

Chapter 14

Noise Versus Outliers

Cátia M. Salgado, Carlos Azevedo, Hugo Proença
and Susana M. Vieira

Learning Objectives

- What common methods for outlier detection are available.
- How to choose the most appropriate methods.
- How to assess the performance of an outlier detection method and how to compare different methods.

14.1 Introduction

An outlier is a data point which is different from the remaining data [1]. Outliers are also referred to as *abnormalities*, *discordants*, *deviants* and *anomalies* [2]. Whereas noise can be defined as mislabeled examples (class noise) or errors in the values of attributes (attribute noise), outlier is a broader concept that includes not only errors but also discordant data that may arise from the natural variation within the population or process. As such, outliers often contain interesting and useful information about the underlying system. These particularities have been exploited in fraud control, intrusion detection systems, web robot detection, weather forecasting, law enforcement and medical diagnosis [1], using in general methods of supervised outlier detection (see below).

Within the medical domain in general, the main sources of outliers are equipment malfunctions, human errors, anomalies arising from patient specific behaviors and natural variation within patients. Consider for instance an anomalous blood test result. Several reasons can explain the presence of outliers: severe pathological states, intake of drugs, food or alcohol, recent physical activity, stress, menstrual cycle, poor blood sample collection and/or handling. While some reasons may point to the existence of patient-specific characteristics discordant with the “average”

Electronic supplementary material The online version of this chapter (doi:[10.1007/978-3-319-43742-2_14](https://doi.org/10.1007/978-3-319-43742-2_14)) contains supplementary material, which is available to authorized users.

patient, in which case the observation being an outlier provides useful information, other reasons may point to human errors, and hence the observation should be considered for removal or correction. Therefore, it is crucial to consider the causes that may be responsible for outliers in a given dataset before proceeding to any type of action.

The consequences of not screening the data for outliers can be catastrophic. The negative effects of outliers can be summarized in: (1) increase in error variance and reduction in statistical power; (2) decrease in normality for the cases where outliers are non-randomly distributed; (3) model bias by corrupting the true relationship between exposure and outcome [3].

A good understanding of the data itself is required before choosing a model to detect outliers, and several factors influence the choice of an outlier identification method, including the type of data, its size and distribution, the availability of ground truth about the data, and the need for interpretability in a model [2]. For example, regression-based models are better suited for finding outliers in linearly correlated data, while clustering methods are advisable when the data is not linearly distributed along correlation planes. While this chapter provides a description of some of the most common methods for outlier detection, many others exist.

Evaluating the effectiveness of an outlier detection algorithm and comparing the different approaches is complex. Moreover, the ground-truth about outliers is often unavailable, as in the case of unsupervised scenarios, hampering the use of quantitative methods to assess the effectiveness of the algorithms in a rigorous way. The analyst is left with the alternative of qualitative and intuitive evaluation of results [2]. To overcome this difficulty, we will use in this chapter logistic regression models to investigate the performance of different outlier identification techniques in the medically relevant case study.

14.2 Part 1—Theoretical Concepts

Outlier identification methods can be classified into supervised and unsupervised methods, depending on whether prior information about the abnormalities in the data is available or not. The techniques can be further divided into univariable and multivariable methods, conditional on the number of variables considered in the dataset of interest.

The simplest form of outlier detection is extreme value analysis of unidimensional data. In this case, the core principle of discovering outliers is to determine the statistical tails of the underlying distribution and assume that either too large or too small values are outliers. In order to apply this type of technique to a multidimensional dataset, the analysis is performed one dimension at a time. In such a multivariable analysis, outliers are samples which have unusual combinations with other samples in the multidimensional space. It is possible to have outliers with reasonable marginal values (i.e. the value appears normal when confining oneself to one dimension), but due to linear or non-linear combinations of multiple attributes

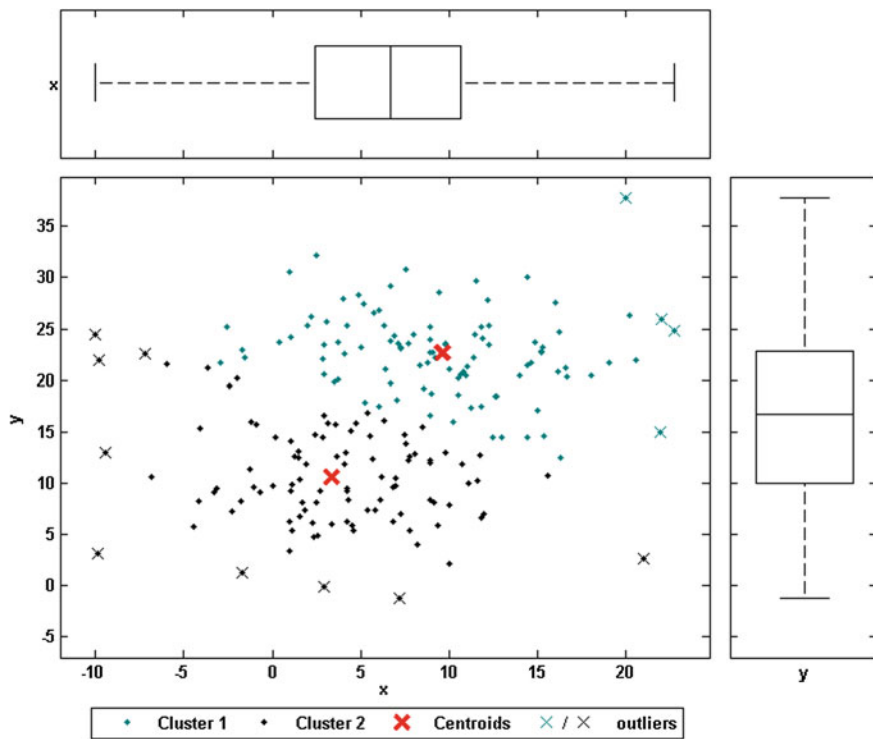


Fig. 14.1 Univariable (*boxplots*) versus multivariable (*scatter plot*) outlier investigation

these observations unveil unusual patterns in regards to the rest of the population under study.

To better understand this, the Fig. 14.1 provides a graphical example of a scenario where outliers are only visible in a 2-dimensional space. An inspection of the boxplots will reveal no outliers (no data point above and below 1.5 IQR (the interquartile range, refer to Chap. 15—Exploratory Data Analysis), a widely utilized outlier identification method), whereas a close observation of the natural clusters present in data will uncover irregular patterns. Outliers can be identified by visual inspection, highlighting data points that seem to be relatively out of the inherent 2-D data groups.

14.3 Statistical Methods

In the field of statistics, the data is assumed to follow a distribution model (e.g., normal distribution) and an instance is considered an outlier if it deviates significantly from the model [2, 4]. The use of normal distributions simplifies the analysis,

as most of the existing statistical tests, such as the Z-score, can be directly interpreted in terms of probabilities of significance. However, in many real world datasets the underlying distribution of the data is unknown or complex. Statistical tests still provide a good approximation of outlier scores, but results of the tests need to be interpreted carefully and cannot be expressed statistically [2]. The next sections describe some of the most widely used statistical tests for outliers identification.

14.3.1 Tukey's Method

Quartiles are the values that divide an array of numbers into quarters. The (IQR) is the distance between the lower (Q1) and upper (Q3) quartiles in the boxplot, that is $IQR = Q3 - Q1$. It can be used as a measure of how spread out the values are. Inner “fences” are located at a distance of $1.5 IQR$ below Q1 and above Q3, and outer fences at a distance of $3 IQR$ below Q1 and above Q3 [5]. A value between the inner and outer fences is a possible outlier, whereas a value falling outside the outer fences is a probable outlier. The removal of all possible and probable outliers is referred to as the Interquartile (IQ) method, while in Tukey's method only the probable outliers are discarded.

14.3.2 Z-Score

The Z-value test computes the number of standard deviations by which the data varies from the mean. It presents a reasonable criterion for the identification of outliers when the data is normally distributed. It is defined as:

$$z_i = \frac{x_i - \bar{x}}{s} \quad (14.1)$$

where \bar{x} and s denote the sample mean and standard deviation, respectively. In cases where mean and standard deviation of the distribution can be accurately estimated (or are available from domain knowledge), a good “rule of thumb” is to consider values with $|z_i| \geq 3$ as outliers. Of note, this method is of limited value for small datasets, since the maximum z-score is at most $n - 1/\sqrt{n}$ [6].

14.3.3 Modified Z-Score

The estimators used in the z-Score, the sample mean and sample standard deviation, can be affected by the extreme values present in the data. To avoid this problem, the

modified z-score uses the median \tilde{x} and the median absolute deviation (MAD) instead of the mean and standard deviation of the sample [7]:

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD} \quad (14.2)$$

where

$$MAD = \text{median}\{|x_i - \tilde{x}|\} \quad (14.3)$$

The authors recommend using modified z-scores with $|M_i| \geq 3.5$ as potential outliers. The assumption of normality of the data still holds.

14.3.4 Interquartile Range with Log-Normal Distribution

The statistical tests discussed previously are specifically based on the assumption that the data is fairly normally distributed. In the health care domain it is common to find skewed data, for instance in surgical procedure times or pulse oxymetry [8]. Refer to Chap. 15-Exploratory Data Analysis for a formal definition of skewness. If a variable follows a log-normal distribution then the logarithms of the observations follow a normal distribution. A reasonable approach then is to apply the \ln to the original data and they apply the tests intended to the “normalized” distributions. We refer to this method as the log-IQ.

14.3.5 Ordinary and Studentized Residuals

In a linear regression model, ordinary residuals are defined as the difference between the observed and predicted values. Data points with large residuals differ from the general regression trend and may represent outliers. The problem is that their magnitudes depend on their units of measurement, making it difficult to, for example, define a threshold at which a point is considered an outlier. Studentized residuals eliminate the units of measurement by dividing the residuals by an estimate of their standard deviation. One limitation of this approach is it assumes the regression model is correctly specified.

14.3.6 Cook’s Distance

In a linear regression model, Cook’s distance is used to estimate the influence of a data point on the regression. The principle of Cook’s distance is to measure the

effect of deleting a given observation. Data points with a large distance may represent outliers. For the i th point in the sample, Cook's distance is defined as:

$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j \hat{y}_{j(i)})^2}{(k+1)s^2} \quad (14.4)$$

Where $\hat{y}_{j(i)}$ is the prediction of y_j by the revised regression model when the i th point is removed from the sample, and s is the estimated root mean square error. Instinctively, D_i is a normalized measure of the influence of the point i on all predicted mean values \hat{y}_j with $j = 1, \dots, n$. Different cut-off values can be used for flagging highly influential points. Cook has suggested that a distance >1 represents a simple operational guideline [9]. Others have suggested a threshold of $4/n$, with n representing the number of observations.

14.3.7 Mahalanobis Distance

This test is based on Wilks method designed to detect a single outlier from a normal multivariable sample. It approaches the maximum squared Mahalanobis Distance (MD) to an F -distribution function formulation, which is often more appropriate than a χ^2 distribution [10]. For a p -dimensional multivariate sample x_i ($i = 1, \dots, n$), the Mahalanobis distance of the i th case is defined as:

$$MD_i = \sqrt{(x_i - t)^T C^{-1} (x_i - t)} \quad (14.5)$$

where t is the estimated multivariate location, which is usually the arithmetic mean, and C is the estimated covariance matrix, usually the sample covariance matrix.

Multivariate outliers can be simply defined as observations having a large squared Mahalanobis distance. In this work, the squared Mahalanobis distance is compared with quantiles of the F -distribution with p and $p - 1$ degrees of freedom. Critical values are calculated using Bonferroni bounds.

14.4 Proximity Based Models

Proximity-based techniques are simple to implement and unlike statistical models they make no prior assumptions about the data distribution model. They are suitable for both supervised and unsupervised multivariable outlier detection [4].

Clustering is a type of proximity-based technique that starts by partitioning a N -dimensional dataset into c subgroups of samples (clusters) based on their similarity. Then, some measure of the fit of the data points to the different clusters is used in order to determine if the data points are outliers [2]. One challenge associated with

this type of technique is that it assumes specific shapes of clusters depending on the distance function used within the clustering algorithm. For example, in a 3-dimensional space, the Euclidean distance would consider spheres as equidistant, whereas the Mahalanobis distance would consider ellipsoids as equidistant (where the length of the ellipsoids in one axis is proportional to the variance of the data in that direction).

14.4.1 k-Means

The k-means algorithm is widely used in data mining due to its simplicity and scalability [11]. The difficulty associated with this algorithm is the need to determine k , the number of clusters, in advance. The algorithm minimizes the within-cluster sum of squares, the sum of distances between each point in a cluster and the cluster centroid. In k-means, the center of a group is the mean of measurements in the group. Metrics such as the Akaike Information Criterion or the Bayesian Information Criterion, which add a factor proportional to k to the cost function used during clustering, can help determine k . A k value which is too high will increase the cost function even if it reduces the within-cluster sum of squares [12, 13].

14.4.2 k-Medoids

Similarly to k-means, the k-medoids clustering algorithm partitions the dataset into groups so that it minimizes the sum of distances between a data point and its center. In contrast to the k-means algorithm, in k-medoids the cluster centers are members of the group. Consequently, if there is a region of outliers outside the area with higher density of points, the cluster center will not be pushed towards the outliers region, as in k-means. Thus, k-medoids is more robust towards outliers than k-means.

14.4.3 Criteria for Outlier Detection

After determining the position of the cluster center with either k-means or k-medoids, the criteria to classify an item as an outlier must be specified, and different options exist:

Criterion 1: The first criterion proposed to detect outliers is based on the Euclidean distance to the cluster centers C , such that points more distant to their center than the minimum interclusters distance are considered outliers:

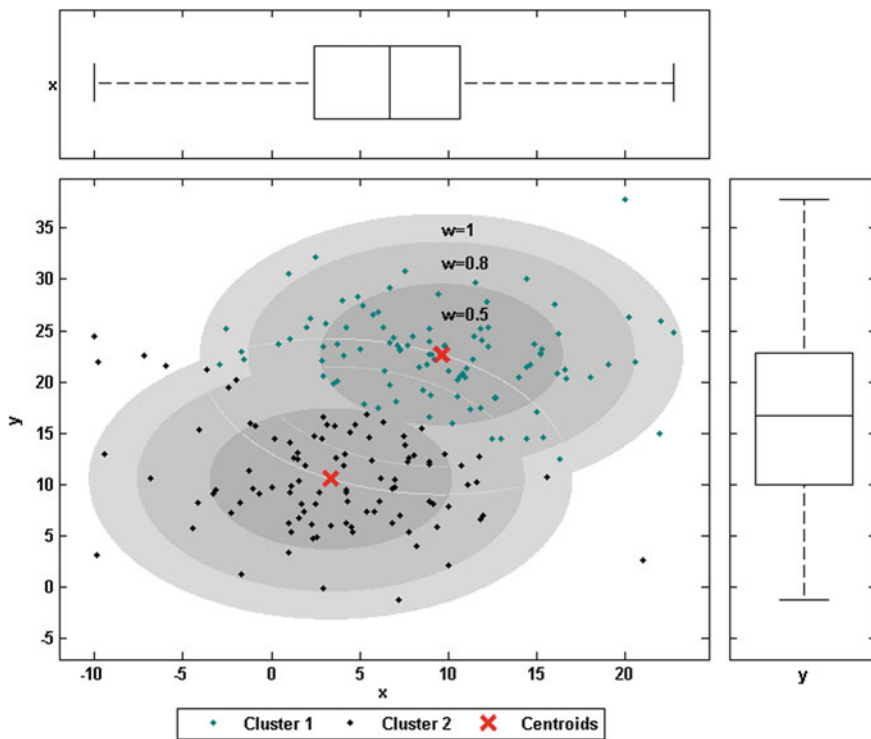


Fig. 14.2 Effect of different weights w in the detection of cluster-based outliers, using criterion 1

$$x \in C_k \text{ is outlier if } d(x, C_k) > \min_{k \neq j} \{\delta(C_k, C_j)\} \times w \quad (14.6)$$

where $d(x, C_k)$ is the Euclidean distance between point x and C_k center, $\delta(C_k, C_j)$ is the distance between C_k and C_j centers and $w = \{0.5, 0.7, 1, 1.2, 1.5, \dots\}$ is a weighting parameter that determines how aggressively the method will remove outliers.

Figure 14.2 provides a graphical example of the effect of varying values of w in the creation of boundaries for outlier detection. While small values of w aggressively remove outliers, as w increases the harder it is to identify them.

Criterion 2: In this criterion, we calculate the distance of each data point to its centroid (case of k-means) or medoid (case of k-medoids) [14]. If the ratio of the distance of the nearest point to the cluster center and these calculated distances are smaller than a certain threshold, than the point is considered an outlier. The threshold is defined by the user and should depend on the number of clusters selected, since the higher the number of clusters the closer are the points inside the cluster, i.e., the threshold should decrease with increasing c .

14.5 Supervised Outlier Detection

In many scenarios, previous knowledge about outliers may be available and can be used to label the data accordingly and to identify outliers of interest. The methods relying on previous examples of data outliers are referred to as supervised outlier detection methods, and involve training classification models which can later be used to identify outliers in the data. Supervised methods are often devised for anomaly detection in application domains where anomalies are considered occurrences of interest. Examples include fraud control, intrusion detection systems, web robot detection or medical diagnosis [1]. Hence, the labels represent what an analyst might be specifically looking for rather than what one might want to remove [2]. The key difference comparing to many other classification problems is the inherent unbalanced nature of data, since instances labeled as “abnormal” are present much less frequently than “normal” labeled instances. Interested readers can find further information about this topic in the textbook by Aggarwal, for instance [2].

14.6 Outlier Analysis Using Expert Knowledge

In univariate analyses, expert knowledge can be used to define thresholds of values that are normal, critical (life-threatening) or impossible because they fall outside permissible ranges or have no physical meaning [15]. Negative measurements of heart rate or body temperatures are examples of impossible values. It is very important to check the dataset for these types of outliers, as they originated undoubtedly from human error or equipment malfunction, and should be deleted or corrected.

14.7 Case Study: Identification of Outliers in the Indwelling Arterial Catheter (IAC) Study

In this section, various methods will be applied to identify outliers in two “real world” clinical datasets used in a study that investigated the effect of inserting an indwelling arterial catheter (IAC) in patients with respiratory failure. Two datasets are used, and include patients that received an IAC (IAC group) and patients that did not (non-IAC). The code used to generate the analyses and the figures is available in the GitHub repository for this book.

Table 14.1 Normal, critical and impossible ranges for the selected variables, and maximum and minimum values present in the datasets

Variable	Reference value			Analyzed data		
	Normal range	Critical	Impossible	IAC	Non-IAC	Units
Age	–	–	<17 (adults)	15.2–99.1	15.2–97.5	Years
SOFA	–	–	<0 and >24	1–17	0–14	No units
WBC	3–10.7	≥ 100	<0	0.3–86.0	0.2–109.8	$\times 10^9$ cells/L
Hemoglobin	Male: 13.5–17.5	≤ 6 and ≥ 20	<0	Male: 3–19.0	4.9–18.6	g/dL
	Female: 12–16			Female: 2.0–18.1		
Platelets	150–400	≤ 40 and ≥ 1000	<0	7.0–680.0	9.0–988.0	$\times 10^9$ /L
Sodium	136–145	≤ 120 and ≥ 160	<0	105–165.0	111.0–154.0	mmol/L
Potassium	3.5–5	≤ 2.5 and ≥ 6	<0	1–9.8	1.9–8.3	mmol/L
TCO ₂	22–28	≤ 10 and ≥ 40 [4]	<0	2.0–62.0	5.0–52.0	mmol/L
Chloride [29]	95–105	≤ 70 and ≥ 120	<0 and ≥ 160	81.0–133.0	78.0–127.0	mmol/L
BUN	7–18	≥ 100 [1]	<0	2.0–139.0	2.0–126.0	mg/dL
Creatinine	0.6–1.2	≥ 10	<0	0.2–12.5	0.0–18.3	mg/dL
PO ₂	75–105	≤ 40	<0	25–594.0	22.0–634.0	mmHg
PCO ₂	33–45	≤ 20 and ≥ 70	<0	8.0–141.0	14.0–158.0	mmHg

14.8 Expert Knowledge Analysis

Table 14.1 provides maximum and minimum values for defining normal, critical and permissible ranges in some of the variables analyzed in the study, as well as maximum and minimum values present in the dataset.

14.9 Univariate Analysis

In this section, univariate outliers are identified for each variable within pre-defined classes (survivors and non-survivors), using the statistical methods described above.

Table 14.2 summarizes the number and percentage of outliers identified by each method in the In-dwelling Arterial Catheter (IAC) and non-IAC groups. Overall, Tukey's and log-IQ are the most conservative methods, i.e., they identify the

Table 14.2 Number and percentage of outliers identified by each method

	IAC									
	Class 0 (811 patients)					Class 1 (163 patients)				
	IQ	Tukey's	log-IQ	Z-score	Mod z-score	IQ	Tukey's	Log-IQ	Z-score	Mod z-score
Age	0 (0.0 %)	0 (0.0 %)	1 (0.1 %)	0 (0.0 %)	0 (0.0 %)	5 (0.6 %)	0 (0.0 %)	8 (1.0 %)	4 (0.5 %)	5 (0.6 %)
SOFA	13 (1.6 %)	0 (0.0 %)	6 (0.7 %)	2 (0.2 %)	20 (2.5 %)	16 (2.0 %)	3 (0.4 %)	8 (1.0 %)	1 (0.1 %)	5 (0.6 %)
WBC	20 (2.5 %)	3 (0.4 %)	21 (2.6 %)	5 (0.6 %)	10 (1.2 %)	6 (0.7 %)	1 (0.1 %)	5 (0.6 %)	1 (0.1 %)	3 (0.4 %)
Hemoglobin	8 (1.0 %)	1 (0.1 %)	13 (1.6 %)	5 (0.6 %)	4 (0.5 %)	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)
Platelets	17 (2.1 %)	1 (0.1 %)	36 (4.4 %)	7 (0.9 %)	7 (0.9 %)	4 (0.5 %)	0 (0.0 %)	2 (0.2 %)	2 (0.2 %)	1 (0.1 %)
Sodium	30 (3.7 %)	8 (1.0 %)	30 (3.7 %)	10 (1.2 %)	26 (3.2 %)	8 (1.0 %)	1 (0.1 %)	8 (1.0 %)	2 (0.2 %)	2 (0.2 %)
Potassium	39 (4.8 %)	10 (1.2 %)	35 (4.3 %)	14 (1.7 %)	26 (3.2 %)	9 (1.1 %)	1 (0.1 %)	7 (0.9 %)	2 (0.2 %)	8 (1.0 %)
TCO ₂	24 (3.0 %)	4 (0.5 %)	31 (3.8 %)	13 (1.6 %)	13 (1.6 %)	9 (1.1 %)	1 (0.1 %)	6 (0.7 %)	2 (0.2 %)	2 (0.2 %)
Chloride	21 (2.6 %)	3 (0.4 %)	24 (3.0 %)	13 (1.6 %)	18 (2.2 %)	4 (0.5 %)	0 (0.0 %)	3 (0.4 %)	1 (0.1 %)	1 (0.1 %)
BUN	72 (8.9 %)	37 (4.6 %)	48 (5.9 %)	20 (2.5 %)	60 (7.4 %)	13 (1.6 %)	9 (1.1 %)	7 (0.9 %)	5 (0.6 %)	13 (1.6 %)
Creatinine	50 (6.2 %)	31 (3.8 %)	43 (5.3 %)	18 (2.2 %)	40 (4.9 %)	11 (1.4 %)	2 (0.2 %)	2 (0.2 %)	2 (0.2 %)	8 (1.0 %)
PO ₂	0 (0.0 %)	0 (0.0 %)	2 (0.2 %)	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)
PCO ₂	53 (6.5 %)	22 (2.7 %)	48 (5.9 %)	19 (2.3 %)	37 (4.6 %)	11 (1.4 %)	4 (0.5 %)	13 (1.6 %)	4 (0.5 %)	9 (1.1 %)
Total patients	220 (27.1 %)	86(10.6 %)	210 (25.9 %)	91 (11.2 %)	165 (20.3 %)	63 (7.8 %)	20 (2.5 %)	47 (5.8 %)	23 (2.8 %)	43 (5.3 %)
	Non-IAC									
	Class 0 (524 patients)					Class 1 (83 patients)				
	IQ	Tukey's	log-IQ	Z-score	Mod z-score	IQ	Tukey's	Log-IQ	Z-score	Mod z-score
Age	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	1 (0.2 %)	0 (0.0 %)	3 (0.6 %)	1 (0.2 %)	1 (0.2 %)
SOFA	51 (9.7 %)	2 (0.4 %)	48 (9.2 %)	2 (0.4 %)	7 (1.3 %)	9 (1.7 %)	1 (0.2 %)	8 (1.5 %)	1 (0.2 %)	3 (0.6 %)
WBC	21 (4.0 %)	4 (0.8 %)	10 (1.9 %)	4 (0.11 %)	11 (2.1 %)	4 (0.8 %)	1 (0.2 %)	4 (0.8 %)	1 (0.2 %)	3 (0.6 %)
Hemoglobin	1 (0.4 %)	0 (0.0 %)	6 (1.1 %)	2 (0.4 %)	2 (0.4 %)	0 (0.0 %)	0 (0.0 %)	2 (0.4 %)	0 (0.0 %)	0 (0.0 %)

(continued)

Table 14.2 (continued)

	Non-IAC						Class 1 (83 patients)					
	Class 0 (524 patients)						Class 1 (83 patients)					
	IQ	Tukey's	log-IQ	Z-score	Mod z-score	IQ	Tukey's	Log-IQ	Z-score	Mod z-score	Mod z-score	
Platelets	15 (2.9 %)	5 (1.0 %)	21 (4.0 %)	5 (1.0 %)	6 (1.1 %)	4 (0.8 %)	1 (0.2 %)	5 (1.0 %)	2 (0.4 %)	2 (0.4 %)		
Sodium	25 (4.8 %)	9 (1.7 %)	25 (4.11 %)	9 (1.7 %)	20 (3.11 %)	5 (1.0 %)	1 (0.2 %)	5 (1.0 %)	1 (0.2 %)	1 (0.2 %)		
Potassium	22 (4.2 %)	2 (0.4 %)	14 (2.7 %)	6 (1.1 %)	14 (2.7 %)	1 (0.2 %)	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)		
TCO ₂	27 (5.2 %)	4 (0.8 %)	31 (5.9 %)	8 (1.5 %)	5 (1.0 %)	4 (0.8 %)	1 (0.2 %)	4 (0.8 %)	2 (0.4 %)	3 (0.6 %)		
Chloride	21 (4.0 %)	4 (0.8 %)	20 (3.11 %)	9 (1.7 %)	11 (2.1 %)	9 (1.7 %)	1 (0.2 %)	9 (1.7 %)	1 (0.2 %)	4 (0.8 %)		
BUN	35 (6.7 %)	20 (3.8 %)	27 (5.2 %)	13 (2.5 %)	34 (6.5 %)	6 (1.1 %)	2 (0.4 %)	2 (0.4 %)	2 (0.4 %)	6 (1.1 %)		
Creatinine	29 (5.5 %)	17 (3.2 %)	25 (4.8 %)	8 (1.5 %)	22 (4.2 %)	7 (1.3 %)	2 (0.4 %)	3 (0.6 %)	2 (0.4 %)	5 (1.0 %)		
PO ₂	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	1 (0.2 %)	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	3 (0.6 %)		
PCO ₂	34 (6.5 %)	11 (2.1 %)	33 (6.3 %)	10 (1.9 %)	28 (5.3 %)	8 (1.5 %)	4 (0.8 %)	6 (1.1 %)	2 (0.4 %)	8 (1.5 %)		
Total patients	176 (33.6 %)	59 (11.3 %)	172 (32.8 %)	56 (10.7 %)	111 (21.2 %)	37 (7.1 %)	11 (2.1 %)	29 (5.5 %)	11 (2.1 %)	28 (5.3 %)		

“Total patients” represents the number of patients identified when considering all variables together. The results in bold highlight the variable with the most outliers in each method, and also the method that removes more patients in total, in each class. Class 0: represents survivors, Class 1: non-survivors

smallest number of points as outliers, whereas IQ identifies more outliers than any other method. With a few exceptions, the modified z-score identifies more outliers than the z-score.

A preliminary investigation of results showed that values falling within reference normal ranges (see Table 14.1) are never identified as outliers, whatever the method. On the other hand, critical values are often identified as such. Additional remarks can be made as in general (1) more outliers are identified in the variable BUN than in any other and (2) the ratio of number of outliers and total number of patients is smaller in the class 1 cohorts (non-survivors). As expected, for variables that approximate more to lognormal distribution than to a normal distribution, such as potassium, BUN and PCO₂, the IQ method applied to the logarithmic transformation of data (log-IQ method) identifies less outliers than the IQ applied to the real data. Consider for instance the variable BUN, which follows approximately a lognormal distribution. Figure 14.3 shows a scatter of all data points and the identified outliers in the IAC group.

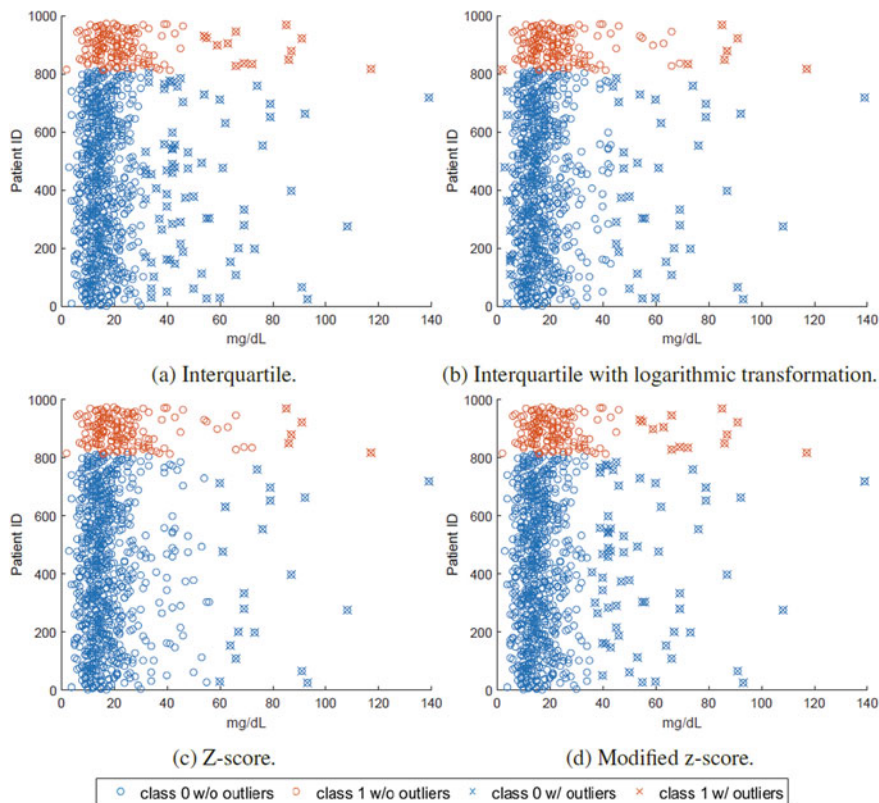


Fig. 14.3 Outliers identified by statistical analysis for the variable BUN, in the IAC cohort. Class 0: survivors; Class 1: non survivors

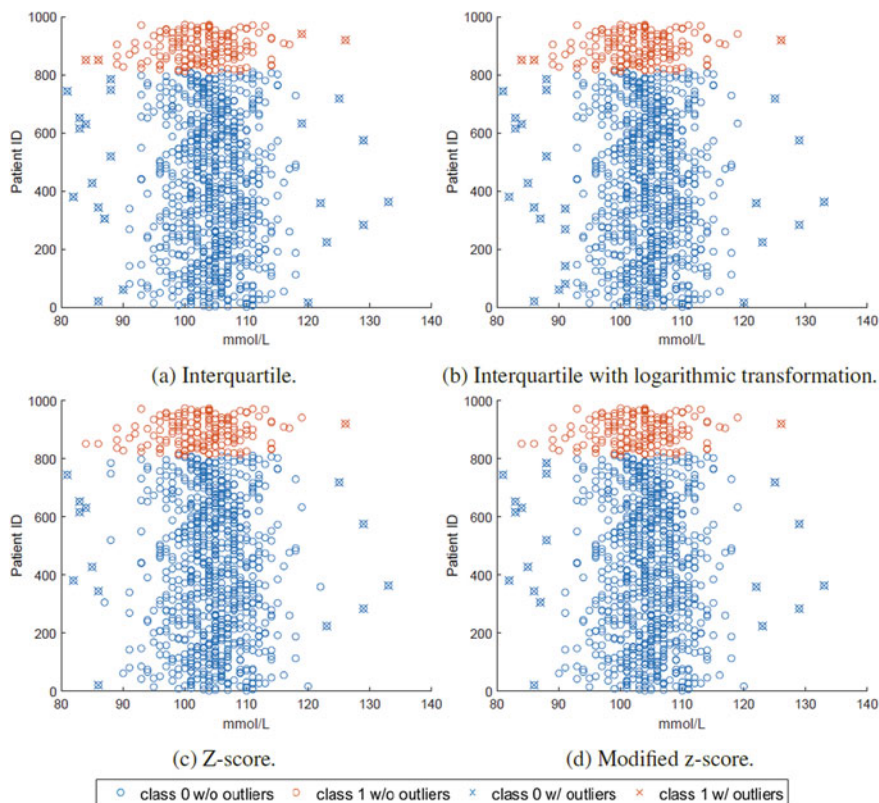


Fig. 14.4 Outliers identified by statistical analysis for the variable chloride, in the IAC cohort. Class 0: survivors; Class 1: non survivors

On the other hand, when the values follow approximately a normal distribution, as in the case of chloride (see Fig. 14.4), the IQ method identifies less outliers than log-IQ. Of note, the range of values considered outliers differs between classes, i.e., what is considered an outlier in class 0 is not necessarily an outlier in class 1. An example of this is values smaller than 90 mmol/L in the modified z-score.

Since this is a univariate analysis, the investigation of extreme values using expert knowledge is of interest. For chloride, normal values are in the range of 95–105 mmol/L, whereas values <70 or >120 mmol/L are considered critical, and concentrations above 160 mmol/L are physiologically impossible [15]. Figure 14.4 confirms that normal values are always kept, whatever the method. Importantly, some critical values are not identified in both z-score and modified z-score (especially in class 1). Thus, it seems that the methods identify outliers that should not be eliminated, as they likely represent actual values in extremely sick patients.

14.10 Multivariable Analysis

Using model based approaches, unusual combination of values for a number of variables can be identified. In this analysis we will be concerned with multivariable outliers for the complete set of variables in the datasets, including those that are binary. In order to investigate multivariable outliers in IAC and non-IAC patients, the Mahalanobis distance and cluster based approaches are tested within pre-defined classes. Table 14.3 shows the average results in terms of number of clusters c determined by the silhouette index, and the percentage of patients identified as

Table 14.3 Multivariable outliers identified by k-means, k-medoids and Mahalanobis distance

	Criterion	Weight	c		% of outliers Class 0	
			Class 0	Class 1	Class 0	Class 1
<i>IAC</i>						
K-means, silhouette index	1	1.2	4 ± 3.1	2 ± 0.0	25.2 ± 7.4	20.9 ± 11.0
	1	1.5	3 ± 2.9	2 ± 0.0	7.9 ± 4.6	3.3 ± 5.9
	1	1.7	3 ± 2.6	2 ± 0.0	3.6 ± 2.5	0.4 ± 2.2
	1	2.0	4 ± 3.1	2 ± 0.0	1.0 ± 1.1	0.1 ± 0.3
K-means, c = 2	2	0.05	2 ± 0.0	2 ± 0.0	28.5 ± 4.8	21.4 ± 11.9
	2	0.06	2 ± 0.0	2 ± 0.0	9.3 ± 4.2	2.9 ± 5.2
K-medoids, silhouette index	1	1.2	4 ± 3.0	2 ± 0.0	4.1 ± 2.2	0.8 ± 3.1
	1	1.5	3 ± 2.6	2 ± 0.0	1.1 ± 1.0	0.1 ± 0.3
	1	1.7	3 ± 2.9	2 ± 0.0	0.2 ± 0.2	0.0 ± 0.0
	1	2.0	4 ± 3.0	2 ± 0.0	0.7 ± 0.4	0.0 ± 0.0
K-medoids, c = 2	2	0.01	2 ± 0.0	2 ± 0.0	34.6 ± 8.6	2.5 ± 0.0
	2	0.02	2 ± 0.0	2 ± 0.0	20.8 ± 6.1	0.0 ± 0.0
Mahalanobis	–	–	–	–	16.7 ± 5.5	0.0 ± 0.0
<i>Non-IAC</i>						
K-means, silhouette index	1	1.2	9 ± 1.8	7 ± 2.4	12.8 ± 4.1	13.0 ± 9.5
	1	1.5	9 ± 1.7	7 ± 2.5	2.8 ± 1.8	1.0 ± 1.7
	1	1.7	9 ± 1.8	7 ± 2.5	0.9 ± 1.2	0.0 ± 0.2
	1	2.0	9 ± 2.4	7 ± 2.5	0.2 ± 0.7	0.0 ± 0.0
K-means, c = 2	2	0.05	2 ± 0.0	2 ± 0.0	25.5 ± 4.5	41.0 ± 11.9
	2	0.06	2 ± 0.0	2 ± 0.0	10.6 ± 2.6	4.8 ± 7.2
K-medoids, silhouette index	1	1.2	9 ± 1.5	7 ± 2.5	3.8 ± 1.6	1.4 ± 1.6
	1	1.5	9 ± 2.0	7 ± 2.4	0.9 ± 1.9	0.0 ± 0.0
	1	1.7	9 ± 2.0	7 ± 2.4	0.3 ± 0.6	0.0 ± 0.0
	1	2.0	9 ± 1.3	7 ± 2.5	0.4 ± 0.9	0.0 ± 0.0
K-medoids, c = 2	2	0.01	2 ± 0.0	2 ± 0.0	19.7 ± 4.0	2.7 ± 8.8
	2	0.02	2 ± 0.0	2 ± 0.0	11.0 ± 2.8	1.0 ± 5.0
Mahalanobis	–	–	–	–	6.8 ± 2.6	0.8 ± 4.0

Results are presented as mean ± standard deviation

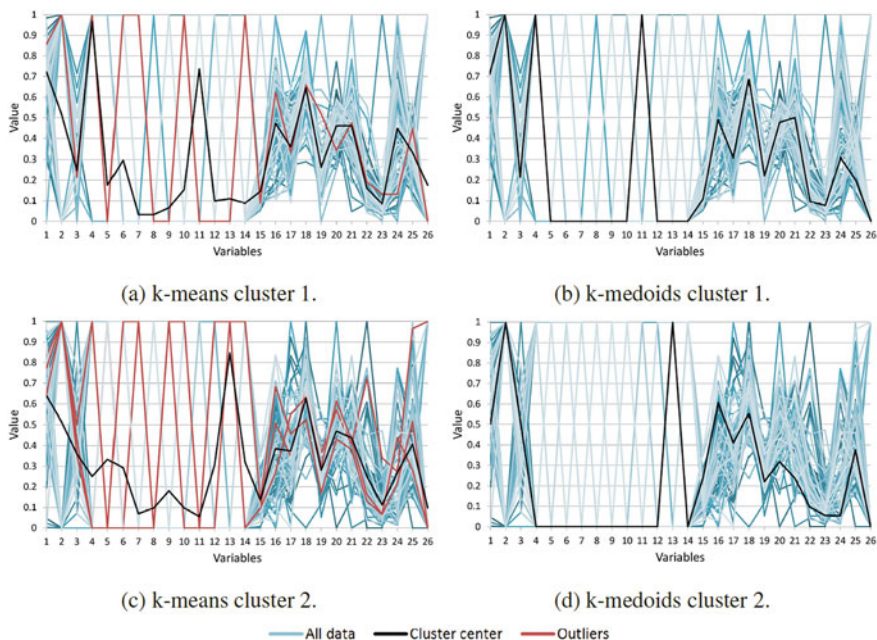


Fig. 14.5 Outliers identified by clustering based approaches for patients that died after IAC. Criterion 1, based on interclusters distance, with $c = 2$ and $w = 1.5$ was used. K-medoids does not identify outliers, whereas k-means identifies 1 outlier in cluster 1 and 2 outliers in cluster 2

outliers. In order to account for variability, the tests were performed 100 times. The data was normalized for testing the cluster based approaches only.

Considering the scenario where two clusters are created for the complete IAC dataset separated by classes, we investigate outliers by looking at multivariable observations around cluster centers. Figure 14.5 shows an example of the outliers detected using k-means and k-medoids with the criterion 1 and weight equal to 1.5. For illustrative purposes, we present only the graphical results of patients that died in the IAC group (class 1). The x-axis represents each of the selected features (see Table 14.1) and the y-axis represents the corresponding values normalized between 0 and 1. K-medoids does not identify any outlier, whereas k-means identifies 1 outlier in the first cluster and 2 outliers in the second cluster. This difference can be attributed to the fact that the intercluster distance is smaller in k-medoids than in k-means.

The detection of outliers seems to be more influenced by binary features than by continuous features: red lines are, with some exceptions, fairly close to black lines for the continuous variables (1 to 2 and 15 to 25) and distant in the binary variables. A possible explanation is that clustering was essentially designed for multivariable continuous data; binary variables produce a maximum separation, since only two values exist, 0 and 1, with nothing between them.

14.11 Classification of Mortality in IAC and Non-IAC Patients

Logistic regression models were created to assess the effect of removing outliers using the different methods in the classification of mortality in IAC and non-IAC patients, following the same rationale as in Chap. 13-Missing Data. A 10-fold cross validation approach was used to assess the validity and robustness of the models. In each round, every outlier identification method was applied separately for each class of the training set, and the results were averaged over the rounds. Before cross-validation, the values were normalized between 0 and 1 using the min-max procedure. For the log-IQ method, the data was log-transformed before normalization, except for variables containing null values (binary variables in Table 14.1, SOFA and creatinine). We also investigate the scenario where only the 10 % worst examples detected by each statistical method within each class are considered, and the case where no outliers were removed (all data is used). In the clustering based approaches, the number of clusters c was chosen between 2 and 10 using the silhouette index method. We also show the case where c is fixed as 2. The weight of the clustering based approaches was adjusted according to the particularities of the method. Since a cluster center in k-medoids is a data point belonging to the dataset, the distance to its nearest neighbor is smaller than in the case of k-means, especially because a lot of binary variables are considered. For this reason, we chose higher values of w for k-means criterion 2.

The performance of the models is evaluated in terms of area under the receiver operating characteristic curve (AUC), accuracy (ACC, correct classification rate), sensitivity (true positive classification rate), and specificity (true negative classification rate). A specific test suggested by DeLong and DeLong can then test whether the results differ significantly [16].

The performance results for the IAC group are shown in Table 14.4, and the percentage of patients removed using each method in Table 14.5. For conciseness, the results for the non-IAC group are not shown. The best performance for IAC is $AUC = 0.83$ and $ACC = 0.78$ (highlighted in bold). The maximum sensitivity is 87 % and maximum specificity is 79 %, however these two do not occur simultaneously. Overall, the best AUC is obtained when all the data is used and when only a few outliers are removed. The worst performances are obtained using the z-score without trimming the results and k-means and k-medoids using $c = 2$, criterion 1 and weight 1.2. As for non-IAC, the best performance corresponds to $AUC = 0.88$, $ACC = 0.84$, sensitivity = 0.85 and specificity = 0.85. Again, the best performance is achieved when all the data is used and in the cases where less outliers are removed. The worst performance by far is obtained when all outliers identified by the z-score are removed. Similarly to IAC, for k-means and k-medoids criterion 1, increasing values of weight provide better results.

Table 14.4 IAC logistic regression results using 10-fold cross validation, after removal of outliers and using the original dataset

Statistical	Cutoff	AUC	ACC	Sensitivity	Specificity
IQ	–	0.81 ± 0.05	0.76 ± 0.05	0.71 ± 0.14	0.76 ± 0.06
	10	0.82 ± 0.06	0.77 ± 0.06	0.76 ± 0.11	0.77 ± 0.07
Tukey's	–	0.82 ± 0.05	0.75 ± 0.06	0.76 ± 0.09	0.75 ± 0.06
	10	0.83 ± 0.06	0.78 ± 0.05	0.75 ± 0.10	0.78 ± 0.06
Log-IQ	–	0.82 ± 0.06	0.76 ± 0.05	0.74 ± 0.14	0.76 ± 0.06
	10	0.83 ± 0.06	0.78 ± 0.04	0.73 ± 0.10	0.79 ± 0.05
Z-score	–	0.78 ± 0.03	0.67 ± 0.06	0.85 ± 0.09	0.64 ± 0.08
	10	0.81 ± 0.07	0.75 ± 0.06	0.74 ± 0.13	0.75 ± 0.07
Modified z-score	–	0.82 ± 0.05	0.76 ± 0.05	0.77 ± 0.14	0.76 ± 0.05
	10	0.82 ± 0.06	0.77 ± 0.06	0.75 ± 0.10	0.77 ± 0.06
Mahalanobis	–	0.81 ± 0.08	0.75 ± 0.06	0.73 ± 0.10	0.76 ± 0.07
Cluster based	Weight	AUC	ACC	Sensitivity	Specificity
K-means silhouette criterion 1	1.2	0.81 ± 0.08	0.72 ± 0.05	0.80 ± 0.12	0.70 ± 0.06
	1.5	0.82 ± 0.05	0.76 ± 0.06	0.76 ± 0.11	0.76 ± 0.06
	1.7	0.83 ± 0.06	0.78 ± 0.05	0.77 ± 0.10	0.78 ± 0.06
	2	0.83 ± 0.06	0.78 ± 0.05	0.74 ± 0.09	0.78 ± 0.06
K-means $c = 2$ criterion 1	1.2	0.79 ± 0.08	0.66 ± 0.05	0.84 ± 0.10	0.63 ± 0.06
	1.5	0.82 ± 0.06	0.73 ± 0.06	0.79 ± 0.09	0.72 ± 0.07
	1.7	0.82 ± 0.06	0.75 ± 0.06	0.78 ± 0.08	0.75 ± 0.08
	2	0.83 ± 0.07	0.78 ± 0.06	0.76 ± 0.09	0.78 ± 0.06
K-means criterion 2	0.05	0.83 ± 0.07	0.77 ± 0.05	0.74 ± 0.09	0.78 ± 0.06
	0.06	0.83 ± 0.06	0.77 ± 0.06	0.75 ± 0.10	0.78 ± 0.06
K-medoids silhouette criterion 1	1.2	0.81 ± 0.04	0.68 ± 0.04	0.85 ± 0.09	0.64 ± 0.05
	1.5	0.83 ± 0.05	0.74 ± 0.04	0.80 ± 0.10	0.73 ± 0.06
	1.7	0.83 ± 0.05	0.75 ± 0.06	0.78 ± 0.10	0.74 ± 0.07
	2	0.83 ± 0.06	0.77 ± 0.05	0.77 ± 0.09	0.77 ± 0.06
K-medoids $c = 2$ criterion 1	1.2	0.78 ± 0.06	0.62 ± 0.07	0.87 ± 0.08	0.57 ± 0.07
	1.5	0.81 ± 0.06	0.70 ± 0.06	0.83 ± 0.10	0.68 ± 0.08
	1.7	0.82 ± 0.06	0.72 ± 0.06	0.80 ± 0.10	0.71 ± 0.08
	2	0.83 ± 0.07	0.76 ± 0.06	0.77 ± 0.10	0.75 ± 0.07
K-medoids criterion 2	0.01	0.83 ± 0.07	0.74 ± 0.07	0.77 ± 0.10	0.74 ± 0.08
	0.02	0.81 ± 0.06	0.67 ± 0.06	0.85 ± 0.09	0.63 ± 0.08
All data	–	0.83 ± 0.06	0.78 ± 0.05	0.76 ± 0.11	0.79 ± 0.06

Results are presented as mean ± standard deviation

Table 14.5 Percentage of IAC patients removed by each method in the train set, during cross-validation

Statistical	Cutoff	Class 0	Class 1	Total	
IQ	–	23.1 ± 1.4	33.3 ± 1.9	24.8 ± 1.4	
	10	3.3 ± 0.2	5.2 ± 0.3	3.6 ± 0.2	
Tukey’s	–	8.7 ± 0.05	10.1 ± 1.1	9.0 ± 0.5	
	10	1.2 ± 0.1	1.3 ± 0.2	1.3 ± 0.1	
Log-IQ	–	22.8 ± 1.1	25.4 ± 2.0	23.2 ± 1.1	
	10	3.1 ± 0.2	3.7 ± 0.5	3.2 ± 0.1	
Z-score	–	35.0 ± 1.6	0.67 ± 0.06	32.6 ± 1.4	
	10	5.3 ± 0.2	2.9 ± 1.3	4.9 ± 0.3	
Modified z-score	–	18.3 ± 0.05	24.5 ± 1.3	19.4 ± 0.5	
	10	2.4 ± 0.1	3.5 ± 0.4	2.6 ± 0.1	
Mahalanobis	–	19.6 ± 9.6	17.4 ± 3.0	19.2 ± 8.1	
Cluster based		Weight	Class 0	Class 1	Total
K-means silhouette criterion 1	1.2	19.6 ± 9.6	17.4 ± 3.0	19.2 ± 8.1	
	1.5	6.1 ± 5.1	1.9 ± 0.5	5.4 ± 4.2	
	1.7	2.5 ± 2.6	0.3 ± 0.3	2.2 ± 2.2	
	2	0.7 ± 0.9	0.0 ± 0.0	0.6 ± 0.8	
K-means $c = 2$ criterion 1	1.2	29.7 ± 3.5	17.4 ± 3.0	27.6 ± 2.9	
	1.5	11.9 ± 3.0	1.9 ± 0.5	10.2 ± 2.5	
	1.7	5.5 ± 2.0	0.3 ± 0.3	4.7 ± 1.6	
	2	1.7 ± 0.8	0.0 ± 0.0	1.4 ± 0.7	
K-means criterion 2	0.05	0.3 ± 0.2	0.0 ± 0.0	0.3 ± 0.2	
	0.06	1.1 ± 0.5	0.0 ± 0.0	0.9 ± 0.4	
K-medoids silhouette criterion 1	1.2	25.0 ± 10.7	3.8 ± 2.0	21.5 ± 8.8	
	1.5	12.9 ± 7.4	0.0 ± 0.0	10.8 ± 6.2	
	1.7	9.5 ± 6.1	0.0 ± 0.0	7.9 ± 5.1	
	2	3.1 ± 2.3	0.0 ± 0.0	2.5 ± 1.9	
K-medoids $c = 2$ criterion 1	1.2	34.7 ± 0.7	3.8 ± 2.0	29.5 ± 0.7	
	1.5	19.6 ± 0.6	0.0 ± 0.0	16.3 ± 0.5	
	1.7	14.9 ± 1.1	0.0 ± 0.0	12.4 ± 0.9	
	2	5.1 ± 0.4	0.0 ± 0.0	4.2 ± 0.4	
K-medoids criterion 2	0.01	8.3 ± 2.1	0.0 ± 0.0	6.9 ± 1.7	
	0.02	28.9 ± 3.9	1.8 ± 3.8	24.4 ± 3.6	

Results are presented as mean ± standard deviation

14.12 Conclusions and Summary

The univariable outlier analysis provided in the case study showed that a large number of outliers were identified for each variable within the predefined classes, meaning that the removal of all the identified outliers would cause a large portion of

data to be excluded. For this reason, ranking the univariate outliers according to score values and discarding only those with highest scores provided better classification results.

Overall, none of the outlier removal techniques was able to improve the performance of a classification model. As it had been cleaned these results suggest that the dataset did not contain impossible values, extreme values are probably due to biological variation rather than experimental mistakes. Hence, the “outliers” in this study appear to contain useful information in their extreme values, and automatically excluding resulted in a loss of this information.

Some modeling methods already accommodate for outliers so they have minimal impact in the model, and can be tuned to be more or less sensitive to them. Thus, rather than excluding outliers from the dataset before the modeling step, an alternative strategy would be to use models that are robust to outliers, such as robust regression.

Take Home Messages

1. Distinguishing outliers as useful or uninformative is not clear cut.
2. In certain contexts, outliers may represent extremely valuable information that must not be discarded.
3. Various methods exist and will identify possible or likely outliers, but the expert eye must prevail before deleting or correcting outliers.

Open Access This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, duplication, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, a link is provided to the Creative Commons license and any changes made are indicated.

The images or other third party material in this chapter are included in the work’s Creative Commons license, unless indicated otherwise in the credit line; if such material is not included in the work’s Creative Commons license and the respective action is not permitted by statutory regulation, users will need to obtain permission from the license holder to duplicate, adapt or reproduce the material.

Code Appendix

The code used in this chapter is available in the GitHub repository for this book: <https://github.com/MIT-LCP/critical-data-book>. Further information on the code is available from this website.

References

1. Barnett V, Lewis T (1994) *Outliers in statistical data*, 3rd edn. Wiley, Chichester
2. Aggarwal CC (2013) *Outlier analysis*. Springer, New York
3. Osborne JW, Overbay A (2004) The power of outliers (and why researchers should always check for them). *Pract Assess Res Eval* 9(6):1–12
4. Hodge VJ, Austin J (2004) A survey of outlier detection methodologies. *Artif Intell Rev* 22(2):85–126
5. Tukey J (1977) *Exploratory data analysis*. Pearson
6. Shiffler RE (1988) Maximum Z scores and outliers. *Am Stat* 42(1):79–80
7. Iglewicz B, Hoaglin DC (1993) *How to detect and handle outliers*. ASQC Quality Press
8. Seo S (2006) A review and comparison of methods for detecting outliers in univariate data sets. 09 Aug 2006 [Online]. Available: <http://d-scholarship.pitt.edu/7948/>. Accessed 07-Feb-2016
9. Cook RD, Weisberg S (1982) *Residuals and influence in regression*. Chapman and Hall, New York
10. Penny KI (1996) Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance. *Appl Stat* 45(1):73–81
11. Macqueen J (1967) Some methods for classification and analysis of multivariate observations. Presented at the proceedings of 5th Berkeley symposium on mathematical statistics and probability, pp 281–297
12. Hu X, Xu L (2003) A comparative study of several cluster number selection criteria. In: Liu J, Cheung Y, Yin H (eds) *Intelligent data engineering and automated learning*. Springer, Berlin, pp 195–202
13. Jones RH (2011) Bayesian information criterion for longitudinal and clustered data. *Stat Med* 30(25):3050–3056
14. Cherednichenko S (2005) *Outlier detection in clustering*
15. Provan D (2010) *Oxford handbook of clinical and laboratory investigation*. OUP Oxford
16. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44(3):837–845