IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 22, 2021, accepted November 18, 2021, date of publication November 23, 2021,
date of current version December 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3130127

Revisiting the Dissimilarity Representation in the Context of Regression

VICENTE GARCÍA[1], (Member, IEEE), J. SALVADOR SÁNCHEZ[2], RAFAEL MARTÍNEZ-PELÁEZ[3], AND LUIS C. MÉNDEZ-GONZÁLEZ[4], (Member, IEEE)

[1]Department of Electrical and Computer Engineering, Universidad Autónoma de Ciudad Juárez, 32310 Ciudad Juárez, México
[2]Department of Computer Languages and Systems, Institute of New Imaging Technologies, Universitat Jaume I, 12071 Castelló de la Plana, Spain
[3]Faculty of Information Technologies, Universidad de La Salle Bajío, 37150 León, México
[4]Department of Industrial Engineering and Manufacturing, Universidad Autónoma de Ciudad Juárez, 32310 Ciudad Juárez, México

Corresponding author: Vicente García (vicente.jimenez@uacj.mx)

This work was supported in part by Universitat Jaume I under Grant UJI-B2018-49.

ABSTRACT In machine learning, a natural way to represent an instance is by using a feature vector. However, several studies have shown that this representation may not accurately characterize an object. For classification problems, the dissimilarity paradigm has been proposed as an alternative to the standard feature-based approach. Encoding each object by pairwise dissimilarities has been demonstrated to improve the data quality because it mitigates some complexities such as class overlap, small disjuncts, and low-sample size. However, its suitability and performance when applied to regression problems have not been fully explored. This study redefines the dissimilarity representation for regression. To this end, we have carried out an extensive experimental evaluation on 34 datasets using two linear regression models. The results show that the dissimilarity approach decreases the error rates of both the traditional linear regression and the linear model with elastic net regularization, and it also reduces the complexity of most regression datasets.

INDEX TERMS Data complexity, dissimilarity representation, linear models, regression.

I. INTRODUCTION

An underlying step in machine learning and pattern recognition is the characterization of objects, where an ideal representation ensures building accurate learning algorithms [1]. Three approaches have emerged to represent a real-world object [2], [3]: the structural or syntactical approach using a symbolic data structure, the statistical approach based on a feature representation, and the class models.

The statistical approach assumes that an object is characterized by an n-dimensional vector $\mathbf{x} = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$, where each x_i is a numeric attribute (feature) whose values are obtained through observation or as samples of the data (e.g., pixels of an image) [3], [4]. However, this representation may not capture the internal structure of some objects that have an intrinsic and detectable organization [5]–[7]. In classification problems, it is often difficult to obtain an appropriate feature-based characterization of objects, leading to a high-dimensional representation with class overlap or also a representation with a mixture of continuous and categorical features [5], [8].

Pękalska and Duin [9] proposed a dissimilarity representation in which objects are characterized by the difference or the dissimilarity to other objects from a representation set. A straightforward method of constructing the new representation is by means of mapping processes that convert a feature vector into a dissimilarity vector using a distance metric. Several studies have demonstrated that this alternative characterization suggests practical advantages over the feature representation, such as: i) it is possible to use a simple linear prediction model [10], ii) it yields a good separability between classes [11], iii) all dimensions in the dissimilarity space are equally relevant [11], and iv) the small disjunct problem is reduced [12].

Dissimilarity representation has been extensively applied to a variety of classification problems. For example, Bruno *et al.* [13] proposed a particular form of dissimilarity space for multimodal information, enabling fast and efficient interactive content-based retrieval of video data. Porro-Muñoz *et al.* [14] concluded that the use of the dissimilarity representation for the classification of chemical spectral data, which is characterized by changes in the shape of the spectra of different classes, outperformed the results achieved on the feature space.

The associate editor coordinating the review of this manuscript and approving it for publication was Lefei Zhang.

Theodorakopoulos *et al.* [15] developed a method for pose-based human action recognition in the dissimilarity space. The problem of corporate bankruptcy prediction was tackled using four linear classifiers designed on the dissimilarity space, showing that their performance was considerably better than that of the models applied onto the feature space [10]. Orozco-Alzate *et al.* [1] investigated the suitability of a dynamic time warping based on the dissimilarity representation for distinguishing among the different seismic volcanic patterns. Classification of time series was carried out by using the dissimilarity representation [16].

Martins *et al.* [17] introduced a framework based on dissimilarity vectors and dynamic classifier selection to identify microscopic images of forest species. A two-stage model that consists of a feature selection algorithm and the dissimilarity-based representation for the classification of microarray gene expression data was proposed by García and Sánchez [18], who reported that the dissimilarity approach appears to be less sensitive to the number of genes than the feature-based representation. Also, the dissimilarity representation was combined with multiple classifier systems for text categorization [19].

To the best of our knowledge, far less research attention has been paid to the applicability of the dissimilarity representation to regression tasks. For instance, Jaramillo-Garzon *et al.* [20] modeled time-frequency representations by means of support vector regression, and the distance between regressions was calculated through dissimilarity measures based on dot products for the classification of phonocardiographic recordings. Silva-Mata *et al.* [21] combined the dissimilarity representation with the classical partial least square regression model for the recognition of substances and their chemical-physical properties in biochemical data. Despite these few works, we argue that the dissimilarity representation has not yet been deeply studied in the framework of regression problems.

This paper offers a large-scale experimental analysis with 34 benchmark regression datasets aiming to compare the performance of two linear models trained using feature and dissimilarity vectors. We intend to shed light on the suitability of the dissimilarity representation by addressing the following questions:

1) How the representation set size affects the predictive performance of regression models?
2) Does the dissimilarity representation reduce the complexity of a regression problem?
3) Do the dissimilarity-based linear regression models perform significantly better than the feature-based ones?

We cope with these issues by evaluating the dissimilarity representation constructed by means of the Euclidean distance and a random selection procedure designed to conform the representation set. To capture the difficulty of a regression problem, we compute a data complexity measure [22] to check whether or not the regression problem is simpler using the dissimilarity strategy than the traditional feature representation.

Henceforth the paper is organized as follows. Section II introduces the basis of the dissimilarity representation, whereas Section III describes the process for adapting the dissimilarity approach to regression problems. Next, Section IV provides the experimental set-up. In Section V, the results are presented and discussed. Finally, Section VI remarks the main conclusions and outlines possible future directions for extending this work.

II. THE DISSIMILARITY REPRESENTATION

The construction of the dissimilarity representation from the feature representation is based on measuring pairwise dissimilarities between an object and a set of prototypes or representative objects of each class $R = \{p_1, \ldots, p_M\}$. This R can be taken as the complete set of objects $T = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ or a subset of T ($R \subseteq T$), or even it can be defined as a set of generated prototypes [23]. The most straightforward method to select M prototypes from T is the random selection, which can be done either by ensuring that R contains prototypes of each class or by a global selection where R may not have examples from all classes. Several works have shown that an appropriate, intelligent selection strategy improves the performance owing to a better transformation of the feature space [24].

For the dissimilarity representation, we need a suitable dissimilarity measure $d(\cdot, \cdot)$ computed or derived from the objects; this dissimilarity measure must be non-negative ($d(\mathbf{x}_i, \mathbf{x}_j) > 0$ if $\mathbf{x}_i$ is distinct from $\mathbf{x}_j$) and obey the reflexivity condition ($d(\mathbf{x}_i, \mathbf{x}_i) = 0$), but it might be non-metric. Common dissimilarity measures include Chi-square, Euclidean distance, Kolmogorov-Smirnov distance, cosine distance, Pearson correlation coefficient, Minkowski distance, and Spearman correlation.

A dissimilarity representation is defined as a data-dependent mapping function $D(\cdot, R)$ from T to the dissimilarity space [9], [11]. This implies that each object $\mathbf{x}_i \in T$ can be represented by an M-dimensional real-valued vector in the dissimilarity space, $D(\mathbf{x}_i, R) = [d(\mathbf{x}_i, p_1), \ldots, d(\mathbf{x}_i, p_M)]$, that is, each dimension corresponds to the dissimilarity computed between $\mathbf{x}_i$ and a prototype $p_j \in R, (j = 1, \ldots M)$. Then, the dissimilarities between all objects in T and the prototypes in R are represented by a matrix $D(T, R)$ of size $N \times M$ [25]:

$$D(T,R) = \begin{bmatrix} d(\mathbf{x}_1, p_1) & d(\mathbf{x}_1, p_2) & \cdots & d(\mathbf{x}_1, p_M) \\ d(\mathbf{x}_2, p_1) & d(\mathbf{x}_2, p_2) & \cdots & d(\mathbf{x}_2, p_M) \\ \vdots & \vdots & \ddots & \vdots \\ d(\mathbf{x}_N, p_1) & d(\mathbf{x}_N, p_2) & \cdots & d(\mathbf{x}_N, p_M) \end{bmatrix}$$

A representation space can now be built from this matrix. The dimensionality of the dissimilarity space is equal to M (the cardinality of R). Thus, each dimension corresponds to the dissimilarities with one of the prototypes in R. Note that the mapping process generates new variables to represent the data, thus changing the meaning of the original attributes.

III. DISSIMILARITY REPRESENTATION FOR REGRESSION

A regression problem involves a pair of measurements $(\mathbf{x}, z)$, where $\mathbf{x}$ is called the independent variable and $z \in \mathbb{R}$ is the dependent variable. The aim of the regression is to find the function $f(.)$ that can predict z from T based on N observations $(\mathbf{x}_i, z_i)$, $i = 1, \ldots, N$.

Similar to the mapping process for converting a feature vector into a dissimilarity vector in classification tasks, in regression problems, the procedure is carried out for the learning and testing phases, as shown in the flowchart of Figure 1.

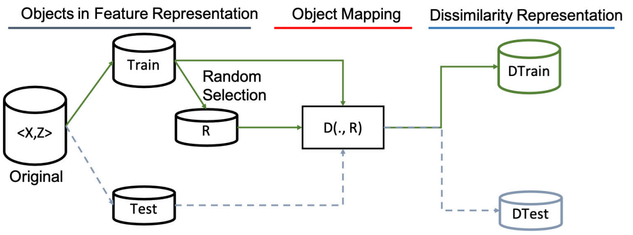


FIGURE 1. The mapping process into a dissimilarity representation. Dotted lines stand for the process to convert a test dataset into a dissimilarity dataset, whereas straight lines correspond to the step for building the dissimilarity regression training set.

The first step in constructing a dissimilarity matrix is to select an R from the training set. Here, this process is performed using a random selection method (Algorithm 1). Therefore, R was constructed by taking M random samples without replacement from T. In classification tasks, several instance selection methods have focused on extracting the most significant samples.

Algorithm 1: Random Selection of the Representation Set

```
Input: Regression training set
        $T = \{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_N, z_N)\}$,
        $K$ Number of samples to select.
Output: Representation set $R \subseteq T$
/* Sampling without replacement        */
$M$ = Generate $K$ random numbers $\in [1, N]$
$R = \{\mathbf{x}_M\}$
```

The dissimilarity matrix is constructed using a dissimilarity measure once R has been selected, as shown in Algorithm 2; we used the Euclidean distance in the experiments. It should be noted that the resulting matrix D is a set that contains the target values taken from T. Remember that this mapping process should be performed in both the training and testing datasets. These dissimilarity sets are passed through the regression model for learning and predicting the independent variable of a new instance $\mathbf{x}'$.

The computational complexity of the proposed algorithm depends on the computational costs associated with the mapping process to construct a dissimilarity matrix D. For the training stage, both the time complexity and space complexity are $\mathcal{O}(N \cdot M)$, where N is the training set size and M is the representation set size. As the testing stage uses R, the time and space complexities of mapping a test example are $\mathcal{O}(M)$.

Algorithm 2: Mapping Process to Construct a Dissimilarity Matrix

```
Input: Regression set $T = \{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_N, z_N)\}$,
       Representation set $R = \{p_1, , \ldots, p_M\}$.
Output: Dissimilarity set $D$
$D = \emptyset$
foreach $\mathbf{x}_i \in T$ do
    $distances = \emptyset$
    foreach $p_j \in R$ do
        $dist_{i,j} = d(\mathbf{x}_i, p_j)$
        $distances = distances \cup \{dist_{i,j}\}$
    $D = D \cup \{distances\}$
$D = \{D, Z\}$
```

IV. EXPERIMENTAL SET-UP

Considering that the ultimate goal of this work is to investigate the benefits of the dissimilarity representation over the feature representation in regression problems, we performed a systematic experimental study using two linear regression models that are tested over a pool of gold-standard datasets. In addition, a Wilcoxon's paired signed-rank test was employed to support the statistical validity of the results. The dissimilarity mapping process was implemented in the mlr3 library [26] and is available at https://github.com/JAIR-VG/dissreg-tools.

A. DATASETS

Experiments were carried out on 34 benchmark small-and medium-sized regression datasets that are commonly used in some studies on regression [27]–[29]. Table 1 summarizes the main characteristics of these datasets, which were obtained from the following sources:

1) Torgo repository (https://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html)
2) Weka dataset repository (https://waikato.github.io/weka-wiki/datasets/)
3) Energy efficiency dataset used in [30]
4) Extrusion diameter dataset used in [31]
5) Residential building dataset used in [32]

For each dataset, the input variables were normalized in the range of [0, 1]. The quality estimation of the linear regression models was generated using a 5-fold cross-validation. The resulting training and testing datasets were transformed by Euclidean distances using R.

The training and testing processes of both regression models were carried out on the original dataset (feature representation) and the transformed dataset (dissimilarity representation). The performance results reported in this paper correspond to the averaged values from the five trials.

TABLE 1. Overview of the databases (superscripts refer to the sources given in the list of database repositories and related articles).

No.	Dataset	Variables	Instances
1	Diabetes-numeric[1]	2	43
2	Quake[2]	3	2178
3	Basketball[2]	4	96
4	DeltaAilerons[1]	5	7129
5	MachineCPU[1]	6	209
6	DeltaElevators[1]	6	9517
7	Cooling[3]	8	768
8	Heating[3]	8	768
9	Bank-8FM[1]	8	8192
10	Kinematics[1]	8	8192
11	Puma8NH[1]	8	8192
12	Calhousing[1]	8	20640
13	House8L[1]	8	22784
14	Stock[1]	9	950
15	PwLinear[2]	10	200
16	2DPlanes[1]	10	40768
17	Friedman[1]	10	40768
18	CPUSmall[1]	12	8192
19	BodyFat[2]	14	252
20	Pollution[2]	15	60
21	ExtrusionOuter [4]	15	260
22	ExtrusionInner [4]	15	260
23	House16H[1]	16	22784
24	Elevators[1]	18	16599
25	CPUAct[1]	21	8192
26	Pyrimidines[1]	27	74
27	Wisconsin[1]	32	194
28	Bank32NH[1]	32	8192
29	Puma32H	32	8192
30	Ailerons[1]	40	13750
31	Pol[1]	48	15000
32	Triazines[1]	60	186
33	SalesPrices[5]	105	372
34	ConstructionCosts[5]	105	372

B. REGRESSION MODELS

The two regression models evaluated in our experiments were the generalized linear model with elastic net regularization (GLM) and a linear regressor (LR). In a linear model, the response variable z_i is modeled by a linear function of explanatory variables $x_j, j = 1, \ldots, n$ plus an error term ϵ_i (typically, it is assumed $\epsilon_i \sim N(0, \sigma^2)$) as follows:

$$z_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni} + \epsilon_i, \tag{1}$$

where β_j are the regression coefficients and x_{ji} are the regression variables.

In contrast with linear regression where the output is assumed to follow a Gaussian distribution, the generalized linear model [33] is a special class of nonlinear models where the response variable y_i does not need to be normally distributed, but it can follow some distribution from the exponential family (Poisson, multinomial, Bernoulli, chi-squared, gamma, and many others). Furthermore, homogeneity of variance does not need to be satisfied and errors need to be independent but not normally distributed.

A GLM is made up of a linear predictor $\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni}$, a smooth and invertible linearizing link function $g(\mu_i) = \eta_i$ that describes how the mean $(E(Y_i) = \mu_i)$ depends on the linear predictor, and a variance function $var(y_i) = \phi V(\mu_i)$ which describes the conditional distribution of the response variable y_i, that is, how the variance depends on the mean μ_i and a dispersion parameter ϕ.

The LR and the GLM with a Gaussian distribution were taken from the mlr3 framework [26] implemented in the R environment [34]. The default parameter values were used so that the results were not affected by a fine-tuning parameter step.

C. EVALUATION CRITERIA

We adopted two performance metrics commonly used in regression problems [29]. Both compute the numerical difference between the prediction of the model $(\hat{z}_i)$ and the true value (z_i) [35]. First, the Root Mean Squared Error (RMSE) was defined as follows:

$$RMSE = \sqrt{\frac{1}{N_{test}} \sum_{i=1}^{N_{test}} (\hat{z}_i - z_i)^2}, \tag{2}$$

where N_{test} is the number of test samples.

The second metric was the Mean Absolute Error (MAE):

$$MAE = \sum_{i=1}^{N_{test}} \mid \hat{z}_i - z_i \mid . \tag{3}$$

D. STATISTICAL TESTS

The Wilcoxon's paired signed-rank test was used to check for statistically significant differences between each pair of models. This statistic ranks the differences in the performances of the two algorithms for each dataset, ignoring the signs, and compares the ranks for the positive and the negative differences. Let d_i be the difference between the performance scores of the two models on i-th out of L datasets. Differences were ranked according to their absolute values. Let R^+ be the sum of ranks for the datasets on which the first model outperforms the second, and R^- the sum of ranks for the opposite. Ranks of $d_i = 0$ are split evenly among the sums; if there is an odd number of them, one is ignored:

$$R^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2} \sum_{d_i=0} rank(d_i), \tag{4}$$

$$R^- = \sum_{d_i<0} rank(d_i) + \frac{1}{2} \sum_{d_i=0} rank(d_i). \tag{5}$$

Let Z be the smaller of the sum, $Z = \min(R^+, R^-)$. If Z is less than or equal to the distribution value of Wilcoxon for L degrees of freedom, the null-hypothesis that both models perform equally well can be rejected.

E. DATA COMPLEXITY IN REGRESSION

Data complexity analysis was proposed in classification tasks as an approach to describing the intrinsic data characteristics [36]. The ultimate aim is to quantify some difficulties such as class ambiguity, boundary complexity, sample sparsity, and feature space dimensionality [36].

Lorena *et al.* [22] proposed several complexity measures to estimate the regression complexity. In this paper, we employed a feature correlation measure that captures the relationship between feature values and outputs. This is called the maximum feature correlation (C_1), where higher values indicate simpler problems, and lower values the opposite situation.

C_1 takes the maximum correlation value over all feature dimensions and can be computed as follows:

$$C_1 = \max_{j=1,\ldots,n} |\rho(x_j, z)|, \tag{6}$$

where ρ is the Spearman correlation, x_j is the feature j, z is the independent variable, and n is the dimensionality.

V. RESULTS

The general objective of the experiments can be divided into a series of more specific purposes. First, we analyzed the effect of selecting different R sizes on the performance of the regression models. Second, we statistically checked whether or not the dissimilarity representation outperformed the feature representation when applied to regression. Finally, we compared the complexity of the dissimilarity-based datasets with that of the feature-based datasets.

A. INFLUENCE OF THE REPRESENTATION SET SIZE

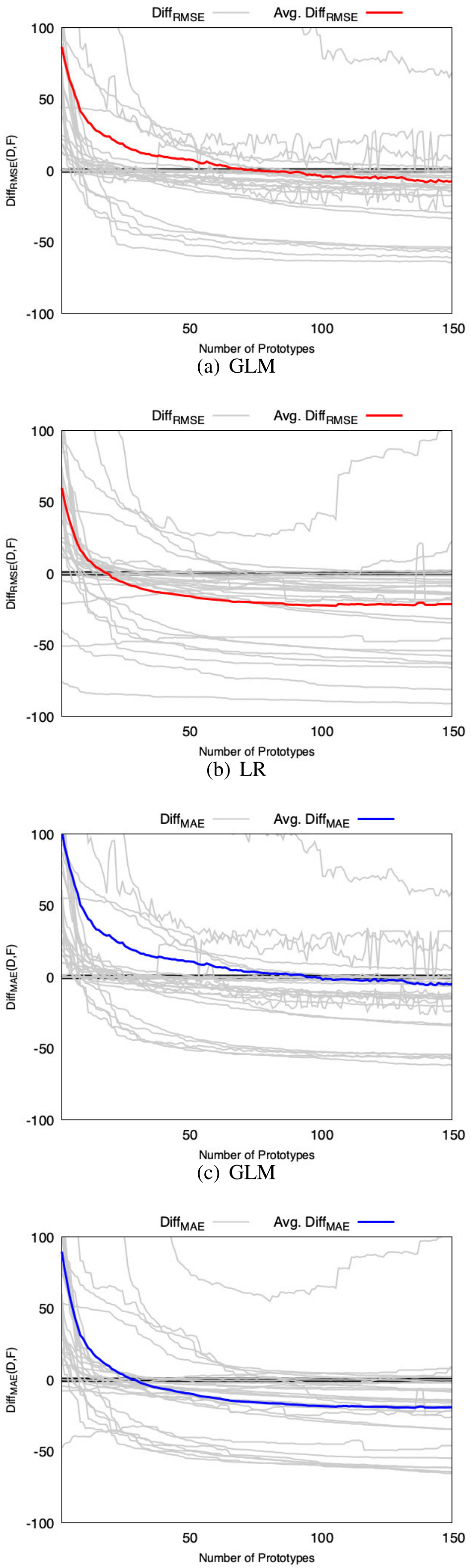
In this experiment, we are interested in the impact of different R sizes over the dissimilarity regression models. We omitted the small-sized databases (Diabetes-numeric, Basketball, Pollution, Pyrimidines, and Triazines) and for the remaining 29 datasets, we randomly selected a number of representative objects ranging from 2 to 150. The upper bound was set to 150 because the random selection process cannot guarantee the optimal number of prototypes, whereas previous studies observed that selecting more than 150 objects did not produce significant differences [18].

When comparing and contrasting two or more datasets, it is important to represent them on comparable scales. Thus, we defined a relative error difference [29], [37] computed for each dataset as follows:

$$Diff(D, F) = \frac{D - F}{F} \times 100, \tag{7}$$

where F and D represent the feature-based regressor and dissimilarity-based regressor results, respectively. Note that this score can be viewed as an indicator of improvement or deterioration of the dissimilarity-based model compared to the feature-based one.

Figure 2 shows the relative error difference of the 29 selected datasets achieved with the dissimilarity mapping process for all R sizes $(2, \ldots, 150)$, where the red and blue lines correspond to the RMSE and MAE values averaged across all datasets, respectively. The x-axis represents the number of objects selected to construct R and the y-axis is the relative error difference in terms of RMSE (Figure 2–a, Figure 2–b) and MAE (Figure 2–c, Figure 2–d). Negative values indicate that the model using the dissimilarity representation was better than that based on the feature representation.



(a) GLM

(b) LR

(c) GLM

(d) LR

FIGURE 2. Relative error difference for both regression models in terms of MRSE and MAE. Red and blue lines correspond to the average values across all datasets.

TABLE 2. Average RMSE and MAE values (for each dataset, the best result is highlighted).

	RMSE				MAE			
Datasets	GLM-F	**GLM-D**	LR-F	**LR-D**	GLM-F	**GLM-D**	LR-F	**LR-D**
Diabetes-numeric	6.6086e−1	6.3962e−1	6.0247e−1	6.4734e−1	5.3554e−1	5.044e−1	4.7050e−1	5.2263e−1
Quake	1.8935e−1	1.8935e−1	1.8887e−1	1.8896e−1	1.4927e−1	1.4927e−1	1.4858e−1	1.4837e−1
Basketball	9.1455e−2	8.9383e−2	8.5302e−2	8.5141e−2	7.2492e−2	7.0860e−2	6.7051e−2	6.7526e−2
DeltaAilerons	1.7484e−4	1.6834e−4	1.7195e−4	1.6490e−4	1.2382e−4	1.1898e−4	1.2367e−4	1.1570e−4
MachineCPU	8.8162e1	6.3543e1	6.9117e1	5.4220e1	4.3905e1	3.2306e1	4.1662e1	3.1331e1
DeltaElevators	1.4608e−3	1.4340e−3	1.4487e−3	1.4252e−3	1.1069e−3	1.0859e−3	1.1025e−3	1.0778e−3
Cooling	3.3464	2.9734	3.0857	2.7977	2.3760	2.0835	2.2980	1.9779
Heating	3.0394	2.6663	2.8781	2.4715	2.1712	1.8259	2.0802	1.6839
Bank-8FM	3.9509e−2	3.7539e−2	3.8820e−2	3.6591e−2	2.9357e−2	2.8804e−2	2.8556e−2	2.8170e−2
Kinematics	2.0360e−1	1.4339e−1	2.0210e−1	1.3837e−1	1.6468e−1	1.1028e−1	1.6232e−1	1.0641e−1
Puma8NH	4.4885	3.9619	4.4639	3.8562	3.7287	3.2126	3.6460	3.1076
Calhousing	7.0222e4	6.3125e4	6.9575e4	6.0647e4	5.1452e4	4.4760e4	5.0780e4	4.2730e4
House8L	4.2291e4	3.4944e4	4.1544e4	3.3935e4	2.4536e4	1.9510e4	2.4349e4	1.8981e4
Stock	2.4112	1.0329	2.3446	8.8450e−1	1.8752	7.9780e−1	1.8404	6.6056e−1
PwLinear	2.3208	1.8942	2.2064	1.7694	1.8519	1.5151	1.7525	1.4019
2DPlanes	2.3909	1.1120	2.3841	1.0964	1.9287	8.8403e−1	1.9225	8.7236e−1
Friedman	2.6432	1.7696	2.6310	1.7277	2.0674	1.3683	2.0384	1.3399
CPUSmall	1.0101e1	3.6057	9.8376	3.3586	6.0274	2.6305	6.1750	2.3795
BodyFat	1.7145	1.7270	1.1956	1.5080	1.1216	1.1992	5.1532e−1	7.9893e−1
Pollution	4.0683e1	4.0872e1	4.7583e1	3.7256e1	3.0584e1	3.1323e1	3.6088e1	2.9796e1
ExtrusionOuter	1.2753e−2	1.2743e−2	2.6372e−2	1.2993e−2	3.8206e−3	3.8206e−3	8.0987e−3	4.1025e−3
ExutrusionInner	1.0727e−3	1.0727e−3	1.3520e−3	1.0748e−3	8.9223e−4	8.9223e−4	9.6523e−4	8.9157e−4
House16H	4.6408e4	4.0053e4	4.5481e4	3.8915e4	2.5741e4	2.1860e4	2.5450e4	2.1680e4
Elevators	2.9514e−3	2.5907e−3	2.9058e−3	2.4896e−3	1.9875e−3	1.8578e−3	1.9910e−3	1.8330e−3
CPUActivity	9.9122	3.8733	9.6236	3.5444	5.8959	2.6599	6.0697	2.3484
Pyrimidines	1.3984e−1	8.8503e−2	2.2168e−1	7.7675e−2	9.6768e−2	6.2905e−2	1.2751e−1	5.4419e−2
Wisconsin	3.3216e1	3.2704e1	3.4254e1	3.1248e1	2.8484e1	2.7799e1	2.8487e1	2.5426e1
Bank32NH	8.4512e−2	8.6267e−2	8.3536e−2	8.4312e−2	5.8151e−2	6.0941e−2	5.8659e−2	6.0076e−2
Puma32H	2.6948e−2	2.6884e−2	2.6763e−2	2.6707e−2	2.1253e−2	2.1238e−2	2.0988e−2	2.0941e−2
Ailerons	1.7584e−4	1.6698e−4	1.7580e−4	1.6373e−4	1.2889e−4	1.2184e−4	1.2913e−4	1.1965e−4
Pol	3.0599e1	1.3767e1	3.0488e1	1.2920e1	2.6732e1	1.0229e1	2.6592e1	9.4926
Triazines	1.5312e−1	1.5009e−1	1.5689e−1	1.4325e−1	1.1519e−1	1.0921e−1	1.1932e−1	1.0317e−1
SalesPrices	1.7665e2	4.6195e2	1.7190e3	3.2565e2	9.7319e1	3.1496e2	3.0033e2	2.2127e2
ConstructionCosts	3.3477e1	5.5326e1	4.7211e2	4.3560e1	2.1410e1	3.3524e1	7.3097e1	2.5363e1

As can be observed from these plots, there appears a general tendency to reduce the error when increasing the R size. In addition, the relative error difference indicates that the regression models performed better when most datasets were transformed into a dissimilarity representation than with the original feature-based datasets. However, it was not possible to determine the optimal R size. Although the dissimilarity mapping process did not yield an error reduction when applied to some databases, we believe that the use of an intelligent prototype selection strategy instead of the random method could lead to the expected good behavior.

B. PERFORMANCE EVALUATION

Table 2 reports the RMSE and MAE values of both regression models, respectively. As the dissimilarity experiments were performed using different representation set sizes, for clarity and conciseness only the best RMSE and MAE values for each pair of dataset and regression model were collected. As can be observed, the best performances were mostly obtained by training the regressor with the dissimilarity-based datasets. In addition, for each performance measure, Table 3 summarizes how many times the regression models built on the dissimilarity representation were better/same/worse than the regressors based on the feature representation.

TABLE 3. Comparison of representation strategies. The three values in the cells show the number of datasets where the dissimilarity-based model of the row was better/same/worse than the feature-based regressor of the column.

		GLM-Feature	LR-Feature
RMSE	GLM-Dissimilarity	27/2/5	
	LR-Dissimilarity		30/0/4
MAE	GLM-Dissimilarity	26/3/5	
	LR-Dissimilarity		30/0/4

To determine whether or not the dissimilarity representation was better than the traditional approach, we ran a Wilcoxon's signed-rank test for detecting statistically significant differences using the results of RMSE and MAE. Table 4 shows the ranks and the p-values when comparing

TABLE 4. Wilcoxon test for dissimilarity-based vs feature-based regressors.

	Comparison	R^+	R^-	p-value
RMSE	**GLM-Dissimilarity** vs GLM-Feature	494.5	100.5	4.4760e−4
	LR-Dissimilarity vs LR-Feature	551.0	44.0	1.7102e−6
MAE	**GLM-Dissimilarity** vs GLM-Feature	482.0	113.0	1.1152e−3
	LR-Dissimilarity vs LR-Feature	547.0	48.0	2.8280e−6

TABLE 5. C_1 values computed in the feature and dissimilarity representations.

Datasets	Feature	Dissimilarity
Diabetes-numeric	0.55	0.60
Quake	0.05	0.06
Basketball	0.55	0.61
DeltaAilerons	0.80	0.75
MachineCPU	0.81	0.89
DeltaElevators	0.69	0.66
Cooling	0.86	0.91
Heating	0.86	0.93
Bank-8FM	0.70	0.18
Kinematics	0.52	0.50
Puma8NH	0.50	0.44
Calhousing	0.68	0.32
House8L	0.62	0.48
Stock	0.69	0.74
PwLinear	0.71	0.71
2DPlanes	0.69	0.61
Friedman	0.58	0.59
CPUSmall	0.73	0.76
BodyFat	0.99	0.79
Pollution	0.61	0.64
ExtrusionOuter	0.18	0.11
ExtrusionInner	0.10	0.12
House16H	0.60	0.43
Elevators	0.50	0.55
CPUAct	0.81	0.83
Pyrimidines	0.61	0.75
Wisconsin	0.33	0.35
Bank32NH	0.50	0.16
Puma32H	0.41	0.13
Ailerons	0.73	0.70
Pol	0.35	0.76
Triazines	0.28	0.31
SalesPrices	0.99	0.77
ConstructionCosts	0.97	0.88

one representation against the other. Considering a significance level of $\alpha = 0.05$, the winner models are highlighted in bold when the associated p-value is lower than α. As can be seen, for all comparisons, the best algorithms were those trained with the data mapped into a dissimilarity space.

C. DATA COMPLEXITY IN DISSIMILARITY REGRESSION DATASETS

C_1 was computed for each dataset in the feature representation as well as for the dissimilarity representation constructed from several R sizes. For the sake of simplicity and clarity, Table 5 summarizes the maximum C_1 values obtained from all dissimilarity datasets. The results show that, for some datasets, the mapping process converted a dataset into a simpler problem. However, this behavior was not observed in all the datasets, despite the fact that some of them achieved better results when using the dissimilarity representation.

VI. CONCLUSION AND FURTHER EXTENSIONS

A good object representation influences the performance of supervised machine learning methods. The dissimilarity representation has been used as an alternative to the feature characterization in classification problems, showing that dissimilarity-based classifiers may improve their accuracy. In addition, this mapping process presents important advantages regarding the reduction of some intrinsic data difficulties that allow the use of single linear models.

In this sense, the dissimilarity mapping process can also be used in the context of regression. Therefore, we performed an extensive experimental study on 34 benchmark regression datasets where each dataset was transformed into a dissimilarity matrix using the Euclidean distance and a set R. The independent variables in the new space correspond to the dissimilarity between the pairs of objects. From the experimental results, it is possible to draw some concluding remarks that support our findings:

1) Through a random selection of R of various sizes, it has been proven that mapping a feature sample into a dissimilarity space improves the performance of the linear regression models. On the other hand, it seems that in some cases, the use of a larger R can result beneficial.
2) The Wilcoxon's signed-rank test with $\alpha = 0.05$ validates our claim that the linear regression models built on the dissimilarity representation perform better than those based on the traditional feature-based representation.
3) Using a data complexity measure, it has been observed that the problems yield high correlation values when the datasets are mapped into a dissimilarity representation, that is, the problems become simpler.

The main criticism that can be made to the present work is the lack of a theoretical analysis. However, the experimental results have demonstrated the potential benefits of using the dissimilarity representation in the context of regression problems, thus opening some avenues for further research. One of them can be the design of systematic methods for the selection of representative objects specifically focused on regression tasks. Although it has been claimed that the Euclidean distance is suitable for dissimilarity transformation, other metrics should also be explored. In this sense, we believe that the adoption of distance metrics for high-dimensional problems could be an interesting direction for extending the present work. Another point to investigate in the future can be to study the behavior of the dissimilarity representation for non-linear regression models, such as XGBoost, K-nearest neighbors, support vector regressor and random forest. In addition, living in the Big Data era where data is growing exponentially, we would like to extend the present work by exploring the performance of the dissimilarity-based regression method on massive datasets, which bring a series of special computational challenges.

Finally, our proposal in its present form could not be applied to real-world applications in a continuous learning system, capable of incrementally storing and discarding streaming data. Thus the design of dissimilarity-based regression models with the power of continuous learning and adaptation during real-time operations under changes in the environment may constitute an interesting open line for further research.

REFERENCES

[1] M. Orozco-Alzate, P. A. Castro-Cabrera, M. Bicego, and J. M. Londoño-Bonilla, "The DTW-based representation space for seismic pattern classification," *Comput. Geosci.*, vol. 85, pp. 86–95, Dec. 2015.

[2] J. Calvo-Zaragoza, J. J. Valero-Mas, and J. R. Rico-Juan, "Prototype generation on structural data using dissimilarity space representation," *Neural Comput. Appl.*, vol. 28, no. 9, pp. 2415–2424, Sep. 2017.

[3] R. P. W. Duin and E. Pęalska, "The dissimilarity space: Bridging structural and statistical pattern recognition," *Pattern Recognit. Lett.*, vol. 33, no. 7, pp. 826–832, 2012.

[4] A. R. Webb and K. D. Copsey, *Statistical Pattern Recognition*, 3rd ed. Hoboken, NJ, USA: Wiley, 2011.

[5] Y. M. G. Costa, D. Bertolini, A. S. Britto, Jr., G. D. C. Cavalcanti, and L. E. S. Oliveira, "The dissimilarity approach: A review," *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 2783–2808, Apr. 2020.

[6] L. Nanni, G. Minchio, S. Brahnam, G. Maguolo, and A. Lumini, "Experiments of image classification using dissimilarity spaces built with Siamese networks," *Sensors*, vol. 21, no. 5, pp. 1–18, 2021.

[7] D. M. J. Tax, M. Loog, R. P. W. Duin, V. Cheplygina, and W.-J. Lee, "Bag dissimilarities for multiple instance learning," in *Similarity-Based Pattern Recognition*, M. Pelillo and E. R. Hancock, Eds. Berlin, Germany: Springer, 2011, pp. 222–234.

[8] R. P. Duin and E. Pękalska, "The dissimilarity representation for non-Euclidean pattern recognition, a tutorial," Dept. Electr. Eng., Math. Comput. Sci., Delft Univ. Technol., Netherlands School Comput. Sci., Univ. Manchester, Manchester, U.K., Tech. Rep., Nov. 2011.

[9] E. Pękalska and R. P. W. Duin, "Dissimilarity representations allow for building good classifiers," *Pattern Recognit. Lett.*, vol. 23, no. 8, pp. 943–956, Jun. 2002.

[10] V. García, A. I. Marqués, J. S. Sánchez, and H. J. Ochoa-Domínguez, "Dissimilarity-based linear models for corporate bankruptcy prediction," *Comput. Econ.*, vol. 53, no. 3, pp. 1019–1031, Mar. 2019.

[11] E. Pękalska, P. Paclik, and R. P. W. Duin, "A generalized kernel approach to dissimilarity-based classification," *J. Mach. Learn. Res.*, vol. 2, pp. 175–211, Mar. 2002.

[12] V. García, J. S. Sánchez, H. J. O. Domínguez, and L. Cleofas-Sánchez, "Dissimilarity-based learning from imbalanced data with small disjuncts and noise," in *Pattern Recognition and Image Analysis*, R. Paredes, J. S. Cardoso, and X. M. Pardo, Eds. Cham, Switzerland: Springer, 2015, pp. 370–378.

[13] E. Bruno, N. Moenne-Loccoz, and S. Marchand-Maillet, "Design of multimodal dissimilarity spaces for retrieval of video documents," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1520–1533, Sep. 2008.

[14] D. Porro-Muñoz, I. Talavera, R. P. W. Duin, N. Hernández, and M. Orozco-Alzate, "Dissimilarity representation on functional spectral data for classification," *J. Chemometrics*, vol. 25, no. 9, pp. 476–486, Sep. 2011.

[15] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, "Pose-based human action recognition via sparse representation in dissimilarity space," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 12–23, Jan. 2014.

[16] S. Mauceri, J. Sweeney, and J. McDermott, "Dissimilarity-based representations for one-class classification on time series," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107122.

[17] J. G. Martins, L. S. Oliveira, A. S. Britto, Jr., and R. Sabourin, "Forest species recognition based on dynamic classifier selection and dissimilarity feature vector representation," *Mach. Vis. Appl.*, vol. 26, nos. 2–3, pp. 279–293, Apr. 2015.

[18] V. García and J. S. Sánchez, "Mapping microarray gene expression data into dissimilarity spaces for tumor classification," *Inf. Sci.*, vol. 294, pp. 362–375, Feb. 2015.

[19] R. H. W. Pinheiro, G. D. C. Cavalcanti, and I. R. Tsang, "Combining dissimilarity spaces for text categorization," *Inf. Sci.*, vol.s. 406–407, pp. 87–101, Sep. 2017.

[20] J. Jaramillo-Garzon, A. Quiceno-Manrique, I. Godino-Llorente, and C. G. Castellanos-Dominguez, "Feature extraction for murmur detection based on support vector regression of time-frequency representations," in *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Vancouver, BC, Canada, Aug. 2008, pp. 1623–1626.

[21] F. J. Silva-Mata, C. Jiménez, G. Barcas, D. Estevez-Bresó, N. Acosta-Mendoza, A. Gago-Alonso, and I. Talavera-Bustamante, "Improving regression models by dissimilarity representation of biochemical data," in *Proc. 23rd Iberoamerican Congr. Pattern Recognit.*, Madrid, Spain, 2018, pp. 64–71.

[22] A. C. Lorena, A. I. Maciel, P. B. C. de Miranda, I. G. Costa, and R. B. C. Prudêncio, "Data complexity meta-features for regression problems," *Mach. Learn.*, vol. 107, no. 1, pp. 209–246, Jan. 2018.

[23] E. Pękalska, R. P. W. Duin, and P. Paclík, "Prototype selection for dissimilarity-based classifiers," *Pattern Recognit.*, vol. 39, no. 2, pp. 189–208, 2006.

[24] E. Pękalska and R. P. W. Duin, "Prototype selection for finding efficient representations of dissimilarity data," in *Proc. 16th Int. Conf. Pattern Recognit.*, vol. 3. Quebec City, QC, Canada: IEEE Computer Society, 2002, pp. 37–40.

[25] E. Pękalska, *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. River Edge, NJ, USA: World Scientific, 2005.

[26] M. Lang, M. Binder, J. Richter, P. Schratz, F. Pfisterer, S. Coors, Q. Au, G. Casalicchio, L. Kotthoff, and B. Bischl, "mlr3: A modern object-oriented machine learning framework in R," *J. Open Source Softw.*, vol. 4, no. 44, pp. 1–2, 2019.

[27] M. Kurzynski, M. Krysmann, and J. Kozerski, "Fuzzy inference systems applied to the combining regression models," in *Proc. Int. Conf. Sustain. Energy, Electron., Comput. Syst.*, Greater Noida, India, 2018, pp. 1–6.

[28] Z. Zhang, J. He, G. Gao, and Y. Tian, "Bi-sparse optimization-based least squares regression," *Appl. Soft Comput.*, vol. 77, pp. 300–315, Apr. 2019.

[29] J. J. Rodríguez, M. Juez-Gil, Á. Arnaiz-González, and L. I. Kuncheva, "An experimental evaluation of mixup regression forests," *Expert Syst. Appl.*, vol. 151, Aug. 2020, Art. no. 113376.

[30] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy Buildings*, vol. 49, pp. 560–567, Jun. 2012.

[31] V. García, J. S. Sánchez, L. A. Rodríguez-Picón, L. C. Méndez-González, and H. D. J. Ochoa-Domínguez, "Using regression models for predicting the product quality in a tubing extrusion process," *J. Intell. Manuf.*, vol. 30, no. 6, pp. 2535–2544, Aug. 2019.

[32] M. H. Rafiei and H. Adeli, "A novel machine learning model for estimation of sale prices of real estate units," *J. Construct. Eng. Manage.*, vol. 142, no. 2, pp. 1–18, 2016.

[33] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *J. Roy. Stat. Soc. A, Gen.*, vol. 135, no. 3, pp. 370–384, 1972.

[34] R Core Team and R Foundation for Statistical Computing, Vienna, Austria. (2020). *R: A Language and Environment for Statistical Computing*. [Online]. Available: https://www.R-project.org/

[35] L. Torgo, *Data Mining With R: Learning With Cases*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2017.

[36] M. Basu and T. K. Ho, *Data Complexity in Pattern Recognition* (Advanced Information and Knowledge Processing). Berlin, Germany: Springer-Verlag, 2006.

[37] V. García, J. S. Sánchez, A. I. Marqués, and R. Martínez-Peláez, "A regression model based on the nearest centroid neighborhood," *Pattern Anal. Appl.*, vol. 21, no. 4, pp. 941–951, Nov. 2018.

VICENTE GARCÍA (Member, IEEE) received the B.E. degree in computer systems from the Instituto Tecnológico de Villahermosa, México, in 2000, the M.S. degree in computer sciences from the Instituto Tecnológico de Toluca, México, in 2002, and the Ph.D. degree in advanced computer science from the Universitat Jaume I, Castelló de la Plana, Spain, in 2010.

From 2010 to 2013, he was a Research Assistant with the Institute of New Imaging Technologies, Universitat Jaume I, Castelló de la Plana, Spain. Since 2014, he has been full-time Professor with the Electrical and Computer Engineering Department, Universidad Autónoma de Ciudad Juárez. His research interests include data preprocessing methods, data complexity, non-parametric classification, performance evaluation, and big data.

Prof. García served as a reviewer for several national and international journals and as a board member of multiple international conferences. He is a member of the Mexican Academy of Computing and the National Mexican Science Council (CONACYT).

J. SALVADOR SÁNCHEZ received the B.Sc. degree in computer science from the Universidad Politécnica de Valencia, Spain, in 1990, and the Ph.D. degree in computer science engineering from the Universitat Jaume I, Castelló de la Plana, Spain, in 1998.

In 2006, he was awarded a two-year research fellowship to work with the University of Wales, Bangor, U.K., with Prof. Ludmila I. Kuncheva. He is currently a Full Professor with the Department of Computer Languages and Systems, Universitat Jaume I, and the Head of the Pattern Analysis and Learning Laboratory. He is author or coauthor of more than 200 scientific publications, co-editor of three books, and a guest editor of several special issues in international journals. His research interests include the fields of pattern recognition, machine learning, and data mining, including classification, feature and prototype selection, ensembles of classifiers, data analysis, and reinforcement learning. He serves as an Associate Editor for *Pattern Analysis and Applications*, *Progress in Artificial Intelligence*, and *Applied Sciences* journals. He was the Former President of AERFAI (the Spanish Association for Pattern Recognition and Image Analysis).

RAFAEL MARTÍNEZ-PELÁEZ received the Ph.D. degree from the Technical University of Catalonia, in 2010. Since 2016, he has been an Associate Professor with the Facultad de Tecnologías de Información, Universidad de la Salle Bajío, León, México. His research interests include human element in cybersecurity, risk management, and machine learning. He has served as a TPC member for many international conferences and workshops. He is a member of the National System of Researchers (SNI) of the National Mexican Science Council (CONACYT).

LUIS C. MÉNDEZ-GONZÁLEZ (Member, IEEE) received the degree in electronic engineering, in 2007, and the master's degree in industrial engineering from the Technological Institute of Ciudad Juárez, in 2011, México, and the Doctorate of Science degree in engineering from the Autonomous University of Ciudad Juárez, in 2015.

He has carried out research and exchange stays at New Mexico State University, Las Cruces, NM, USA. He has more than ten years in software and hardware design, applied statistics, measurement system analysis, reliability engineering, and quality engineering, among others in the industry. His research aims to develop reliability models, stochastic processes, hardware design, electronics, and robotics. He has presented papers and research in different parts of the world, such as in the USA, Italy, Colombia, and México. He is currently teaching reliability, hardware design, robotics, maintenance, and control as a full-time Professor with the Autonomous University of Ciudad Juárez. His current research interests include time-varying voltage scenarios induced on power lines and their effects on electrical and electronic devices.