed.

2. Concepts

Following Lin (1998), with the purpose of providing a formal definition of the intuitive concept of similarity between two entities X and Y, the intuitions about similarity must be clarified. Thus: i) The similarity is related to their commonality in that the more commonality they share, the more similar they are; ii) The similarity is related to the differences between them, in that the more

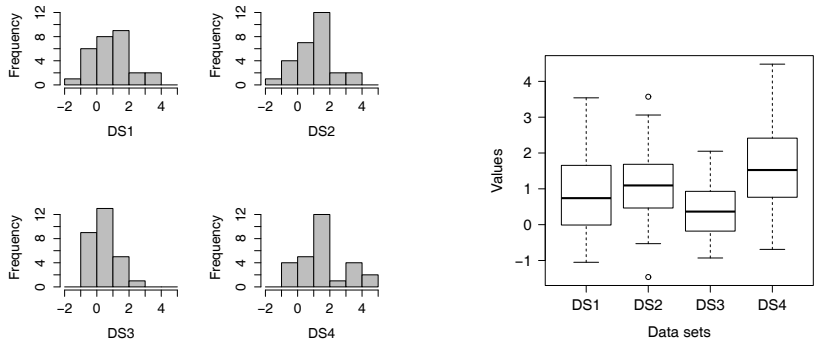
differences they have, the less similar they are; and iii) The maximum similarity is reached when X and Y are identical.

Let us denote a similarity measure between X and Y by $K(X, Y)$. Ideally this function must satisfy the following properties:

1. Identity: $K(X, Y)$ at its maximum corresponds to the fact that the two entities are identical in all respects;
2. Distinction: $K(X, Y) = 0$ corresponds to the fact that the two entities are distinct in all respects; and
3. Relative Ordinality: If $K(X, Y) > K(X, Z)$, then it should imply that X is more similar to Y than it is to Z.

Hence, certain similarities are defined in this paper which are consistent with the above intuitions and properties.

Let us consider four one-dimensional data sets, DS_1, DS_2, DS_3 and DS_4 (see the Appendix), where the DS_1 and DS_2 data sets are taken from a N(1,1) distribution, the DS_3 data set from a N(0.5,1) distribution, and the DS_4 data set from a N(1.5, 1.25) distribution, where a N(μ, σ) distribution is a Normal distribution with mean μ and standard deviation σ. In practice, comparison of these data sets involves: a) plotting graphical summaries, such as histograms and boxplots, next to each other; b) simply comparing the means and variances (see Figure 1); or



	DS_1	DS_2	DS_3	DS_4
Mean	0.989	1.046	0.427	1.632
Variance	1.221	1.152	0.569	1.956

FIGURE 1: Histograms, boxplots, means, and variances of the data sets of the Appendix.

c) calculating correlation coefficients (if items of data are appropriately paired). These methods are straightforward to interpret and explain. Nevertheless, these approaches contain a major drawback since the interpretation is subjective and the similarities are not quantified.

Let us introduce the concept of interval distance. Given an open interval (similarly for another kind of interval) of finite length, there are two main ways

to represent that interval: using the extreme points as *(a,b)* (classic notation) or as an open ball $B_r(c)$ (Borelian notation) where $c = (a+b)/2$ (centre) and $r = (b-a)/2$ (radius). Using Borelian notation, the following distance between intervals given in González, Velasco, Angulo, Ortega & Ruiz (2004) is considered:

Definition 1. Let $I_1 = (c_1 - r_1, c_1 + r_1)$ and $I_2 = (c_2 - r_2, c_2 + r_2)$ be two real intervals. A distance between these intervals is defined as follows:

$$d_{\mathbf{W}}(I_1, I_2) = \sqrt{(\Delta c\ ,\ \Delta r)\mathbf{W}\begin{pmatrix} \Delta c \\ \Delta r \end{pmatrix}} \tag{1}$$

where $\Delta c = c_2 - c_1$, $\Delta r = r_2 - r_1$, and $\mathbf{W}$ is a symmetrical and positive-defined 2×2 matrix, called weight-matrix.

It is clear from matrix algebra that $\mathbf{W}$ can be written as[1] $\mathbf{W} = \mathbf{P}^t\mathbf{P}$, where $\mathbf{P}$ is a non-singular 2×2 matrix, and hence $d_{\mathbf{W}}(I_1, I_2) = \|\mathbf{P}\,(\Delta c\,,\,\Delta r)^t\|$, where $\|\cdot\|$ is the quadratic norm in $\mathbb{R}^2$, and therefore $d_W(\cdot,\cdot)$ is an ℓ_2-distance. It can be observed that, from the matrix $\mathbf{W}$, the weight assigned to the position of the intervals c, and to their size r, can be controlled. Furthermore, the distance (1) provides more information on the intervals than does the Hausdorff distance (González et al. 2004).

From the distance given in (1), three new similarity measures are defined in this paper.

3. A First Similarity

Definition 2. Given a data set $X = \{x_1, \ldots, x_n\}$ and a parameter $\ell > 1$, the ℓ-associated interval of X, denoted by I_X^ℓ, is defined as follows:

$$I_X^\ell = (\overline{X} - \ell \cdot S_X, \overline{X} + \ell \cdot S_X)$$

where $\overline{X}$ and S_X are the mean and the standard deviation of X, respectively.

It is worth noting that Chebyshev's inequality states that there are at least a $(1 - 1/\ell^2)$ proportion of observations x_i in the interval I_X^ℓ. Hence, the similarity between two data sets X and Y can be quantified from the distance between the intervals I_X^ℓ and I_Y^ℓ. However, it is possible that some instances $z \in X \cup Y$ exist such that $z \notin I_X^\ell \cup I_Y^\ell$. Thus, a penalizing factor (the proportion of instances within I_X^ℓ and I_Y^ℓ) is taken into account in the following similarity measure.

Definition 3. Given two data sets $X = \{x_1, \ldots, x_n\}$, $Y = \{y_1, \ldots, y_m\}$ and a parameter $\ell > 1$, a similarity measure between X and Y, denoted by $K_{\mathbf{W}}^\ell(X, Y)$, is defined as follows:

$$K_{\mathbf{W}}^\ell(X, Y) = \frac{\#((X \cup Y) \cap (I_X^\ell \cap I_Y^\ell))}{\#(X \cup Y)} \cdot \frac{1}{1 + d_{\mathbf{W}}(I_X^\ell, I_Y^\ell)} \tag{2}$$

where $\#A$ denotes the cardinality of set A.

[1] The notation "$\mathbf{u}^t$" denotes the transposed vector of $\mathbf{u}$.

The function defined, $K^{\ell}_{\mathbf{W}}$, is a similarity measure (Cristianini & Shawe-Taylor 2000) which has been proposed based on distance measurements in Lee, Kim & Lee (1993) and Rada, Mili, Bicknell & Blettner (1989). Furthermore, for any ℓ and $\mathbf{W}$, $K^{\ell}_{\mathbf{W}}$ is a positive, symmetrical function since it is a radial basis function (Skhölkopf & Smola 2002).

It can be proved from (1) that $d^2_{\mathbf{W}}(I^{\ell}_X, I^{\ell}_Y) = w_{11}(\Delta\overline{X})^2 + 2\ell w_{12}\Delta\overline{X}\Delta S + \ell^2 w_{22}(\Delta S)^2$, where $\Delta\overline{X} = \overline{X} - \overline{Y}$, $\Delta S = S_X - S_Y$, and the weight-matrix is $\mathbf{W} = \{w_{ij}\}^2_{i,j=1}$ with $w_{ij} = w_{ji}$ when $i \neq j$.

Thus, the $K^{\ell}_{\mathbf{W}}$ similarity takes into account the following characteristics: i) The position of the whole data set on the real line given by the mean; ii) The spread of the data set around its mean given by the standard deviation multiplied by a parameter $\ell > 1$; iii) The weighted importance of the mean and the standard deviation of each data set, given in the weight-matrix $\mathbf{W}$; and iv) A factor which quantifies, from the number of outlying values, the goodness of fit of the associated intervals.

Example 1. For the data sets of the Appendix, $\ell = 2$ and $\mathbf{W} = \mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, the similarity $K^{\ell=2}_{\mathbf{I}}$ is given in Table 4. It can be seen that the similarities obtained are consistent with the distributions generating the data sets.

After having experimented with different choices of ℓ and $\mathbf{W}$, it is observed that the numerical results differ slightly but the conclusions on their similarities remain the same. ‖

4. A Second Similarity

When the size of the data set is large, consideration of only the number of outlying values and the mean and the standard deviation is grossly insufficient to obtain meaningful results. Furthermore, it is clear that these features are not likely to be very helpful outside a normal distribution family (the mean and variance are highly sensitive to heavy tails and outliers, and are unlikely to provide good measures of location, scale or goodness-of-fit in their presence). Hence more characteristics which summarize the information of each data set must be taken into account.

In this framework, the percentiles of the data set are used. Let

$$Q_X = \{p_{1X}, \cdots, p_{qX}\}$$

be a set of q percentiles of a data set X with $p_{iX} \leq p_{(i+1)X}$ and $q \geq 2$. Hence, $q-1$ intervals, denoted by I_{iX}, are considered as follows: $I_{iX} = (p_{iX}, p_{(i+1)X})$, for $i = 1, \ldots, q-1$.

Example 2. Given the DS_1 data set, an example of the Q_{DS_1} set is given by

$$Q_{DS_1} = \{-0.8926,\ -0.0099,\ 0.7376,\ 1.6571,\ 3.5146\}$$

where these values are the percentiles 2.5, 25, 50, 75, and 97.5, respectively, and $q = 5$.

Definition 4. Given a weight-matrix $\mathbf{W}$ and two sets of q percentiles, Q_X and Q_Y, of the data sets X and Y, respectively, a similarity between X and Y, denoted by $K^Q_{\mathbf{W}}(X, Y)$, is defined as follows:

$$K^Q_{\mathbf{W}}(X, Y) = \frac{1}{1 + \frac{1}{q-1}\sum_{i=1}^{q-1} d_{\mathbf{W}}(I_{iX}, I_{iY})} \tag{3}$$

The $K^Q_{\mathbf{W}}$ similarity has the following properties: i) This function is positive and symmetrical; ii) If $X = Y$ then $K^Q_{\mathbf{W}}(X, Y) = 1$; iii) The similarity is low if the percentiles are far from each other; and iv) It is a radial basis function.

In Table 1, several examples of $d_{\mathbf{W}}(I_{iX}, I_{iY})$ can be seen whereby the symmetrical weight-matrix $\mathbf{W} = \{w_{ij}\}_{i,j=1}^2$ is varied. In cases 1 and 2, $\mathbf{W}$ is a non-regular matrix ($det(\mathbf{W}) = 0$) and therefore this situation is inadequate. In case 3, $\mathbf{W} = \mathbf{I}$ is the identity matrix, and case 4 provides a straightforward weight-matrix which presents the cross product between the percentiles.

Table 1: Distance between intervals for different weight-matrices $\mathbf{W}$.

Case	w_{11}	w_{12}	w_{22}	$d^2_{\mathbf{W}}(I_{iX}, I_{iY})$
1	1	1	1	$(p_{(i+1)X} - p_{(i+1)Y})^2$
2	1	-1	1	$(p_{iX} - p_{iY})^2$
3	1	0	1	$\frac{1}{2}((p_{(i+1)X} - p_{(i+1)Y})^2 + (p_{iX} - p_{iY})^2)$
4	$\frac{3}{4}$	0	$\frac{1}{4}$	$\frac{1}{4}((p_{(i+1)X} - p_{(i+1)Y})^2 + (p_{iX} - p_{iY})^2 + ..$ $... + (p_{(i+1)X} - p_{(i+1)Y})(p_{iX} - p_{iY}))$

On the other hand, there are many different ways to choose the Q_X set for a fixed data set X; in this paper the discretization process[2] based on equal-frequency intervals (Chiu, Wong & Cheung 1991) is used. Furthermore, in order to obtain a specific value of q, there are several selections based on experience such as Sturges' formula, $q_1 = Int\left[\frac{3}{2} + \frac{Log(n)}{Log(2)}\right]$ and $q_2 = Int[\sqrt{n}]$, where the operator $Int[\cdot]$ is the integer part and n is the size of the data set. Henceforth, $q \equiv q_1$ is considered with $n = \max\{\#X, \#Y\}$ and the set of percentiles Q is obtained such that in each interval $I_{i\cdot}$ there is the same quantity of items of data.

In the following section, an example is presented to show how these similarities could be used.

5. The Glass Identification

The Glass Identification is obtained from the UC Irvine Machine Learning Repository (Bache & Lichman 2013). This database is often used to study the

[2]A discretization process converts continuous attributes into discrete attributes by yielding intervals in which the attribute value can reside instead of being represented as singleton values.

performance between different classifiers. Its main properties are: 214 instances, 9 continuous attributes and 1 attribute with 6 classes (labels). The number of instances in each class is 70, 76, 17, 13, 9 and 29, respectively.

Suppose that a preliminary analysis of this data set is desired before applying a classifier. Firstly, $K^Q_{\mathbf{W}}$ similarities between continuous attributes are given in Table 2 for $\mathbf{W} = \mathbf{Id}$.

TABLE 2: $K^Q_{\mathbf{W}}$ similarities between continuous attributes of the Glass data set.

Attr.	2	3	4	5	6	7	8	9
1	0.0777	0.3365	**0.7733**	0.0139	0.4778	0.1209	0.4000	0.4034
2	1	0.0862	0.0771	0.0166	0.0717	0.1780	0.0697	0.0697
3		1	0.3450	0.0141	0.2802	0.1424	0.2523	0.2528
4			1	0.0138	0.5008	0.1195	0.4182	0.4194
5				1	0.0136	0.0154	0.0136	0.0136
6					1	0.1068	**0.6871**	**0.7089**
7						1	0.1026	0.1026
8							1	**0.9153**

It is observed that attributes 1 and 4 are very similar to each other; and attributes 6, 8 and 9 are also very similar, particularly attributes 8 and 9. Hence, it may be a good idea to eliminate some attributes before the implementation of the classifier, for instance attributes 4, 6 and 8.

Let us study the attributes to determine similarities between the same attributes but with different labels. Hence, if the similarity obtained is low, then the classification is straightforward.

The number of instances with label 1 is 70, and with label 2 this is 76, and $K^Q_{\mathbf{W}}$ similarities between the nine attributes are given in Table 3. It can be seen that these values are very high, which indicates that the discrimination between these two labels is not easy.

On the other hand, the number of instances with label 3 is 17, and with label 4 this is 13, and $K^\ell_{\mathbf{W}}$ similarities between the nine attributes are given in Table 3 for $\ell = 2$. Hence, attributes 3 and 7 are the best in order to separate labels 3 and 4. However the main problem with respect to labels 3 and 4 is that there are very few instances.

TABLE 3: $K^\ell_{\mathbf{W}}$ and $K^Q_{\mathbf{W}}$ similarities between different labels of the Glass data set.

Labels	Attr.	1	2	3	4	5	6	7	8	9
1 – 2	$K^Q_{\mathbf{W}}$	0.9984	0.8991	0.6175	0.8123	0.8710	0.9088	0.6290	1	0.9746
3 – 4	$K^\ell_{\mathbf{W}}$	0.8333	0.8844	0	0.7733	0.7488	0.7991	0.4898	0.8807	0.8986

The main conclusion in this brief preliminary analysis is that the classes of the Glass Identification Database are difficult to separate based only on individual features for the given instances. A good classifier is therefore necessary in order to obtain acceptable accuracy for this classification problem.

An experiment[3] is carried out to show that the conclusions of this brief analysis are correct. Thus, the algorithm considered is the standard 1-v-r SVM formulation (Vapnik 1998), by following the recommendation given in Salazar, Vélez & Salazar (2012), and its performance, (in the form of accuracy rate), has been evaluated using the Gaussian kernel, $k(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ where two hyperparameters must be set: the regularization term C and the width of the kernel σ. This space is explored on a two-dimensional grid with the following values: $C = \{2^0, 2^1, 2^2, 2^3, 2^4, 2^5\}$ and $\sigma^2 = \{2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1\}$. The criterion used to estimate the generalized accuracy is a ten-fold cross-validation on the whole training data. This procedure is repeated 10 times in order to ensure good statistical behaviour. The optimization algorithm used is the exact quadratic program-solver provided by Matlab software.

The best cross-validation mean rate among the several pairs (C, σ^2) is obtained for $C = 1$ and $\sigma^2 = 1$ with 70.95% accuracy rate when all attributes are used and, when attributes 4, 6 and 8 are eliminated, then the best cross-validation mean rate is obtained for $C = 16$ and $\sigma^2 = 1$ with 68.38% accuracy rate. This experiment indicates that the Glass Identification Database is difficult to separate and that the elimination of attributes 4, 6 and 8 only slightly modifies the accuracy rates.

In the following section, a new hypothesis test is designed and is compared with other similar hypothesis tests.

6. Hypothesis Testing

Definition 5. Let $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ be two data sets. Two further data sets X^c and Y^c, called the quasi-typified data sets of X and Y, respectively, are defined as follows:

$$x_i^c = \frac{S_Y}{S_X^2}\left(x_i - \overline{Z}\right), \qquad y_i^c = \frac{S_X}{S_Y^2}\left(y_i - \overline{Z}\right),$$

where $Z = \{x_1, \ldots, x_n, y_1, \ldots\ y_m\}$. This process is called quasi-typification.

It is straightforward to prove that $\overline{X^c} = mS_Y(\overline{X} - \overline{Y})/(S_X^2(n+m))$, $\overline{Y^c} = nS_X(\overline{Y} - \overline{X})/(S_Y^2(n+m))$, $S_{X^c} = S_Y/S_X$, and $S_{Y^c} = S_X/S_Y$. Therefore, if $\overline{X} = \overline{Y}$ and $S_X = S_Y$, then $\overline{X^c} = \overline{Y^c} = 0$ and $S_{X^c} = S_{Y^c} = 1$.

From Definition 5, a third similarity measure between data sets is given as follows:

Definition 6. Let X and Y be two data sets. A measure of similarity between these sets is defined as: $KC_{\mathbf{W}}^Q(X,Y) = K_{\mathbf{W}}^Q(X^c, Y^c)$ provided that X^c and Y^c are the quasi-typified data sets of X and Y, $\mathbf{W}$ is a weight-matrix, and the sets of q percentiles are Q_X and Q_Y of the data sets X and Y, respectively.

Example 3. $KC_{\mathbf{I}}^Q$ similarities between the data sets DS_1, DS_2, DS_3 and DS_4 are given in Table 4. In Figure 2, each subplot depicts the boxplot of data DS_i,

[3]Most results have been obtained following the experimental framework proposed by Hsu & Lin (2002) and continued in Anguita, Ridella & Sterpi (2004).

DS_j, T_i and T_j where the T_i's are the quasi-typified data sets of DS_i and DS_j for $i, j = 1, 2, 3, 4$ and $i \neq j$.

TABLE 4: $K^{\ell=2}_{\mathbf{Id}}$, $K^{Q}_{\mathbf{Id}}$ and $KC^{Q}_{\mathbf{Id}}$ similarities between the data sets in the Appendix.

$K^{\ell=2}_{\mathbf{Id}}$	DS_2	DS_3	DS_4
DS_1	0.928	0.880	0.862
DS_2	——	0.862	0.863
DS_3	——	——	0.781

$K^{Q}_{\mathbf{Id}}$	DS_2	DS_3	DS_4
DS_1	0.728	0.646	0.557
DS_2	——	0.595	0.623
DS_3	——	——	0.441

$KC^{Q}_{\mathbf{Id}}$	DS_2	DS_3	DS_4
DS_1	0.897	0.574	0.600
DS_2	——	0.495	0.592
DS_3	——	——	0.347

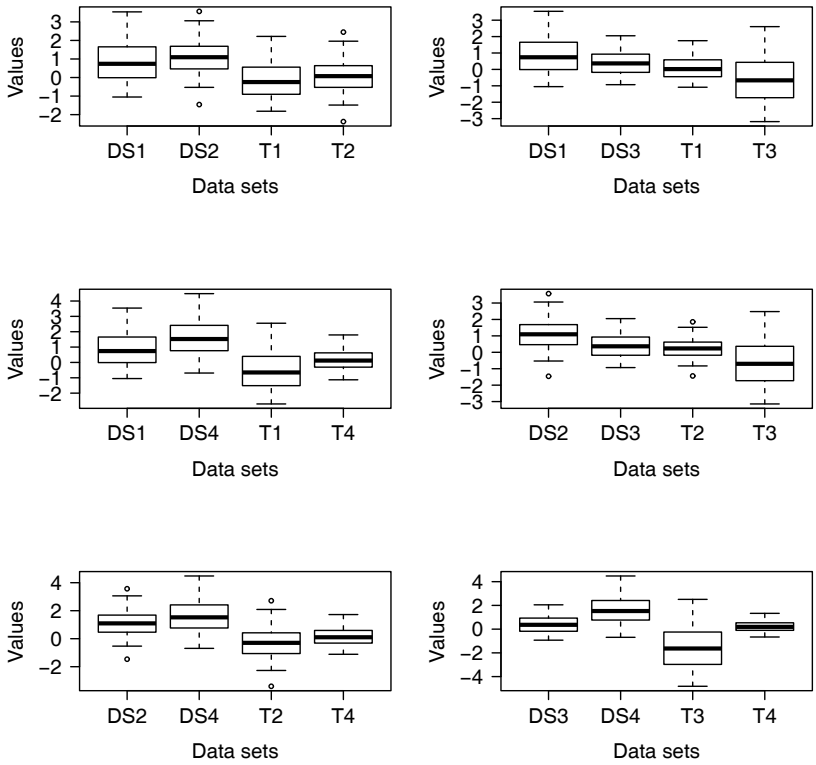


FIGURE 2: Boxplots of each pair of data sets of the Appendix before (DS_i data sets) and after (T_i data sets) applying the quasi-typification.

It is worth noting that all three similarities verify that the similarity between DS_1 and DS_2 is the highest and the similarity between DS_3 and DS_4 is the lowest similarity. Thus, the similarities obtained are consistent with the distribution that generates the data sets. ‖

Several percentiles are obtained from $KC_{\mathbf{I}}^{Q}$ similarities of a simulated distribution between random samples of size 100 from two $N(0,1)$ distributions. The results are shown in Table 5. It is important to point out that the thresholds have been simulated $1,000,000$ times and it is observed that the sensitivity of the thresholds is very low (less than 10^{-5} units). Hence, it is now possible to use these

TABLE 5: Percentiles of the simulated distribution $KC_{\mathbf{I}}^{Q}$ between two $N(0,1)$ distributions for $n = 100$.

α	0.001	0.01	0.025	0.05	0.10
$P(100,\alpha)$	0.60110	0.65353	0.67917	**0.70166**	0.72802

percentiles to construct a hypothesis test.

Definition 7. Let $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_m\}$ be two random samples from populations $\mathcal{F}$ and $\mathcal{F}'$. Let a hypothesis test be $H_0 : \mathcal{F}' = \mathcal{F}$ versus $H_1 : \mathcal{F}' \neq \mathcal{F}$. Let $RC = \{(X,Y) : KC(X,Y) < P(n^*,\alpha)\}$ be the critical region of size α where $P(n^*,\alpha)$ is the percentile α of the simulated distribution $KC_{\mathbf{I}}^{Q}$ between two $N(0,1)$ distributions for $n^* = \min(n,m)$. Henceforth, this test is denoted as the GA-test.

***Note* 1.** It is worth noting that this test is valid for normal or similar populations. If another type of population is given, then the corresponding percentiles should be calculated.

6.1. Comparison with a Parametric Test

Let the following test be: $H_0 : \mathcal{F}' = \mathcal{F}$ versus $H_1 : \mathcal{F}' \neq \mathcal{F}$, where $\mathcal{F} = N(\mu_1, \sigma_1)$, $\mathcal{F}' = N(\mu_2, \sigma_2)$ and where μ_1, μ_2, σ_1 and σ_2 are unknown parameters. In this case, the null hypothesis states that the two normal populations have both identical means and variances.

Let $X = \{X_1, \ldots, X_{100}\}$ and $Y = \{Y_1, \ldots, Y_{100}\}$ be two random samples from $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$ distributions. A classic test (C-test) is considered, which is a union of two tests. Firstly, a test is performed to determine whether two samples from a normal distribution have the same mean when the variances are unknown but assumed equal. The critical region of size 0.025 is

$$RC_1 = \left\{ (X,Y) : \left|\overline{X} - \overline{Y}\right| > 2.2586 \sqrt{(S_1^2 + S_2^2)/99} \right\}$$

where 2.258 is the percentile 0.9875 of Student's t distribution with 198 degrees of freedom. Another test is also performed to determine whether two samples from a normal distribution have the same variance. The critical region of size 0.025 is

$$RC_2 = \left\{ (X,Y) : S_1^2/S_2^2 < 0.6353,\ S_1^2/S_2^2 > 1.5740 \right\}$$

where 0.6353 and 1.5740 are the percentiles 0.0125 and 0.9875 of Snedecor's F distribution, both with 99 degrees of freedom.

In this framework, a comparison is made between the classic test for Normal populations whose critical region of size 0.0494 ($= 1-0.975^2$) is $RC = RC_1 \cap RC_2$, versus the GA-test whose critical region of size 0.05 is $\{(X,Y) : KC(X,Y) < 0.70166\}$ (see Table 5). For this comparison, it is considered that one population is $N(20,4)$ and the other population is $N(\mu,\sigma)$, and the hypothesis test is carried out for 100,000 simulations for each value $\mu = 18, 19, 20, 21, 22$ and $\sigma = 3, 3.5, 4, 4.5, 5$. The results of the experiment are given in Table 6, where the percentage of acceptance of the null hypothesis is shown for the two tests. The best result for each value of the parameters is printed in bold, that is, the minimum of the two values except for the case $\sigma = 4$ and $\mu = 20$ in which the null hypothesis is true and then the maximum of the two values is printed in bold.

The first noteworthy conclusion is that there are no major differences between the two methods and therefore the results of the GA-test are good. As expected, the results are almost symmetrical for equidistant values from the true mean and variance. When only one of the two parameters is the actual value, then the classic test behaves better in general, possibly due to the fact that the classic test is sequential and the other is simultaneous. However, when both parameters only slightly differ from the actual values, then the GA-test performs better. The same holds true for values of the mean that differ from the actual value and for great differences in the variance.

Table 6: Acceptance percentage of the null hypothesis when comparing the $N(20,4)$ and the $N(\mu,\sigma)$ distributions for different values of the mean and of the standard deviation using the classic test (C) and the GA-test. The desired level of significance is 5% and the best result in each case is printed in bold.

σ	3		3.5		4		4.5		5	
μ	GA	C	GA	C	GA	C	GA	C	GA	C
18	**00.88**	01.03	06.56	**05.47**	13.55	**09.72**	14.44	**12.26**	**08.94**	09.56
19	**14.06**	16.06	**51.94**	52.58	70.32	**66.97**	**61.56**	61.77	**35.94**	38.17
20	27.79	**26.92**	**79.85**	79.93	**94.97**	94.84	83.77	**83.24**	51.02	**49.18**
21	**13.73**	15.71	**52.33**	52.86	70.67	**67.32**	**61.35**	61.53	**35.99**	38.16
22	**00.89**	01.06	06.58	**05.43**	13.95	**10.11**	14.42	**12.34**	**08.80**	09.53

6.2. Comparison with a Non-Parametric Test

In this section, the GA-test is used with non-normal distributions than remains similar to a Normal distribution. At this point, the interest lies in testing $H_0 : \mathcal{F}' = \mathcal{F}$ versus $H_1 : \mathcal{F}' \neq \mathcal{F}$ for a number of populations $\mathcal{F}$ and $\mathcal{F}'$. The GA-test is compared against the Kolmogorov-Smirnov test. In both cases the desired level of significance is 0.05, the hypothesis test is carried out for 10,000 simulations where the populations are $Bi(100, 0.2)$ (Binomial), $Po(20)$ (Poisson) and $N(20,4)$ (Normal). Figure 3 shows that these distributions are very similar and the size of random samples is 100.

The results of the experiment are given in Table 7 in the form of percentage of acceptance of the null hypothesis. Again, the best result in each case is printed in bold, that is, the minimum of the two values when the null hypothesis is false

(values outside the diagonal) and the maximum of the two values when the null hypothesis is true (values in the diagonal). It can be seen that the GA-test can differentiate between the Poisson distribution and the other two better than can the Kolmogorov-Smirnov test. Nevertheless, the Kolmogorov-Smirnov test behaves better than the GA-test in Binomial and Poisson populations under the null hypothesis (the opposite is true for the normal distribution) and when distinguishing between the normal and the binomial distributions.

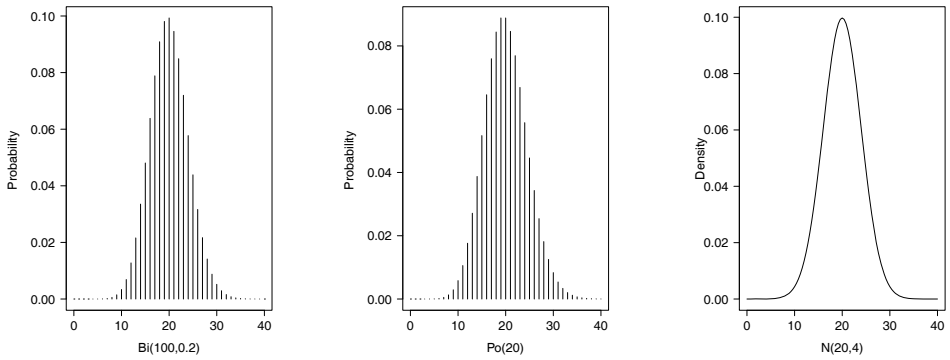


FIGURE 3: Graphical representation of the probability mass function of the $Bi(100, 0.2)$ distribution and the $Po(20)$ distribution, and probability density function of $N(20, 4)$ distribution.

TABLE 7: Acceptance percentage of the null hypothesis in the comparison between the Kolmogorov-Smirnov test and the GA-test for various populations. The desired level of significance is 5% and the best result in each case is printed in bold.

	Bi(100,0.2)		Po(20)		N(20,4)	
	GA	K-S	GA	K-S	GA	K-S
Bi(100,0.2)	94.36	**97.56**	**83.46**	96.75	94.94	**86.44**
Po(20)	——	——	94.76	**97.59**	**84.40**	84.95
N(20,4)	——	——	——	——	**95.50**	95.00

A final comparison is carried out with Student's t distributions with several degrees of freedom since these distributions are similar to a standard Normal distribution. The desired level of significance is 0.05, the size of random samples is 100 and the hypothesis test is carried out for 10,000 simulations. The results of the experiment are given in Table 8. Again, the best result in each case is printed in bold, that is, the minimum of the two values when the null hypothesis is false (values outside the diagonal) and the maximum of the two values when the null hypothesis is true (values in the diagonal). It is important to point that the GA-test tends to provide smaller values and therefore tends to accept the null hypothesis less frequently than does the classic test (the classic test therefore tends to be more conservative). As a consequence, the GA-test has a better behaviour when the null hypothesis is false (values outside the diagonal), by differentiating

between Student's t distributions with different degrees of freedom better than does the Kolmogorov-Smirnov test, and a worse behaviour (but not much worse) when the null hypothesis is true (values of the diagonal), that is, the Kolmogorov-Smirnov test behaves slightly better under the null hypothesis.

TABLE 8: Acceptance percentage of the null hypothesis in the comparison between the Kolmogorov-Smirnov test and the GA-test for Student's t distributions. The desired level of significance is 5% and the best result in each case is printed in bold.

	$t(10)$		$t(20)$		$t(30)$		$t(40)$		$t(50)$	
	GA	K-S	GA	K-S	GA	K-S	GA	K-S	GA	K-S
$t(10)$	92.71	**94.56**	**91.03**	94.53	**89.47**	94.55	**88.56**	94.63	**87.78**	94.37
$t(20)$	—	—	94.22	**94.53**	**94.13**	94.85	**93.59**	94.60	**93.97**	94.89
$t(30)$	—	—	—	—	94.53	94.53	**94.33**	94.57	**94.49**	94.84
$t(40)$	—	—	—	—	—	—	94.66	**94.77**	**94.47**	94.57
$t(50)$			—	—	—	—	—	—	**94.66**	94.42

7. Conclusions and Future Work

Several similarity measures between one-dimensional data sets have been developed which can be employed to compare data sets, and a new hypothesis test has been designed. Two comparisons of this test with other classic tests have been made under the null hypothesis that two populations are identical. The main conclusion is that the new test performs reasonably well in comparison with the classic tests considered, and, in certain circumstances, performs even better than said classic tests.

With the distance developed in this paper, various classifications of a data set can be carried out, either by applying the neural network technique, SVM, or via other procedures available.

Although there are other approaches to the choice of the set Q of the percentiles for the $K_{\mathbf{W}}^{Q}$ function from a data set X, such as for example the equal-width interval (Chiu et al. 1991), k-mean clustering (Hartigan 1975), cumulative roots of frequency (González & Gavilan 2001), Ameva (González-Abril, Cuberos, Velasco & Ortega 2009), and the maximum entropy marginal approach (Wong & Chiu 1987), these have not been considered in this paper and will be studied in future papers.

Only the one-dimensional setting is considered in this paper; the possible correlations that can exist between features of multi-dimensional data sets lie outside the scope of this paper and will constitute the focus of study in future work.

Another potential line of research involves the improvement of the design of our hypothesis-testing procedures by using these similarity measures, and the execution of comparisons with other existing methods. For example, the chi-squared test on quantiled bins, or the Wald-Wolfowitz runs test can be tested under the null hypothesis that the two samples come from identical distributions.

Acknowledgements

This research has been partly supported by the Andalusian Regional Ministry of Economy project Simon (TIC-8052).

[Recibido: julio de 2013 — Aceptado: enero de 2014]

References

Anguita, D., Ridella, S. & Sterpi, D. (2004), A new method for multiclass support vector machines, *in* 'Proceedings of the IEEE IJCNN2004', Budapest, Hungary.

Bach, F. R. & Jordan, M. I. (2003), 'Kernel independent component analysis', *Journal of Machine Learning Research* **3**, 1–48.

Bache, K. & Lichman, M. (2013), 'UCI machine learning repository', http://archive.ics.uci.edu/ml, University of California, Irvine, School of Information and Computer Sciences.

Burrell, Q. L. (2005), 'Measuring similarity of concentration between different informetric distributions: Two new approaches', *Journal of the American Society for Information Science and Technology* **56**(7), 704–714.

Chiu, D., Wong, A. & Cheung, B. (1991), Information discovery through hierarchical maximum entropy discretization and synthesis, *in* G. Piatetsky-Shapiro & W. J. Frawley, eds, 'Knowledge Discovery in Databases', MIT Press, pp. 125–140.

Cristianini, N. & Shawe-Taylor, J. (2000), *An introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University press.

González-Abril, L., Cuberos, F. J., Velasco, F. & Ortega, J. A. (2009), 'Ameva: An autonomous discretization algorithm', *Expert Systems with Applications* **36**(3), 5327 – 5332.

González-Abril, L., Velasco, F., Gavilán, J. & Sánchez-Reyes, L. (2010), 'The similarity between the square of the coeficient of variation and the Gini index of a general random variable', *Revista de métodos cuantitativos para la economía y la empresa* **10**, 5–18.

González, L. & Gavilan, J. M. (2001), Una metodología para la construcción de histogramas. Aplicación a los ingresos de los hogares andaluces, *in* 'XIV Reunión ASEPELT-Spain'.

González, L., Velasco, F., Angulo, C., Ortega, J. & Ruiz, F. (2004), 'Sobre núcleos, distancias y similitudes entre intervalos', *Inteligencia Artificial* **8**(23), 113–119.

González, L., Velasco, F. & Gasca, R. (2005), 'A study of the similarities between topics', *Computational Statistics* **20**(3), 465–479.

Hartigan, J. (1975), *Clustering Algorithms*, Wiley, New York.

Hsu, C.-W. & Lin, C.-J. (2002), 'A comparison of methods for multiclass support vector machine', *IEEE Transactions on Neural Networks* **13**(2), 415–425.

Lee, J., Kim, M. & Lee, Y. (1993), 'Information retrieval based on conceptual distance in is-a hierarchies', *Journal of Documentation* **49**(2), 188–207.

Lin, D. (1998), An information-theoretic definition of similarity, *in* 'Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998)', pp. 296–304.

Nielsen, J., Ghugre, N. & Panigrahy, A. (2004), 'Affine and polynomial mutual information coregistration for artifact elimination in diffusion tensor imaging of newborns', *Magnetic Resonance Imaging* **22**, 1319–1323.

Parthasarathy, S. & Ogihara, M. (2000), 'Exploiting dataset similarity for distributed mining', http://ipdps.eece.unm.edu/2000/datamine/18000400.pdf.

Rada, R., Mili, H., Bicknell, E. & Blettner, M. (1989), 'Development and application of a metric on semantic nets', *IEEE Transaction on Systems, Man, and Cybernetics* **19**(1), 17–30.

Salazar, D. A., Vélez, J. I. & Salazar, J. C. (2012), 'Comparison between SVM and logistic regression: Which one is better to discriminate?', *Revista Colombiana de Estadística* **35, 2**, 223–237.

Sheridan, R., Feuston, B., Maiorov, V. & Kearsley, S. (2004), 'Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR', *Journal of Chemical Information and Modeling* **44**, 1912–1928.

Skhölkopf, B. & Smola, A. J. (2002), *Learning with Kernel*, MIT Press.

Vapnik, V. (1998), *Statistical Learning Theory*, John Wiley & Sons, Inc.

Wong, A. & Chiu, D. (1987), 'Synthesizing statistical knowledge from incomplete mixed-mode data', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **9**(6), 796–805.

Appendix. Data sets of Section 2

The DS_1 and DS_2 data sets are taken from a N(1,1) distribution, the DS_3 data set from a N(0.5,1) distribution, and the DS_4 data set from a N(1.5, 1.25) distribution.

$DS_1 = \{1.47, 0.01, 1.29, -0.27, -0.23, 0.54, -0.20, 3.54, 0.59, -0.03, 2.41, -0.08, 1.48, 1.02, 0.71, 3.40, 0.67, 0.74, 1.64, 2.41, 1.67, 0.74, -0.10, 1.76, 1.82, -1.05, 0.54, 1.20\}$

$DS_2 = \{-0.29, 0.02, 0.84, 1.66, 1.04, 0.69, 2.04, 1.21, 1.71, 1.75, 1.64, 1.10, -1.46, 1.25, -0.10, 1.74, 3.06, -0.53, 0.84, 1.09, 1.26, -0.39, 0.88, 2.15, 1.59, 0.56, 0.37, 3.57\}$

$DS_3 = \{0.22, 0.87, -0.11, 0.29, -0.93, -0.25, 2.05, -0.53, -0.51, 0.80, 0.65, 0.99, 1.28, 0.85, 0.00, -0.28, 0.55, 0.27, -0.68, 1.08, 1.20, 0.44, 0.20, 0.66, 0.29, -0.46, 1.02, 1.99\}$

$DS_4 = \{1.81, -0.41, 1.25, 3.12, 1.91, 1.99, 1.75, 0.93, -0.39, 3.68, -0.69, 1.57, 1.48, 3.59, 0.60, 2.84, 0.37, 1.26, 1.94, -0.19, 1.77, 3.20, 1.11, 4.24, 0.16, 4.48, 0.98, 1.34\}$