Revisiting the Dissimilarity Representation in the Context of Regression

VICENTE GARCÍA1, (Member, IEEE), J. SALVADOR SÁNCHEZ2, RAFAEL MARTÍNEZ-PELÁEZ3, AND LUIS C. MÉNDEZ-GONZÁLEZ4, (Member, IEEE)

1Department of Electrical and Computer Engineering, Universidad Autónoma de Ciudad Juárez, 32310, Ciudad Juárez, México (e-mail: vicente.jimenez@uacj.mx)
2Institute of New Imaging Technologies and Department of Computer Languages and Systems, Universitat Jaume I, 12071 Castelló de la Plana, Spain (e-mail: sanchez@uji.es)
3Faculty of Information Technologies, Universidad de La Salle Bajío, 37150, León, México (e-mail: rmartinezp@delasalle.edu.mx)
4Department of Industrial Engineering and Manufacturing, Universidad Autónoma de Ciudad Juárez, 32310, Ciudad Juárez, México (e-mail: luis.mendez@uacj.mx)

Corresponding author: Vicente García (e-mail: 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 a feature vector. However, several studies have shown that this representation may not characterize an object accurately. 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 demonstrated to improve the data quality because it mitigates some complexities such as the class overlap, the small disjuncts, or the lack of samples. However, it has not been fully explored its suitability and performance when applied to regression problems. This paper redefines the dissimilarity representation for regression. To this end, we have carried out an extensive experimental evaluation on 34 data sets with 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 data sets.

INDEX TERMS  Data complexity, Dissimilarity representation, Linear models, Regression

I. INTRODUCTION  In machine learning, a natural way to represent an instance is by a feature vector. However, several studies have shown that this representation may not characterize an object accurately. 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 demonstrated to improve the data quality because it mitigates some complexities such as the class overlap, the small disjuncts, or the lack of samples. However, it has not been fully explored its suitability and performance when applied to regression problems. This paper redefines the dissimilarity representation for regression. To this end, we have carried out an extensive experimental evaluation on 34 data sets with 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 data sets.
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.

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], reporting that the dissimilarity representation 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 data sets 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 strategy designed to conform the representation set. To capture the difficulty of a regression problem, we compute a data complexity measure [22] with the purpose of checking whether or not the regression problem is simpler using the dissimilarity representation than the traditional feature representation.

Henceforth the paper is organized as follows. Section II introduces the basics of the dissimilarity representation, whereas Section III describes the process for adapting the dissimilarity representation to regression problems. Next, Section IV provides the experimental set-up. In Section V, the experimental results are presented and discussed. Finally, Section VI remarks the main conclusions and outlines possible future directions to extend 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 for each class $R = \{p_1, \ldots, p_M\}$. This set $R$ can be taken as the complete set of objects $T = \{x_1, \ldots, x_N\}$ 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 due 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 ($d(x_i, x_j) > 0$ if $x_i$ is distinct from $x_j$) and obey the reflexivity condition ($d(x_i, x_i) = 0$), but it might be non-metric. Among others, common dissimilarity measures are 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 $x_i \in T$ can be represented by an $M$-dimensional real-valued vector in the dissimilarity space, $D(x_i, R) = [d(x_i, p_1), \ldots, d(x_i, p_M)]$, that is, each dimension corresponds to the dissimilarity computed between $x_i$ 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]:

$$D(T, R) = \begin{bmatrix} d(x_1, p_1) & d(x_1, p_2) & \cdots & d(x_1, p_M) \\ d(x_2, p_1) & d(x_2, p_2) & \cdots & d(x_2, p_M) \\ \vdots & \vdots & \ddots & \vdots \\ d(x_N, p_1) & d(x_N, p_2) & \cdots & d(x_N, p_M) \end{bmatrix}$$
Now a representation space can be built from this matrix. The dimensionality of the dissimilarity space is equal to \( M \) (the cardinality of \( R \)). In this way, each dimension corresponds to the dissimilarities with one of the prototypes in the representation set. 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 \((x, z)\), where \( x \) is the independent variable and \( z \in \mathbb{R} \) is the dependent variable. The aim of regression is to find a function \( f(.) \) that can predict \( z \) from \( T \) based on \( N \) observations \((x_i, z_i), i = 1, \ldots , N\).

Alike the mapping process to convert a feature vector into a dissimilarity vector in classifications tasks, in regression problems the process is carried out for the learning and testing phases as can be observed in the flowchart of Fig. 1.

The first step for the construction of a dissimilarity matrix is to select a representation set from the training set. In this paper, this process is performed by a random selection method (Algorithm 1). The representation set is constructed taking \( M \) random samples without replacement from \( T \). In classifications tasks, there exist several instance selection methods focused on extract the most significant samples.

### Algorithm 1: Random selection of the representation set

- **Input:** Regression training set \( T = \{(x_1, z_1), \ldots , (x_N, z_N)\} \),
- **Output:** Representation set \( R \subseteq T \)
- \( M \) Generate \( K \) random numbers \( \in [1, N] \)
- \( R = \{x_M\} \)

Once \( R \) has been selected, the dissimilarity matrix is constructed using some dissimilarity measure as shown in Algorithm 2. Note that we will use the Euclidean distance in the experiments. It has to be remarked 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 training and testing data sets. These dissimilarity sets are passed through the regression model for learning and predicting the independent variable of a new instance \( x' \).

### Algorithm 2: Mapping process to construct a dissimilarity matrix

- **Input:** Regression set \( T = \{(x_1, z_1), \ldots , (x_N, z_N)\} \),
- **Output:** Dissimilarity set \( D \)
- \( D = \emptyset \)
- foreach \( x_i \in T \) do
  - \( \text{distances} = \emptyset \)
  - foreach \( p_j \in R \) do
    - \( \text{dist}_{i,j} = d(x_i, p_j) \)
    - \( \text{distances} = \text{distances} \cup \{\text{dist}_{i,j}\} \)
  - \( D = D \cup \{\text{distances}\} \)
- \( D \)

The computational complexity of the proposed algorithm depends on the computational costs associated with the mapping process to construct a dissimilarity matrix \( D \). Thus, for the training stage, both the time complexity and space complexity are \( O(N \cdot M) \), being \( N \) the training set size and \( M \) the representation set size. As the testing stage makes use of the representation set, the time and space complexities of mapping a test example are \( O(M) \).

### IV. EXPERIMENTAL SET-UP

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

#### A. DATA SETS

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

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

For each data set, the input variables were normalized in the range \([0, 1]\). The quality estimation of the linear regression models was carried out using 5-fold cross-validation.
The resulting training and testing data sets were transformed by Euclidean distances using a representation set. The training and testing processes of both regression models were carried out over the original data set (feature representation) and the transformed data set (dissimilarity representation). The performance results reported in this paper correspond to 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. | Data set          | Variables | Instances |
|-----|-------------------|-----------|-----------|
| 1   | Diabetes-numeric  | 2         | 43        |
| 2   | Quake             | 3         | 2178      |
| 3   | Basketball        | 4         | 96        |
| 4   | DeltaAilerons     | 5         | 7129      |
| 5   | MachineCPT        | 6         | 209       |
| 6   | DeltaElevators    | 6         | 9517      |
| 7   | Cooling           | 8         | 768       |
| 8   | Heating           | 8         | 768       |
| 9   | Bank-SFM          | 8         | 8192      |
| 10  | Kinematics        | 8         | 8192      |
| 11  | Puma8NH          | 8         | 8192      |
| 12  | Calhousing        | 8         | 20640     |
| 13  | House8L          | 8         | 22784     |
| 14  | Stock            | 9         | 950       |
| 15  | PwLinear         | 10        | 200       |
| 16  | 2DPlanes         | 10        | 40768     |
| 17  | Friedman          | 10        | 40768     |
| 18  | CPU8Small        | 12        | 8192      |
| 19  | BodyFat          | 14        | 252       |
| 20  | Pollution         | 15        | 60        |
| 21  | ExtrusionOuter   | 15        | 260       |
| 22  | ExtrusionInner   | 15        | 260       |
| 23  | House16H         | 16        | 22784     |
| 24  | Elevators         | 18        | 16599     |
| 25  | CPU3Act          | 21        | 8192      |
| 26  | Pyrimidines      | 27        | 74        |
| 27  | Wisconsin        | 32        | 194       |
| 28  | Bank32NH         | 32        | 8192      |
| 29  | Puma32H          | 32        | 8192      |
| 30  | Ailerons          | 40        | 13750     |
| 31  | Pol              | 48        | 15000     |
| 32  | Triazines        | 60        | 186       |
| 33  | SalesPrices       | 105       | 372       |
| 34  | ConstructionCosts | 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 $x_{ji}$, $j = 1, \ldots, n$ plus an error term $\epsilon_i$ (typically, it is assumed $\epsilon_i \sim N(0, \sigma^2)$) as follows:

$$z_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in} + \epsilon_i,$$

where $\beta_j$ 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 non-linear 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 $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in}$, 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 $\text{var}(\mu_i) = \phi V(\mu_i)$ that describes the conditional distribution of the response variable $y_i$, that is, how the variance depends on the mean $\mu_i$ and a dispersion parameter $\phi$.

The LR and the GLM with a Gaussian distribution were taken from the mlr3 framework [26] implemented in the R environment [34] using the default parameter values 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 numeric difference between the prediction of the model ($\hat{z}_i$) and the true value ($z_i$) [35]. First the Root Mean Squared Error (RMSE) is defined as follows:

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

where $N_{\text{test}}$ is the number of test samples.

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

$$\text{MAE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} | \hat{z}_i - z_i |.$$

### 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 performances of two algorithms for each data set, 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 $L$ data sets. The differences are ranked according to their absolute values. Let $R^+$ be the sum of ranks for the data sets 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:

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

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

Let $Z$ be the smaller of the sums, $Z = \min(R^+, R^-)$. If $Z$ is less than or equal to the value of the distribution 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 of the feature values with the outputs. It 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:

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

where $\rho$ is the Spearman correlation, $x_j$ the feature $j$, $z$ the independent variable, and $n$ 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 representation set sizes on the performance of the regression models. Second, we statistically checked whether or not the dissimilarity representation outperforms the feature one when applied to regression. Finally, we compared the complexity of the dissimilarity-based data sets against that of the feature-based data sets.

A. INFLUENCE OF THE REPRESENTATION SET SIZE

In this experiment, we evaluated how the different representation set sizes could affect the performance of the regression models. We omitted the small-sized databases (Diabetes-numeric, Basketball, Pollution, Pyrimidines, and Triazines) and for the remaining 29 data sets, 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 data set series, it is important to represent them on comparable scales. Thus, we defined a relative error difference [29], [37] computed for each data set as follows:

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

where $F$ and $D$ are the result of RMSE (or MAE) for the feature-based regressor and the result of RMSE (or MAE) for the dissimilarity-based regressor, 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.

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

As can be observed from these plots, there appears a general tendency to reduce the error when increasing the representation set size. Also, the relative error difference indicates that the regression models performed better with most data sets transformed into a dissimilarity representation than with the original feature-based data sets. However, it was not possible to determine the optimal size of the representation set. Although the dissimilarity mapping process did not yield an error reduction when applied to some databases, we believe that the use of some 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 data set and regression model were collected. As can be observed, the best performances were mostly achieved by training the regressor with the dissimilarity-based data sets. 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.

To state 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 one represen-
TABLE 3. Comparison of representation strategies. The three values in the cells show the number of data sets 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     | 30/0/4     |
| LR-Dissimilarity  |           |            |
| **MAE**    |             |            |
| GLM-Dissimilarity | 26/3/5     | 30/0/4     |
| LR-Dissimilarity  |           |            |

tation against the other. Considering a level of significance \( \alpha = 0.05 \), we highlight in bold the winner models when the associated \( p \)-value was lower than \( \alpha \). As can be seen, for all comparisons, the best algorithms were those trained with the data mapped into a dissimilarity space.

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  |

C. DATA COMPLEXITY IN DISSIMILARITY REGRESSION DATA SETS

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

VI. CONCLUSIONS 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 representation in classification problems, showing that dissimilarity-based classifiers may improve their accuracy. In addition, this representation presents important advantages regarding the reduction of some intrinsic data difficulties that allows 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 data sets where each one was transformed into a dissimilarity matrix using the Euclidean distance and a representation set. 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 a representation set of various sizes, it has been proved 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 representation set can result beneficial.
2) The Wilcoxon’s signed-rank test with \( \alpha = 0.05 \) validates our claims 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 data sets 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

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

| Data sets          | 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          |
| Callhousing        | 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          |
the selection of representative objects specifically focused on regression tasks. Although it has been claimed that the Euclidean distance is suitable for the 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 to extend 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 data sets, 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,” Computers & Geosciences, vol. 85, pp. 86–95, 2015.

[2] J. Calvo-Zaragoza, J. Valero-Mas, and J. Rico-Juan, “Prototype generation on structural data using dissimilarity space representation,” Neural Computing and Applications, vol. 28, pp. 2415–2424, 2017.

[3] R. P. Duin and E. Pękalska, “The dissimilarity space: Bridging structural and statistical pattern recognition,” Pattern Recognition Letters, vol. 33, no. 7, pp. 826–832, 2012.

[4] A. R. Webb and K. D. Copsey, Statistical Pattern Recognition, 3rd ed. John Wiley & Sons, 2011.

[5] V. M. G. Costa, D. Bartolini, A. S. Britto Jr, G. D. C. Cavalcanti, and L. E. S. Oliveira, “The dissimilarity approach: a review,” Artificial Intelligence Review, vol. 53, pp. 2783–2808, 2020.

[6] L. Nanni, G. Minchic, S. Brahman, G. Maguolo, and A. Luminii, “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 dissimilarity: multiclass learning in similarity-based pattern recognition,” in Similarity-Based Pattern Recognition, M. Pelillo and E. R. Hancock, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 222–234.

[8] R. P. Duin and E. Pękalska, “The dissimilarity representation for non-euclidean pattern recognition, a tutorial,” Electrical Engineering, Mathematics and Computer Sciences, Delft University of Technology, The Netherlands and School of Computer Science, University of Manchester, UK. Tech. Rep., November 2011.

[9] E. Pękalska and R. P. W. Duin, “Dissimilarity representations allow for building good classifiers,” Pattern Recognition Letters, vol. 23, no. 8, pp. 943–956, 2002.

[10] V. García, A. I. Marqués, and J. S. Sánchez, “Dissimilarity-based linear models for corporate bankruptcy prediction,” Computational Economics, vol. 53, pp. 1019–1031, 2018.

[11] E. Pękalska, P. Paclik, and R. P. W. Duin, “A generalized kernel approach to dissimilarity-based classification,” Journal of Machine Learning Research, vol. 2, pp. 175–211, 2002.

[12] V. García, J. S. Sánchez, H. J. Ochoa Domínguez, and L. Cleofas-Sánchez, “Dissimilarity-based learning from imbalanced data with small dissimilarities and noise,” in Pattern Recognition and Image Analysis, R. A. Paredes, J. S. Caudillo, and X. M. Pardo, Eds. Cham: Springer International Publishing, 2015, pp. 370–378.

[13] E. Bruno, N. Moenene-Loccoz, and S. Marchand-Maillet, “Design of multimodal dissimilarity spaces for retrieval of video documents,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 9, pp. 1520–1533, 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,” Journal of Chemometrics, vol. 25, no. 9, pp. 476–486, 2011.

[15] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, “Pose-based human action recognition via sparse representation in dissimilarity space,” Journal of Visual Communication and Image Representation, vol. 25, no. 1, pp. 12–23, 2014.

[16] S. Mauzer, J. Sweeney, and J. McDermott, “Dissimilarity-based representations for one-class classification on time series,” Pattern Recognition, vol. 100, p. 107122, 2020.

[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,” Machine Vision and Applications, vol. 26, pp. 279–293, 2015.

[18] V. García and J. S. Sánchez, “Mapping microarray gene expression data into dissimilarity spaces for tumor classification,” Information Sciences, vol. 294, pp. 362–375, 2015.

[19] R. H. Pinheiro, G. D. Cavalcanti, and I. R. Tsang, “Combining dissimilarity spaces for text categorization,” Information Sciences, vol. 406–407, pp. 87–101, 2017.

[20] J. Jaramillo-Garzon, A. Quiceno-Manrique, I. Godino-Llorente, and C. Castellanos-Dominguez, “Feature extraction for murmur detection based on support vector regression of time-frequency representations,” in Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vancouver, Canada, 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 bio-chemical data,” in Proceedings of the 23rd Biennial American Congress on Pattern Recognition, Madrid, Spain, 2018, pp. 64–71.

[22] A. C. Lorena, A. I. Maciel, P. B. C. De Miranda, I. G. Costa, and Prudêncio, “Data complexity meta-features for regression problems,” Machine Learning, vol. 107, pp. 299–246, 2018.

[23] E. Pękalska, R. P. W. Duin, and P. Paclík, “Prototype selection for dissimilarity-based classifiers,” Pattern Recognition, 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 Proceedings of the 16th International Conference on Pattern Recognition, vol. 3. Quebec, Canada: IEEE Computer Society, 2002, pp. 37–40.

[25] V. García and J. S. Sánchez, “The Dissimilarity Representation for Pattern Recognition: Foundations and Applications,” River Edge, NJ: World Scientific Publishing, 2005.

[26] M. Lang, M. Binder, J. Richter, P. Schratz, F. Pfisterer, S. Coors, Q. Au, G. Casalichio, 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 Proceedings of the International Conference on Sustainable Energy, Electronics, and Computing Systems, Greater Noida, India, 2018, pp. 1–6.

[28] Z. Zhang, J. He, G. Gao, and Y. Tian, “Bi-sparse optimization-based least squares regression,” Applied Soft Computing, vol. 77, pp. 300–315, 2019.

[29] J. J. Rodríguez, M. Juez-Gil, A. Arnaiz-González, and L. I. Kuncheva, “An experimental evaluation of mixup regression forests,” Expert Systems with Applications, vol. 151, p. 113736, 2020.

[30] A. Tsanas and A. Xifara, “Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools,” Energy and Buildings, vol. 49, pp. 560–567, 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,” Journal of Intelligent Manufacturing, vol. 30, pp. 2535–2544, 2019.

[32] M. Mas, J. J. Rodríguez, A. Ariza-González, and L. I. Kuncheva, “An experimental evaluation of mixup regression forests,” Expert Systems with Applications, vol. 151, p. 113736, 2020.

[33] M. H. Rafei and H. Adeli, “A novel machine learning model for estimation of sale prices of real estate units,” Journal of Construction Engineering and Management, vol. 142, pp. 1–18, 2016.

[34] J. A. Nelder and R. W. M. Wedderburn, “Generalized linear models,” Journal of the Royal Statistical Society. Series A (General), vol. 135, no. 3, pp. 370–384, 1972.
Vicente García et al.: Revisiting the Dissimilarity Representation in the Context of Regression

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

[35] L. Torgo, *Data Mining with R - Learning With Cases*, 2nd ed. CRC Press, 2017.

[36] M. Basu and T. K. Ho, *Data Complexity in Pattern Recognition (Advanced Information and Knowledge Processing)*. Berlin, Heidelberg: 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 Analysis and Applications*, vol. 21, pp. 941–951, 2018.

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

VICENTE GARCÍA (M’16) received a B.E. degree in Computer Systems from the Instituto Tecnológico de Villahermosa, México in 2000, a M.S. degree in Computer Sciences from the Instituto Tecnológico de Toluca, México, in 2002 and a 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 interest includes 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 a B.Sc. degree in Computer Science from the Universidad Politécnica de Valencia, Spain, in 1990, and a Ph.D. degree in Computer Science Engineering from Universitat Jaume I, Castelló de la Plana, Spain, in 1998. In 2006, he was awarded a two-year research fellowship to work at University of Wales, Bangor, UK 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 co-author 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 focus in 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).

**Luis C. Méndez-González** (M’20) has a degree in Electronic Engineering (2007) and a Master’s degree in Industrial Engineering from the Technological Institute of Ciudad Juárez (2011), México. In 2015, He received a Doctorate of Science degree in Engineering from the Autonomous University of Ciudad Juárez. He has carried out research and exchange stays at New Mexico State University in 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. His current research is related to time-varying voltage scenarios induced on power lines and their effects on electrical, electronic devices. Now, he is teaching reliability, hardware design, robotics, maintenance, and control as a full-time Professor at the Autonomous University of Ciudad Juárez.

*****

LUIS C. MÉNDEZ-GONZÁLEZ (M’20) received a Ph.D. degree from the Technical University of Catalonia in 2010. Since 2016, he is an associate professor in the Facultad de Tecnologías de Información, Universidad de la Salle Bajo, León, México. He has served as TPC member of many international conferences and workshops. He is a member of the National System of Researchers (SNI) of the National Mexican Science Council (CONACYT). His research interests include cyber security, risk management and machine learning.