Supervised and Unsupervised Pattern Recognition, and their Performance

Luciano da Fontoura Costa

Supervised and Unsupervised Pattern Recognition, and their Performance

Luciano da Fontoura Costa
luciano@ifsc.usp.br

São Carlos Institute of Physics – DFCM/USP

25th May 2022

Abstract

Pattern recognition, be in supervised or not, has motivated growing attention because of its several important applications. One issue of particular importance concerns the validation of the quality, e.g. in terms of correct classifications and stability, which is often estimated by performing cross-validation methods. A model-based approach is adopted, in which the data categories are understood statistically in terms of respective random variables, associated to the features, as well as the associated density probability functions. This allows both the supervised and unsupervised pattern recognition cases to be addressed in a principled manner while the important issues of bias, undersampling, underlearning and overfitting are all addressed and revisited. Several important and even surprising results are reported, including the interpretation of overfitting as not being necessarily unwanted, the characterization of the phenomenon of underlearning, in which several unstable working decision boundaries can be obtained, as being a consequence of biased sampling and/or undersampling, as well as the approach to unsupervised learning as involving two related but not necessarily identical issues, namely choosing how to interrelated the clusters and deciding whether a group could be considered as a cluster. To complement this development, we briefly consider the application of the coincidence similarity index to some of the covered problems, as well as present the possibility to use the important problem of image segmentation as a laboratory for better understanding and developing pattern recognition concepts and methods.

1 Introduction

Pattern recognition means the action of, given a set of entities represented by respective measurements (or *features*), to respectively assign existing (supervised recognition) or novel (unsupervised recognition) categories. The already substantial importance of this area (e.g. [1, 2, 3, 4, 5, 6]) has increased steadily along the last decades as a consequence of respective performance advancements combined with an expansion and intensification of respective applications in the most diverse scientific and technological areas. In particular, several of the activities traditionally performed by humans have been progressively assisted or even substituted by artificial intelligence resources, which rely intensely on pattern recognition.

By *entity* it is henceforth meant the objects, individuals, or any other type of patterns to be identified. The collection of the properties (features) characterizing the entities will be referred to as the respective *dataset*, with its respective *data elements* corresponding to the entities to be studied/classified.

Despite its relatively simple conceptual characterization, pattern recognition involves several concepts and methods, extending from multivariate statistics (e.g. [7, 8]) to neuronal networks (e.g. [9]). In addition, several sequential and/or parallel processing stages are typically involved in the implementation of pattern recognition systems. Simplistically, the basic steps of a basic pattern recognition pipeline are shown in Figure 1. These include: (a) acquisition of measurements (features) of the entities to be recognized; (b) pre-processing, possibly involving *normalization*; and (c) the recognition proper; (d) respective validation, often performed by cross-validation methods.

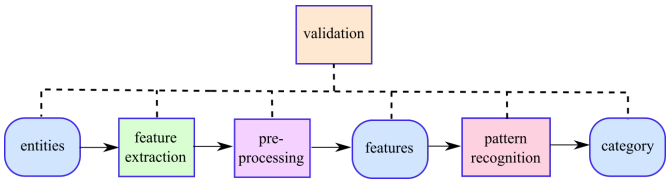


Figure 1: The pattern recognition pipeline. *Feature extraction*: A set of measurements are taken from the entities to be recognized, yielding the respective *features*. *Pre-processing*: the features are pre-processed in order to curate their quality and also for *normalization* purposes. *Pattern recognition*: methods are applied in order to assign categories to each entity. *Validation*: The validation of the whole approach is then performed.

Each of these three main stages are characterized by substantial challenges. At the acquisition level, one problem of particular relevance concerns which measurements are to be adopted for characterizing the entities. At the pre-processing stage, approaches have to be chosen that are able to improve the data quality (e.g. remove noise) as well as means to properly normalize the measurements. Major issues related to the third stage include the choice of recognition methods to be applied. Then, at the validation stage, metrics and approaches have to be defined and adopted in order to validate the recognition results according to the obtained results.

While all the main issues involved in all the above discussed pattern recognition stages have received substantial attention from the respective literature, the problem of characterizing the performance of the obtained recognition framework and results, so as to *validate* the adopted approach, remains an important issue worth continuing attention.

The validation of a pattern recognition approach depends on whether it is supervised or unsupervised, and both cases are considered in the present work.

In the former case, this has been typically performed by using cross-validation approaches (e.g. [5, 1]). In their most simple implementations, this type of validation involves separating the available data with identified categories into a training and a testing sets. The supervised recognition approach is then optimized for the identification of the training set and its performance is then quantified from the results obtained by its application to the test set. There are several variations of this basic principle aimed at improving the validation comprehensiveness and/or accuracy, such as considering several training and test sets, dividing the groups in non-equal proportions, including additional sets and validation stages, etc.

The result from cross-validations on supervised pattern recognition approaches indicate how many correct and incorrect classifications were obtained respectively to each of the involved categories. Ideally, there should be no incorrect classifications, with all the test data elements being correctly identified. When a pattern recognition approach passes through a strict validation, we have an *indication* that it may work properly in identifying the categories of new data.

The failing of an approach in the respective cross-validation indicates that there could be problems virtually anywhere in the framework shown in Figure 1. Table 1 summarizes the main aspects that typically play an important role in defining the performance of a pattern recognition approach.

The relative large number of aspects of different nature involved in the performance of pattern recognition (already hinted by the validation task encompassing all stages in Figure 1), allied to the fact that these aspects tend to influence one another, provides a cogent indication about the complexity of the validation problem. In this work, we develop a model-based approach to studying the several performance limitations involved in supervised and unsupervised pattern recognition while trying to consider in an integrated manner all the aspects identified in Table 1. Special attention is placed on the statistical modeling of the categories in respective feature spaces, which paves the way to identifying several important concepts and potential issues in pattern recognition, including the potentially dramatic effect of the increase of the feature space dimensionality on the recognition results, specially from the perspective of the *curse of the dimensionality*. The issue of *undersampling* therefore receives special attention along this work, from which we

Aspect	*Main characteristics*
Classification Regions	Densities specifying the categories within the feature space.
Uniform / non-uniform regions	The type of the density.
Sampling	Used to represent the category regions, can be sparse or biased.
Wrong samples	Some samples can have wrong categories.
Statistical fluctuations	Deviations from the respective densities caused by sampling.
Dimensionality	Determined by the number of required features.
Decision boundaries	Defined by the recognition methodology.
Confidence anchor	How accurate the regions, samples, boundaries and features are.
Adopted features (a kind of sampling)	The set of features adopted representing the entities.
Supervised / unsupervised	The type of pattern recognition method.
Cross-validation	Performed to quantify the performance of the recognition.
Clustered or not	Categories can be clustered or not.

Table 1: A glossary of the many important aspects influencing pattern recognition.

characterize the phenomenon of *underlearning*, namely the obtention of several provisionally working but unstable recognition configurations that do not withstand systematic cross-validations. The phenomenon commonly known as *overfitting* also receives special attention, and it is argued that it does not intrinsically constitute a shortcoming, but actually an asset in a pattern recognition approach.

An example of the important interrelationship between aspects involved in pattern recognition, we have that the two most often sought properties, namely *selectivity* is generally obtained at the expense of *generalization*. Usually, a balance needs to be achieved regarding these two requirements, which should take into account the nature of the data and questions of interest.

Supervised and unsupervised pattern recognition are both approached in the present work, with the former being addressed first. The case of supervised recognition is developed while considering the above outlined concepts and aspects, with emphasis on undersampling, underlearning, and overfitting, helping us to identify and approach some the main reasons that can undermine supervised pattern recognition, with several important and potentially surprising results. Unsupervised recognition is then treated with emphasis on two key issues: (a) the quantification of the separation between groups of data elements; and (b) the criterion adopted for deciding on the existence of clusters.

To complement our development, we also propose the consideration of image segmentation, an important and challenging issue on itself, as a laboratory for better understanding, developing, and evaluating pattern recognition approaches and systems.

Though focus is kept on presenting the several concepts and methods in a relatively accessible manner, enhanced understanding of this work will be helped by some previous experience in pattern recognition and/or related areas, particularly multivariate statistics (e.g. [7]), stochastic geometry (e.g. [10]), and set/multiset theory (e.g. [11, 12, 13, 14, 15]). In order to emphasize the main concepts and results along the development of the present work, several snippets have been respectively included.

It should also be kept in mind that the presented concepts and methods are still subject to further complementation and validations, so that they should be treated as preliminary. In addition, the application of any pattern recognition approach to real-world data as approached here should be understood mostly as means for providing insights on the data and groups interrelationships to be further investigated and validated, not providing a basis for absolute decisions regarding the separation or existence of clusters.

2 Categories, Statistical Modeling and Sampling

Groups (ensembles) of entities can be modeled in terms of their respective measurements, which are considered as random variables. In these cases, the respective joint probability density function (or field), or densities for brief, provides all available statistical information about the properties of those variables, and therefor about the entities as far as their features are concerned.

Densities can be understood as mappings from each point (entity represented in terms of its feature vector) in a given support (a region of the feature space) into respective non-negative values. In addition, the sum of all densities in the support needs to be identical to one. In a given pattern recognition problem, it is also important to identify from the outset the boundings of the respectively feature space, which we will henceforth refer to as the respective *universe* Ω. This set can be determined from the minimum and maximum values of each involved feature. Observe that each feature defines an associated axis in the respective feature space where the entities are to be represented.

There are two main types of probability density functions: (i) *uniform*; and (ii) *non-uniform*. The first case is characterized by constant values assigned for all points in the whole support. Non-uniform densities have varying values assigned to those points. An example of non-uniform density is the normal distribution. Figure 2(a) illustrates a uniform density defined on a disk on the $\mathcal{R}^2$ space.

From the perspective of this article, uniform densities can be treated in simplified manner, as corresponding to all the points in their respective support. As such, uniform densities provide a particularly effective means for approaching several of the intricacies of pattern recognition and its performance characterization. For instance, in Figure 2(a), it is enough to represent the region associated to the support of a uniform density instead of a three dimensional representation where the constant density would also be shown.

It is not always the case that the normal density and its support are available. Indeed, oftentimes we only have samples, obtained from inaccessible density formulae, from a given density. This is illustrated in Figure 2 in terms of three possible samplings of the density in (a): *sparser* (b); *denser* (c); and *biased* (d).

The amount and quality of samples is of critical importance in pattern recognition. Even if all samples are correct, in practice they will always be available in limited numbers, implying that the original density will never

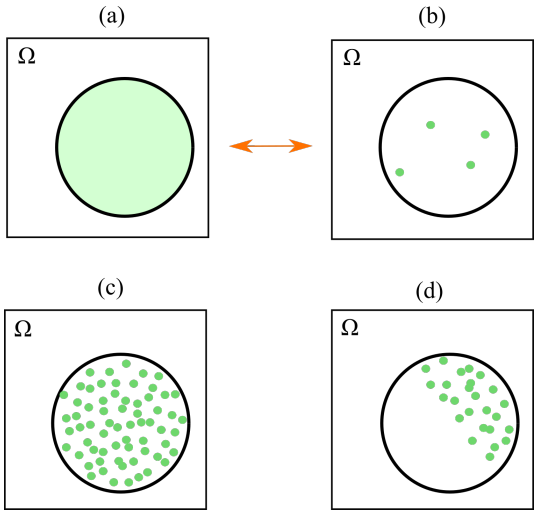


Figure 2: A uniform density, continuous on a disk support (a), and possible respective discrete samplings characterized by being relatively *sparser* (b), *denser* (c), and *biased* (d).

be perfectly sampled and represented. This loss of information impacts the characterization of the density in several manners, including unavoidable *statistical fluctuations*, i.e. the fact that (typically) small scale spatial differences will be always found between among the sample distribution. As illustrated in Figure 3, these fluctuations can lead to respective patterns.

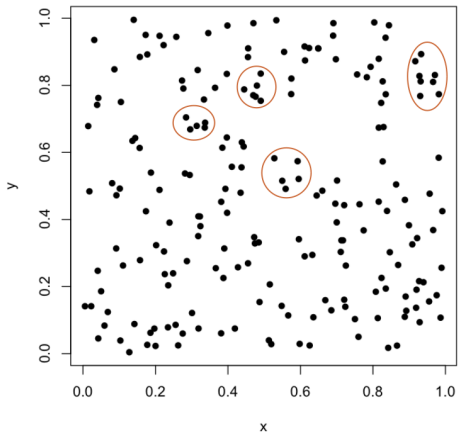


Figure 3: A uniform random field of points. As a consequence of random fluctuations implied by the undersampling of the otherwise completely uniform density, random fluctuations appear that can eventually be taken for clusters.

Generally speaking, the larger the number of samples, the better. The situations in which the number of samples is not enough for proper representation of the original continuous densities are henceforth called *undersampling*. It is also possible that the available samples are biased in several manners, such as that depicted in Figure 2(d). Needless to say, biases can have critical impact on the recognition results.

Sampling is also required for approximating non-uniform densities, as illustrated in Figure 4, where a varying density on a disk support (a) is sampled by a limited number of samples (b). Observe that the density of the samples tends to reflect the respective original density at each of the points in the support.

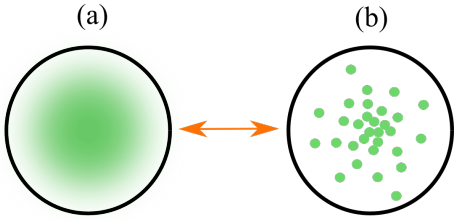


Figure 4: A non-uniform density on a disk support (a), and a possible respective sampling (b).

Another critical influence of sampling on pattern recognition concerns the fact that the higher the dimensionality of the feature space, the larger the number of points that are required for a relatively dense and significant representation of the densities. Actually, it is relatively straightforward to infer that, in the case of M features, the number of samples should increase with the respective $M-$power. Therefore, the densities involved in pattern recognition problems involving a large number of features are often undersampled because the very large number of samples may not be available, or cannot be computationally handled, which gives rise to the so-called *curse of dimensionality*.

We conclude this section with our first snippet:

1 - The approximation of continuous regions by respective samples depends strongly on the dimensionality and is crucial for pattern recognition. The higher the dimensionality, much more samples are needed. Biasing and undersampling undermines the representation of the densities and can lead to recognition mistakes.

3 Supervised Pattern Recognition

As implied in its own name, *supervised* pattern recognition refers to the assignment of categories to data elements under supervision of several types, including sets of *pre-classified data elements*, or *prototypes* of each category (e.g. center of mass of the groups). In this section, we will approach this type of recognition as well as its performance issues mainly from the perspective of the concepts

presented in Section 2.

Basically, supervised pattern recognition involves two stages: (i) training, and (ii) application to classification of new categories. It is the former stage that makes this type of recognition supervised. Let us illustrate this basic principle in supervised recognition with the help of the example in Figure 5, which involves two uniform category regions.

In this particular case, the optimal decision boundary resulted aligned to the own original boundaries, which is characteristic of *adjacent* category regions. Provided there are no errors in new data elements, perfect performance will characterize subsequent classifications.

Another, more frequent, supervised classification situation involving non-adjacent regions is presented in Figure 6.

Some additional examples of relatively compact, well-separated category regions are illustrated in Figure 7.

Of particular importance is the fact that compact, well-separated regions makes the classification much easier, while also reducing the chances of underlearning as implied by undersampling.

Figure 8 presents another supervised recognition example involving non uniform regions in a one-dimensional space.

Provided the densities are fully known, Bayesian decision theory indicates the means for identifying the *optimal* decision boundaries respectively to minimizing the number of misclassifications, whose probability corresponds to the areas where one density overlaps the other. More specifically, let M categories $c = 1, 2, \ldots, M$ be represented by respective conditional densities $p(\vec{x} \mid c)$. Let the mass probability of each category be $P(c)$. Then, the Bayesian classification criterion consists of applying:

$$C(\vec{x}) = c \mid \max_{c=1,M} \{P(c)\ p(\vec{x} \mid c)\} \qquad (1)$$

In the case of Figure 8, this criterion yields the optimal border as corresponding to the intersection between the two densities, i.e. $x = b$.

When the region densities cannot be accurately determined, other approaches need to be applied, and that is precisely where the recognition problems start because the respective loss of information. There are two main situations yielding inaccurate densities: we do not know them, have only approximations or hypothesis, or only respective samples are available, possibly in limited numbers. At least the two following two approaches can be attempted in the latter case: (a) estimate the densities from the samples; and (b) use the samples directly for the recognition.

In both cases, when only a limited number of samples are available, there will always be the possibility of *undersampling*, which can have critical impacts on the classification.

However, before addressing undersampling in a more systematic way, it is interesting to discuss the frequently considered problem of *overfitting*. Basically, this phenomenon consists of the obtained decision boundaries being 'too' closely adapted to the samples, as illustrated in Figure 9(a).

Here, we have two adjacent category regions that have been fully separated at the expense of the use of a relatively intricate decision boundary. Observe that all points have been correctly classified in this case. Now, if a new data element becomes available and is mapped into the previous space as shown in Figure 9(b). Given that this new element resulted within the blue region, a misclassification will be respective implied. However, the system can be retrained so that the new boundary region shown in (c) is obtained, again ensuring correct classifications throughout, but at the cost of an even more intricate decision boundary, therefore enhancing the overfitting. Interestingly, it is possible to show that decision boundaries can be found in any supervised recognition problem that will yield full adherence to the involved categories, therefore implying no classification errors.

Now, an important point concerns the fact that, provided there are no errors in the supplied categories of all the samples, the phenomenon of overfitting is not intrinsically unwanted, but actually necessary to properly represent the categories in the feature space. Actually, in case the configuration in Figure 9(c) corresponds to all possible samples constituting the respective category, the decision boundary in that figure actually corresponds to an optimal solution. In conclusion, generally speaking, overfitting does not constitute a shortcoming of the approach, but actually one respective asset. In summary, we may conclude that:

2 - The overfitting respectively to the correct classification of the whole set of correctly labels samples is not necessarily unwanted, but actually welcomed, irrespectively of the level of intricacy or tight adherence implemented by the respective decision boundaries.

Figure 9 also illustrates some possible results of applying cross-validation to the situation in (a). Figure 9(d) presents the same case, but after removal of some of its points. In this specific case, the retraining under these circumstances will yield a decision boundary similar to

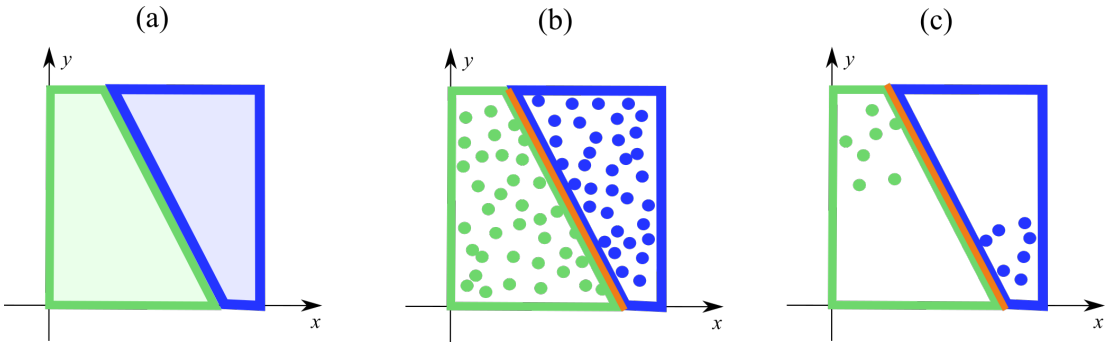


Figure 5: The basic principle underlying supervised pattern recognition, respective to two adjacent uniform regions in a two-dimensional feature space: (a) objects belong to two categories, defined by their respective regions delimitated by the blue and green contours, are to be recognized; (b) samples of the two categories are taken and used to train a respective classifier, which yields the decision boundary shown in orange; and (c) new data can now be classified depending on which region they fall into.

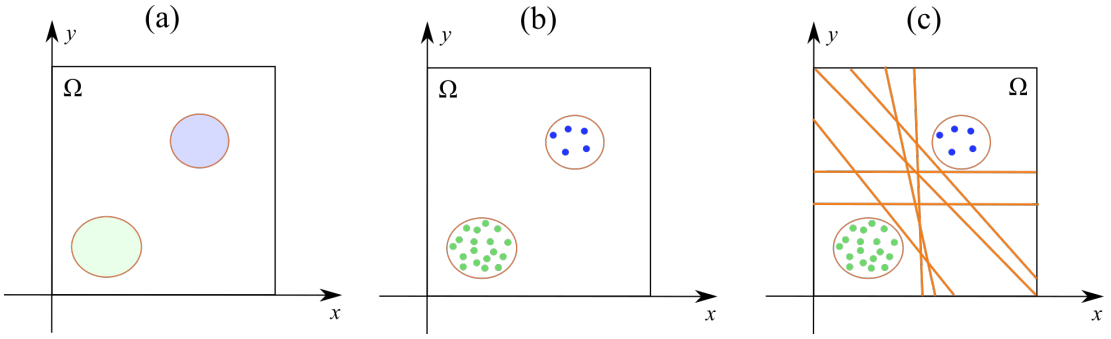


Figure 6: The basic principle underlying supervised pattern recognition, respective to two non-adjacent uniform regions in a two-dimensional feature space: (a) objects belong to two categories, defined by their respective regions delimitated by the blue and green contours, are to be recognized; (b) samples of the two categories are taken and used to train a respective classifier, which yields the decision boundary shown in orange; and (c) new data can now be classified depending on which region they fall into. Remarkably, a wide range of possible boundary decisions, instead of the single one obtained in the example in Fig. 5 are now possible. This does not represent neither *underlearning* nor *overfitting*.

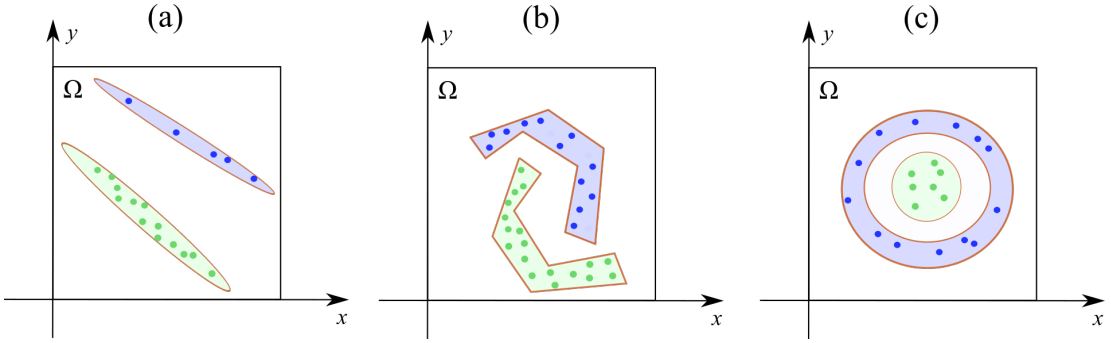


Figure 7: Additional examples of relatively compact, well-separated category regions. Many decision boundaries can be found in these cases that ensures fully correct classifications.

the previous one, yielding no classification errors. Figure 9(e) presents another example in which several points were removed from (a), but now a new decision boundary is obtained. In case the removed points are now tested in this new region, misclassifications will take place (f). It is important to identify what can be learnt from these cross-validation experiments as applied to the specific overfitted situation in (a). The key aspect here is that the decision boundary in (b) are in fact not accurate, hence the misclassification errors implied. In other words, the failing of this approach under cross-validation only indicates that the boundary obtained with fewer points is less precise. To any extent, adjacent category regions with intricate interrelationships will tend to be identified as overfitted,

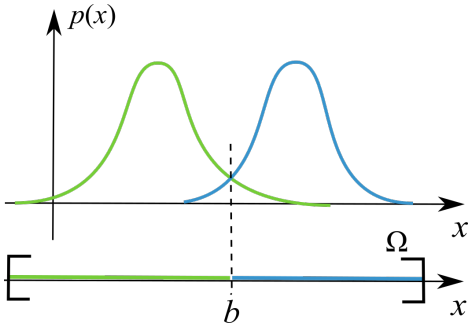


Figure 8: Entities belonging to two categories statistically modeled by the densities shown in this figure are to be recognized. In the case the two categories have the same mass probability (i.e. the two types of entities are equiprobable), Bayesian decision theory specifies the optimal decision boundary in terms of the category corresponding to the maximum density values of the respective universe points. In the case of this example, the optimal decision boundary is defined by the value $x = b$.

but this is as it should be.

We can infer from the above considerations that:

3 - The highest the required performance in terms of accuracy and correct classifications, the higher the overfitting, implying more intricate decision boundaries especially in the case of adjacent regions with jagged interrelationship and also as a consequence of the sampling of the category regions or densities. The identification of this phenomenon by cross-validation does not necessarily imply a shortcoming, though it can provide useful information about the structure of the categories and samples.

Now, let us address another problem, namely *sample biasing*, as illustrated in Figure 10, which shows two biased samplings of the situation shown in Figure 5(a). As a consequence, the two samples became well-separated, allowing a wide range of exact possible decision boundaries, a few of which are illustrated in orange in the figure. Though these boundaries work for the given samples, it will soon fail when more samples are drawn from the respective regions.

The possibility to have several provisionally adequate decision boundaries that are prone to become unstable with new samples (or under cross-validation) is henceforth called *underlearning*, in the sense that the recognition system has not yet reached its proper training as a consequence of biased sampling, which leads us to the next snippet:

4 - *Underlearning*, characterized by many provisionally working decision boundaries that are not similar to the correct one, happens when the sampling does not properly represent the regions.

Now, let us return to the *undersampling* problem briefly introduced above. We will start by performing an experimental study of how uniformly random points randomly separated into two groups relate one another, in terms of Euclidean distances between their samples, as the dimensionality of the feature space is increased. First, all features will constrained within the interval $[0, 1]$, as is the case of features pre-processed by minmax normalization. Figure 11(a) presents the average $\pm$ standard deviations of the minimum distance between the 1000 randomly generated pairs of random categories.

Figure 11(b) presents the average $\pm$ standard deviations of the Euclidean distances between two randomly assigned groups of $20, 25, \ldots, 50$ points in feature spaces of dimension D, with all features varying from -2 to 2, as is typically the case with standardized features. Interestingly, a much wider artificial separation can be respectively observed.

Of critical interest in the obtained results is the fact that a non-null separation is observed between the two randomly assigned groups of uniform samples, and that this separation tends to increase in the average with the dimensionality of the feature space. At the same time, relatively comparable standard deviations have been observed for most dimensions, except for the smallest cases. Observe also that the average distances tend to decrease slowly with the increase of available samples, but they fall short of the distances observed for 1 or 2 dimensions. Ultimately, these results are a consequence of the curse of dimensionality, which implies sparse representation, by samples, of the category regions.

This result plainly indicates that relatively well-separated groups can be obtained out of uniformly random points, especially for highly dimensional feature spaces and when relatively few samples are available. A similar phenomenon can take place for samples obtained from generic respective category regions, even if they are actually not well separated in the original, continuous representation. Given that artificially well-separated groups can appear in these cases, the phenomenon of *underlearning* directly analogous to our discussion regarding biased samples will occur. Actually, the undersampling that often takes place in highly dimensional feature spaces can be considered as a kind of biased sampling.

Cross-validation provide a valuable means for identifying underlearning as a consequence of high dimensionality-related undersampling, because distinct

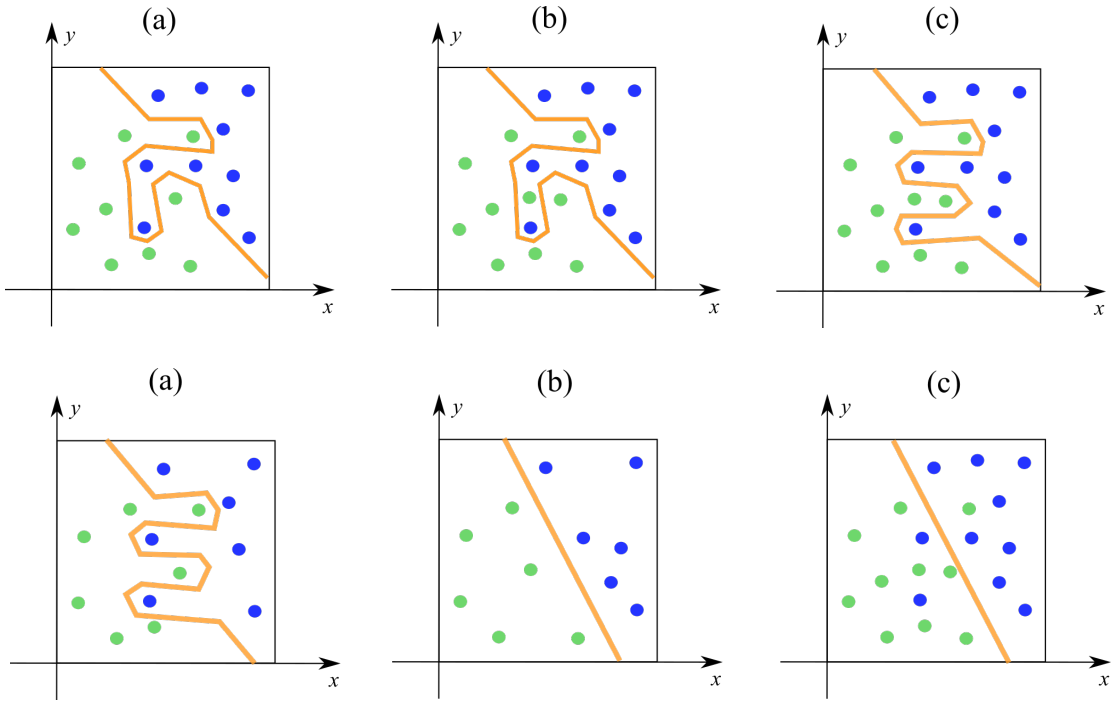


Figure 9: Illustration of the phenomenon of *overfitting* and its relationship to cross-validations.

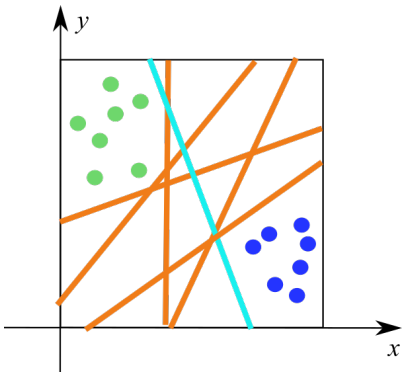


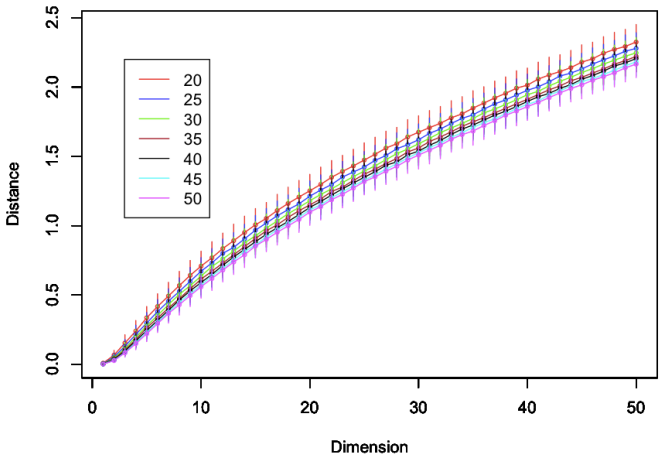
Figure 10: Example of biased sampling of the regions in Figure 5(a) leading underlearning, in the sense that a wide range of decision boundaries can be obtained that implement fully correct classification for this particular sample configuration, but which are unstable and bound to misclassify new samples. The correct decision boundary, as implied by the original category regions in Figure 5(a) is shown in cyan.

choices of the two randomly assigned groups will yield distinct decision boundaries, which leads us to the next snippet:

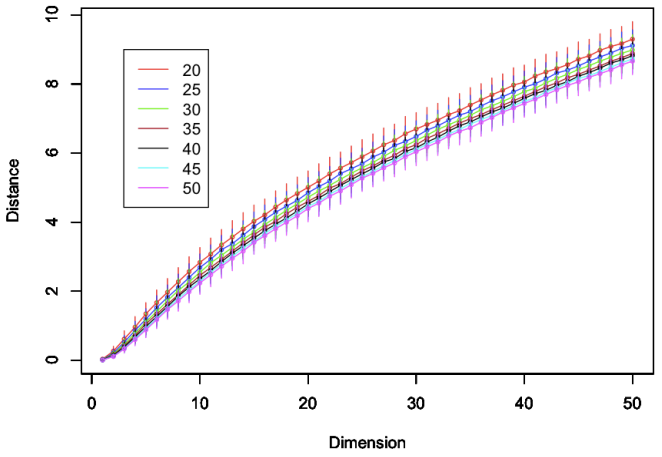
5 - Cross-validation can reveal underlearning caused by sparse sampling (curse of dimensionality), especially in the case when many features are adopted.

However, there is an important exception to the influence of undersampling on supervised recognition in high dimensions, and it has to do with a phenomenon that we shall call *compact sampling*, in order to refer to sampling of relatively compact, well-separated original category regions. Interestingly, under these circumstances, the sampling, even if sparse, becomes restricted to the original regions, therefore reflecting the distances between the original regions. Provided those distances, e.g. in the average, are larger than the distances between the groups induced as a consequence of respective high dimensionality, the effect of underlearning can become substantially minimized. It then becomes possible, provided all the other involved aspects (e.g. classification method, quality of samples, etc.) are proper, to infer effective decision boundaries even in the case of highly dimensional feature spaces.

There are at least two possible means to identify compact sampling: (i) apply cross-validation; and (ii) to compare the distances between the given sampled groups with those obtained for similar configurations (i.e. number of samples, normalization, dimensionality). In the latter case, a significant distinction between the distances implied by the original data and those obtained by randomly assigned groups will suggest that the supervised recognition can proceed with relatively little underlearning. An important conclusion to be drawn from the above reasonings therefore can be summarized as:



(a)



(b)

Figure 11: The average ± standard deviations of the Euclidean distances between two groups of $20, 25, \ldots, 50$ points in feature spaces of dimension D, with all coordinates varying from 0 to 1 (a) ad from -2 to 2 (b), increases steadily, though in sublinear manner. Also interesting is that the standard deviations do not tend to very significantly with the dimensionality and that similar shapes have been obtained in (a) and (b). Results obtained from 1000 random experiments.

6 - In the case of highly dimensional feature spaces, is possible to have underlearning minimized provided the original regions are relatively compact and well-separated. This can be verified by cross-validation or by comparison between the original and random distances.

A relatively simple and quick approximated method for investigating underlearning in high dimensions is as follows. Given samples of M categories, obtain the average ± standard deviations of the distances between each pair of group. The overall interrelationship between the original groups can then be roughly estimated by inspecting the respective representation of the distance as a network where each node corresponds to a category, and the links between the nodes are proportional to the respectively obtained average distances. Another reference network is obtained by randomized groups with the same number of elements and dimensionality. These two networks can then be compared qualitative and/or quantitatively. In case the two networks are similar, it is very likely that undersampling may be taking place, which can be further investigated by cross-validation. Given that the minimum distance between the samples in each pair of clusters is too strict and relatively unstable (the change of a single sample can strongly impact on the result), it is also interesting to consider the distances between the center of mass of the real and random groups.

Let us illustrate this method respectively to a dataset containing 3 types of handwritten characters (''c, 'e', and 'o') [16], each being represented by 50 samples. Each data element is characterized in terms of four respective features, which are henceforth taken in their *standardized* version. The average distances obtained from the randomly assigned pairs of groups with the same dimensionality and number of samples was $\langle d_r \rangle = 0.580765942$. Figure 12 shows the principal component (e.g. [17]) projection of the handwritten characters dataset (a), as well as the distances networks respectively to the real data (b) and randomly assigned simulation (c).

The results of the same experiment as above, but now performed respectively to the features normalized in the interval $[0, 1]$ are presented in Figure 13, being characterized by similar ratios between the real and random average distances.

7 - If cross-validation does not hold, then: (a) the selectivity needs to be increased; (b) the sampling is not enough to properly represent the reference regions; (c) the data has low quality; (d) the recognition method is unstable/unsuitable; or (e) the groups of samples cannot be separated.

It is interesting to observe that, though the above discussion focused on the effect of biased sampling and undersampling in relatively high dimensions, the obtained distance values for the random configurations obtained even for dimensions as smalls 2 or 3 indicate that artifacts in data separation can take place even in these situations.

4 Clustering, or Unsupervised Pattern Recognition

Having discussed supervised pattern recognition from a model-based perspective with special attention on perfor-

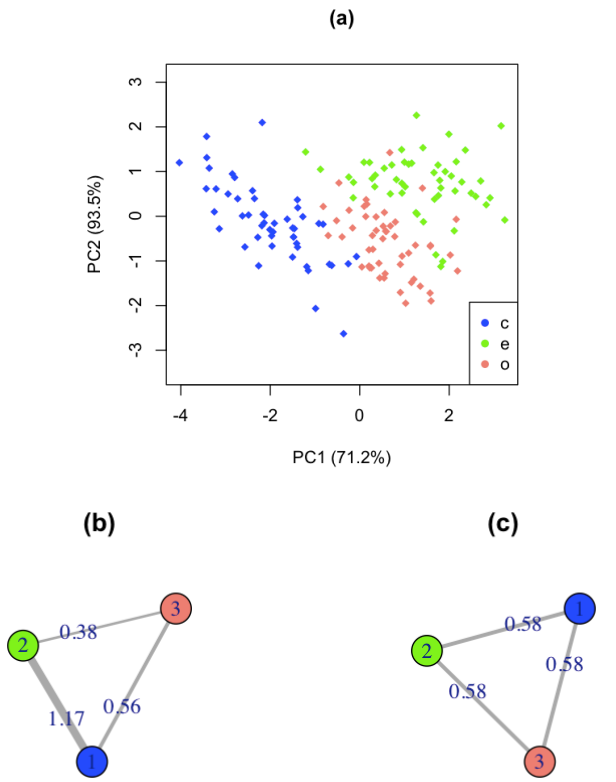


Figure 12: The handwritten characters database presented in terms of its PCA (a), and respective networks of average distance for the real data (b) and respective randomly assigned simulation (c). One of the real-data links resulted about twice the random counterpart, another is comparable, and the third is about half. Given that the minimum distance between clusters is a too strict indication of the separation between two groups, the original real data can be considered as being relatively far from underlearning.

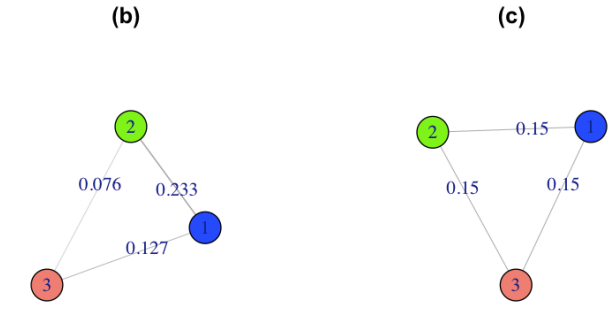


Figure 13: The comparison between the real and random distances for the handwritten characters dataset considering the interval $[0, 1]$. These results have proportions similar to those in Fig. 12.

mance, we now turn our attention to the relatively more challenging task of unsupervised pattern recognition, or *clustering* for short, which is characterized by absence of prototypes or even information about the number of expected clusters.

The following snippet provides an intuitive definition of a cluster:

8 - Given samples in a feature space, a respective *cluster* is a subset of these samples which are more similar one another than to the remainder samples.

One of the first important aspects to be observed concerning clustering is that it involves two related, but distinct, requirements: (a) how to quantify the separation between the clusters; and (b) how to decide on the existence of one or more clusters. This critically important aspect is not always observed, and can lead to respective misunderstandings. So, we have the following snipped:

9 - Unsupervised classification requires a choice of how to quantify the separation between the clusters, as well a decision on the existence of clustering.

Pertaining the issue (a) above, there are several possible approaches that can be used for that finality, including the minimal, maximum, distance between centers of mass, among other possibilities, of metrics and indices (e.g. [18, 19, 20, 21]) including but not being limited to: Euclidean distance, cosine distance, Pearson correlation coefficient, Mahalanobis distance, Manhattan distance, as well as similarity indices including the Jaccard, Interiority, and coincidence. Observe that we distinguish between *metrics* and *index* in order to indicate that the former obeys the metric requirements, while the latter does not.

For simplicity's sake, the present work concentrates on the Euclidean distance agglomeration (*single-linkage*), though other agglomerative approaches including the *complete-linkage*, *average-linkage* and *Ward* methods are also illustrated, and the coincidence similarity is considered, for comparison purposes, in a subsequent section.

In the case of agglomerative clustering approaches, the successive merging of the clusters gives rise to a respective *dendrogram*, which provides a comprehensive graphical representation of the interrelationships between the unfolding clustering respectively to the adopted metric or similarity index. Figure 14 presents the dendrograms obtained from the handwritten characters dataset by using *single-linkage*), *complete-linkage*, *average-linkage* and *Ward* agglomerative clustering.

Observe that the $y-$axes in each of the dendrograms in Figure 14 corresponds to the respective adopted linkage criterion and metrics/index. In the case of metrics, the $y-$axis corresponds to a distance that increases from the bottom to the top. Interestingly, quite distinct clustering structures have been obtained by each method, which motivates the question of which of them could be more relevant for this specific dataset.

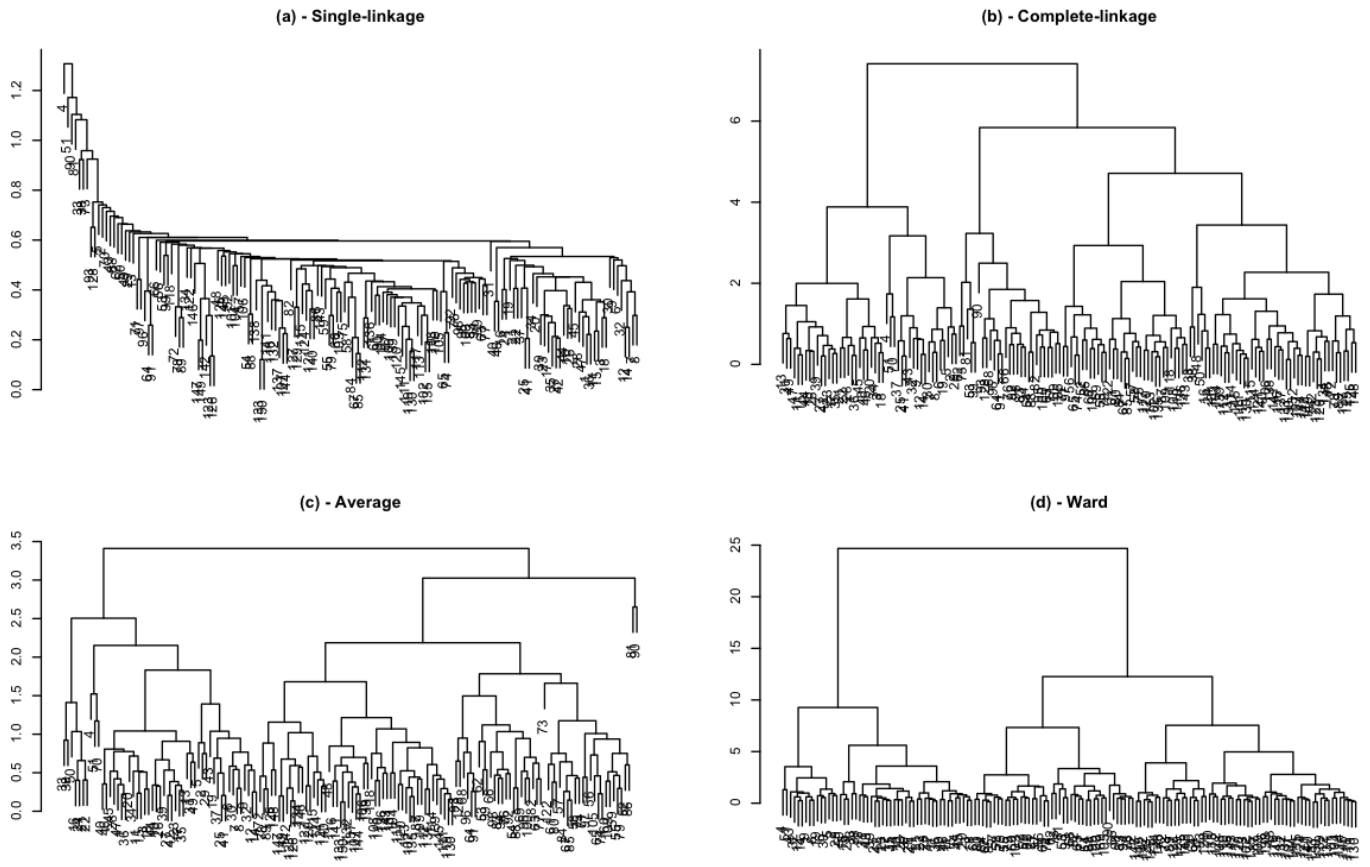


Figure 14: Dendrograms obtained from the handwritten characters dataset by using *single-linkage*, *complete-linkage*, *average-linkage* and *Ward* agglomerative clustering. Observe the completely distinct clusters interrelationships suggested by each of these distinct approaches. What is the most adequate for the handwritten characters dataset?

The issue (b) above, namely *deciding on the existence of one or more cluster* is directly related to the interrelationship, especially the separation, between the candidate groups, and can be approached in those terms. For instance, it is possible (e.g. [22]) to consider the length of the branches leading to a branch, multiplied by the number of samples in that branch as an indication of how much that possible cluster stands out among the others.

While the above mentioned type of approach is interesting and often leads to suitable results, there is an important issue that is not so often realized or discussed, and it has to do with the fact that the scaling of the $y-$axis variable, henceforth referred to as y, has a somewhat arbitrary nature. For instance, in the case of the average-linkage method, instead of taking the respective average Euclidean distances between the groups as y, it would be also possible to consider any monotonic transformation of y, for instance by taking it to the 5-th power, to the 0.2 power, or taking a sharp sigmoid, as depicted in Figure 15.

It is particularly interesting to compared these transformed dendrograms with those in Figure 14. Though they are completely distinct as far as the relative positions of the vertical axes where the mergings occur, the merging *sequence* is *completely identical* in all average-linkage cases presented above. At the same time, the type of illustrated transformations constitute a particularly useful resource for *zooming* in and out of the several scales along the $y-$axis. For instance, in case we are especially interested in studying the clusters relationship at the finer merging scale, we could resource to a transformation similar to that obtained by the sharp sigmoid, and so on. Another relevant observation about the dendrograms obtained for the handwritten characters dataset consists in the fact that none of them, original or transformed, provided a pronounced indication, as far as the relative lengths and widths of their branches are concerned, about the original separation between three main types of characters in this specific dataset.

The above example illustrates the difficulty in using the lengths of the branches as a criterion for deciding on the whether the involved clusters should be separated or not. Actually, there is an alternative approach that does *not*

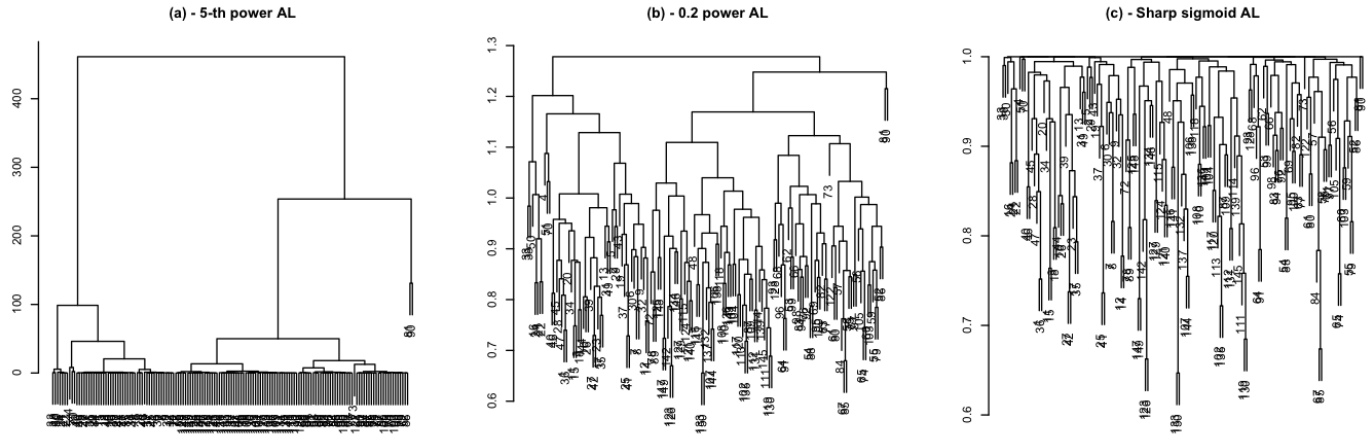


Figure 15: Monotonic transformations of the average-linkage dendrogram obtained for the handwritten characters, but taking the by taking the average distances y to the 5-th power (a), to the 0.2 power (b), or through a sharp sigmoid (c). Completely distinct dendrograms can therefore be obtained, emphasizing respectively the large, medium, and small scales of the clustering structure of the dataset. Importantly, the sequence of merging of each of these transformations is completely preserved, while only the y-axes is 'elastically' modified.

depend on the length of the branches. It consists of using *other* criteria for that purpose, in particular one of the several approaches to quantifying the separation between clusters, including those based on scatter matrices (e.g. [4]) or even network modularity (e.g. [16]).

Now that we have considered some of the most basic aspects of unsupervised clustering, we can attempt to approach the issue of their respective performance. There is a particularly direct manner to do this, by using a set of classified samples that are then treated as it they were not, being therefore classified in an unsupervised manner. The original categories can then be taken into account in order to evaluate the recognition results in terms of misclassifications, as well as all the aspects discussed respectively to supervised pattern recognition. Actually the effects of most of those aspects are precisely the same whatever we are dealing with supervised or unsupervised learning. For instance, the presence of biased samples in relatively high dimensions, and/or undersampling, are likely to induce underlearning, implying clusters to be found where there are none.

One important difference respectively to unsupervised recognition is that cross-validation, e.g. by k-folding, cannot generally be performed in the same way, as there is no training stage involved in that case. Those methods need to be adapted, for instance by identifying clusters respective to a portion of the reference samples, and then comparing them with other the results obtained by the same unsupervised respectively to other sets of samples. However, it should be observed that this is not a complete test, because the procedure may induced similarly biased results in all cases. More comprehensive validation approaches need to consider previously labelled sets of samples, so that the obtained clusters can be confronted with the original ones. As a summary of our brief discussion of unwanted effects on the performance of unsupervised recognition methods and their respective validation, it can be said that:

10 - Overfitting, undersampling, and underlearning phenomena equally affect the supervised or unsupervised case.

5 Similarity-Based Pattern Recognition

In this section we consider some possibilities of using the real-valued coincidence similarity index [23, 21, 16, 24] as the basis for comparing and interrelating groups of samples in both supervised and unsupervised pattern recognition.

Basically, the coincidence similarity corresponds to the product between the real-valued Jaccard and interiority indices, which are based on multiset theory (e.g. [11, 12, 13, 14, 15]). It is primarily aimed at performing similarity comparisons between two patterns, e.g. as represented by respective feature vectors. In the present work we limit our attention to the parameterless version of the coincidence similarity index which, in this particularly case, has results comprised in the interval $[-1, 1]$. The higher the coincidence similarity value, the most similar two patterns can be said to be. So far, the coincidence index has been successfully applied to several applications, including template matching [21] and translation of datasets into respective complex networks [16].

First, we present in Figure 17 the average ± standard deviation of the coincidence values between random groups for $20, 25, \ldots, 50$ points in feature spaces of dimension D, with all features varying from 0 to 1.

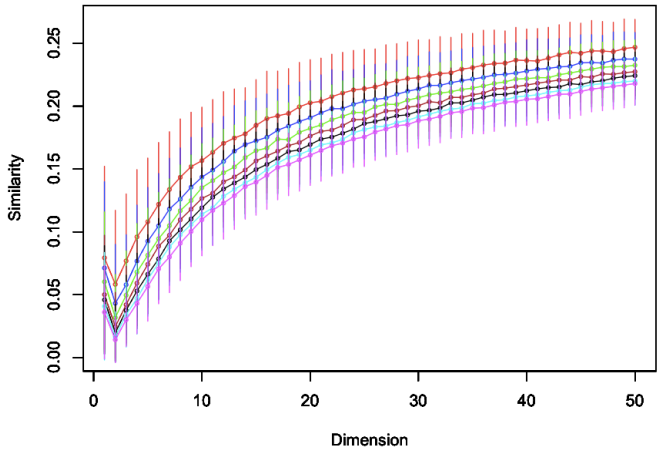


Figure 16: The average ± standard deviations of the coincidence values between two groups of $20, 25, \ldots, 50$ points in feature spaces of dimension D, with all features varying from 0 to 1. Comparatively to the respective Euclidean distance counterpart in Fig. 11, it can be said that the obtained coincidence values increase more steeply along the smaller dimensions, become relatively stable for the larger dimensions.

A similar shape of the coincidences in terms of the dimensions can be verified for when the features comprised in the inverval $[-2, 2]$, shown in Figure 17.

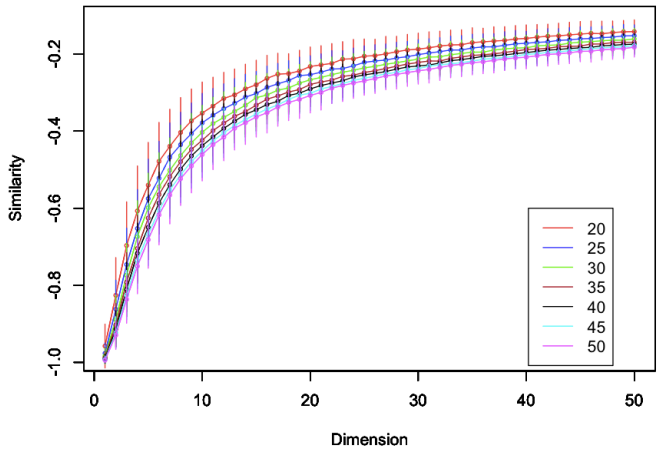


Figure 17: The average ± standard deviations of the coincidence values between two groups of $20, 25, \ldots, 50$ points in feature spaces of dimension D, with all features varying from -2 to 2.

Figure 18 depicts the networks of average similarity obtained respectively to the handwritten characters datased. Interestingly, a more uniform distribution of coincidences was obtained for the real data, two of which are higher (in absolute value) than the random reference, while the remainder distance is similar. This suggests that, at least for this specific example, the coincidence similarity can lead to less intense underlearning as a consequence of biased sampling or undersampling.

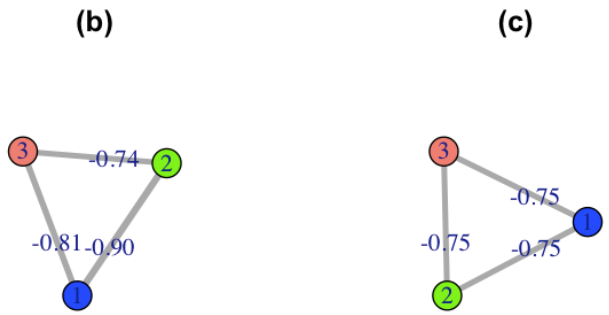


Figure 18: Networks of average similarity for the real data (a) and respective randomly assigned simulation (b) respectively to the handwritten characters database. Comparatively to the Euclidean distance based results shown in Fig. 12, two of the real pairwise distances resulted larger than the random counterparts, while the other distance result very similar. This suggest a better underlearning resilience of the coincidence similarity representation of the data elements, at least for the case of the specific data in this example. Observe that the magnitude of negative coincidence similarity quantifies the dissimilarity between the respective groups.

To conclude this section, we present in Figure 19 the dendrogram obtained for the handwritten characters dataset by the single-linkage method adapted to the coincidence similarity. More specifically, this index is used instead of the Euclidean distance or the other options already discussed. As a consequence, the $y-$axis has to be modified so as to have the dendrogram comparable to those obtained by the other methods. In this work, this has been done by taking the complement of the coincidence values along the respective $y-$axis, i.e.:

$$\tilde{y} = \max\{y\} - y \tag{2}$$

Interestingly, a well-balanced dendrogram has been obtained in which not only details can be appreciated about the pattern relationships at fine and medium comparison scale, while the intrinsic subdivision into three categories corresponding to the three types of handwritten characters can be more effectively perceived in the relative long branches leading to the three groups. Remarkably, this coincidence-based single-linkage does not suffer from the same level of chaining (successive incorporations of samples into a same group) as its respective Euclidean distance-based counterpart.

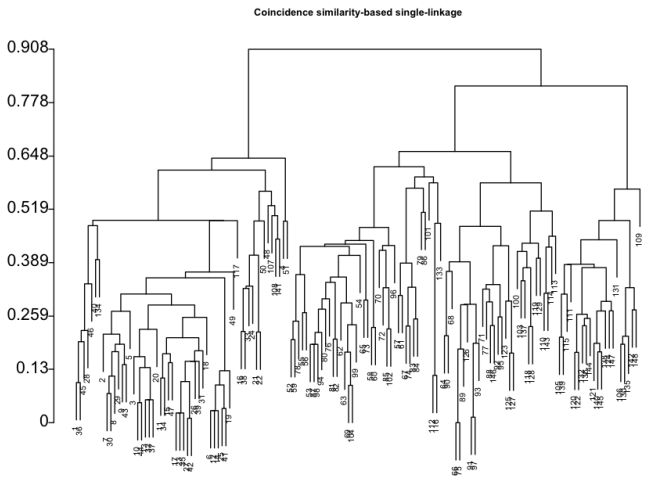


Figure 19: Dendrogram obtained for the handwritten characters dataset through single-linkage of coincidence similarities between the clusters. A well-balanced distribution of mergings is obtained at all scales while moderately emphasizing the intrinsic subdivision into three respective types of handwritten characters. Remarkably, virtually none of the intense chaining characterizing the Euclidean distance-based counter part can be observed.

It should be observed that the coincidence similarity index can also be adapted to the other clustering approaches, substituting the Euclidean distance or correlations whenever necessary.

6 Image Segmentation as a Laboratory

Image analysis and computer vision constitute important branches of artificial intelligence (e.g. [25, 5, 4, 6]) as a consequence of their impressive potential for automating and enhancing activities typically performed by humans, including prospection, surveillance, quality control, astronomy, to name but a few possibilities.

One of the first steps along the image analysis pipeline, the task of *image segmentation* (e.g. [25, 4]) is as critical as it is challenging. Basically, given an image, to segment it typically means identifying its portions of special relevance for being possibly related to specific objects in the image, or portions of these objects. This seemingly simple endeavor is complicated by several effects including noise, shadows, reflections, occlusions, and transparency, among several other unwanted interferences. The importance and challenge of image segmentation has been directly reflected to a so large number of related studies, based on the most varied areas and concepts.

As a consequence of some special characteristics, we argue here that the problem of image segmentation can provide a particularly interesting and effective laboratory not only for better understanding supervised and unsupervised pattern recognition, but also for developing and comparing respective concepts and methods. That is believed to be so as a consequence of the following aspects: (i) the original data to be classified (pixels) can be immediately inferred from the images; (ii) in the case of supervised recognition, the choice of prototypes can be easily performed, e.g. by clicking on specific points of the image; (iii) images in general have great complexity and intricacy, providing a comprehensive resource for testing methods; (iv) the effects of the pattern recognition can be immediately perceived in terms of the highlighted segmented regions, especially the identification of possible underlearning caused by biased sampled and undersampling.

Figure 20 illustrates the above possibilities with respect to the supervised segmentation of a color image of a landscape (a) including natural and human made objects and structures, while also incorporating varying levels of luminosity, shadows, and diverse types of backgrounds and textures. More specifically, segmentation results obtained by using Pearson correlation coefficient and coincidence similarity are shown respectively in (b) and (c), respectively to five prototype points marked with red crosses. In the former case, the segmentation generalized too much in detriment of the selectivity, which implied several structures and textures to be merged. The results obtained by the coincidence similarity resulted substantially more adherent to the structures from which the prototypes were taken with a moderate loss of generalization. In addition, given that a relatively high number of features was involved, namely 25 RGB pixels sorted by intensity within each color channel, the good adherence to the respective objects can be taken as an indication that underlearning is not taking place.

7 Concluding Remarks

Pattern recognition has progressed all the way from early promising approaches toward becoming one of the central current research subjects. This has been motivated by the many important applications to virtually every scientific and technological area and aspect. Yet, the question of evaluating the performance of pattern recognition while also identifying the main causes that can undermine it and devising possibilities of improvements, remains an ever important subject.

It should be kept in mind that all results in the present work are preliminary and still being complemented and evaluated. In addition, the application of pattern recognition as approached here should be taken as a resource for

(a) (b) (c)

Figure 20: Original image (a) and respective segmentations (b) by using the Pearson correlation coefficient between the selected features of the prototypes and those of all pixels in the image. The prototypes, are marked by red crosses, refer to the sandstone wall (2 samples) and the further away bridge (3 samples). The obtained regions, delimitated by respective red contours, can be observed not to adhere selectively to any of the types of structures in this image. In this case, the generalization prevailed strongly in detriment of the selectivity to the types of structures. The results obtained by the coincidence similarity (c) are characterized by a precise adherence to the types of structures in the image while maintaining an excellent generalization ability. These enhanced results are a direct consequence of important specific properties of the coincidence similarity operation, including its high selectivity/sensitivity while being substantially robust to localized feature perturbations [24].

gathering insights about the analyzer problem from the point of view of the interrelationship between its components that can lead to insights and better understanding, not as an absolute or definitive result. Indeed, the application and interpretation of pattern recognition should closely take into account the nature of the data, the questions to be worked, as well as the limitations of the features, classification methods as well as all other involved aspects.

This work addressed the issue of identifying the aspects that influence the performance of supervised and unsupervised pattern recognition from the perspective of statistical modeling of the original categories. Several related factors were addressed, with special attention given to the phenomena of biased sampling, undersampling, unerlearning caused by the former, as well as overfitting. Several important effects were identified and discussed with the help of some real-world data examples. Snippets have also been included in order to emphasize 10 main points discussed and addressed here, provide a good concluding remarks summary.

The developed concepts and methods as reported in the present work pave the way to several future developments. While the range of possibilities is particularly ample, some of the potentially mostly promising prospects are briefly presented in the following. Even though we considered several possible aspects influencing the performance of supervised and unsupervised recognition, it would be interesting to approach the issue of features normalization to greater depth, as this aspect can also strongly influence the recognition results. Several possibilities have also been established respectively to the phenomenon of underlearning, which has been argued to play a critically important role especially not only in the case of highly dimensional feature spaces, but even for moderate dimensions. In particular, it would be interesting to derive more complete tables of the artifact distances not only in terms of addition numbers of samples and dimensions, but also respectively to other normalizing intervals. Regarding the identification of clustering as involving two related tasks that can perhaps be performed more effectively in separate, it would be interesting to evaluate in a more systematic and comparative fashion how it would perform respectively to several types of synthetic and real-world datasets, including diverse types of noise and interferences. Another promising research line consists in considering further multiset-based similarity indices, and especially the coincidence approach, respectively to several other types of data and possible applications to supervised and unsupervised pattern recognition. Among many other related developments, it would be interesting to consider the parametric version of the

coincidence index, which allows for enhanced versatility in its applications.

Acknowledgments.

Luciano da F. Costa thanks CNPq (grant no. 307085/2018-0) and FAPESP (grant 15/22308-2).

References

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

[2] K. Fukunaga. *Statistical Pattern Recognition*. Morgan Kaufmann, San Diego, 1990.

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

[4] L. da F. Costa. *Shape Classification and Analysis: Theory and Practice*. CRC Press, Boca Raton, 2nd edition, 2009.

[5] B. K. P. Horn. *Robot Vision*. McGraw Hill, Cambridge, 1986.

[6] E. R. Davies. *Machine Vision*. Morgan Kaufmann, Amsterdam, 2005.

[7] R. A. Johnson and D.W. Wichern. *Applied multivariate analysis*. Prentice Hall, 2002.

[8] N. Mukhopadhyay. *Probability and Statistical Inference*. CRC Press, New York, 2000.

[9] S. Haykin. *Neural Networks And Learning Machines*. McGraw-Hill Education, 9th edition, 2013.

[10] D. Stoyan, W. S. Kendall, J. Mecke, and L. Ruschendorf. *Stochastic geometry and its applications*. Wiley Chichester, 1995.

[11] W. D. Blizard. Multiset theory. *Notre Dame Journal of Formal Logic*, 30:36—66, 1989.

[12] P. M. Mahalakshmi and P. Thangavelu. Properties of multisets. *International Journal of Innovative Technology and Exploring Engineering*, 8:1–4, 2019.

[13] D. Singh, M. Ibrahim, T. Yohana, and J. N. Singh. Complementation in multiset theory. *International Mathematical Forum*, 38:1877–1884, 2011.

[14] J. Hein. *Discrete Mathematics*. Jones & Bartlett Pub., 2003.

[15] D. E. Knuth. *The Art of Computing*. Addison Wesley, 1998.

[16] L. da F. Costa. Coincidence complex networks. https://iopscience.iop.org/article/10.1088/2632-072X/ac54c3, 2022. J. Phys.: Complexity, (3): 015012.

[17] 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 1-Oct-2020.

[18] S.-H. Cha. Comprehensive survey on distance/similarity measures between probability density functions. *Intl. J. Math. Models and Meths. in Appl. Sci.*, 1(4):300–307, 2007.

[19] C. E. Akbas, A. Bozkurt, M. T. Arslan, H. Aslanoglu, and A. E. Cetin. L1 norm based multiplication-free cosine similiarity measures for big data analysis. In *IEEE Computational Intelligence for Multimedia Understanding (IWCIM)*, France, Nov. 2014.

[20] M. K. Vijaymeena and K. Kavitha. A survey on similarity measures in text mining. *Machine Learning and Applications*, 3(1):19–28, 2016.

[21] L. da F. Costa. On similarity. https://www.sciencedirect.com/science/article/pii/S037843712200334X, 2022. Physica A: Statistical Mechanics and its Applications, 127456.

[22] Eric K. Tokuda, Cesar H. Comin, and Luciano da F. Costa. Revisiting agglomerative clustering. https://www.sciencedirect.com/science/article/pii/S0378437121007068, 2022. Physica A, 585: 26433.

[23] L. da F. Costa. Further generalizations of the Jaccard index. https://www.researchgate.net/publication/355381945_Further_

Generalizations_of_the_Jaccard_Index, 2021. [Online; accessed 21-Aug-2021].

[24] L. da F. Costa. Multiset neurons. https://www.researchgate.net/publication/356042155_Multiset_Neurons, 2021.

[25] R. C. Gonzalez and R. E. Woods and. *Digital Image Processing*. Pearson, New York, 2018.