Features Transformation and Normalization: A Visual Approach (CDT-24)

Preprint · March 2020

DOI: 10.13140/RG.2.2.33287.34727

1 author:

Luciano da F. Costa
University of São Paulo
734 PUBLICATIONS **13,304** CITATIONS

SEE PROFILE

Features Transformation and Normalization: A Visual Approach (CDT-24)

Luciano da Fontoura Costa
luciano@ifsc.usp.br

São Carlos Institute of Physics – DFCM/USP

v.2 : 28th March 2020
v.1 : 23rd March 2020

Abstract

Pattern recognition involves extracting features, transforming/normalizing them, and them performing classification in order to assign new or predefined categories to the entities of interest. In this work, we focus on the transformation and normalization of features. The concept of morphing a set of points and visualizations of respective displacement fields are introduced and applied in order to visualize and better understand the possible effects of feature transformation and normalization. When applied with care, these operations have the potential to enhance the classification stage. However, we illustrate that features transformations and normalizations can also create false clusters as well as merge existing clusters, so that special attention is required when performing these operations. Some important features transformations and normalizations, including the standardization procedure as well as principal component analysis and linear discriminant analysis, are also presented and briefly discussed.

"L'occhio non vede cose ma figure di cose che significano altre cose."

Italo Calvino.

1 Introduction

Pattern recognition (e.g. [1, 2, 3]). is much more general and widespread than often realized, as it underlies most of human (as well as other living beings) intelligence and activities. As a consequence, pattern recognition has an ample range of applications, extending from quality control to text and image interpretation.

Figure 1 depicts the main data and operations involved in pattern recognition. First, the patterns (or entities) to be recognized have to be somehow *generated* (e.g. [4]) with basis on respective parameters. Then, a selected set of *features* are extracted from the patterns so as to provide a quantitative characterization of the entities to be recognized (e.g. [3]). This step is particularly challenging, as the selection of features is not straightforward and depends on previous experience with the data, measurements, and classification methods (e.g. [3]). The extracted features can then be *transformed* or *normalized* with the objective of deriving new features capable of improving the entities characterization. Features transformations can take each feature independently or combined with other features. For instance, transformations can be used to remove noise from features, to make them smoother, to obtain more discriminative measurements, to reduce the dimensionality of the feature space, etc. Normalizations are often required in order to remove translation, rotation or scaling effects from the features, providing a more standardized set of features. The final step in our diagram consists of *classifying*, i.e. assigning previously defined or new categories, to the entities with based on their transformed/normalized features.

The present work focuses on the important stage of *features transformation and normalization*. These two tasks, which can have important effects (wanted and unwanted) in pattern recognition, are indeed unavoidable as the very implementation of any measurement implies some choice related to transformation and normalization. For instance, in case we are measuring the weight of fruits to be considered as respective features, we are immediately faced with the question of which *unit* to adopt, such as grams, ounces, etc. In addition, the very definition of

features often involved transformations combining other features, such as the ratio between the standard deviation and mean of a given measurement (known as variation coefficient).

We will start by introducing the concept of morphing a set of points, which will be subsequently applied to illustrate the concept and effects of feature transformation and normalization. Indeed, most feature transformations and normalizations can be conceptualized in terms of morphing and respective vector fields, which help to visualize and understand the respective effects, such as creating false clusters that did not originally exist and merging clusters when they should be separated. The provided examples motivate the need of special care and attention when transforming and normalizing features.

Next, we will discuss independent feature transformation, in which each feature is transformed into a respective new feature without considering other features. Transformations involving the combination of features are covered subsequently. The specially important normalizations involving the minimum/maximum values of the features, as well as the standardization procedure, are then presented and illustrated. The interesting feature normalizations known as Principal Component Analysis (PCA), which is an unsupervised methodology, as well as Linear Discriminant Analysis (LDA), a supervised method, are then introduced and discussed. Both the latter normalizations involve linear combination of the original features.

2 Morphing a Set of Points

The term *morphing* can be understood as changing the *shape* of a given geometric structure, such as the mirrors in amusement parks. In this section we present the concept of morphing a set of N discrete points in $\Re^2$, each of them represented in terms of the respective coordinates (x, y). The extension to higher dimensions is straightforward, though respective visualization of the morphing operation becomes more challenging.

Each given point (x, y) is transformed into a new point $(\tilde{x}, \tilde{y})$ by the morphing operation. We will consider that the morphing is implemented through two respective scalar fields S_x and S_y which are both functions of x and y, i.e.

$$\begin{aligned} \tilde{x} &= S_x(x, y) \\ \tilde{y} &= S_y(x, y) \end{aligned} \tag{1}$$

The above concept can be directly extended to points defined by higher dimensional feature spaces, as is typically the case in pattern recognition:

$$\tilde{S}_{f_i} = S_{f_i}(f_1, f_2, \ldots, f_M) \tag{2}$$

for $i = 1, 2, \ldots, M$ features. For simplicity's sake, we will illustrate the morphing approach with respect to points $[x, y]$ in $\Re^2$.

In other words, the morphing operation moves each original point to a new position in the feature space. The morphing operation can be more conveniently visualized by decomposing it in terms of the original position of the point, i.e. $[x, y]$, and respective displacement $[D_x(x, y), D_y(x, y)]$:

$$\begin{aligned} \tilde{x} &= x + D_x(x, y) \\ \tilde{y} &= y + D_y(x, y) \end{aligned} \tag{3}$$

This can be understood as transforming the action of the scalar fields $S_x(x, y)$ and $S_y(x, y)$ into the effect of a respective *vector field* $[D_x(x, y), D_y(x, y)]$.

Figure 2 illustrates this decomposition.

It follows from Equations 1 and 4 that:

$$\begin{aligned} D_x(x, y) &= S_x(x, y) - x \\ D_y(x, y) &= S_y(x, y) - y \end{aligned} \tag{4}$$

Interestingly, the displacement vector $[D_x(x, y), D_y(x, y)]$ provides an interesting conceptual visualization of the transformation effect, which will be used in this work in order to illustrate the effect of several feature transformation/normalization operations. The following seciton provides some interesting examples of how the application of these operations can substantially change the cluster structure in feature spaces.

3 Features Transformations

Let the points $[x, y]$ be transformed by the following displacement field:

$$\begin{aligned} \tilde{x} &= D_x(x, y) = -x^2 \; sgn(x) \\ \tilde{y} &= D_y(x, y) = -y^2 \; sgn(y) \end{aligned} \tag{5}$$

Figure 3 depicts the action of the above point transformation on a circularly bound set of uniformly distributed points. As summarized by the visualization of the applied displacement field, the original points moved toward the center of the coordinate system. This happened more noticeably with the points near the circular border, as the magnitude of the applied field is stronger at those positions, see Fig. 3(b). As a consequence, the density

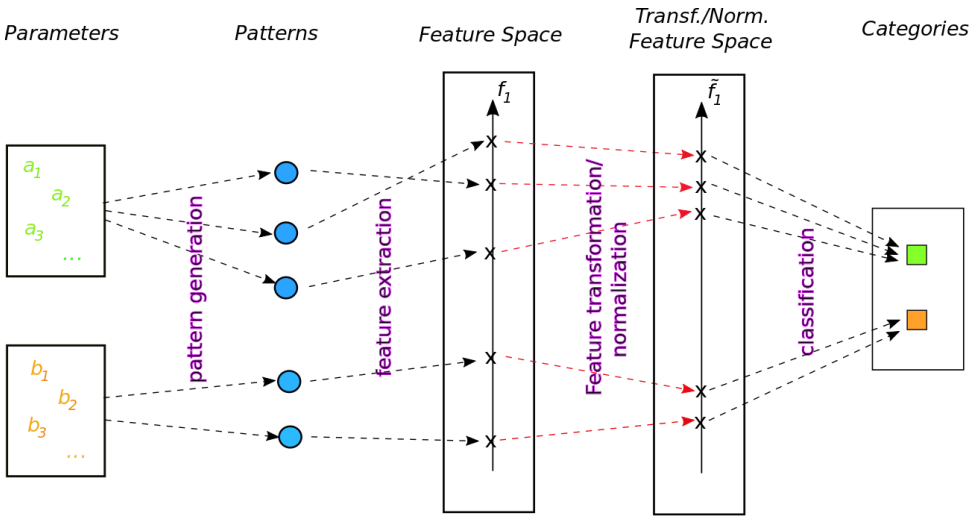

Figure 1: The several stages involved in the endeavor of pattern recognition. First, patterns are generated according to respective parameter configurations. Then, features are extracted from those patterns in order to provide respective quantitative representations. Though an one-dimensional feature space f_1 is shown in this diagram for simplicity's sake, higher dimensional spaces are typically involved in pattern recognition. These features can subsequently be transformed and/or normalized (red arrows), which can involve combinations of the previous features, yielding a new feature space $\tilde{f}_1$. The transformed/ normalized features are then fed to a classification method, which will then hopefully provide the respective correct categories. The present work focuses on the feature transformation and normalization stage.

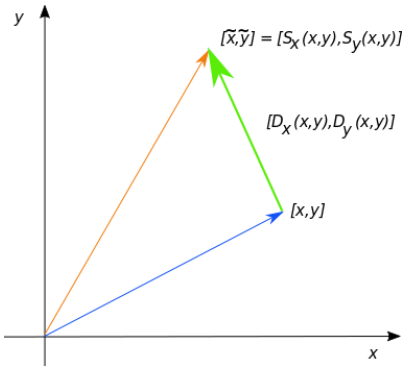

Figure 2: The morphing from point $[x, y]$ (blue vector) into the point $[\tilde{x}, \tilde{y}]$ (orange vector) can be decomposed in terms of the vector sum of the original point position and the respective displacement vector $[D_x(x, y), D_y(x, y)]$ (green vector).

near the the border of the obtained point distribution is found to be higher than that near the origin of the coordinate system, which had its density nearly preserved. The own border of the region where the points were originally contained changed shape as a consequence of the transformation.

This first example corroborates the fact that applying transformations to features can have substantial effects on the obtained density and distribution of points.

Figure 4 illustrates another important effect that can be observed when applying feature transformations. In this case, the original points were subjected to a displacement field that forced convergence at two distinct centers, namely $[-3, -3]$ and $[3, 3]$. Each of these centers were implied by a respective gaussian distribution of displacement magnitudes.

This examples illustrates the situation in which false clusters can arise as a consequence of feature transformations.

Another effect to be avoided is shown in Figure 5, in which a feature space containing two well-defined clusters has been mapped into a single cluster by applying a displacement field – Fig. 5(c) – that does not depend on y and that vectors with magnitudes defined by a one-dimensional gaussian centered at the origin of the coordinate system.

This situation illustrates that feature transformation/normalization can impact the separation and shape of clusters.

Despite the possible problems illustrated above, features transformations and normalizations are often very useful when applied with due care and attention, as they can help to emphasize clusters, control noise and bias/distortions in the original data, as well as to reduce the dimension of the feature space.

The remainder of this work presents some of feature transformations and normalizations often adopted in pattern recognition, but before that we characterize two main groups of transformations: those depending only on each feature (independent), and those involving combination of features. The general forms of these two types of transformations are given in Equations 6 and 7, respectively.

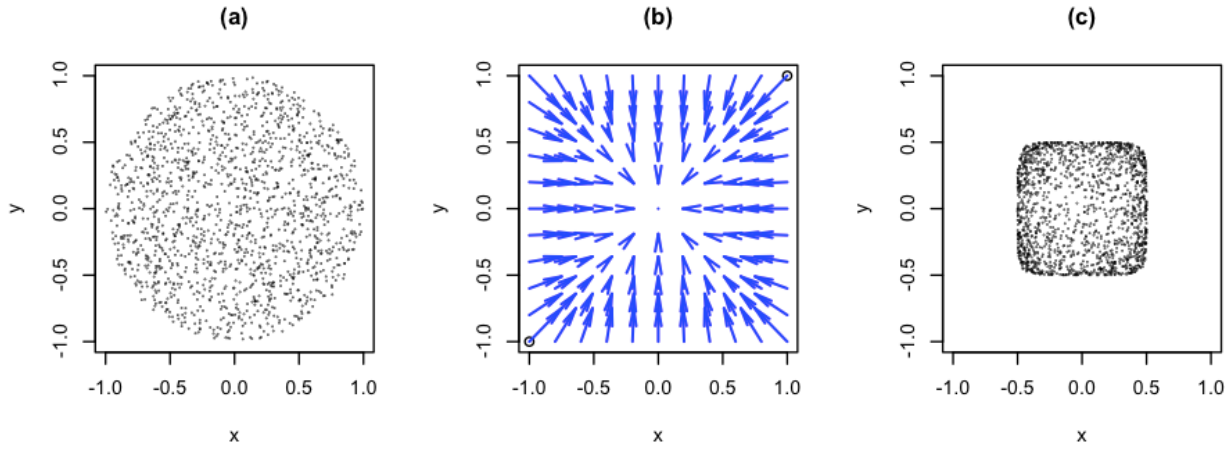

Figure 3: Action of the transformation defined by Eq. 5. The initial set of uniformly distributed points (a) is transformed through the displacement field in (b) into a new set of points (c) that exhibits enhanced density at the respective borders. Observe that this transformation also acts in changing the initially circular shape of points distribution border. The magnitude of the displacement field has been shown to a fraction (0.5) of its original value in order not to clutter the visualization.

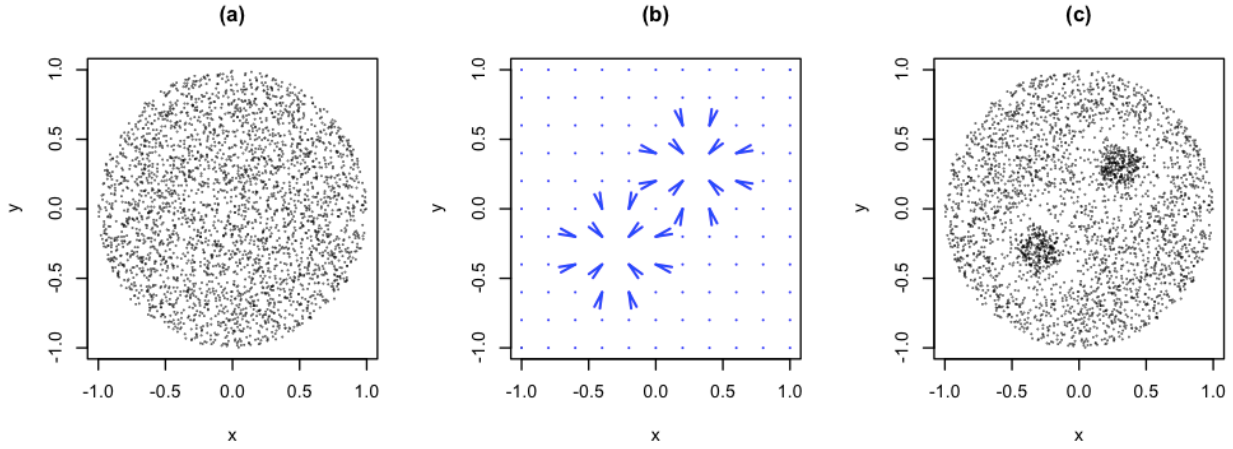

Figure 4: Feature transformations, as illustrated here, can inadvertently create clusters that were not present in the uniformly distributed original data. Each of the concentration centers were implied by respective radial displacement fields (pointing toward the respective center) with gaussian magnitudes.

$$\begin{bmatrix} \tilde{f}_1 \\ \tilde{f}_2 \\ \vdots \\ \tilde{f}_M \end{bmatrix} = \begin{bmatrix} S_1(f_1) \\ S_2(f_2) \\ \vdots \\ S_M(f_M) \end{bmatrix} \tag{6}$$

$$\begin{bmatrix} \tilde{f}_1 \\ \tilde{f}_2 \\ \vdots \\ \tilde{f}_M \end{bmatrix} = \begin{bmatrix} S_1(f_1, f_2, \ldots, f_M) \\ S_2(f_1, f_2, \ldots, f_M) \\ \vdots \\ S_M(f_1, f_2, \ldots, f_M) \end{bmatrix} \tag{7}$$

For instance, the transformation in Equation 5 is an independent transformation. Observe also that the transformation scalar fields, e.g. S_x and S_y, will necessarily be of the same type as the respectively associated displacement vector fields, e.g. D_x and D_y.

4 MinMax Normalization

Given a feature (or measurement) f_i varying in the interval $[f_{i,min}, f_{i,max}]$, a new respective version of this feature $\tilde{f}_i$ varying in the interval $[0, 1]$ can be obtained by applying the following independent feature transformation:

$$\tilde{f}_i = \frac{f_i - f_{i,min}}{f_{i,max} - f_{i,min}} \tag{8}$$

For simplicity's sake, this normalization transformation will be henceforth called *minmax normalization* in the present work.

Figure 6 illustrates the minmax feature normalization with respect to two clusters of uniformly distributed points in (a). The respective displacement, shown in (b), implies a vertical expansion of the clusters, resulting in

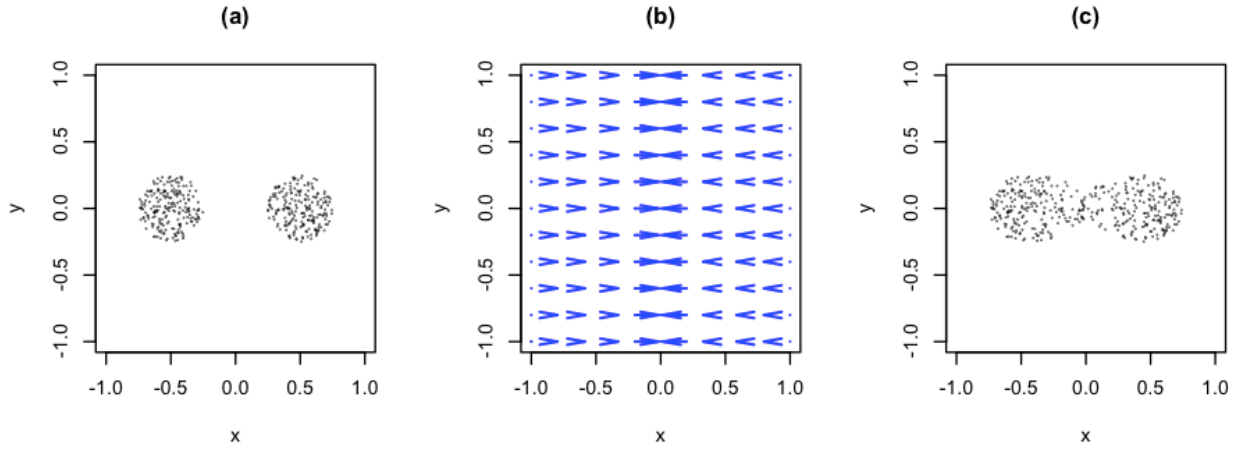

Figure 5: Two well-separated clusters can be merged as a consequence of certain feature transformations such as that applied in this example.

the elongated clusters in (c). This new shape of the two clusters can have impact in the subsequent classification stage.

5 Unit Transformations

We have already briefly discussed in the introduction of this work that the units in which the features are taken can influence the respective classification, eventually requiring transformation and normalization. Given a measurement f_i varying in the interval $[f_{i,min}, f_{i,max}]$, it is possible to *linearly* transform it into another measurement f_j varying in the interval $[f_{j,min}, f_{j,max}]$ by applying the following equation:

$$f_j = (f_{j,max} - f_{j,min}) \frac{f_i - f_{i,min}}{f_{i,max} - f_{i,min}} + f_{j,min} \qquad (9)$$

As illustrated in Figure 7, ihis transformation can be understood as a minmax normalization, yielding the new variation interval $[0, 1]$, followed by a scaling product by $(f_{j,max} - f_{j,min})$ and a translating subtraction by $f_{j,min}$. Observe that this transformation assumes a *linear* relationship between the two features of interest.

For instance, the conversion from Celsius (C) to Fahrenheit (F) can be obtained by considering respective intervals $[0, 100]$ and $[32, 212]$ as:

$$F = (212 - 32) \frac{C - 0}{100 - 0} + 32 = 1.8C + 32 \qquad (10)$$

Similarly, the conversion from Fahrenheit to Celsius can be immediately obtained as:

$$C = (100 - 0) \frac{F - 32}{212 - 32} + 0 = \frac{100}{180}(F - 32) \qquad (11)$$

6 Standardization

Standardization of a measurement f_i involves subtracting its mean μ_{f_i} followed by a division by its respective standard deviation σ_{f_i}, i.e.:

$$\tilde{f}_i = \frac{f_i - \mu_{f_i}}{\sigma_{f_i}} \qquad (12)$$

This *statistical transformation* of a feature yields a new feature $\tilde{f}_i$ that is dimensionless and that has zero means and unit standard deviation (e.g. [5]). In addition, a great deal of the instances of the new measurement $\tilde{f}_i$ will tend to fall within the interval $[-2, 2]$.

The standardization of a feature f_i can be understood as moving the center of the respective distribution to zero (a translation in the feature space), accompanied by a scale normalization in which the dispersion of the measurements becomes fixed or standardized.

Because standardization of a features yields a dimensionless respective feature, this operation is often applied in order to normalize the influence of unit choices on the respective classification.

Figure 8 illustrates the standardization of two clusters composed of uniformly distributed points in a two-dimensional respective feature space (the same situation as in the previous example). The respective displacement field, shown in (b), is similar but with more intense magnitudes, as the points undergo larger movements. Observe, however, that the magnitude of the displacements should often be considered in relative terms between the involved displacements (e.g. even larger magnitudes would be obtained if the clusters were further away from the coordinate origin, but the result would still be the same). As with the minmax normalization, the two clusters resulted with an elongated shape that can impact the subsequent processing.

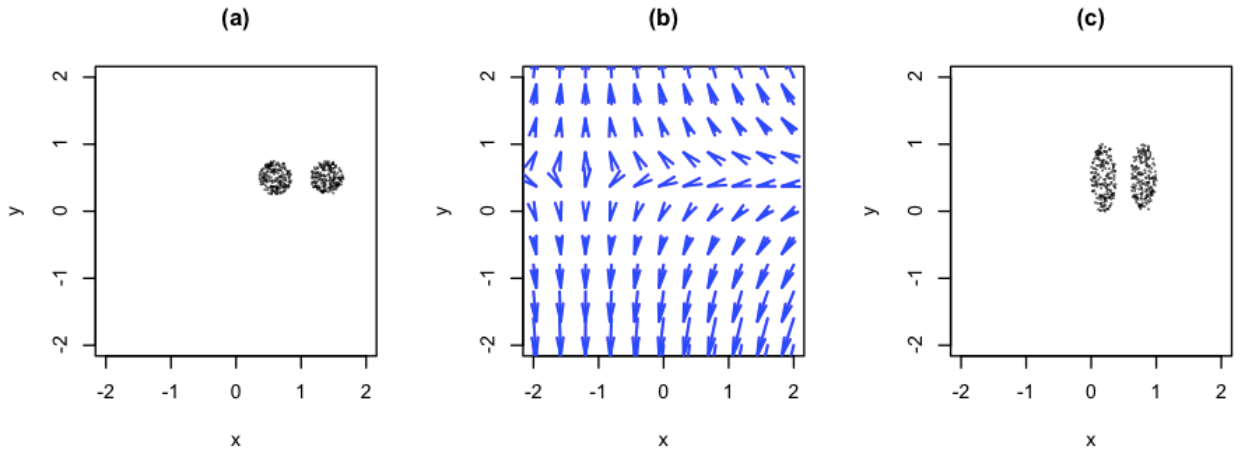

Figure 6: The minmax normalization of two well-separated clusters of uniformly distributed points (a) . Elongated clusters are obtained (c) as a consequence of the vertical expansion implied by the respective displacement field (b).

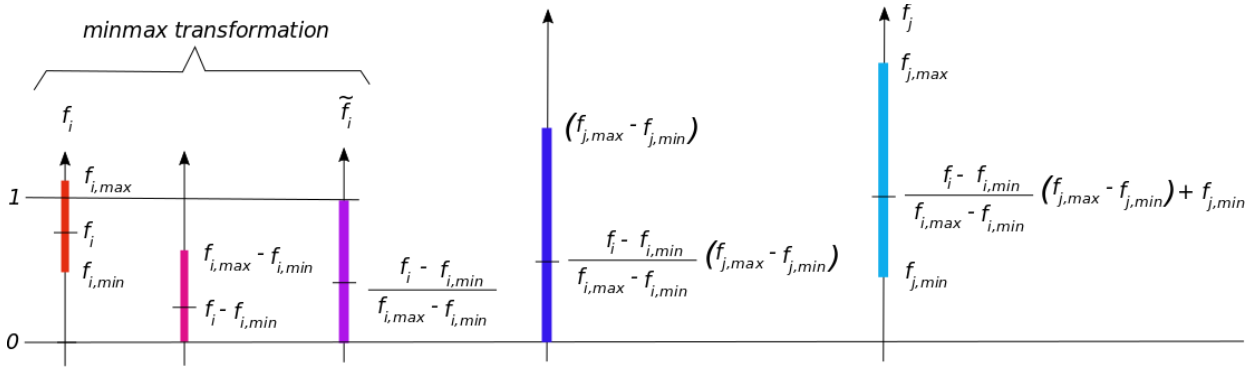

Figure 7: Generic unit transformations from a measurement f_i ranging in the interval $[f_{i,min}, f_{i,max}]$ into another measurement f_j varying in the interval $[f_{j,min}, f_{j,max}]$ can be conceptually understood as a minmax transformation followed by a scaling by $(f_{j,max} - f_{j,min})$ and a translation by $f_{j,min}$.

7 Principal Component Analysis (PCA)

Given a set of N entities, each represented by a respective feature vector $\vec{f_p} = [f_{1,p}; f_{2,p}; \ldots; f_{M,p}]^T$ (M features), with $p = 1, 2, \ldots, N$, it is possible to apply a transformation that completely decorrelates these features, allowing a possible dimensionality reduction, in the sense of yielding a new set of m features such that $m < M$ while implying in little loss of variation. This can be achieved by using the statistical transformation typically known as principal component analysis (PCA, e.g. [6]).

We start by deriving the covariance matrix K of the feature vectors, which corresponds to a random vector, which in many cases can be estimated as

$$K_{i,j} = covariance(f_i, f_j) \approx \frac{\sum_{p=1}^{N}(f_{i,p} - \mu_{f_i})(f_{j,p} - \mu_{f_j})}{(N-1)} \tag{13}$$

Once this covariance matrix is obtained, its eigenvalues and eigenvectors are obtained. The eigenvalues are then sorted in decreasing order, yielding $\lambda_1, \lambda_2, \ldots, \lambda_M$, with respectively associated eigenvectors $\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_M$. The latter are stacked as lines of an $M \times M$ matrix Q, i.e.:

$$Q = \begin{bmatrix} \leftarrow \vec{v_1} \rightarrow \\ \leftarrow \vec{v_2} \rightarrow \\ \cdots \\ \leftarrow \vec{v_M} \rightarrow \end{bmatrix} \tag{14}$$

The new feature vectors, $\tilde{f}_i$, can now be obtained by the PCA transformation as:

$$\vec{\tilde{f}} = Q\vec{f_i} \tag{15}$$

This corresponds to a linear transformation, implying that the new feature vectors are obtained as linear combinations of the original ones. Furthermore, this transformation can be understood as a rotation of the original feature space so that the data variance is concentrated

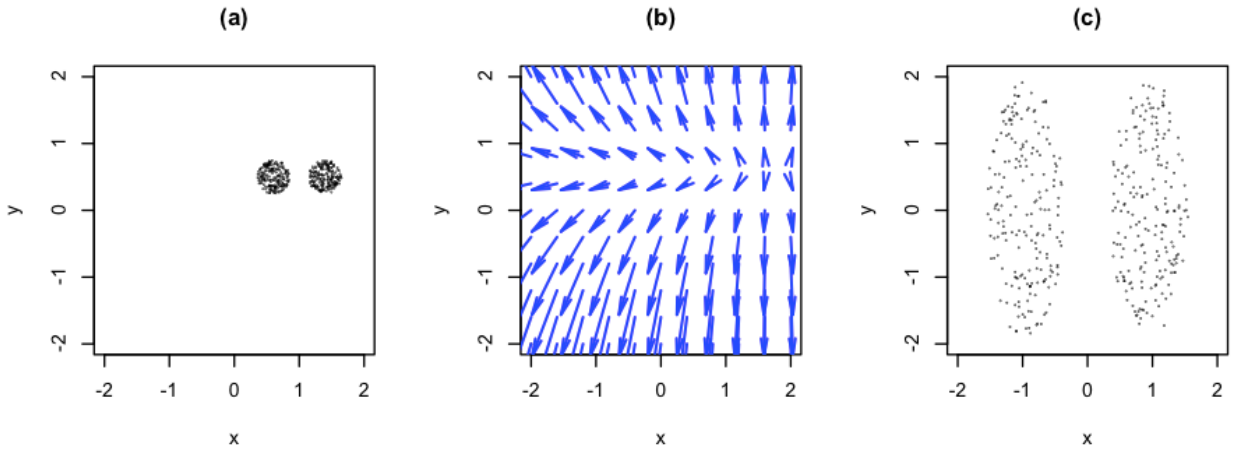

Figure 8: The standardization of twowell-separated clusters of uniformly distributed points (a). As with the minmax normalization, elongated clusters are obtained (c) as a consequence of the vertical expansion implied by the respective displacement field (b).

along the first new axes, corresponding to the *principal* new features. The variance of the new features is given by the eigenvalues associated to the respective axes.

It can be shown (e.g. [6]) that PCA yields a new set of features that are fully *uncorrelated*, implying in respective *redundancy reduction*. As a consequence, the new covariance matrix will necessarily be *diagonal*.

We can define the *covariance explanation index* η provided by the first m new features as:

$$\eta = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{M} \lambda_i} \tag{16}$$

If we set a value for η, such as 90% ,we can keep only the m features required for achieving at least that covariance explanation index, thus achieving a reduction of dimensionality from M to m. This is possible because, by removing covariance, PCA also decreases the redundancy of the features.

It is often interesting to standardize the features (see Section 6) prior to PCA.

Figure 9 illustrates the effect of PCA of an elongated set of points. Observe that the first new axis, $\tilde{x}$ results aligned with the direction of largest variation in the original set of points.

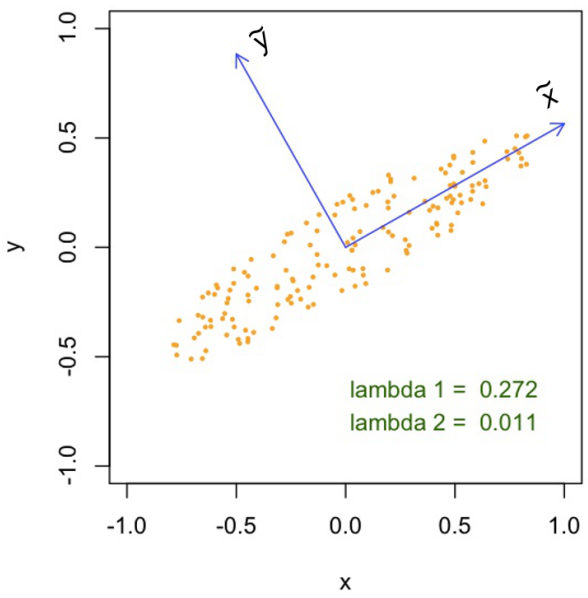

Figure 9: Illustration of PCA action on an elongated set of uniformly distributed points. Obseve that the first new axis, $\tilde{x}$, aligns itself along the direction of maximum variation in the original data. The eigenvalues associated to the two new axes are also shown, which yields a variance explanation of $\eta = 0.272/(0.272 + 0.011) = 96\%$ when keeping only the first new axis.

8 Concluding Remarks

Pattern recognition involves several stages, including features transformation and normalization. In this work, we addressed the latter operations with the help of the concept of morphing a set of points and visualization of respective displacement fields. In addition to revising some of the main transformations and normalizations, it has also been shown that these operations can substantially influence the subsequent stage of classification. Depending on the type of data and transformation/normalization, we can both enhance and undermine cluster identification. Therefore, great care should be taken while transforming and normalizing features.

Acknowledgments.

Luciano da F. Costa thanks CNPq (grant no. 307085/2018-0) for sponsorship. This work has benefited from FAPESP grant 15/22308-2.

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, 2000.

[2] K. Koutrombas and S. Theodoridis. *Pattern Recognition*. Academic Press, 2008.

[3] L. da F. Costa. Pattern cognition, pattern recognition. Researchgate, Dec 2019. https://www.researchgate.net/publication/338168835_Pattern_Cognition_Pattern_Recognition_CDT-19. Online; accessed 29-Feb-2020.

[4] L. da F. Costa. Where do patterns to be recognized come from? Researchgate, 2020. https://www.researchgate.net/publication/339599069_Where_Do_Patterns_To_Be_Recognized_Come_From_CDT-22. Online; accessed 09-March-2020.

[5] L. da F. Costa. Statistical modeling. https://www.researchgate.net/publication/334726352_Statistical_Modeling_CDT-13, 2019. [Online; accessed 22-Dec-2019.].

[6] F. Gewers, G. R. Ferreira, H. F. Arruda, F. N. Silva, C. H. Comin, D. R. Amancio, and L. da F. Costa. Principal component analysis: A natural approach to data exploration. Researchgate, 2019. https://www.researchgate.net/publication/324454887_Principal_Component_Analysis_A_Natural_Approach_to_Data_Exploration. accessed 25-Dec-2019.

Costa's Didactic Texts – CDTs

CDTs intend to be a halfway point between a formal scientific article and a dissemination text in the sense that they: (i) explain and illustrate concepts in a more informal, graphical and accessible way than the typical scientific article; and (ii) provide more in-depth mathematical developments than a more traditional dissemination work.

It is hoped that CDTs can also incorporate new insights and analogies concerning the reported concepts and methods. We hope these characteristics will contribute to making CDTs interesting both to beginners as well as to more senior researchers.

Each CDT focuses on a limited set of interrelated concepts. Though attempting to be relatively self-contained, CDTs also aim at being relatively short. Links to related material are provided in order to complement the covered subjects.

Observe that CDTs, which come with absolutely no warranty, are non distributable and for non-commercial use only.

The complete set of CDTs can be found at: https://www.researchgate.net/project/Costas-Didactic-Texts-CDTs.