Unsupervised learning

Dirk Valkenborg,[a,b] Axel-Jan Rousseau,[a,b] Melvin Geubbelmans,[a,b] and Tomasz Burzykowski[a,b,c]
Hasselt, Belgium, and Bialystok, Poland

As mentioned in the previous article,[1] unsupervised learning involves using datasets without clear notice of the dependent (response) variable. Unsupervised means that the machine or computer should learn patterns from the data without referring to any specific response. Unsupervised learning aims to explore the data structure and generate a hypothesis rather than to test any hypothesis by statistical methods or to construct prediction or classification models on the basis of a set of conditions and a specified response. Algorithms for unsupervised learning can be subdivided into 2 categories: (1) clustering algorithms and (2) informative data transformations

To better illustrate the concepts, we will use the dataset of Konstantonis et al[2] to investigate decisions about extraction and identification of treatment predictors in Class 1 malocclusions. The dataset comprises 542 randomly selected records of patients with a Class I relationship observed in a university graduate program and 5 private orthodontic offices. For each participant, several variables are observed: 26 cephalometric variables, 6 model measurements, 2 demographic variables (gender, age), and the type of treatment: nonextraction (397) or extraction of the 4 first premolars (145). More details about the dataset can be found in Konstantonis et al.[2] The scope of this study is evident as the authors want to predict the optimal treatment (response) given the set of explanatory (predictor) variables. Furthermore, they wanted to identify essential variables in predicting the treatment. The data can be presented in a tabular format that organizes all the information, as depicted in Table I.

Clustering

A clustering task can be best defined by an example. Consider the image in Figure 1. A task related to this image could be determining how many herds of animals with different genera are visible in this picture. On the basis of the physical characteristics of each animal, you could try to lump them into homogenous clusters (groups). In this example, you could cluster (group) the animals with black and white stripe patterns and place the horned animals with brownish fur in another cluster. To execute this task, it is not necessary to be an expert in wildebeest or zebra, nor is it required to have these animals tagged by a label that explains the genus of the animal. Clustering algorithms can discover this structure in a dataset without any prior knowledge. Toward this aim, a clustering algorithm will compute a distance measure to quantify similarity or dissimilarity between different subjects in the dataset. On the basis of this measure, subjects will be clustered (grouped) or split from each other to yield clusters (groups) that have the highest similarity within the cluster and the largest differences between the clusters.

Typically, a clustering method has 3 key elements: (1) a distance measure to quantify the similarity or dissimilarity between subjects; (2) an additional distance measure to quantify the difference between clusters or between a cluster and a subject (ie, linkage); and (3) a computer algorithm that maximizes the similarity within a cluster and the dissimilarity between the clusters. The variance is often used to measure the heterogeneity in a dataset. In this case, clustering will minimize the variance within the clusters and maximize the variance between the clusters.

The distance is a number that tells us how far 2 subjects are separated by considering the difference for each observed variable. In the next example, the Frankfort mandibular incisor angle (FMIA) and the incisor mandibular plane angle (IMPA) are examined for 3 patients in the dataset of Konstantonis et al.[2] Two patients exhibit an FMIA and IMPA combination of (41.8/113.0) and (52.0/114.5). These patients seem very alike when considering these 2 covariates, especially when contrasting these observations with our third patient, who has an FMIA and IMPA of (89.1/76.0). Intuitively, the third

[a]Center for Statistics, Hasselt University, Hasselt, Belgium.
[b]Data Science Institute, Hasselt University, Belgium.
[c]Department of Biostatistics and Medical Informatics, Medical University of Bialystok, Bialystok, Poland.
This work was supported by the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program.
Address correspondence to: Dirk Valkenborg, Hasselt University - Data Science Institute, Agoralaan 1, Building D, B-3590 Diepenbeek, Belgium; e-mail, dirk.valkenborg@uhasselt.be.
Am J Orthod Dentofacial Orthop 2023;163:877-82
0889-5406/$36.00

https://doi.org/10.1016/j.ajodo.2023.04.001

Table I. Data layout of Konstantonis et al[2]

Subject ID	*Cephalometric (26)*	*Model (6)*	*Demographic (2)*	*Treatment*
1	…	…	…	extraction
…	…	…	…	…
542	…	…	…	nonextraction

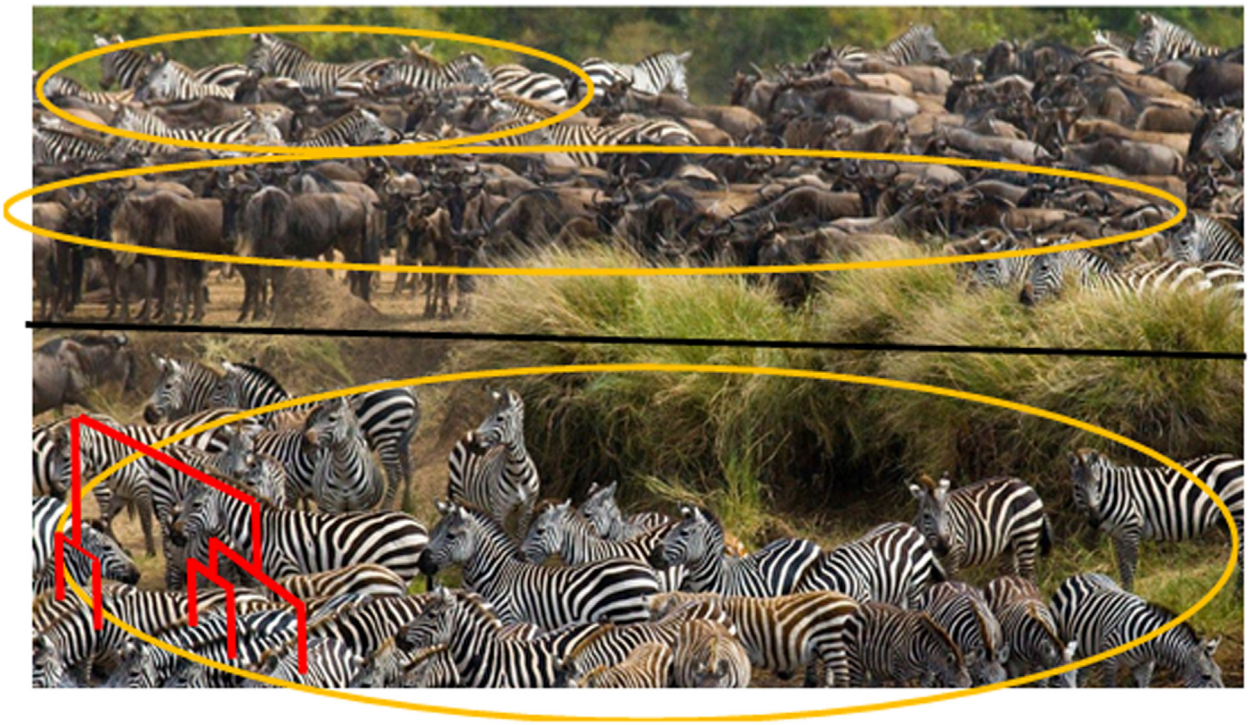

Fig 1. A mock example that illustrates the different clustering paradigms. K-means clustering is depicted in *yellow* as animals are grouped on the basis of their visual characteristics. Agglomerative hierarchical clustering is denoted in *red* as this method first looks for animals which are closest together in terms of visual approaches before extending the formed subgroups. *Black* indicates divisive hierarchical clustering as it will first split the herd into 2 subgroups before further splitting the subgroups up to the level of the individual animal.

patient is more distant from the other 2 subjects by eyeballing these covariates because of the large differences in FMIA and IMPA. However, when many variables are recorded in a study, we must compute a formal number expressing the dissimilarity between patients. Depending on the nature and the scale of the variables, a distance can be defined for each variable separately, and those distances can be subsequently combined in a single summary measure. A popular method to combine the computed distances across the variables is the Euclidean distance, the square root of the sum of squared distances. In the case of patients 1 and 2, the distance would be equal to the following: $D = \sqrt{(41.8-52)^2+(113-114.5)^2} = 10.31$. The Euclidean distance between patients 1 and 3 would become $D = \sqrt{(41.8-89.1)^2+(113-76)^2} = 60.05$. Alternatively, other measures (eg, Pearson's correlation) can also be used. Elaborate lists of distance measures exist, and they will yield different clustering results that will depend on the nature of the data.

Some clustering methods require a linkage measure to link a subject to a cluster or link 2 clusters with each other. Again, several measures are available: average, centroid, complete, single, or silhouette linkage. A default choice is the centroid linkage which considers the distance between the centroids (eg, center of gravity)

of the clusters. A cluster centroid can be easily computed by taking the mean of each variable over the subjects that compose the cluster. Again, the optimal choice for the linkage method depends on the nature of the data.

The clustering algorithm is a computer procedure that defines how the subjects and clusters are merged and split up. These algorithms can be divided into 2 groups: partitioning and hierarchical clustering methods.

Partitioning methods aggregate the data in a prespecified number of clusters, in which k-means clustering[3] is among the most common approaches. The main principle is that the cluster is nonoverlapping and nonempty (ie, every subject is a member of only 1 cluster), and every cluster has at least 1 subject as a member. The partitioning algorithm is an iterative procedure that first allocates subjects into nearby clusters on the basis of a distance measure. Next, the centroids of each cluster are recomputed on the basis of the membership of the subjects. The previous 2 steps are iterated until the centroids converge to a particular location in the data space. For instance, the yellow ellipsoids in Figure 1 indicate a possible positioning of the centroids if 3 clusters are specified. Partitioning methods have 2 disadvantages. First, the number of clusters is unknown, but it must be defined before the initiation of the algorithm, contributing to uncertainty in the result. Second, a stochastic mechanism is involved in the procedure, as the clustering will be initiated by randomly assigning the membership between subjects and clusters. As such, another algorithm run on the same dataset can yield different results because of this random component.

Hierarchical clustering methods[4] are deterministic algorithms that generate a dendrogram. A dendrogram is a tree-like structure drawn with the root on top and branches developing underneath, further splitting up the clusters until arriving at the leaves, as indicated in Figure 2. The leaves at the bottom of a dendrogram present the patients in our dataset. The dendrogram reflects a hierarchy of clusters on the basis of a degree of similarity. The branches (vertical lines from leaves or clusters) are combined into pairwise nodes (horizontal lines). The branch's length indicates the distance between 2 subjects, a subject, and a cluster, or between 2 clusters. A short branch means a high similarity, whereas a long branch means a low similarity. The dissimilarity increases with the vertical distance from the leaves. The top node is called the root and represents the entire dataset. A dendrogram or cluster tree can be generated in 2 manners:

1. Agglomerative or bottom-up clustering: This method is the most popular as it is the least computationally demanding. The algorithm starts by combining the 2 leaves closest to each other. Next, another leaf can be merged with the cluster, or 2 other leaves can be combined. Leaves and clusters are combined until the root node entails the entire dataset. The *red* connecting lines in Figure 1 visually depict the first steps of agglomerative clustering.
2. Divisive or top-down clustering: This method considers every possible split in the dataset and splits the root to maximize the distance between the 2 new clusters. This procedure is repeated until all clusters are split-up to the level of the individual leaves. The *black* line in Figure 1 indicates how a first split could look for divisive clustering, maximally separating the species into 2 distinct herds.

An example of agglomerative clustering is presented in Figure 2 on the 6 model variables from the Konstantonis et al[2] dataset. The dendrogram on the basis of 542 subjects seems cluttered; however, some structure can be added by coloring the obtained clusters. These clusters result from applying a user-defined threshold for the similarity at the y-axis. This threshold will cut the branches of the tree in such a way that clusters are formed. Figure 2 applies an arbitrary threshold of 5 to limit the number of clusters. When looking at these clusters, one can notice that the highlighted clusters have an overrepresentation of patients with the 4 first premolar extractions.

In contrast, the large yellow cluster on the left, which encompasses almost two-thirds of the dataset, contains only 55 patients with premolar extractions. These may be exciting observations but should not be given too much attention. Proper statistical or predictive methods should investigate whether the cooccurrence of clustered patients and the type of treatment they received (extraction or nonextraction) is meaningful. A common mistake is to overinterpret these types of dendrograms. For example, the proximity of leaves in the x-axis is often interpreted, but this direction has no meaning in hierarchical cluster analysis.

Informative data transformations

Another unsupervised learning technique involves a meaningful transformation of a high-dimensional data object into a lower dimension object such that the relevant information is preserved. The mathematics behind data transformations are complex and often involve concepts from statistics (variance-covariance matrix), calculus (Lagrange multiplier and constrained optimization), and linear algebra (eigenvalues or singular value decompositions). However, the basic idea can be easily explained philosophically.

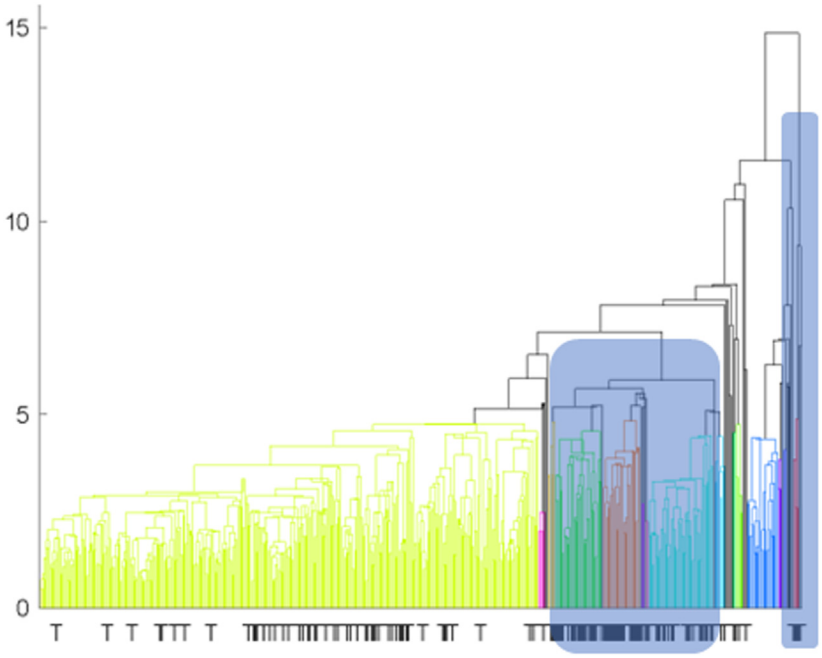

Fig 2. Result of agglomerative clustering on the dataset of Konstantonis et al[2] on the basis of the 6 model covariates. The patients that received treatment (ie, extraction) are indicated by the label "T" on the x-axis. This figure shows that some color-coded clusters are enriched or depleted with this label, which could trigger further research to investigate the correlation between the label and the respective patient groups.

The best way to start this explanation is with the famous Greek philosopher Plato and his allegory of the cave. The allegory discusses prisoners trapped inside a cave, forced to watch shadows projected on a wall from real-world 3-dimensional objects. These shadows are the prisoner's reality, ignoring the existence of a higher dimension of reality. When confronted with them outside the cave, would these prisoners recognize real-world objects? The correct answer is that it depends on the projection and the shape of the wall. If Plato knew about informative data transformations, he would have positioned the 3-dimensional objects in such a way that the projection maximally would reflect the shape of the object on a flat wall, as indicated in Figure 3. The projection on the right-hand side of the figure is an inadequate representation of how a ring that works in the real world would look.

In contrast, the projection on the left-hand side is informative, allowing a viewer to infer details about the shape of the ring. In this case, the shadow only omits information about the width of the ring in the higher dimension, which is information of little importance. There exist other objects that could cast similar shadows. Because our objective is to reduce the dimensionality, we are ignorant about which information is left out exactly by the data projection. However, some techniques allow us to quantify the amount of information loss.

An important concept in data transformation and dimensionality reduction is meaningful information. Which information would you like to have retained in the projection? For example, principal component analysis (PCA)[5] is one of the most popular data transformations that aims at preserving the variance of the original data. For this purpose, PCA computes principal components (PC) that are linear combinations of the variables in the dataset. The linear combination is optimized so that the newly formed PC maximizes the variance when the data are projected on this PC. After computing the PCs, we could restrict the representation to several components such that a prespecified amount of the original data variance is retained. The plot in Figure 4 displays the result of a PCA applied to the 6 model variables from the Konstantonis et al[2] datasets. Every point represents a subject in the study and is color-coded according to the patient's treatment status. *Red* indicates the patients that have undergone 4 first premolar extractions, and *blue* means no extraction. Instead of presenting the data in 6 dimensions corresponding to the 6 model variables, we can use the first 2 PCs, as shown in Figure 4. It can be observed

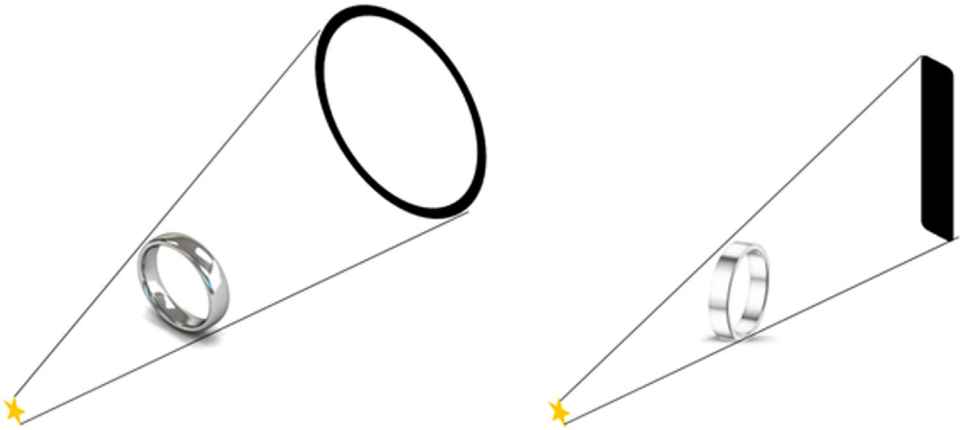

Fig 3. A hypothetical example of informative data transformations. The figure on the left provides a more informative projection of the ring than on the right.

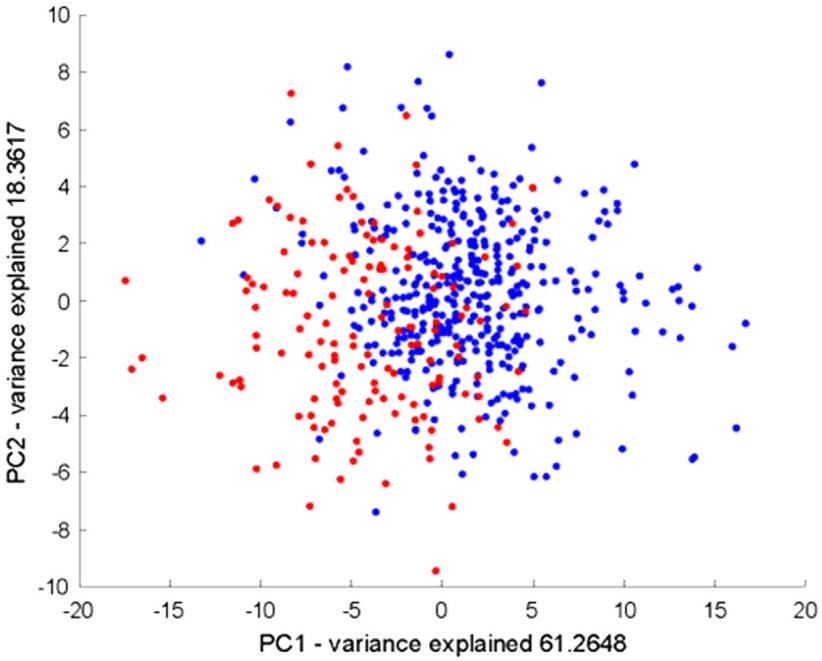

Fig 4. Result of a PCA on the dataset of Konstantonis et al[2] on the basis of the 6 model covariates. Based on this simple transformation, mild separation of the patient with and without treatment can be observed.

from the figure that the first PC (in the horizontal axis) represents 61.26% of the original data variance. Some separation of treated and nontreated patients along this axis seems possible. Adding more dimensions in the form of PCs and adding more variables to form linear combinations could improve the split between treated and nontreated patients.

Many other techniques exist that can reduce the dimensions of the data. Examples are self-organizing maps, autoencoders, t-distributed stochastic neighbor embedding, and Uniform manifold approximation and projection. Canonical correlation analysis focuses on maximizing the correlation between sets of variables. The mathematical details of these techniques are

complicated and beyond the scope of this series of articles on machine learning.

CONCLUSIONS

Unsupervised learning techniques like clustering and data transformation methods are often confused because they share the same objective (ie, finding patterns and structures in a dataset) and often present the results in a similar format that allows discovering some grouping of the subjects. However, the mathematical concepts underlying clustering and data transformations are very different.

Without a well-defined response, the challenge of unsupervised learning is to interpret the obtained clusters or informative transformations. Toward this aim, different variables in the dataset are often overlaid with the result of the unsupervised analysis to cherry-pick a promising response. This abductive reasoning is not a good practice. Instead, proper inferential techniques should be used to falsify the generated hypothesis. When a response is present in a dataset, as with the treatment variable in the Konstantonis et al[2] datasets, it is common to indicate this label in the resulting plots, as shown in Figures 2 and 4. This visualization verifies whether the unsupervised analysis can recover the dataset patterns associated with the response. There is nothing wrong with such an overlay. However, when the discovered patterns do not agree with the response, it is tempting to adjust the hyperparameters of the unsupervised learner so that the results align with the response labels. This is a dangerous practice that is no longer considered unsupervised learning. By providing feedback, the human-in-the-loop directs the unsupervised learner to focus on particular aspects of the dataset. To be clear, let us not forget that the goal of unsupervised learning is explorative and that obtained results should not be overinterpreted.

An important aspect of this data exploration is to realize that mathematical engines used for unsupervised learning are agnostic to the scale or unit represented by the numeric variables. Concepts such as distance measure or variance are influenced by these numeric values. For example, expressing a distance of 1000 mm or 1 meter is the same when considering the actual distance, but using different numeric scales can lead to different results from the unsupervised method. Therefore, it is good practice to center (subtract the mean) and standardize (divide by standard deviation) the data when variables mix different units. As a result, the mean of the variable is centered at 0 with a standard deviation of 1, ensuring that every variable is treated equally by the unsupervised learner.

In contrast, neither centering nor standardization is necessary when the variables are presented on the same scale. For example, when performing an unsupervised analysis of gene expression data, it could be the goal to focus more on abundant genes than on low-expression genes. In this case, centering and standardization would remove this information. In other words, when variables are on the same scale, it depends on whether variable centering and standardization is needed. It is worth keeping in mind that more weight is given to variables with large numeric values in an unsupervised analysis.

As a last remark, unsupervised learners are flexible when interpreting the data. In the analysis of the Konstantonis et al[2] dataset (Figs 2 and 4), clustering and data transformation were used to search for clusters (or groups) of individual patients. This viewpoint is inspired by the associated response variable (the treatment) present in the dataset. However, instead of focusing on the patient, an additional analysis could have been applied to the variables in the dataset to discover possible correlations. Such flexibility is not allowed in supervised learning, which focuses more on the response.

REFERENCES

1. Burzykowski T, Rousseau A-J, Geubbelmans M, Valkenborg D. Introduction to machine learning. Am J Orthod Dentofacial Orthop 2023;163:732-4.
2. Konstantonis D, Anthopoulou C, Makou M. Extraction decision and identification of treatment predictors in Class I malocclusions. Prog Orthod 2013;14:47.
3. Steinhaus H. Sur la division des corps matériels en parties. Bulletin L'Académie Polonaise des Science 1957;4:801-4. French.
4. Bridges CC. Hierarchical cluster analysis. Psychol Rep 1966;18:851-4.
5. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. The London Edinburgh and Dublin Philosophical Magazine and Journal of Science 1901;2:559-72.