Chemometrics and Intelligent Laboratory Systems 157 (2016) 50–57

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A new concept of higher-order similarity and the role of distance/similarity measures in local classification methods

Roberto Todeschini *, Davide Ballabio, Viviana Consonni, Francesca Grisoni

Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano-Bicocca, Piazza della Scienza, 1, 20126 Milan, Italy

ARTICLE INFO

Article history:
Received 2 May 2016
Received in revised form 16 June 2016
Accepted 18 June 2016
Available online 23 June 2016

Keywords:
Classification
Similarity measures
Distance measures
Meta-distances
KNN
N3
BNN

ABSTRACT

In this paper, a new concept of similarity is introduced with the aim of detecting higher-order similarities among objects, and meta-distances and meta-similarities are derived from it. A total of 100 meta-distances were obtained from a set of ten classical distances and were compared, in terms of classification performances, against classical distance measures. Classification methods based on local similarity analysis and several benchmark datasets were used. In several cases, the non-error rate (*NER*) of classifiers based on the new meta-distances significantly increased with respect to that of the classical Euclidean distance.

1. Introduction

The concept of similarity represents how close two entities are according to the analogy of their features. Despite its being very intuitive, a formal definition of similarity (and of its counterpart, the distance) is fundamental for the mathematical treatment of analogous/different entities. Starting from Euclid and later with the rise of the twentieth century mathematics, many measures of similarity and distance have been proposed [1–3], so that every definition enables capturing analogy/difference under different points of view.

Because of their ability to quantitatively grasp the analogy between entities, distance/similarity metrics have been playing a crucial role in many fields, such as Quantitative Structure-Activity Relationship (QSAR) [5,6,29], food authentication [7–9], outlier detection [10–13] and drug discovery [14–16]. This underscores the potential of introducing and studying new metrics.

This work, in particular, leverages well-established distance metrics into a novel definition of meta-distance, able to capture higher orders of diversity. The proposed measure melds first-order diversity measures, which detect the basic information about diversity relationships, with a second-order measure, which encodes the respective (dis)similarity of two entities with all the remaining ones. The new measure is able to encode surprisingly different aspects with respect to the traditional metrics.

In particular, ten well-known distances were used to generate 100 meta-distances, whose potential was tested on 30 benchmark datasets of different nature. The effect of the newly introduced similarity/distance concept was tested on three local classifiers (KNN [17], N3 [18] and BNN [18]), which use a certain number of similar objects (neighbors) to predict a new object's class as a majority vote of the neighbors. As they are local methods, the dissimilarity measure used to select the nearest neighbors plays a fundamental role in determining the classification outcomes. Moreover, their differences in how the neighborhood contributes to the final predictions were used to obtain additional insights into the role of the distance measures.

After presenting the theory, this work investigates the role of meta-distances on all the chosen datasets and methods, using the ten classical distances as the benchmark.

2. Theory

A meta-distance is obtained by combining traditional distance measures (here named primary distances) with a newly proposed adjunct dissimilarity, which acts as a smoothing factor of the primary distance. The rationale and the mathematical definitions of these concepts are here outlined: after summarizing the concepts regarding the well-established (primary) distances, the meta-distances are introduced focusing on the role played by the adjunct dissimilarity factor.

* Corresponding author.
E-mail address: roberto.todeschini@unimib.it (R. Todeschini).
URL: http://michem.disat.unimib.it (R. Todeschini).

http://dx.doi.org/10.1016/j.chemolab.2016.06.013
0169-7439/© 2016 Elsevier B.V. All rights reserved.

Table 1
List of the ten primary distances. D_{xy} is the distance measure between objects x and y according to their values of p variables.

ID	Name	Acronym	Distance
1	Average Euclidean	*Euc*	$D_{xy}^{\mathrm{Euc}} = \sqrt{\frac{\sum_{j=1}^{p} (x_j - y_j)^2}{p}}$
2	Average Canberra	*Can*	$D_{xy}^{\mathrm{Can}} = \frac{1}{p} \cdot \sum_{j=1}^{p} \frac{\lvert x_j - y_j \rvert}{\lvert x_j + y_j \rvert}$
3	Lance-Williams	*LW*	$D_{xy}^{\mathrm{LW}} = \frac{\sum_{j=1}^{p} \lvert x_j - y_j \rvert}{\sum_{j=1}^{p} \lvert x_j + y_j \rvert}$
4	Average Manhattan	*Man*	$D_{xy}^{\mathrm{Man}} = \frac{1}{p} \cdot \sum_{j=1}^{p} \lvert x_j - y_j \rvert$
5	Lagrange	*Lag*	$D_{xy}^{\mathrm{Lag}} = \mathrm{max}_j \lvert x_j - y_j \rvert$
6	Average Clark	*Cla*	$D_{xy}^{\mathrm{Cla}} = \frac{1}{p} \cdot \sqrt{\sum_{j=1}^{p} (\frac{\lvert x_j - y_j \rvert}{x_j + y_j})^2}$
7	Average Matusita	*Mat*	$D_{xy}^{\mathrm{Mat}} = \sqrt{\frac{\sum_{j=1}^{p} (\sqrt{x_j} - \sqrt{y_j})^2}{p}}$
8	Soergel	*Soe*	$D_{xy}^{\mathrm{Soe}} = \frac{\sum_{j=1}^{p} \lvert x_j - y_j \rvert}{\sum_{j=1}^{p} \max(x_j, y_j)}$
9	Average Wave-Edges	*WE*	$D_{xy}^{\mathrm{WE}} = \frac{1}{p} \cdot \sum_{j=1}^{p} (1 - \frac{\min(x, y)}{\max(x, y)})$
10	Jaccard-Tanimoto	*JT*	$D_{xy}^{\mathrm{JT}} = 1 - \frac{\sum_{j=1}^{p} x_j \cdot y_j}{\sum_{j=1}^{p} x_j^2 + \sum_{j=1}^{p} y_j^2 - \sum_{j=1}^{p} x_j \cdot y_j}$

2.1. Primary distance measures

A distance is a numerical description of how far apart entities are. In particular, given two objects x and y described by a set of p variables, their distance (D_{xy}) is calculated starting from the differences in their variable values: the higher the differences, the more different the two objects. Obviously, these differences can be mathematically quantified in many ways [19], and, in this work, a set of ten classical distance measures was used (Table 1). As they are calculated with different algorithms, all the variables were previously range-scaled between 0 and 1, allowing for a direct comparison.

Since all the distances are in the range [0, 1], the calculation of a corresponding similarity measure (S_{xy}) is as follows:

$$S_{xy} = 1 - D_{xy} \tag{1}$$

where S_{xy} ranges from 0 to 1.

2.2. Meta-distance

Distance functions of Table 1 between two objects measure diversity and, thus, the greater the distance the more different the objects. These functions are first-order diversity measures, that is, they detect the basic information about diversity relationships. However, two objects can be compared also on a relative scale by observing their respective (dis)similarity with all the remaining objects of the set they belong to: the more comparable their similarity with the remaining objects, the more similar the objects.

Let x and y be two objects belonging to a set of n objects, then the meta-distance $D_{xy}^{(M)}$ here proposed is defined as the product between a primary distance D_{xy} and a correction factor α, hereinafter referred to as adjunct dissimilarity, which takes into account higher-order (dis)-similarities:

$$D_{xy}^{(M)} = \alpha \cdot D_{xy} \tag{2}$$

The adjunct dissimilarity α is defined as:

$$\alpha = e^{-2 \cdot P_{xy}} \qquad 0.1353 \le \alpha \le 1 \tag{3}$$

where the term P_{xy}, which quantifies higher-order similarities, is calculated as:

$$P_{xy} = \frac{\sum_{\substack{z \neq x \\ z \neq y}}^{n} \delta(z)}{n-2} \qquad \delta(z) = \begin{cases} 1 & \mathit{if} \quad \frac{1 + \min(D_{xz}, D_{yz})}{1 + \max(D_{xz}, D_{yz})} \geq t \\ 0 & \mathit{otherwise} \end{cases} \tag{4}$$

where D_{xz} and D_{yz} are the primary distances between any z of the set, and x and y, respectively; t is a threshold that defines the range in which D_{xz} and D_{yz} are considered as equivalent. In this work, t was set to a value of 0.97. Note that D_{xz} and D_{yz} can be any chosen primary distance (e.g. from Table 1). The resulting P_{xy} is the proportion of the *n-2* objects sharing similar distances with both x and y. Consequently, the higher this proportion, the more similar x and y to all of the other objects are. In other words, the lower the adjunct dissimilarity, the more similar x and y are from the viewpoint of the other objects. The exponent −2 of Eq. (3) was defined in an empirical way and further studies could be done evaluating the influence of other types of exponent on the primary distance.

The adjunct dissimilarity α acts as a smoothing parameter of the primary distance: in fact, if x and y have a high proportion of common similar z objects, their α will be low and $D_{xy}^{(M)} < D_{xy}$; on the contrary, if no common similar objects are found, α will be equal to one and, thus,