Knowledge-Based Systems 217 (2021) 106830

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Archetype analysis: A new subspace outlier detection approach

Ismael Cabero [a], Irene Epifanio [b,*], Ana Piérola [c], Alfredo Ballester [c]

[a] Dept. Educació i Didàctiques Específiques. Universitat Jaume I, 12071 Castelló, Spain
[b] Dept. Matemàtiques - IF. Universitat Jaume I, 12071 Castelló, Spain
[c] Institut de Biomecànica de València, 46022 València, Spain

ARTICLE INFO

Article history:
Received 12 January 2020
Received in revised form 27 January 2021
Accepted 2 February 2021
Available online 6 February 2021

Keywords:
Archetypal analysis
Unsupervised anomaly detection
Nearest neighbors
Ensembles
Multivariate outlier detection
Footwear

ABSTRACT

The problem of detecting outliers in multivariate data sets with continuous numerical features is addressed by a new method. This method combines projections into relevant subspaces by archetype analysis with a nearest neighbor algorithm, through an appropriate ensemble of the results. Our method is able to detect an anomaly in a simple data set with a linear correlation of two features, while other methods fail to recognize that anomaly. Our method performs among top in an extensive comparison with 23 state-of-the-art outlier detection algorithms with several benchmark data sets. Finally, a novel industrial data set is introduced, and an outlier analysis is carried out to improve the fit of footwear, since this kind of analysis has never been fully exploited in the anthropometric field.

1. Introduction

Nowadays, we tend to work with enormous amounts of data and variables, which greatly hinders their analysis. It is necessary to perform a quality analysis to avoid making wrong decisions. One of the possible causes of such decisions is outliers. A classic definition of an outlier given by [1], is "an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism". Outliers can also be defined as observations whose characteristics differ significantly from the normal profile.

Detection of outliers is an early and necessary step in any data analysis application. Although outliers are considered in many cases as noise or errors, they often incorporate vital information. But it is clear that what is noise for one person can be a focus of interest for another [2], and depending on what you are studying, an outlier (e.g. in the detection of credit card fraud) can be of great importance. Failure to look for them and study them can lead to poor specification of the model and an incorrect estimate of its parameters. It is therefore important to identify them prior to modeling and analysis [3].

Identifying outliers in univariate data is relatively simple because it is easy to find the extreme cases. However, in the multivariate case, the detection of outliers is more difficult because multidimensional outliers are observations that are considered strange not because of the value they take in a certain variable, but due to the value in all of them [4].

Many techniques have been proposed for the detection of outliers along time (see [5] for a detail explanation of many of them for different types of data). In the case of multivariate data, [6] reviewed and compared many of the most standard unsupervised anomaly detection algorithms in a set of benchmark data sets. A similar study was carried out by [7] and [8]. [6] proposed a taxonomy of unsupervised anomaly detection algorithms, which are divided into four categories: (1) Nearest-neighbor (NN) based techniques, (2) Clustering-based methods, (3) Statistical algorithms, and (4) Subspace techniques.

We propose a new method for unsupervised (no labels are available) detection of outliers in continuous multivariate data. It can be categorized into several of those categories, mainly (1) and (4), because it uses an unsupervised learning technique (a subspace technique), which can also be used as a clustering technique [9], and it also relies on NN-based techniques. Note that techniques based on distances are very popular due to their good results, conceptual simplicity and interpretability. However, when the number of features is high, these techniques can fail because of the curse of dimensionality. A key point to solve this problem would be to eliminate the dimensions and project the data into subspaces, where outliers can be easily revealed. Projection into appropriate subspaces can also improve distance-based techniques. This is the idea of the proposed method: first to project the data into the relevant subspaces and then to use proximity-based techniques to detect outliers in those subspaces.

The proposed method, which we refer to as AA + k-NN, is based on Archetype Analysis (AA), the objective of which is to

* Corresponding author.
E-mail addresses: icabero@uji.es (I. Cabero), epifanio@uji.es (I. Epifanio), ana.pierola@ibv.org (A. Piérola), alfredo.ballester@ibv.org (A. Ballester).

https://doi.org/10.1016/j.knosys.2021.106830

represent the observations by means of a mixture of archetypes, which are a mixture of observations. Archetypes lie on the boundary of the convex hull of the data, meaning that they are extreme profiles. This makes AA sensitive to outliers, and we will take advantage of this in order to detect outliers. AA is not a parametric technique, it is a data-driven method, so we do not have to make any assumption about data distribution. Furthermore, the results returned by AA are easily interpretable, even for non-experts. The combination of AA together with proximity-based methods therefore results in a non-parametric method with a high level of interpretability, which is very important in many applications.

AA was defined by [10] and has been applied in a broad spectrum of fields, such as biology [11], developmental psychology [12], didactics [13], engineering [14–20], finance [21], genetics [22], global development [23], image processing [24], machine learning problems [25], market research [26], multi-document summarization [27], neuroscience [28,29] and sports [30–32]. With AA we can see all samples by looking at a few based on extreme profiles, but these extreme profiles should not be outliers. In fact, there are several works that try to robustify AA in order to make AA immune to outliers [21,33]. In this work, we go against the trend: we do not care that AA returns outliers among the archetypes, indeed this can be good for detecting outliers. We exploit AA's weakness (being sensitive to outliers) with respect to its original objective, and convert it into a strength for finding outliers. To the best of our knowledge, this is the first time that AA is used for finding outliers in multivariate data.

The main contributions of this paper are as follows: we present a new method for unsupervised detection of outliers in multivariate data. We conduct an experimental evaluation with a large number of well-known data sets and standard algorithms. In this comparison our new proposal provides very favorable results. Furthermore, we apply the new method to an original data set of foot measurements, which is used in an engineering problem that we introduce here. This is our motivating problem. Outlier detection in Anthropometry has only been used as a cleansing technique for correcting or removing the outliers before analyzing data [34,35]. However, outliers report very valuable information in the footwear design process, since they can show which kinds of feet are more different from the rest and may therefore pose fitting problems in footwear if the design is not appropriate.

The rest of this article is organized as follows. Section 2 presents the standard methods used in the comparison, while AA is reviewed in Section 3. In Section 4, we introduce our method. Section 5 presents the advantages of our method and the results of the comparison. The new methodology is applied to a new data set in an engineering problem in Section 6. Finally, we finish with some conclusions and future prospects for further research in Section 7.

2. Related work: Unsupervised anomaly detection algorithms

There are a huge number of unsupervised anomaly detection algorithms. Let us take a quick look at the most widely used in practice and those used by [6]. Furthermore, these algorithms will be used in the comparison.

k-NN Anomaly Detection**:** This algorithm searches for the nearest k-neighbors for every element in the data set and calculates the average distance of the k-neighbors. This procedure returns outlier scores, which depends on the selection of k. In the experiments, we follow the same strategy as in [6]: values from $k = 10$ to 50 are considered and averaged in order to achieve a fair evaluation when comparing algorithms. It focuses on global outliers.

*k*th-NN Global Anomaly Detection : As above, but once we have the nearest k-neighbors, only the distance of the kth nearest neighbor is considered. It also focuses on global outliers.

Local Outlier Factor (LOF) and LOF-upper bond (LOF − UB)**:** This algorithm is designed to find local outliers. It follows these steps: (1) search for the k-NN for each observation; (2) compute the local density for each observation; (3) the LOF score is computed by comparing the local densities of each observation with those of its k neighbors. See [36] for details. This algorithm finds local outliers and also global ones, but if we are only interested in global outliers, we will have a lot of false alarms. The choice of k will have a great influence on the results. Therefore, we will follow the same strategy as in [6]: scores for different k's up to an upper bound are calculated and the maximum of these scores is considered. This strategy is referred to as LOF. However, we can also consider different upper bounds and average the results. This strategy is referred to as LOF-UB.

Connectivity-Based Outlier Factor (COF)**:** This algorithm, proposed by [37] works like LOF except that instead of using the Euclidean distance, COF uses the "chaining distance"; this distance is the minimum sum of all the distances connecting all the k-neighbors and the case. The objective of changing the distance is to avoid the lack of precision of LOF when the density of the data that is around the observation has some kind of linear correlation, which is not appreciated with the inherent sphericity of the Euclidean distance.

Influenced Outlierness (INFLO)**:** If the data set has close clusters with very different densities, it is possible that an algorithm such as LOF may identify the border points between cluster as outliers. To avoid this, [38] proposed to work just like LOF but also taking into account the "reverse neighbors". This makes it possible to calculate the outlier scores of that kind of points more precisely.

Local Outlier Probability (LoOP)**:** LoOP [39] instead of assigning a score to the outlier, it gives the probability that each element is an outlier. LoOP also studies the local density of the neighbors around each element, but it assumes that the distances to the nearest neighbors follow a Gaussian distribution.

Local Correlation Integral (LOCI)**:** This algorithm tries to eliminate the difficulty of choosing the best k (number of neighbors). To do this, instead of looking at the nearest k-neighbors, it takes a circle of radius r around the case and study the density that exists in it. This radius r expands over time and, like LoOP, it also calculates the density using a Gaussian average distribution and compares two neighborhoods of different sizes instead of the ratio of local densities. A parameter α controls the relationship between the different neighborhoods. See [40] for details. Computationally, it is a very expensive algorithm and is too slow for large data sets so, as stated by [41], it can only be applied to very small data sets (at most 3000 observations).

Approximate Local Correlation Integral (aLOCI)**:** In order to reduce the computational cost of LOCI, this algorithm uses quad trees and some restrictions on α. However, due to these approximations, its performance can sometimes be very poor [6,41].

Cluster-Based Local Outlier Factor (CBLOF/uCBLOF): The algorithm proposed by [42] no longer uses the NNs to estimate density, but divides the data into different clusters and determines the density of each cluster. The most commonly used algorithm is k-means due to its small computational cost. Then clusters are classified as large and small. For the large ones, a weighted distance from the center to each element of the cluster is calculated and for the small ones the distance to the nearest large cluster is calculated. However, the weighting strategy used can lead to an incorrect density estimation [6]. A modified version to solve this problem is uCBLOF, which simply neglects the weighting. Algorithms based on cluster analysis still have a problem similar to those of k-neighbors, because they have to choose the number k of clusters.

Local Density Cluster-based Outlier Factor (LDCOF): LDCOF [43] is analogous to the previous procedure, but now for each cluster the average distance of all cluster members to the centroid is calculated. Then the score is calculated by dividing the distance of an observation to its cluster center by the average distance.

Clustering-based Multivariate Gaussian Outlier Score (CMGOS-Red, Reg and MCD): This algorithm works like the previous ones, finding clusters with the k-means and separating them into large and small ones. For each cluster, the covariance matrix $\boldsymbol{\Sigma}$ is robustly estimated by three different procedures that give rise to the CMGOS-Red, CMGOS-Reg and CMGOS-MCD algorithms (see [6] for details). The outlier score is calculated by dividing the Mahalanobis distance from one observation to its nearest cluster center by the certain percentile of the chi-square distribution.

Histogram-based Outlier Score (HBOS): HBOS [44] is an algorithm that assumes the independence of the variables. For each variable, a histogram is computed and normalized, and the height of each bin is used to compute the outlier scores. The histogram can be created in different ways, and the number k of bins also influences the results.

Robust Principal Component Analysis (rPCA): [45] use robust principal component analysis, and in particular, the major and minor components.

One-Class Support Vector Machine (oc-SVM and ν-oc-SVM): one -class SVM with robust techniques and a modification in the objective function (a ν parameter is included) is trained using the data set and afterwards, each observation is scored by a normalized distance to the determined decision boundary [46].

2.1. Recent techniques

Besides comparison with the previous state-of-the-art algorithms, we compare our technique with recent techniques:

Relative Density-based Outlier Score (RDOS): It was proposed by [47]. This procedure uses a density-based outlier detection approach with local kernel density estimation, and instead of using only k nearest neighbors, they also consider reverse nearest neighbors and shared nearest neighbors of a case for density distribution estimation. We use the implementation of the R package **DDoutlier** [48].

Virtual Outlier Score (VOS): It was proposed by [49]. In Section 5.3.3 we also compare our proposal with VOS. This technique is based on graphs. They compared their method with other outlier detection techniques, such as Isolation Forest (IForest) [50], an improvement of this technique (OIF) [51], and other techniques based on graphs, such as Outlier Detection using Indegree Number (ODIN) [52], OutRank [53,54] and Hierarchical Contextual Outlier Detection (HCOD) [55].

3. Background: fundamentals of archetype analysis

AA is an unsupervised statistical learning technique [56, Chapter 14]. AA seeks out extreme profiles called archetypes, which are restricted to being a convex combination of the elements of the data set. Also, these archetypes will represent each individual in our data set as a convex combination of the archetypes. Expressing the observations as mixtures of extremes profiles facilitates comprehension of the data. Humans understand the data better when the instances are shown through their extreme constituents [57] or when features of one instance are shown as opposed to those of another [58].

AA lies somewhere in between two well-known unsupervised statistical techniques: Principal Component Analysis (PCA) and cluster analysis. Those techniques are also data decomposition techniques, where a data matrix is decomposed as a linear combination of several factors to find the latent components. Depending on the decomposition, different techniques are obtained. A table summarizing the relationship between several unsupervised techniques is provided by [25] and [15]. With PCA, factors are linear combinations of features; therefore, they are the least restrictive. On the other hand, PCA bases are difficult to interpret as unsupervised learning tool, while the factors of clustering techniques, such as the centroids (averages of groups of data) of k-means, have more restrictions in their set-up, but their interpretation is very easy. However, their modeling flexibility is compromised due to the binary assignment of data to the clusters. AA lies in between PCA and cluster tools, with higher modeling flexibility than cluster techniques but without losing the interpretability of their factors (see [17,19,25] for seeing examples where PCA, cluster analysis and AA are compared).

3.1. AA definition

Let us review AA definition. Let $\mathbf{X}$ be an $n \times m$ matrix that represents a data set with n observations and m features. AA goal is to find a $p \times m$ matrix $\mathbf{Z}$ that characterizes the archetypal patterns of the data so that each data point can be represented as a mixture of these archetypes. Specifically, AA tries to obtain the two $n \times p$ matrices of the coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that minimize the residual sum of the squares (RSS) that arise from the equation that shows x_i as an approximation of a convex combination of archetypes z_j and the equations that show z_j as a convex combination of data:

$$RSS = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{p} \alpha_{ij} z_j\|^2 = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{p} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} x_l\|^2, \qquad (1)$$

with two conditions: (1) $\sum_{j=1}^{p} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ and $i = 1, \ldots, n$, and (2) $\sum_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \geq 0$ and $j = 1, \ldots, p$.

Therefore from (1) the approximations of x_i are a finite archetypal mixture $\hat{x}_i = \sum_{j=1}^{k} \alpha_{ij} z_j$ and the α_{ij} will indicate the weight of each archetype z_j for the element x_i. On the other hand, restriction (2) will show that the archetypes z_j are convex combinations of the data, $z_j = \sum_{l=1}^{n} \beta_{jl} x_l$.

AA is an exploratory data analysis (EDA) tool that is based on a geometric formulation (no distribution of data is assumed). [10]

showed that archetypes are on the boundary of the convex hull of the data if $p > 1$ (the archetype coincides with the mean for $p = 1$).

3.2. AA computation

[10] developed an alternating minimizing algorithm to compute the matrices in the AA problem, where the best α for given archetypes $\mathbf{Z}$ and the best archetypes $\mathbf{Z}$ for a given α are estimated by turns. A penalized version of the non-negative least squares algorithm by [59] is used to solve the convex least squares problems. That algorithm was implemented in the R package **archetypes** by [60], although with some modifications. For example, the spectral norm in Eq. (1) is used instead of the Frobenius norm for matrices. We reverted those modifications in our R implementation, i.e. the objective function to minimize is defined by Eq. (1). This algorithm is not deterministic, AA is run beginning from 20 random initializations, and the best model is selected for each p.

Archetypes are not necessarily nested or orthogonal to one another, so the selection of p is an important issue. A simple but effective heuristic tool for choosing p, which has been used elsewhere [10,15,60,61], is the elbow criterion. With the elbow criterion, we plot the RSS for different p values and the value of p is selected as the point where the elbow is located. Nevertheless, in our case, once the elbow has been identified, the selection of p is not as critical as in problems with merely EDA objectives, where only one p needs to be selected for interpretative purposes and where we want to avoid outliers being selected as archetypes. However, for our purposes, this is not the case; we can consider different p values, since we prefer to better capture the shape of the data set by changing the resulting archetypes and collect the information for different p values. For us, it is not a problem that an archetype is an outlier, since our objective is to detect them and, in fact, this can facilitate the mission. Note that the determination of e does not introduce any computational burden, since we only need to display the screeplot.

4. The AA + k-NN method for detecting anomalies

As explained in Section 1, the idea of our proposal is to project the continuous multivariate data into the relevant subspaces and then to use proximity-based techniques to detect outliers in those subspaces. Feature extraction is a well-known and powerful method for improving the performance of learning algorithms [56]. In the same way, a sensible tactic in the outlier detection field is to identify relevant subspaces where outlier analysis can be honed, i.e. where outliers deviate clearly from the normal observations after projection on the relevant subspaces, and then combine the results from different subspaces in order to create a more robust ensemble [5].

An overall picture of our method is to compute AA for a certain p and project the data. We then apply the k-NN method to the α values. Note that AA actually seeks extreme profiles, so we can take advantage of this fact. If we repeat the procedure for different p values, we will have different explanations of the data, and we can use independent ensembles (the combined procedures are independent) to combine the results.

The outline of the procedure is as follows:

Step 1 Min–max normalize or standardize the data.

Step 2 Compute AA from $p = 1$ to $p = P$, and determine the value $p = e$ where the elbow is found.

Step 3 Apply k-NN (sum of distance to k nearest neighbors) for a certain k for the α matrices from $p = e$ to $p = P$. Then, the $P - e + 1$ outlier scores obtained in each subspace are merged by a cumulative-sum approach, which is equivalent to averaging the scores.

Note that the bias–variance trade-off in anomaly detection is almost identical to that in classification [5], so it follows that averaging also reduces variance in anomaly detection.

This procedure returns outlier scores; as usual, the highest score denotes the highest degree of outlierness. A way to establish a binary decision about whether or not to label a point as an outlier, is to use a box-plot with the outlier scores and to consider the points detected as outliers by the box-plot as anomalies. Obviously, this hardening method will work well if the outlier scores corresponding to true outliers are well separated from those of the normal cases.

Let us give the details of each step. In Step 1, we consider both alternatives in the experiments: min–max normalization and standardization. In Step 2, we consider two values in the experiments, $P = 10$ and $P = 15$. In Step 3, we begin with e, since it is expected to be the first value for which archetypes explain the data well. The following values from e to P should also describe the data well, but may give different descriptions. This can be desirable, since diversity and accuracy are two key factors in the success of ensembles. The aggregation ensemble of Step 3 is valid since the scale of the outlier scores is the same, as we use the same k each time. Also note that the α values always add 1, for any p.

In the experiments, instead of considering a single k, we evaluate the procedure for different k values, from $k = 10$ to 50, as in [6]. Then the summary, mean and standard deviation for all these AUCs are calculated. If we wanted to report the results in terms of binary labels, i.e. to convert the scores into binary labels, we could use the box plot-based hardening strategy explained above for each k, then the final decision can be given by aggregation and majority voting [62]. In other words, we have the binary labels for each k, from $k = 10$ to 50, and finally, we consider the points that are labeled as anomalies at least 50% of the times as outliers.

Our proposal is composed of two major computational parts. The computational complexity of the algorithm used to compute AA has been analyzed in detail by [60]. It is a compute intensive method. More efficient alternative algorithms for computing AA have been proposed, especially for large data sets, such as the implementation by [25,63,64] and [65]. On the other hand, k-NN may require $O(n^2)$ time to compute all k-nearest neighbor distances [5]. Therefore, AA + k-NN is not a computationally efficient method, but this may be compensated for by its ease of interpretability and intuitive analysis, and its mathematical precision (effectiveness).

5. Results and discussion

5.1. Evaluation measures

In order to assess the algorithms for detection of outliers with unsupervised data, apart from taking into account the accuracy, the order of the outliers must be considered, especially because outlier scores are available. Therefore, we reproduce the same strategy followed by [6], which consists of ranking the outlier scores and iteratively applying a threshold from the first to the last rank. In this way, n tuple values (true positive rate and false positive rate) are obtained, and a single receiver operator characteristic (ROC) is generated. As an assessment measure we use the integral of the ROC, i.e. the area under the curve (AUC).

Note that the AUC value can be interpreted as the probability that an outlier detection algorithm will assign a lower score to a randomly chosen normal observation than to a randomly chosen anomalous observation [66]. On the other hand, note that many algorithms depend on a parameter, e.g. k for all the NN or clustering-based algorithms, which can be critical. In order to ensure fair comparisons, we also follow the same strategy as [6] and compute the AUC from $k = 10$ to $k = 50$ in Section 5.3. Then the AUC results are averaged and the standard deviation is also computed.

5.2. Artificial data sets

The proposed procedure is illustrated with two toy data sets. In the first example, the whole procedure is illustrated and compared with the k-NN method. In the second example, we show how our procedure can effectively detect the outliers, unlike other techniques.

5.2.1. First synthetic data set

The first example is shown in Fig. 1(a). The plot consists of two Gaussian clusters, whose data points are represented by solid green circles, plus five uniformly sampled outliers that are represented by red unfilled circles. If we directly apply the k-NN method to these data with $k = 5$ and a box-plot to the outlier scores, the five outliers are detected, but another five points are also falsely labeled as outliers.

We compute the archetypes from 1 to 6 (P) and the screeplot is displayed in Fig. 1(b). The elbow is found at $p = 2$ (e), so we consider the alpha values of AA from $p = 2$ to 6. As an example, we display the α values for $p = 3$ in a ternary plot in Fig. 1(c), where the outliers are represented by red triangles, while the rest of the points are represented by black circles. The archetypes for $p = 3$ are represented as black crosses in Fig. 1(a). We apply the k-NN method to the α values with $k = 5$, from $p = e = 2$ and $p = P = 6$, and the outlier scores are the sum of these results. The outlier scores are visualized by the bubble-size of each case in Fig. 1(a). Then we apply a box-plot to the outlier scores: the five outliers are detected and only one point is falsely labeled as an outlier. In summary, our procedure gives only one error, unlike the five errors given by k-NN in this example. AA + k-NN therefore manages to distinguish between anomalies and normal instances better than simply using k-NN. This also happens if we change the k value, for example for k from 1 to 7 and 10 to 14. For k values that are higher than 14 the same number of errors, 2, are obtained for AA + k-NN and k-NN.

Note that AA is a very intuitive tool and its results are easily interpretable, much more so than a PCA transformation. The data in Fig. 1(a) are expressed as a mixture (the α values represented in Fig. 1(c)) of the archetypes (the black crosses). For example, the outlier located at (2, 1.8) is expressed as 26% of archetype 1, plus 44% of archetype 2, plus 30% of archetype 3. Interpretability is a valuable factor for the analyst [5]. Knowing the reasons why a particular data point is labeled as an outlier, i.e. discovering the intensional knowledge about the outliers [67], can be a great help in real applications, as will be shown in Section 6.

5.2.2. Second synthetic data set

In the second example, we show a simple two-dimensional data set, where the features have a linear dependency, except one point that does not follow the linear relationship of the other points. We apply eight different procedures, and the outlier scores are shown in Fig. 2 as above, by the bubble-size of each case, with filled green circles denoting normal cases and an unfilled red circle denoting the outlier. We consider $k = 3$ for the NN algorithms. For our procedure, the elbow is at $p = e = 2$, and we consider $p = 2$ and 3 for the α computation, since the convex hull of this data set is formed by three vertices. We consider a new procedure here. In RPCA + k -NN, we follow a similar procedure as in our proposal, but instead of using AA, robust PCA is considered and k-NN is used with the PC scores. This alternative method is considered to show that AA is more useful than (robust) PCA for this situation (in fact, the same result as k-NN is obtained, since PCA rotates the data).

Table 1
The AUC results for example 2 (first row) and the rank of outlierness of the anomalous point (second row). As there are 11 points, the highest possible rank is 11, which corresponds with the lowest degree of outlierness.

AA + k-NN	k-NN	LOF	COF	LoOP	HBOS	RPCA + k-NN	RDOS
1	0.3	0	0.5	0.7	0.2	0.3	0.6
1	8	11	6	4	9	8	5

We also compute the AUC values, which are shown in Table 1 together with the rank of outlierness of the anomalous point, i.e. the highest rank denotes the highest degree of outlierness. For AA + k-NN, the outlier score of the anomalous point is more than double the next highest score. For k-NN the outlier score of the anomalous point is the fourth lowest (remember that the data set is composed of 11 points). For LOF, the outlier score of the anomalous point is the lowest (0.86, below 1, when 1 is supposed to be the score for normal cases). For COF, the outlier scores are the same for all the points. For LoOP, the probability that the anomalous point is an outlier is only 0.17. In fact, the probabilities are higher for three other points. For HBOS, the outlier score of the anomalous point is the third lowest. For RPCA + k-NN, the outlier score of the anomalous point is the fourth lowest. Finally, the outlier score of the anomalous point is the fifth highest for RDOS. In summary, our procedure provides the highest AUC of all the algorithms in this example. Note that although COF was designed to detect this kind of anomaly, it is not able to identify it, because in this case the anomaly is too close to the other points. With the same distribution of points, if the anomaly was $(5, 5 + \epsilon)$, with $\epsilon > 0$, instead of (5,5), COF would be able to detect it. However, our procedure can detect any anomalous point outside the pattern of the linear relationship, since that anomalous point would not belong to the convex hull of the rest of the data, i.e it is a vertex of the convex hull of the whole data set and, therefore, it is used as an archetype in AA for $p = 3$.

5.3. Real data sets

5.3.1. Benchmark data sets

We use all the data sets employed by [6] that contain continuous numerical features. Remember that our proposed method is appropriate for continuous numerical data. The data sets have been obtained from multiple sources, such as [68], and can be found in [69]. The data sets contain a variable with labels that indicate whether or not an observation is an outlier. However, we work with an unsupervised anomaly detection method and labels will not be used, except at the end for assessing the results. Details about the construction of the data sets can be found in [6], but a brief summary of them is given below.

Breast Cancer Wisconsin (Diagnostic)**:** This data set is composed of 367 individuals with 30 different features and 10 anomalies, which represent 2.72%. This data set focuses on the diagnosis of breast cancer to discriminate between benign and malignant tumors. It includes a set of features of cell nuclei from a digitized image of a fine needle aspirate (FNA) from a breast mass. The anomalies correspond to malignant instances, while the rest of the data set consists of benign instances.

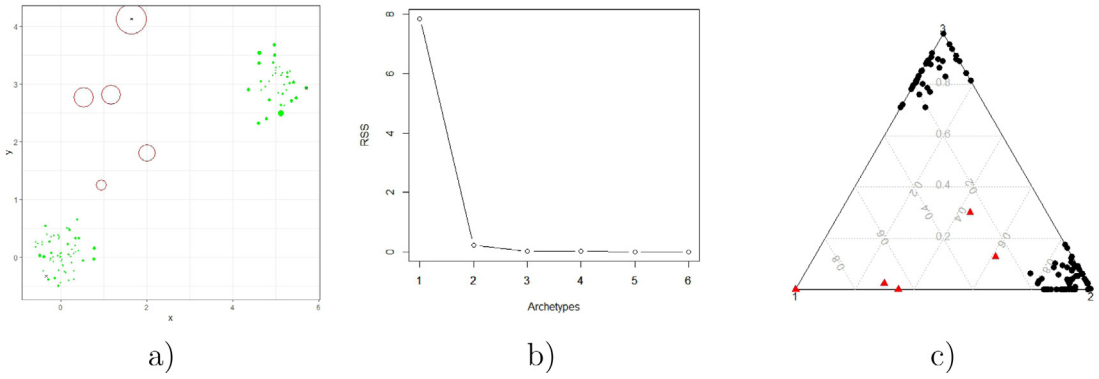

Fig. 1. Example 1: (a) plot of the data set (see the text for details); (b) Screeplot; (c) Ternary plot.

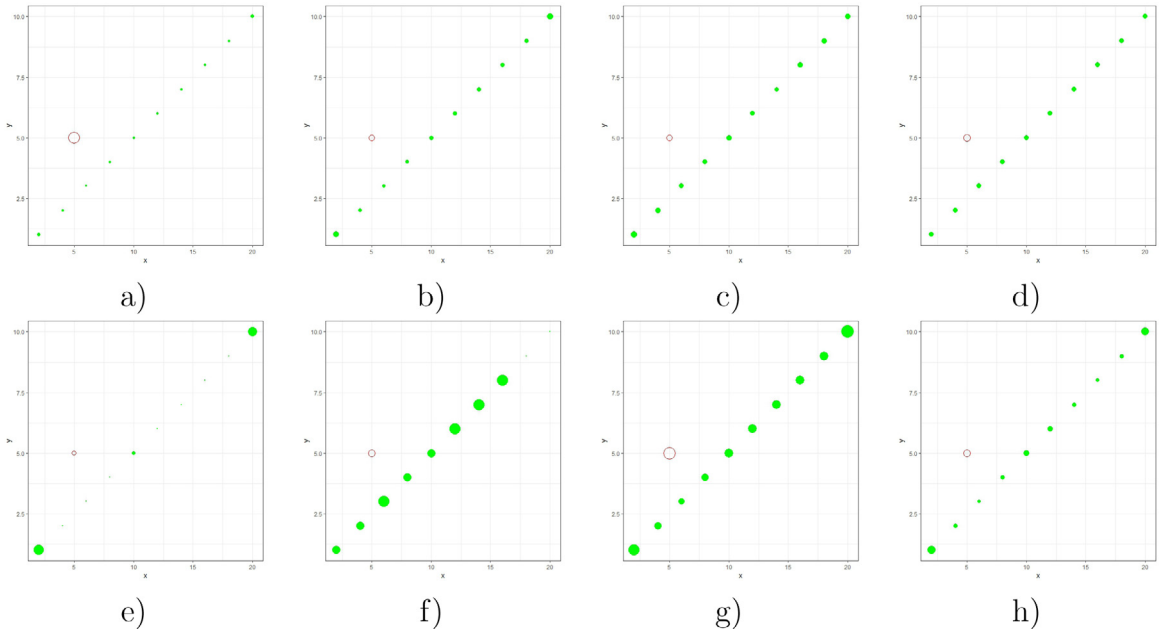

Fig. 2. Example 2, plot of the anomaly scores for different algorithms (see the text for details): (a) AA + k-NN; (b) k-NN; (c) LOF; (d) COF; (e) LoOP; (f) HBOS; (g) RPCA + k-NN; (h) RDOS.

Pen-Based Recognition of Handwritten Text (global): This set has 16 features and 809 observations, 11.1% of which are anomalies. This data set contains handwritten digits; in particular, the digit 8 is considered as the normal class and a sample of 10 digits from all of the other classes are considered anomalies. Therefore, there is a large normal class, and the anomalies are very different from each other.

Pen-Based Recognition of Handwritten Text (local): This data set also contains 0–9 handwritten digits. Specifically, it has 6724 cases with 16 variables. There are 9 large clusters corresponding to all digits, except the digit 4. For this class, only 10 cases are considered and they are therefore considered outliers, representing 0.15% of the data.

Before applying any method for detecting anomalies, the data should be preprocessed so that the features have equal weights. [6] use classic min–max normalization (the range transformation scales the data to be within [0, 1]). With our proposal, besides min–max normalization, we also use standardization (the mean is subtracted and values are divided by the standard deviation), since it is the common preprocessing procedure in AA.

5.3.2. Results of the benchmark data sets

Table 2 shows the results for all the algorithms and benchmark data sets. Our proposal has been run using min–max normalization as in [6], for fair comparison, and $e = 4$ and $P = 10$ in all cases. The best result for each data set is highlighted in bold font. Note that our proposal is the best for the breast cancer and pen local data sets. For pen global, the result of AA + k-NN is also very competitive, it is the third best among the twenty algorithms.

Nevertheless, AA was frequently used with standardized data so Table 3 reports the results of our proposal when the two options for Step 1 (normalization or standardization) and Step 2 ($P = 10$ or $P = 15$) are used. Note that the good results for the breast cancer and pen local data sets are improved if the data are standardized rather than normalized.

We also analyze the stability of the results of our proposal if e is changed and also if a different range of k-values are used. The breast-cancer data set is used as an illustration and the results are shown in Table 4. We see that the best results are achieved with $e = 2$ and with small k values. With this option, we can convert the outlier scores into binary levels as explained above. Table 5 shows the confusion matrices with min–max normalization, k from 5 to 15, and different e. Zero indicates a normal case, whereas one indicates an outlier. All the outliers (10) are correctly identified,

Table 2
Average AUC together with the standard deviation for each algorithm and benchmark data set. Due to the computational complexity, LOCI could not be computed for larger data sets, as pen local data set.

Method	Breast cancer	Pen global	Pen local
AA + k-NN	**0.9851** ±0.0030	0.9634 ±0.0030	**0.9915** ±0.0013
k-NN	0.9791 ±0.001	**0.9872** ±0.0055	0.9837 ±0.0018
kth -NN	0.9807 ±0.0008	0.9778 ±0.0142	0.9757 ±0.0069
LOF	0.9816 ±0.0024	0.8495 ±0.0679	0.9877 ±0.0016
LOF-UB	0.9805 ±0.0020	0.8541 ±0.0777	0.9876 ±0.0013
COF	0.9518 ±0.0054	0.8695 ±0.1261	0.9513 ±0.0134
INFLO	0.9642 ±0.0171	0.7887 ±0.0540	0.9817 ±0.0024
LoOP	0.9725 ±0.0123	0.7684 ±0.0994	0.9851 ±0.0068
LOCI	0.9787	0.8877	–
aLOCI	0.8105 ±0.0883	0.6889 ±0.0345	0.8011 ±0.0615
CBLOF	0.2983 ±0.1492	0.3190 ±0.1155	0.6995 ±0.1407
uCBLOF	0.9496 ±0.0390	0.8721 ±0.0511	0.9555 ±0.0109
LDCOF	0.7645 ±0.1653	0.5948 ±0.0825	0.9593 ±0.0145
CMGOS-Red	0.9140 ±0.0815	0.5693 ±0.1000	0.9727 ±0.0141
CMGOS-Reg	0.8992 ±0.0643	0.6994 ±0.0681	0.9449 ±0.0510
CMGOS-MCD	0.9196 ±0.0830	0.6265 ±0.0969	0.9038 ±0.0511
HBOS	0.9827 ±0.0016	0.7477 ±0.0206	0.6798 ±0.0249
rPCA	0.9664 ±0.0000	0.9375 ±0.0001	0.7841 ±0.0151
oc-SVM	0.9721 ±0.0102	0.9512 ±0.0436	0.9543 ±0.0130
ν-oc-SVM	0.9581 ±0.0311	0.8993 ±0.0387	0.9236 ±0.0140
RDOS	0.9783 ±0.0025	0.8944 ±0.0787	0.9868 ±0.0042

Table 3
Mean AUC and the standard deviation for different option of AA + k-NN.

Method	Breast cancer	Pen global	Pen local
AA + k-NN min–max, $P = 10$	0.9851 ±0.0030	0.9634 ±0.0030	0.9915 ±0.0013
AA + k-NN standardization, $P = 10$	**0.9862** ±0.0023	0.9812 ±0.0025	0.9944 ±0.0008
AA + k-NN min–max, + 15	0.9807 ±0.0027	0.9798 ±0.0041	0.9943 ±0.0007
AA + k-NN standardization, $P = 15$	0.9782 ±0.0024	**0.9818** ±0.0028	**0.9962** ±0.0005

but some cases are erroneously labeled as outliers. We also show the results of hardening with k-NN, which returns a high number of errors, not only as false positives, but also as false negatives. We have obtained good results with all the different combinations of parameters. Therefore, it does not seem very sensitive to parameter choice.

5.3.3. Comparison with recent techniques

As previously commented, besides comparison with well-known algorithms, we compare our proposal with the recent techniques RDOS and VOS.

Table 4
Mean AUC and the standard deviation for different options of AA + k-NN with breast cancer.

e	Min–max $k = 5$ to 15	Standardization $k = 5$ to 15	Min–max $k = 10$ to 50	Standardization $k = 10$ to 50
$e = 2$	**0.9930** ±0.0016	0.9913 ±0.0007	0.9882 ±0.0018	0.9892 ±0.0012
$e = 3$	0.9928 ±0.0016	0.9913 ±0.0005	0.9873 ±0.0024	0.9884 ±0.0017
$e = 4$	0.9916 ±0.0013	0.9900 ±0.0006	0.9851 ±0.0030	0.9862 ±0.0023
$e = 5$	0.9895 ±0.0017	0.9886 ±0.0003	0.9832 ±0.0030	0.9849 ±0.0026

Table 5
Confusion matrices with binary labels for different options of AA + k-NN and k-NN with breast cancer (min–max normalization and k from 5 to 15).

True labels	$e = 2$		$e = 3$		$e = 4$		$e = 5$		k-NN	
Predictions	0	1	0	1	0	1	0	1	0	1
0 (normal)	341	0	343	0	344	0	343	0	338	2
1 (outlier)	16	10	14	10	13	10	14	10	19	8

We consider the same real data sets with continuous numerical features used by [49]. Those data sets with other kind of features are discarded. In particular, the data set Glass and Stamps are considered, which are completely described by [7]. Glass has 7 features and 214 observations, 4.2% of which are anomalies, while Stamps has 9 features and 340 observations, 9.1% of which are outliers. Data are normalized for ranging between 0 and 1 as in [49]. Table 6 shows the AUC results for the same state-of-the-art algorithms used in Table 1. For computing the AUC, we consider k values from 10 to 30 for Glass, since the number of records of Glass is small. Moreover, we consider $e = 3$ and $P = 10$ for Glass, while $e = 9$ and $P = 15$ for Stamps. Our procedure provides the best results.

In [49], instead of considering average AUC, the best AUC is reported, which will be an overly optimistic estimate, an overestimate. Furthermore, the variation due to selection of parameters is ignored by using the best combination of parameters. The strategy of selecting the best combination of parameters cannot be used in real applications, where true labels are unknown. Although this strategy suffers from much bias and should be avoided, we consider it only for being able to compare our proposal to that introduced by [49]. The best AUC for VOS is higher than the best AUC for IForest, OIF, ODIN, OutRank and HCOD for Glass and Stamps, according to the results in [49]. The best AUC for Glass with VOS is 0.864, which is worse that the mean AUC obtained with our proposal, 0.8846 (see Table 6). The best AUC for Stamps with VOS is 0.929. This result is improved by our proposal. The best mean AUC for Stamps with our proposal is 0.9354 for $e = 15$ (remember that we are averaging from $k = 10$ to $k = 50$). If instead averaging, we select the best AUC among $k = 10$ to $k = 50$, the best AUC 0.9416 is obtained for $k = 27$.

In summary, we have compared our proposal with a very high number of different methods in six data sets: one artificial data set (Table 1) and five real data sets (Tables 2 and 6). In five of the six data sets our proposal provides the best result.

6. Application

Knowledge of foot shape is of great importance for the appropriate design of footwear. It is a crucial issue for manufacturing shoes, since a proper fit is a key factor in the decision to buy, besides the fact that poorly fitting footwear can cause foot pain and deformity, especially in women. For these reasons, there are a large number of studies on foot shapes, such as [70–72], etc. In

Table 6
Average AUC together with the standard deviation for glass and stamps.

Data set	AA + k-NN	k-NN	LOF	COF	LoOP	HBOS	RPCA + k-NN	RDOS
Glass	**0.8846** ±0.0019	0.8666 ±0.0012	0.7999 ±0.0627	0.8280 ±0.0292	0.8008 ±0.0328	0.7312	0.8666 ±0.0012	0.7210 ±0.0127
Stamps	**0.8914** ±0.0296	0.8834 ±0.0119	0.7032 ±0.0443	0.6582 ±0.0801	0.6340 ±0.0586	0.8890	0.8834 ±0.0119	0.7313 ±0.0795

many of these studies, and in anthropometric studies devoted to product design in general, or apparel design in particular, data are studied without carrying out an outlier analysis, as in [73] or [74]. However, this is crucial, not only for data cleansing [75–77], which is a classical application of outlier analysis [5], but also to take advantage of the information that outliers can provide with regard to the design of shoes that fit well for a high percentage of the population. For example, in the apparel industry, many brands offer special sizes. However, outlier detection in the field of anthropometry is usually carried out by means of very simple procedures, as is the case in [34] or [35], where they look at extreme values in individual variables, or two-dimensional plots are inspected, which are the recommendations given in [78] for cleansing anthropometric data sets. Obviously, a somewhat more sophisticated method can be more effective, and better advantage can be taken of the information.

Therefore, the purpose of this Section is to detect the outliers in an anthropometric foot data set, before form analysis is carried out. Here, we restrict and focus on the outlier analysis part only. We carry out a separate analysis for men and women, since gender foot shape differences are well-known [71,72]. Furthermore, footwear designers usually propose different designs for women and men. We apply our proposal, which also helps us to understand why those points are labeled as outliers. Note that this a real-world problem, where we do not know which points are anomalies or not.

6.1. Foot data set

As described by [19], 22 foot measurements have been extracted from an anthropometric data set of 775 3D right foot scans representing the Spanish adult female and male population, 382 corresponding to women and 393 to men. The data were collected in different regions across Spain at shoe shops and workplaces using an INFOOT[1] laser scanner.

The 22 foot measurements are used in product design and in clinical assessment. All 3D registered feet were digitally measured with the algorithms developed by the IBV (Biomechanics Institute of Valencia). In contrast to body measurements, foot measurements are not standardized. Only Foot Length, Ball Girth and Ball Width are considered in [79,80] and [81]. The definitions are those used by the Human Shape Lab of the IBV, which comply with standards and are compatible with the accepted definitions found in the literature [82–84].

6.2. Outlier analysis

Instead of the whole set of 22 variables, in interest of brevity only the 4 features that could most influence shoe fitting according to shoe design experts are analyzed. Specifically, these features are: Foot Length, FL (distance between the rear and foremost point the foot axis); Ball Girth, BG (perimeter of the ball section), Ball Width, BW (maximal distance between the extreme points of the ball section projected onto the ground plane); and Instep Height, IH (maximal height of the instep section, located at 50% of the foot length).

[1] http://www.i-ware.co.jp/.

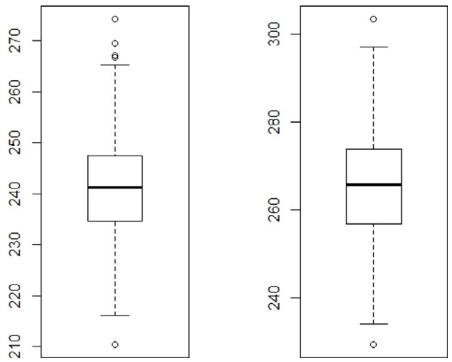

Fig. 3. Box-plots for women (left-handed) and men (right-handed), respectively.

In this application, size and shape is important. According to shoe experts, FL is the variable that best describes the size of the foot; in fact, this variable has great importance for shoe size. Therefore, we consider the size, represented by FL, and the shape, as explained by [85], separately. Shape corresponds to the geometrical information that remains once the scale is removed. Therefore, to describe the shape, we consider the rest of the features after removing the scale by dividing each of the features by FL: BG/FL, BW/FL and IH/FL.

For FL we can use simple box-plots to determine the outlier in size. Fig. 3 shows the different size ranges for women and men, with a different number of outliers. For women, five outliers are detected: one due to a very small size and four due to very large sizes. However, for men, the number of extreme sizes is smaller: there is one outlier corresponding to a very small size and another corresponding to a very large size. The variation in men, in terms of both range and interquartile range, is greater than in women. This could explain the fact that more outliers are found in women.

We apply AA + k-NN to the previously standardized foot shape features, with $e = 3$, since the elbow appears at this value for both men and women, and with k from 10 to 50. We convert the outlier scores into binary labels for the shake of brevity in the illustration.

In literature on AA, archetypes are usually displayed by the percentile values of each variable as compared to the data. We consider the same strategy here to interpret the outliers found. Tables 7 and 8 show the percentile profiles of the outliers found in foot shape features for women and men, respectively. This information is useful not only for cleansing, but also for shoe designers to know which shapes are "not normal". For that reason, we also include the percentile of FL for the outliers, although this variable is not used in the outlier detection of shapes.

For women, a total of 14 outliers are found, more than in the group of men, where 8 outliers are detected. One type of outlier detected in both men and women corresponds to points with very high percentiles in all three shape features. We refer to these as type 1 outliers. Another type of outlier, type 2, is the points with a high percentile in BG/FL and IH/FL, but a medium percentile in BW/FL. This kind of outlier is mainly found in women. For men, two outliers could also be included in this type, but their BG/FL percentiles are not as high as in the case of women. In women, we find another type of outlier, type 3, with very low percentiles

Table 7
Percentile profiles of outliers of foot shape features for women.

BG/FL	BW/FL	IH/FL	FL
87	40	94	48
90	80	100	4
96	82	97	28
90	63	89	8
2	3	86	10
99	100	83	35
97	88	94	36
93	62	91	51
99	99	71	55
0	0	44	62
96	95	100	17
99	99	99	2
73	39	96	87
100	100	98	2

Table 8
Percentile profiles of outliers of foot shape features for men.

BG/FL	BW/FL	IH/FL	FL
0	0	26	31
31	3	96	29
64	75	3	42
37	5	95	69
70	60	99	39
100	99	100	6
100	100	97	8
84	82	100	53
82	86	1	43
56	31	99	3

for BG/FL and BW/FL. Only one man is an outlier of this type. In men, we find another two types of outliers that do not appear in women: type 4 are outliers with a very low percentile in BW/FL, but a very high percentile in IH/FL, whereas type 5 are outliers with high percentiles in BG/FL and BW/FL and very low percentiles in IH/FL. Note that the majority of outliers have one or more features with high percentiles, more so than with low percentiles, so they are due to excess, especially for women.

In summary, for women the outliers are grouped into three sets: one from type 1 (2nd, 3rd, 6th, 7th, 9th, 11th, 12th and 14th), one from type 2 (1st, 4th, 8th and 13th) and one from type 3 (5th and 10th), while for men the outliers are from type 1 (6th, 7th, 8th), a variation of type 2 (5th and 10th), type 3 (1st), type 4 (2nd and 4th) and type 5 (3rd and 9th). Type 1 outliers have small-size feet, i.e. their FL percentiles are small, although in the case of women, some of them are not excessively small, and for one man it is medium. Type 2, 3 and 4 outliers are found in feet of all sizes. The only two type 5 outliers correspond to medium-size feet.

7. Conclusions

We have proposed a method to detect outliers in multivariate continuous data based on projection into relevant subspaces by means of AA, applying a k-NN algorithm to these subspaces and combining the results. This method returns outlier scores and we have also proposed a procedure to binarize the scores. We have illustrated their advantages in two simple examples. Our method is able to detect an anomaly in a simple data set with a linear correlation of two features, while other methods fail to recognize that anomaly. We have compared our proposal with 23 anomaly detection algorithms, in several benchmark data sets. In fact, it is really compared with more than 23 methods, since in Section 5.3.3 we compare the results with five other methods used by [49]. AA + k-NN returns very competitive results. Our proposal obtains the best results in five of six data sets. Our proposal is the third best in the other data set. In other words, it worked well with data sets with global and local anomalies, with continuous numerical features, such as pen global and pen local, respectively. We have also seen that changing the normalization procedure for standardization and also the k values used in the second part of our method can improve the results. Nevertheless, we have obtained good results with all the different combinations of parameters. Therefore, it does not seem very sensitive to parameter choice.

As discussed in Section 4, its weak point is its computational inefficiency, but new AA implementations could improve its speed. On the other hand, it has the advantage of its effectiveness (accuracy) and interpretability, which has been shown in the illustrative examples of Section 4 and the application of Section 6. Although our AA implementation is not deterministic, its solutions are stable [60]. Another advantage of our method is that it does not need clean (without outliers) training data for detecting anomalies, unlike other methods [86]. This is very convenient for our application, where we do not know if there are or not anomaly data. When we cannot establish a priori if a sample is from the normal class or not, methods that need data samples from the normal class for training models are not useful.

We applied AA + k-NN to a novel industrial data set and outliers were detected and interpreted. There are more outliers in women's feet than in men's. For example, in the case of women, there are more outliers due to a very long FL than to a short FL; and as regards their shapes, many of the outliers are due to large dimensions in BG, BW and IH relative to their small FLs. This information can be taken into account in the design process, or also to propose a range of shoes of special lengths or shapes for women. However, for men there are fewer shape outliers but with more different typologies. In summary, detecting the outliers in this kind of data sets can help shoe designers adjust their designs to a larger part of the population and be aware of the characteristics of the users that will make them uncomfortable to wear, whether when considering a range of special sizes or modifying any shoe feature to fit more customers.

We have used AA, but in future work a variant of AA such as archetypoid analysis [15] could be tested. The hardening process could also be improved by changing simple box-plots for other alternatives, such as those proposed by [87]. The speed of AA could be improved by using alternative AA algorithms such as those discussed in Section 5. On the other hand, we have considered only complete instances, but in real problems not all the cases are complete. We could easily extend the methodology for data sets with missing data, taking into account the proposal of [9]. We could also extend the methodology to other kind of data, such as multivariate time series and compare with recent literature on deep learning based methods for outlier detection in this field [88,89]. Finally, as regards the application, we have focused on the most important features in shoe design, but a more complete and exhaustive study could be carried out by considering other important features.

CRediT authorship contribution statement

Ismael Cabero: Methodology, Software, Formal analysis, Writing - original draft, Writing - review & editing, Visualization. **Irene Epifanio:** Conceptualization, Methodology, Software, Formal analysis, Resources, Writing - original draft, Writing - review & editing, Visualization, Supervision, Funding acquisition. **Ana Piérola:** Formal analysis, Investigation, Data curation, Resources, Writing - original draft, Writing - review & editing. **Alfredo Ballester:** Formal analysis, Investigation, Resources, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data statement

The code of our procedure and data sets for reproducing the results are available at http://www3.uji.es/~epifanio/RESEARCH/AAout.rar.

Acknowledgments

The authors would like to thank IMPIVA for having promoted "3D anthropometric and morphological study of the feet of the Spanish population for its application to the design of footwear and components" (IMPRDA/2005/38). This work is supported by the following grants: DPI2017-87333-R from the Spanish Ministry of Science, Innovation and Universities (AEI/FEDER, EU) and UJI-B2017-13 and UJI-B2020-22 from Universitat Jaume I, Spain.

References

[1] D.M. Hawkins, Identification of Outliers, Vol. 11, Springer, 1980.
[2] T. Johnson, I. Kwok, R.T. Ng, Fast computation of 2-dimensional depth contours, in: KDD, 1998, pp. 224–228.
[3] G. Williams, R. Baxter, H. He, S. Hawkins, L. Gu, A comparative study of RNN for outlier detection in data mining, in: IEEE International Conference on Data Mining, 2002, pp. 709–712.
[4] R. Gnanadesikan, J.R. Kettenring, Robust estimates, residuals, and outlier detection with multiresponse data, Biometrics 28 (1) (1972) 81–124.
[5] C.C. Aggarwal, Outlier Analysis, second ed., Springer, 2017.
[6] M. Goldstein, S. Uchida, A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data, PLOS ONE 11 (4) (2016) e0152173.
[7] G.O. Campos, A. Zimek, J. Sander, R.J.G.B. Campello, et al., On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study, Data Min. Knowl. Discov. 30 (4) (2016) 891–927.
[8] R. Domingues, M. Filippone, P. Michiardi, J. Zouaoui, A comparative evaluation of outlier detection algorithms: Experiments and analyses, Pattern Recognit. 74 (2018) 406–421.
[9] I. Epifanio, M.V. Ibáñez, A. Simó, Archetypal analysis with missing data: see all samples by looking at a few based on extreme profiles, Amer. Statist. 74 (2) (2020) 169–183.
[10] A. Cutler, L. Breiman, Archetypal analysis, Technometrics 36 (4) (1994) 338–347.
[11] M.R. D'Esposito, F. Palumbo, G. Ragozini, Interval archetypes: A new tool for interval data analysis, Stat. Anal. Data Min. 5 (4) (2012) 322–335.
[12] G. Ragozini, F. Palumbo, M.R. D'Esposito, Archetypal analysis for data-driven prototype identification, Stat. Anal. Data Min.: ASA Data Sci. J. 10 (1) (2017) 6–20.
[13] I. Cabero, I. Epifanio, Finding archetypal patterns for binary questionnaires, SORT 44 (1) (2020) 39–66.
[14] I. Epifanio, G. Vinué, S. Alemany, Archetypal analysis: contributions for estimating boundary cases in multivariate accommodation problem, Comput. Ind. Eng. 64 (3) (2013) 757–765.
[15] G. Vinué, I. Epifanio, S. Alemany, Archetypoids: A new approach to define representative archetypal data, Comput. Statist. Data Anal. 87 (2015) 102–115.
[16] G. Vinué, Anthropometry: An r package for analysis of anthropometric data, J. Stat. Softw. 77 (6) (2017) 1–39.
[17] I. Epifanio, M.V. Ibáñez, A. Simó, Archetypal shapes based on landmarks and extension to handle missing data, Adv. Data Anal. Classif. 12 (3) (2018) 705–735.
[18] L. Millán-Roures, I. Epifanio, V. Martínez, Detection of anomalies in water networks by functional data analysis, Math. Probl. Eng. 2018 (13) (2018) 5129735.
[19] A. Alcacer, I. Epifanio, M.V. Ibáñez, A. Simó, A. Ballester, A data-driven classification of 3D foot types by archetypal shapes based on landmarks, PLOS ONE 15 (1) (2020) e0228016.
[20] G. Vinué, I. Epifanio, Robust archetypoids for anomaly detection in big functional data, Adv. Data Anal. Classif. (2020) 1–26.
[21] J. Moliner, I. Epifanio, Robust multivariate and functional archetypal analysis with application to financial time series analysis, Physica A 519 (2019) 195–208.
[22] J.C. Thøgersen, M. Mørup, S. Damkiær, S. Molin, L. Jelsbak, Archetypal analysis of diverse pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways, BMC Bioinformatics 14 (2013) 279.
[23] I. Epifanio, Functional archetype and archetypoid analysis, Comput. Statist. Data Anal. 104 (2016) 24–34.
[24] I. Cabero, I. Epifanio, Archetypal analysis: an alternative to clustering for unsupervised texture segmentation, Image Anal. Stereol. 38 (2019) 151–160.
[25] M. Mørup, L.K. Hansen, Archetypal analysis for machine learning and data mining, Neurocomputing 80 (2012) 54–63.
[26] G.C. Porzio, G. Ragozini, D. Vistocco, On the use of archetypes as benchmarks, Appl. Stoch. Models Bus. Ind. 24 (2008) 419–437.
[27] E. Canhasi, I. Kononenko, Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization, Expert Syst. Appl. 41 (2) (2014) 535–543.
[28] A. Tsanousa, N. Laskaris, L. Angelis, A novel single-trial methodology for studying brain response variability based on archetypal analysis, Expert Syst. Appl. 42 (22) (2015) 8454–8462.
[29] J.L. Hinrich, S.E. Bardenfleth, R.E. Roge, N.W. Churchill, K.H. Madsen, M. Mørup, Archetypal analysis for modeling multisubject fMRI data, IEEE J. Sel. Top. Signal Process. 10 (7) (2016) 1160–1171.
[30] M.J.A. Eugster, Performance profiles based on archetypal athletes, Int. J. Perform. Anal. Sport 12 (1) (2012) 166–187.
[31] G. Vinué, I. Epifanio, Archetypoid analysis for sports analytics, Data Min. Knowl. Discov. 31 (6) (2017) 1643–1677.
[32] G. Vinué, I. Epifanio, Forecasting basketball players' performance using sparse functional data, Stat. Anal. Data Min.: ASA Data Sci. J. 12 (2019) 534–547.
[33] M.J.A. Eugster, F. Leisch, Weighted and robust archetypal analysis, Comput. Statist. Data Anal. 55 (3) (2011) 1215–1225.
[34] M. Kouchi, 3 - anthropometric methods for apparel design: body measurement devices and techniques, in: Anthropometry, Apparel Sizing and Design, Woodhead Publishing, 2014, pp. 67–94.
[35] A. Kuehnapfel, P. Ahnert, M. Loeffler, A. Broda, M. Scholz, Reliability of 3D laser-based anthropometry and comparison with classical anthropometry, Sci. Rep. 6 (2016) 26672.
[36] M. Breunig, H.-P. Kriegel, R.T. Ng, J. Sander, Lof: Identifying density-based local outliers, in: Proceeding of the 2000 ACM Sigmoid international conference on management data, 2000, pp. 93–104.
[37] J. Tang, Z. Chen, A.W.-C. Fu, D.W. Cheung, Enhancing effectiveness of outlier detections for low density patterns, in: Advances in Knowledge Discovery and Data Mining, Springer, Berlin Heidelberg, 2002, pp. 535–548.
[38] W. Jin, A.K.H. Tung, J. Han, W. Wang, Ranking outliers using symmetric neighborhood relationship, in: Advances in Knowledge Discovery and Data Mining, Springer, Berlin Heidelberg, 2006, pp. 577–593.
[39] H.-P. Kriegel, P. Kröger, E. Schubert, A. Zimek, Loop: Local outlier probabilities, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009, pp. 1649–1652.
[40] S. Papadimitriou, H. Kitagawa, P.B. Gibbons, C. Faloutsos, LOCI: Fast outlier detection using the local correlation integral, in: ICDE, 2003, pp. 315–326.
[41] M. Hofmann, R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, CRC Press, 2013.
[42] Z. He, X. Xu, S. Deng, Discovering cluster-based local outliers, Pattern Recognit. Lett. 24 (9) (2003) 1641–1650.
[43] M. Amer, M. Goldstein, Nearest-neighbor and clustering based anomaly detection algorithms for RapidMiner, in: Proceedings of the 3rd RapidMiner Community Meeting and Conference, 2012, pp. 1–12.
[44] M. Goldstein, A. Dengel, Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm, in: KI-2012: Poster and Demo Track, 2012, pp. 59–63.
[45] M.-L. Shyu, S.-C. Chen, K. Sarinnapakorn, L. Chang, A novel anomaly detection scheme based on principal component classifier, in: Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, 2003, pp. 171–179.
[46] M. Amer, M. Goldstein, S. Abdennadher, Enhancing one-class support vector machines for unsupervised anomaly detection, in: Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description, 2013, pp. 8–15.
[47] B. Tang, H. He, A local density-based approach for outlier detection, Neurocomputing 241 (2017) 171–180.
[48] J.H. Madsen, Ddoutlier: Distance & density-based outlier detection, 2018, R package version 0.1.0.
[49] C. Wang, Z. Liu, H. Gao, Y. Fu, VOS: A new outlier detection model using virtual graph, Knowl.-Based Syst. 185 (2019) 104907.
[50] F.T. Liu, K.M. Ting, Z. Zhou, Isolation forest, in: 2008 Eighth IEEE International Conference on Data Mining, 2008, pp. 413–422.
[51] Z. Liu, X. Liu, J. Ma, H. Gao, An optimized computational framework for isolation forest, Math. Probl. Eng. (2018) 2318763.
[52] V. Hautamaki, I. Karkkainen, P. Franti, Outlier detection using k-nearest neighbour graph, in: Proceedings of the 17th International Conference on Pattern Recognition, Vol. 3, Vol. 3, 2004, pp. 430–433.

[53] H.D.K. Moonesignhe, P. Tan, Outlier detection using random walks, in: 2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06), 2006, pp. 532–539.
[54] H.D.K. Moonesignhe, P. Tan, Outrank: A graph-based outlier detection framework using random walk, Int. J. Artif. Intell. Tools 17 (01) (2008) 19–36.
[55] X. Wang, I. Davidson, Discovering contexts and contextual outliers using random walks in graphs, in: 2009 Ninth IEEE International Conference on Data Mining, 2009, pp. 1034–1039.
[56] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. Data Mining, Inference and Prediction, second ed., Springer-Verlag, New York, 2009.
[57] T. Davis, B.C. Love, Memory for category information is idealized through contrast with competing options, Psychol. Sci. 21 (2) (2010) 234–242.
[58] C. Thurau, K. Kersting, M. Wahabzada, C. Bauckhage, Descriptive matrix factorization for sustainability adopting the principle of opposites, Data Min. Knowl. Discov. 24 (2) (2012) 325–354.
[59] C.L. Lawson, R.J. Hanson, Solving Least Squares Problems, Prentice Hall, Englewood Cliffs, 1974.
[60] M.J. Eugster, F. Leisch, From spider-man to hero - archetypal analysis in R, J. Stat. Softw. 30 (8) (2009) 1–23.
[61] S. Seth, M.J.A. Eugster, Probabilistic archetypal analysis, Mach. Learn. 102 (1) (2016) 85–113.
[62] H.V. Nguyen, H.H. Ang, V. Gopalkrishnan, Mining outliers with ensemble of heterogeneous detectors on random subspaces, in: Database Systems for Advanced Applications, Springer, Berlin Heidelberg, 2010, pp. 368–383.
[63] Y. Chen, J. Mairal, Harchaoui. Z., Fast and robust archetypal analysis for representation learning, in: CVPR 2014 - IEEE Conference on Computer Vision & Pattern Recognition, 2014, pp. 1478–1485.
[64] C. Bauckhage, K. Kersting, F. Hoppe, C. Thurau, Archetypal analysis as an autoencoder, in: Workshop New Challenges in Neural Computation, 2015, pp. 8–15.
[65] S. Mair, A. Boubekki, U. Brefeld, Frame-based data factorizations, in: International Conference on Machine Learning, 2017, pp. 2305–2313.
[66] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.
[67] E.M. Knorr, R.T. Ng, Finding intensional knowledge of distance-based outliers, in: Proceedings of the 25th International Conference on Very Large Data Bases, 1999, pp. 211–222.
[68] D. Dheeru, E. Karra Taniskidou, UCI machine learning repository, 2017.
[69] M. Goldstein, Unsupervised Anomaly Detection Benchmark, Harvard Dataverse, 2015.
[70] L. Delgado-Abellán, X. Aguado, E. Jiménez-Ormeño, L. Mecerreyes, L.M. Alegre, Foot morphology in spanish school children according to sex and age, Ergonomics 57 (5) (2014) 787–797.
[71] I. Krauss, C. Langbein, T. Horstmann, S. Grau, Sex-related differences in foot shape of adult caucasians – a follow-up study focusing on long and short feet, Ergonomics 54 (3) (2011) 294–300.
[72] M. Saghazadeh, N. Kitano, T. Okura, Gender differences of foot characteristics in older Japanese adults using a 3D foot scanner, J. Foot Ankle Res. 8 (1) (2015) 29.
[73] K. Jung, O. Kwon, H. You, Evaluation of the multivariate accommodation performance of the grid method, Applied Ergon. 42 (1) (2010) 156–161.
[74] S. Alemany, A. Ballester, E. Parrilla, A. Pierola, J. Uriel, B. Nacher, A. Remon, A. Ruescas, J.V. Durá, P. Piqueras, Solves. C., 3D body modelling and applications, in: Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018), 2019, pp. 623–636.
[75] M.V. Ibáñez, G. Vinué, S. Alemany, A. Simó, I. Epifanio, J. Domingo, G. Ayala, Apparel sizing using trimmed PAM and OWA operators, Expert Syst. Appl. 39 (12) (2012) 10512–10520.
[76] A. Pierola, I. Epifanio, S. Alemany, An ensemble of ordered logistic regression and random forest for child garment size matching, Comput. Ind. Eng. 101 (2016) 455–465.
[77] L. Markiewicz, M. Witkowski, R. Sitnik, E. Mielicka, 3D anthropometric algorithms for the estimation of measurements required for specialized garment design, Expert Syst. Appl. 85 (2017) 366–385.
[78] ISO 15535: 2012, General requirements for establishing anthropometric databases, 2012.
[79] ISO 8559-1: 2017, Size designation of clothes - Part 1: Anthropometric definitions for body measurement, 2017.
[80] ASTM D52 19-15, StandArd Terminology Relating to Body Dimensions for Apparel Sizing, 2015, ASTM International, West Conshohocken, PA, 2015.
[81] ISO 7250-1:2008, Basic human body measurements for technological design - part 1, 2008.
[82] W.A. Rossi, R. Tennant, Professional Shoe Fitting, National Shoe Retailers Association, 2013.
[83] J. Ramiro, E. Alcántara, A. Forner, R. Ferrandis, García-Belenguer, et al., Guía de Recomendaciones Para el Diseño de Calzado, Instituto de Biomecánica de Valencia, 1995, pp. 135–151.
[84] A. Luximon, Handbook of Footwear Design and Manufacture, Elsevier, 2013.
[85] I.L. Dryden, K.V. Mardia, Statistical Shape Analysis: With Applications in R, John Wiley & Sons, Chichester, 2016.
[86] B. Zong, Q. Song, M.R. Min, W. Cheng, C. Lumezanu, D. Cho, H. Chen, Deep autoencoding Gaussian mixture model for unsupervised anomaly detection, in: International Conference on Learning Representations, 2018.
[87] H.-P. Kriegel, P. Kröger, E. Schubert, A. Zimek, Interpreting and unifying outlier scores, in: Proceedings of the SIAM International Conference on Data Mining, 2011, pp. 13–24.
[88] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, D. Pei, Robust anomaly detection for multivariate time series through stochastic recurrent neural network, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, 2019, pp. 2828–2837.
[89] J. Audibert, P. Michiardi, F. Guyard, S. Marti, Zuluaga. M.A., USAD: Unsupervised anomaly detection on multivariate time series, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20, 2020, pp. 3395–3404.