Unsupervised Parameter Estimation of Non Linear Scaling for Improved Classification in the Dissimilarity Space

Mauricio Orozco-Alzate[1(✉)], Robert P.W. Duin[2], and Manuele Bicego[3]

[1] Departamento de Informática y Computación, Universidad Nacional de Colombia, Sede Manizales, km 7 vía al Magdalena, Manizales 170003, Colombia
morozcoa@unal.edu.co
[2] Pattern Recognition Laboratory, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
r.p.w.duin@tudelft.nl
[3] Dipartimento di Informatica, Università degli Studi di Verona, Cá Vignal 2, Strada le Grazie 15, Verona 37134, Italy
manuele.bicego@univr.it

Abstract. The non-linear scaling of given dissimilarities, by raising them to a power in the (0,1) interval, is often useful to improve the classification performance in the corresponding dissimilarity space. The optimal value for the power can be found by a grid search across a leave-one-out cross validation of the classifier: a procedure that might become costly for large dissimilarity matrices, and is based on labels, not permitting to capture the global effect of such a scaling. Herein, we propose an entirely unsupervised criterion that, when optimized, leads to a sub-optimal but often good enough value of the scaling power. The criterion is based on a trade-off between the dispersion of data in the dissimilarity space and the corresponding intrinsic dimensionality, such that the concentrating effects of the power transformation on both the space axes and the spatial distribution of the objects are rationed.

Keywords: Dissimilarity space · Intrinsic dimensionality · Dispersion · Non linear scaling · Nearest neighbor classification · Power transformation

1 Introduction

In statistical pattern recognition, an object is conventionally represented as a vector whose entries correspond to numerical values of its features. Therefore, in such a representation, objects are points in a vector space: the well-known *feature space*. However, this conventional representation is often inconvenient, particularly when the extraction of features from symbolic data (such as graphs and grammars) or from raw sensor measurements (such as signals and images) is difficult or even when it is not clear how to do it in the first place. As an alternative, Pekalska and Duin proposed [13] the option of measuring dissimilarities

A. Robles-Kelly et al. (Eds.): S+SSPR 2016, LNCS 10029, pp. 74–83, 2016.
DOI: 10.1007/978-3-319-49055-7_7

between pairs of objects and organizing them as vectors such that each object is represented as a point in the so-called *dissimilarity space* [7] where any classifier can be trained and applied. The dissimilarity representation is within the field of (dis)similarity pattern recognition that has been actively researched during the last years [14,15].

In many pattern classification problems it is mandatory to normalize the feature space, i.e. to make comparable the ranges of the different features that can derive from different measures/sensors: the typical approach is to apply a *linear* scaling to the axes of the vector space; this operation guarantees that the classifier decision equally takes into account values in all directions, once the unwanted influences of their original dynamic ranges have been removed. In the dissimilarity space, in contrast, range differences among the directions tend to be less notorious because all the features are of the same nature, i.e. they are all distances to the objects of the reference group, formally called the *representation set*. For this reason linear scaling is less crucial. However, other more complex scaling operations, such as those involving *non linear transformations*, can be very useful and lead to improvements in the classification – this has been suggested also for classical feature spaces [1,4,5,11,17]. For dissimilarity spaces, Duin et al. [6] found that the non-linear scaling of given dissimilarities by their power transformation appears to be useful for improving the nearest neighbor performance in the dissimilarity space. They studied its behavior in terms of classification error and found that raising dissimilarities to powers less than 1 often contributes to such an improvement. When trying to explain the phenomenon, they suggested that the benefits derive from the following three properties: when applying a power transformation with power less than 1, (i) objects tend to be equally distant from the others, (ii) distances to outliers are shrunk, and (iii) the neighborhood of each object is enlarged by emphasizing distances between close objects.

In their study, as well as in the others related to classical feature spaces cited above [4,5,11,17], the estimation of the proper power parameter represents a crucial open issue; typically such parameter is set by hand, or found by an exhaustive search; in [6] it is estimated via the computationally prohibitive cross validation. In this paper we propose a novel unsupervised criterion which can guide the selection of the parameter of the power transformation: this criterion tries to find a compromise – as the power parameter approaches to zero – between the reduction in the dispersion in the data and the increase in the intrinsic dimensionality of the resulting dissimilarity space (if a too small power is applied all points are converging around 1). This criterion is unsupervised – since it does not require labels – and computationally more feasible than cross validation – since it does not require repeated training of classifiers. A thorough experimental evaluation on several different datasets shows that by applying the power transformation with the best parameter according to the proposed criterion we obtain accuracies which are (i) almost always significantly better than those obtained in the space without the preprocessing and (ii) many times equivalent or better than those obtained by the computationally expensive cross validation procedure.

The rest of the paper is organized as follows: in Sect. 2 we briefly summarize the dissimilarity space and the non linear scaling by power transformation; then, in Sect. 3 we detail the proposed approach; the experimental evaluation is presented in Sect. 4; finally, in Sect. 5, conclusions are drawn and future perspectives are envisaged.

2 Background

2.1 The Dissimilarity Space

The vector arrangement of the dissimilarities computed between a particular object x and other objects from a set $\mathcal{R}$ allows representing x as a point in a vector space. Such a space is called the *dissimilarity space*, having in principle as many dimensions as the cardinality of $\mathcal{R}$, which is known as the *representation set*. For a set of training objects $\mathcal{T}$, the set $\mathcal{R}$ builds a so-called *dissimilarity representation* in the form of a dissimilarity matrix $\mathbf{D}(\mathcal{T}, \mathcal{R})$. The representation set is often the same as the training set, so $\mathbf{D}(\mathcal{T}, \mathcal{R}) = \mathbf{D}(\mathcal{T}, \mathcal{T})$. For notation simplicity, hereafter we simply use $\mathbf{D}$ to refer to the square dissimilarity matrix $\mathbf{D}(\mathcal{T}, \mathcal{T})$.

Several studies [7,13] have shown the possibilities of training classifiers in the dissimilarity space, such that a test object represented in terms of its dissimilarities to $\mathcal{R}$ can be classified by a more sophisticated rule than the nearest neighbor classifier on the given dissimilarities (i.e. template matching, denoted as 1-NN). The classifier in the dissimilarity space can even be the same nearest neighbor rule but now based on distances between points in the dissimilarity space; here we denote that case as 1-NND in order to distinguish it from template matching.

2.2 Non Linear Scaling

Raising all dissimilarities to the same power is a simple and straightforward non linear scaling. For a dissimilarity matrix $\mathbf{D}$, such a transformation can be written as follows:

$$\mathbf{D}^{\star\rho} = (d_{ij}^{\rho}), \qquad \rho > 0 \tag{1}$$

where each entry, $d_{ij} = d(x_i, x_j)$, of the matrix denotes the dissimilarity between two objects x_i and x_j and * denotes the entrywise (Hadamard) power function [9]. There exists an optimal value for ρ that provides the best 1-NND classification performance. Let's denote it as ρ^*. In most cases, ρ^* is lower than 1. This is reasonable, since with $\rho < 1$ we have a concave function that raises low values and shrinks high values: for dissimilarities, this may have a good impact on the representation in the dissimilarity spaces, since it reduces the impact of outliers (large distances are reduced) and increases the importance of the neighborhood (small distances are increased).

Therefore, we only consider to search for an estimate $\widehat{\rho^*}$ in the interval $(0, 1]$. Below we explain the existing method to estimate ρ^* by cross validation, followed in Sect. 3 by the explanation of our proposed estimation via the optimization of an unsupervised criterion.

Optimization via Cross Validation. A typical procedure to optimize the value of a parameter is by searching over the parameter domain for the lowest cross validation classification error. This strategy was the one used by Duin et al. [6] for finding the best parameter, which we call in this case $\widehat{\rho}^*_{cv}$, as follows:

$$\widehat{\rho}^*_{cv} = \arg\min_{\rho \in (0,1]} \epsilon_{1-NND}(\mathbf{D}^{\star\rho}) \tag{2}$$

where ϵ_{1-NND} denotes the leave-one-out cross validation error of 1-NND. Even though experiments in [6] suggested that this optimization permits a good classification performance, it might become computationally prohibitive for large datasets. Moreover, such criterion does not permit to understand what is happening with the non linear scaling, i.e. it does not provide an explanation of the topological effect of the parameter value in the space.

3 The Proposed Criterion

As introduced before, when applying a power transformation with $\rho < 1$, we obtain a two-fold effect on data in the dissimilarity space. First, the dispersion of the values in each dimension of the space is shrunk (by raising small distances and reducing large distances); second, the neighborhood of each point is highly emphasized (raising small distances). This behaviour is becoming more and more extreme when ρ approaches zero. Clearly, up to some extent these effects are desirable, in order to reduce the impact of outliers (distances to far away points are reduced) and to better characterize the neighborhood of each object (distances to nearby points are raised); however, after a certain point such positive effects are lost, since all points tend to be equally spaced in the space, thus loosing all the information contained in the original dissimilarity matrix. This effect can be monitored by looking at the intrinsic dimensionality of the data, which increases when points tend to be more equally spaced. Therefore, using a criterion that optimizes a trade-off between those two effects (reduction of dispersion and increase of the intrinsic dimensionality) seems a reasonable way to find $\widehat{\rho}^*$.

Among the available dispersion measures, the quartile coefficient of dispersion (*qcd*) [10, p. 15] is a robust statistical estimator that gives a scale-free measure of data spread. It is given as:

$$qcd = \frac{Q_3 - Q_1}{Q_3 + Q_1}, \tag{3}$$

where Q_3 and Q_1 are the third and first quartiles, respectively. In our case, they are computed as follows: for each column (dimension) of $\mathbf{D}^{\star\rho}$, we find the median of the upper half of the values (which is Q_3, also called the 75th percentile) and the median of the lower half of them (which is Q_1, also called the 25th percentile).

Similarly, there are many methods to estimate the intrinsic dimensionality (*id*) of a dataset, see for instance the reviews by Camastra [2,3]. We have a

adopted the one described in [13, p. 313] which directly computes the estimation from dissimilarity data:

$$\widehat{id}(\mathbf{D}) = \left\lceil 2\frac{(\mathbf{1}^\top \mathbf{D}^{\star 2}\mathbf{1})^2}{n(n-1)\mathbf{1}^\top \mathbf{D}^{\star 4}\mathbf{1} - (\mathbf{1}^\top \mathbf{D}^{\star 2}\mathbf{1})^2} \right\rceil \tag{4}$$

where $\mathbf{D}^{\star 2} = (d_{ij}^2)$, $\mathbf{D}^{\star 4} = (d_{ij}^4)$ and n is the number of columns (and rows) of the square matrix $\mathbf{D}$.

Given these definitions, our criterion tries to determine the best parameter (which we call $\widehat{\rho}^*_{nlm}$) by optimizing the compromise between (i) the average – or, better, its robust estimate, the median – of the dispersion (3) per dimension and (ii) the intrinsic dimension of (4) computed for the pairwise distances in the dissimilarity space, that is, for a matrix of Euclidean distances $\mathbf{D}_{DS}$ between pairs of points in the dissimilarity space. The final criterion can be written as:

$$\widehat{\rho}^*_{nlm} = \underset{\rho \in (0,1]}{\arg\min} \left[\underset{1 \le i \le n}{\text{median}} \left(qcd_i\right) \times \widehat{id}\left(\mathbf{D}^{\star\rho}_{DS}\right) \right] \tag{5}$$

Notice that, even though there are several alternatives to define a compromise between two variables, we have chosen to minimize the product between them. A multiplicative criterion has also been adopted in other scenarios [8,12] where the two variables of interest are related in a non-trivial way.

3.1 Inductive and Transductive Versions

The criterion introduced in the previous section is completely unsupervised: exploiting this property, we investigate its usefulness in two different flavours, which we called "Version 1" and "Version 2", respectively:

1. Version 1 ($\widehat{\rho}^*_{nlm1}$): the best parameter is the one optimizing the proposed criterion on the training set: this represents the classical learning, also known as inductive inference [16, p. 577], where the criterion is determined by using only the training objects.
2. Version 2 ($\widehat{\rho}^*_{nlm2}$): the best parameter is the one optimizing the proposed criterion on the whole dataset, clearly by ignoring the labels. This represents the so called transductive learning [18] where all the available objects are used: the training objects, for which we can employ the labels, and the testing objects, for which labels are unknown. Since the proposed criterion does not take into account the labels, the transductive learning can be applied.

4 Experimental Results

The proposed approach has been tested using a set of public domain datasets[1] (also employed in [6]) – see Table 1. Most of them are derived from real objects (images, text, protein sequences). The Chickenpieces dataset consists out of 44

[1] More information on datasets can be found at http://37steps.com/prdisdata.

Table 1. Datasets employed for empirical evaluation.

Name	Objects	Classes
(1) Catcortex	65	4
(2) Coildelftdiff	288	4
(3) Coildelftsame	288	4
(4) Coilyork	288	4
(5) Delftgestures	1500	20
(6) Flowcytodis1	612	3
(7) Flowcytodis2	612	3
(8) Flowcytodis3	612	3
(9) Flowcytodis4	612	3
(10) Newsgroups	600	4
(11) Prodom	2604	4
(12) Protein	213	4
(13) Woodyplants50	791	14
(14) Zongker	2000	10
(15) Chickenpieces (44 sets)	446	5
(16) Polydish57	4000	2
(17) Polydism57	4000	2

dissimilarity matrices: in the tables, the average characteristics are shown. In our empirical evaluation we compared the errors made by the Nearest Neighbor rule[2] (errors of 1-NND) in four different versions of the dissimilarity space:

1. *Original*: this is unprocessed case (no transformation is applied), i.e. the dissimilarity space is built using the original dissimilarity matrix $\mathbf{D}$.
2. *NL-Cross Val*: in this case the dissimilarity space is built starting from $\mathbf{D}^{\star\widehat{\rho^*_{cv}}}$, i.e. after applying a non linear transformation where the optimal parameter is chosen by optimizing the LOO error on the training set. As said before, this represents the criterion proposed in [6].
3. *NL-Disp (ver. 1)*: in this case the dissimilarity space is built starting from $\mathbf{D}^{\star\widehat{\rho^*_{nlm1}}}$, i.e. after applying non linear transformation with parameter chosen by optimizing the proposed criterion on the training set.
4. *NL-Disp (ver. 2)*: in this case the dissimilarity space is built starting from $\mathbf{D}^{\star\widehat{\rho^*_{nlm2}}}$, i.e. after applying a non linear transformation with parameter chosen by optimizing the proposed criterion on the whole dataset (in a transductive way, see previous section).

[2] We restrict ourselves to using a parameterless classifier – the nearest neighbor rule – because we are interested in judging the potential improvement of the data representation after the power transformation, independently from the influence of any classifier parameter.

Table 2. 1NN-D errors for the different datasets. Between brackets we reported the standard errors of the mean.

Dataset	Original	NL-Cross Val	NL-Disp (v1)	NL-Disp (v2)
Catcortex	0.1067(7e-03)	0.1057(7e-03)	0.1012(6e-03)	0.0981(7e-03)
Coildelftdiff	0.4611(2e-03)	0.4498(2e-03)	0.4575(2e-03)	0.4528(2e-03)
Coildelftsame	0.4181(2e-03)	0.4130(2e-03)	0.4158(2e-03)	0.4102(2e-03)
Coilyork	0.3948(2e-03)	0.3265(2e-03)	0.3532(2e-03)	0.3371(2e-03)
Delftgestures	0.0949(2e-04)	0.0526(2e-04)	0.0599(2e-04)	0.0563(2e-04)
Flowcytodis1	0.3857(9e-04)	0.3797(9e-04)	0.3781(9e-04)	0.3770(9e-04)
Flowcytodis2	0.3827(9e-04)	0.3749(1e-03)	0.3754(9e-04)	0.3730(1e-03)
Flowcytodis3	0.4077(9e-04)	0.3911(9e-04)	0.3890(9e-04)	0.3850(9e-04)
Flowcytodis4	0.4251(9e-04)	0.4127(9e-04)	0.4109(8e-04)	0.4083(9e-04)
Newsgroups	0.2960(9e-04)	0.2915(9e-04)	0.2887(9e-04)	0.2887(9e-04)
Prodom	0.0193(9e-05)	0.0072(6e-05)	0.0065(6e-05)	0.0065(6e-05)
Protein	0.0059(6e-04)	0.0063(7e-04)	0.0062(6e-04)	0.0055(6e-04)
Woodyplants50	0.1617(5e-04)	0.1188(5e-04)	0.1379(5e-04)	0.1292(5e-04)
Zongker	0.0529(1e-04)	0.0408(2e-04)	0.0377(2e-04)	0.0377(2e-04)
Chickenpieces	0.1543(1e-04)	0.1252(1e-04)	0.1307(1e-04)	0.1263(1e-04)
Polydish57	0.0306(5e-05)	0.0166(4e-05)	0.0233(4e-05)	0.0233(4e-05)
Polydism57	0.0153(4e-05)	0.0135(3e-05)	0.0226(5e-05)	0.0226(5e-05)

Errors have been computed using averaged hold out cross validation, i.e. by using half of the dataset for training (and representation) and the remaining half for testing. In order to ensure robust estimation of errors, this procedure has been repeated 200 times, and results are averaged. For criteria 2–4, the best value has been chosen in the range $1.25^{-15}, 1.25^{-14.5}, 1.25^{-14}, ..., 1$ for the exponent. Averaged errors, together with standard errors of the mean, are reported in Table 2. In order to get a more direct view on the results, we reported in Table 3 an improvement/degradation table, as resulting from several different pairwise statistical tests. In particular, we compared errors obtained with the proposed criterion (NL-Disp in both versions v1 and v2) with those obtained without transforming the space (Original) and with the parameter chosen via Cross Validation (NL-Cross Val). As statistical test we employed the paired t-test, comparing the 200 errors obtained with the 200 repetitions of the cross validation. In the table, we used five different symbols:

– the symbols "↑" and "↑↑" indicate a statistically significant improvement (results with our criterion are better): the former indicates that the test passed with a p-value less than 0.05 but greater than 0.001, whereas in the latter case the p-value was less than 0.001;

Table 3. Pairwise statistical comparisons: "↑" indicates a statistically significant improvement (results with our criterion are better), "↓" a statistically significant degradation (results with our criterion are worst), whereas "≈" indicates that the two methods are equivalent (i.e. there is no statistically significant difference).

Dataset	NL-Disp (v1) *vs* Original	NL-Disp (v2) *vs* Original	NL-Disp (v1) *vs* NL-Cross Val	NL-Disp (v2) *vs* NL-Cross Val
Catcortex	↑	↑↑	↑	↑↑
Coildelftdiff	↑↑	↑↑	↓↓	≈
Coildelftsame	↑	↑↑	↓	↑
Coilyork	↑↑	↑↑	↓↓	↓↓
Delftgestures	↑↑	↑↑	↓↓	↓↓
Flowcytodis1	↑↑	↑↑	≈	↑
Flowcytodis2	↑↑	↑↑	≈	≈
Flowcytodis3	↑↑	↑↑	≈	↑↑
Flowcytodis4	↑↑	↑↑	≈	↑↑
Newsgroups	↑↑	↑↑	↑	↑
Prodom	↑↑	↑↑	↑↑	↑↑
Protein	≈	≈	≈	≈
Woodyplants50	↑↑	↑↑	↓↓	↓↓
Zongker	↑↑	↑↑	↑↑	↑↑
Chickenpieces	↑↑	↑↑	↓↓	↓↓
Polydish57	↑↑	↑↑	↓↓	↓↓
Polydism57	↓↓	↓↓	↓↓	↓↓

– "↓" and "↓↓" indicate a statistically significant degradation (results with our criterion are worst); also in this case the former indicates that the test passed with a p-value less than 0.05 but greater than 0.001, whereas in the latter case the p-value was less than 0.001;
– "≈" indicates that the two methods are equivalent (i.e. there is no statistically significant difference).

From the table different observations can be derived. First, as expected, the transductive version (version 2) of our criterion is almost always slightly better than version 1; this interesting result is possible thanks to the unsupervised nature of the proposed criterion. Reasonably, this does not hold if the dataset is large enough (as for Zongker, Polydish57 and Polydism57). Second, non linearly preprocessing the dissimilarity matrix by choosing the parameter with our criterion almost always results in a statistically significant improvement in the classification performances with respect to the original space. The only exceptions are for the protein and the Polydism57 datasets, for which, however, an almost zero error was already achieved in the original space, leaving small room for

improvements. This is coherently true for both version 1 and version 2. Finally, the proposed criterion also compares reasonably well with the cross validation approach: if we consider the version 2, in 11 cases out of 17 our results are better or equivalent (in 8 cases they are significantly better), whereas only in 6 cases they are worst. In these latter cases, however, degradations are very small: ≈0.01 for CoilYork, WoodyPlants50, Polydism57 and Polydish57, ≈0.004 for DelftGestures, and ≈0.001 for ChickenPieces. We are convinced that these represent really promising results, also considering that our criterion is completely unsupervised.

5 Conclusions

In this paper a novel unsupervised criterion to tune the parameter of the power transformation (non-linear scaling) of dissimilarities has been proposed. The new tuning criterion is based on a trade-off between the median dispersion per dimension in the dissimilarity space (measured in terms of the quartile coefficient of dispersion) and the intrinsic dimension of the resulting dissimilarity space. The idea behind our approach is that a good performance of the nearest neighbor classifier in the dissimilarity space is associated to such a compromise between how much we shrink the data at the cost of increasing the intrinsic dimensionality – the shrinking is desirable because, by reducing the range, we can potentially reduce the influence of the outliers since we are largely reducing high distances (i.e. the distances to – possible – outliers) more than reducing short distances.

The proposed criterion is unsupervised and, therefore, can be even applied in a transductive learning setting. Empirical results on many different datasets partially support our intuitions. As a future work, we would like to study the properties of the proposed criterion also in classical feature based problems [4,5,11,17]. Moreover, we aim at providing a more formal – theoretical or numerical – explanation: one possibility is to try to bridge our experimental evidence with the theory on Hadamard powers [9].

Acknowledgments. Discussions for the proposal in this paper started while Mauricio Orozco-Alzate and Manuele Bicego visited the Pattern Recognition Laboratory, Delft University of Technology (Delft, The Netherlands) in September 2015 by a kind invitation from Robert P.W. Duin to attend the "Colors of dissimilarities" workshop.

This material is based upon work supported by Universidad Nacional de Colombia under project No. 32059 (Code Hermes) entitled "*Consolidación de las líneas de investigación del Grupo de Investigación en Ambientes Inteligentes Adaptativos GAIA*" within "Convocatoria interna de investigación de la Facultad de Administración 2015, para la formulación y ejecución de proyectos de consolidación y/o fortalecimiento de los grupos de investigación. Modalidad 1: Formulación y ejecución de proyectos de consolidación".

The fist author also acknowledges travel funding to attend S+SSPR 2016 provided by Universidad Nacional de Colombia through "Convocatoria para la Movilidad Internacional de la Universidad Nacional de Colombia 2016–2018. Modalidad 2: Cofinanciación de docentes investigadores o creadores de la Universidad Nacional de Colombia

para la presentación de resultados de investigación o representaciones artísticas en eventos de carácter internacional, o para la participación en megaproyectos y concursos internacionales, o para estancias de investigación o residencias artísticas en el extranjero".

References

1. Bicego, M., Baldo, S.: Properties of the Box-Cox transformation for pattern classification. Neurocomputing (2016, in press)
2. Camastra, F.: Data dimensionality estimation methods: a survey. Pattern Recogn. **36**(12), 2945–2954 (2003)
3. Camastra, F., Staiano, A.: Intrinsic dimension estimation: advances and open problems. Inf. Sci. **328**, 26–41 (2016)
4. Carli, A.C., Bicego, M., Baldo, S., Murino, V.: Non-linear generative embeddings for kernels on latent variable models. In: Proceedings of ICCV 2009 Workshop on Subspace Methods, pp. 154–161 (2009)
5. Carli, A.C., Bicego, M., Baldo, S., Murino, V.: Nonlinear mappings for generative kernels on latent variable models. In: Proceedings of International Conference on Pattern Recognition, pp. 2134–2137 (2010)
6. Duin, R.P.W., Bicego, M., Orozco-Alzate, M., Kim, S.-W., Loog, M.: Metric learning in dissimilarity space for improved nearest neighbor performance. In: Fränti, P., Brown, G., Loog, M., Escolano, F., Pelillo, M. (eds.) S+SSPR 2014. LNCS, vol. 8621, pp. 183–192. Springer, Heidelberg (2014). doi:10.1007/978-3-662-44415-3_19
7. Duin, R.P.W., Pekalska, E.: The dissimilarity space: bridging structural and statistical pattern recognition. Pattern Recogn. Lett. **33**(7), 826–832 (2012)
8. Fahmy, A.A.: Using the Bees Algorithm to select the optimal speed parameters for wind turbine generators. J. King Saud Univ. Comput. Inf. Sci. **24**(1), 17–26 (2012)
9. Guillot, D., Khare, A., Rajaratnam, B.: Complete characterization of Hadamard powers preserving Loewner positivity, monotonicity, and convexity. J. Math. Anal. Appl. **425**(1), 489–507 (2015)
10. Kokoska, S., Zwillinger, D.: CRC Standard Probability and Statistics Tables and Formulae, Student edn. CRC Press, Boca Raton (2000)
11. Liu, C.L., Nakashima, K., Sako, H., Fujisawa, H.: Handwritten digit recognition: investigation of normalization and feature extraction techniques. Pattern Recogn. **37**(2), 265–279 (2004)
12. Mariani, G., Palermo, G., Zaccaria, V., Silvano, C.: OSCAR: an optimization methodology exploiting spatial correlation in multicore design spaces. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **31**(5), 740–753 (2012)
13. Pekalska, E., Duin, R.P.W.: The Dissimilarity Representation for Pattern Recognition: Foundations and Applications (Machine Perception and Artificial Intelligence), vol. 64. World Scientific, Singapore (2005)
14. Pelillo, M. (ed.): Similarity-Based Pattern Analysis and Recognition. Advances in Computer Vision and Pattern Recognition. Springer, London (2013)
15. Pelillo, M., Hancock, E.R., Feragen, A., Loog, M. (eds.): SIMBAD. LNCS, vols. 7005, 7953, 9370. Springer, Heidelberg (2011, 2013, 2015)
16. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 4th edn. Academic Press, London (2009)
17. Van Der Heiden, R., Groen, F.C.A.: The Box-Cox metric for nearest neighbour classification improvement. Pattern Recogn. **30**(2), 273–279 (1997)
18. Vapnik, V.N.: Statistical Learning Theory. Adaptive and Learning Systems for Signal Processing, Communication and Control. Wiley, New York (1998)