Atypical (Rare) Elements Detection – A Conditional Nonparametric Approach

Piotr Kulczycki[1,2(✉)], Malgorzata Charytanowicz[1,3], Piotr A. Kowalski[1,2], and Szymon Lukasik[1,2]

[1] Centre of Information Technology for Data Analysis Methods, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
kulczycki@ibspan.waw.pl
[2] Division for Information Technology and Systems Research, Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, Kraków, Poland
kulczycki@agh.edu.pl
[3] Institute of Mathematics and Computer Science, Faculty of Mathematics, Computer Science and Landscape, Catholic University of Lublin, Lublin, Poland

Abstract. This paper presents a ready-to-use procedure for detecting atypical (rarely occurring) elements, in one- and multidimensional spaces. The issue is considered through a conditional approach. The application of nonparametric concepts frees the investigated procedure from distributions of describing and conditioning variables. Ease of interpretation and completeness of the presented material lend themselves to the use of the worked out method in a wide range of tasks in various applications of data analysis in science and practice, engineering, economy and management, environmental and social issues, biomedicine, and related fields.

Keywords: Rare element · Atypical element · Outlier · Outlier detection · Conditional approach · Distribution free method · Numerical algorithm

1 Introduction

Atypical elements (often casually referred to as outliers) can intuitively be considered as significantly differing from the general population [1, 3, 5, 6]. Their occurrence most commonly results from considerable ("gross") errors arising during the measurement, collection, storage, and processing of data [17]. In practice they hinder the correct utilization of knowledge available and their elimination or correction enables the use of more convenient and more effective methods at later stages of analysis and exploration [2]. What is more, in marketing, atypical elements may represent cases so different from the majority of the population, that any individual decision based on such a group – so different and insignificant – often turns out to be economically unviable. In engineering, the presence of atypical states in dynamic systems may be an evidence of malfunction of a component or the entire device, and proper reaction usually enables any serious consequences to be avoided. The detection of an atypical element may also

R.P. Barneva et al. (Eds.): CompIMAGE 2016, LNCS 10149, pp. 56–64, 2017.
DOI: 10.1007/978-3-319-54609-4_4

signify an attempt to hack into a computer system. On the other hand, in many social and economic problems the appearance of this element could be a positive trait, as it may characterize completely new trends or uncommon phenomena, and their quick discovery allows the appropriate specific action to be taken in anticipation. Therefore, the detection of atypical elements constitutes a natural cognitive challenge of great scientific and practical meaning.

The task of detecting atypical elements is one of very difficult conditioning. Above all most often there is no definition or even criterion indicating which elements should be considered atypical. What is more, we do not have a pattern of atypical elements, and even if we do, it would be – by its nature – small in number, strongly unbalanced with respect to the typical set of elements. For an illustrative example, in the simplest one-dimensional case, where data distribution is unimodal, atypical elements can be considered to be elements distant (according to the basic meaning of the term "outlier") from a median – defining the "center" of a data set – of more than 3/2 of the interquartile range; see [20; Sect. 2.7]. However, a similar approach cannot be taken concerning complex multimodal distributions. In particular, when specific modes are significantly distanced from each other, elements lying in the center between them should be considered as atypical, although they may be located very near to the median, definitely closer than 3/2 of the interquartile range.

This paper assumes as atypical those elements occurring rarely in the population. Thus, having a representative set of data, we highlight regions of lowest distribution density, such that the common probabilities of the elements appearing in those regions are equal to the assumed value, e.g. 0.01, 0.05, 0.1. Such locations can be of any shape, location and number.

In many practical tasks, the data possessed can be significantly refined through the measurement and introduction to a model of the current value of quantity considerably influencing the subject of investigation. In engineering practice, such a factor may often be the current temperature. From a formal point of view, the above aim can be realized by using a conditional probabilistic approach [4]. In this case, the basic attributes, called describing, become dependent on the conditioning attributes, the measured and introduced specific values of which can make substantially more precise the information related to the object under research. This approach is the subject of the present paper.

For defining data characteristics, the nonparametric methodology of kernel estimators is used, which frees the investigated procedures from forms of distributions characterizing both the describing and conditioning quantities. The presented material is complete and ready-to-use without laborious investigations. Particularly valuable is its easy, illustrative interpretation. A broader description of the material investigated here is presented in the paper [15], currently in press.

2 Mathematical Preliminaries: Kernel Estimators

Let the n-dimensional continuous random variable X be given, with a distribution characterized by the density f. Its kernel estimator $\hat{f} : \mathsf{R}^n \to [0, \infty)$, calculated using experimentally obtained values for the m-element random sample

$$x_1,\ x_2,\ \ldots,\ x_m, \tag{1}$$

in its basic form is defined as

$$\hat{f}(x) = \frac{1}{mh^n} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right), \tag{2}$$

where $m \in \mathrm{N}\backslash\{0\}$, the coefficient $h > 0$ is called a smoothing parameter, while the measurable function $K : \mathrm{R}^n \to [0, \infty)$ of unit integral $\int_{\mathrm{R}^n} K(x)\mathrm{d}x = 1$, symmetrical with respect to zero and having a weak global maximum in this place, takes the name of a kernel. The choice of form of the kernel K and the calculation of the smoothing parameter h is made most often with the criterion of the mean integrated square error [10, 21, 22].

Thus, the choice of the kernel form has – from a statistical point of view – no practical meaning and thanks to this, it becomes possible to take primarily into account properties of the estimator obtained or calculational aspects advantageous from the point of view of the applicational problem under investigation; for broader discussion see the books [10 – Sect. 3.1.3; 22 – Sects. 2.7 and 4.5]. In practice, for the one-dimensional case (i.e., when $n = 1$), the function K is assumed most often to be the density of a common probability distribution. In the multidimensional case, two natural generalizations of the above concept are used: radial and product kernels. However, the former is somewhat more effective, although from an applicational point of view, the difference is insignificant, and the product kernel – significantly more convenient for analysis – is often favored for practical problems. The n-dimensional product kernel K can be expressed as

$$K(x) = K\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}\right) = K_1(x_1)\ K_2(x_2) \ldots K_n(x_n), \tag{3}$$

where K_i denotes the previously-mentioned one-dimensional kernels, while the expression h^n appearing in the basic formula (2) should be replaced by the product of the smoothing parameters for particular coordinates $h_1 \cdot h_2 \cdot \ldots \cdot h_n$. For further investigations, the (one-dimensional) Cauchy kernel

$$K(x) = \frac{2}{\pi} \frac{1}{(1 + x^2)^2} \tag{4}$$

is applied.

The fixing of the smoothing parameter h has significant meaning for quality of estimation. Fortunately many suitable procedures for calculating the value of the parameter h on the basis of random sample (1) have been worked out. For broader discussion of this task see the books [10, 21, 22]. In particular, for the one-dimensional case, the effective plug-in method [10 – Sect. 3.1.5; 22 – Sect. 3.6.1] is especially

recommended. Of course this method can also be applied in the n-dimensional case when product kernel (3) is used, sequentially n times for each coordinate. One can also apply the simplified method [10 – Sect. 3.1.5; 21 – Sect. 3.4.1; 22 – Sect. 3.2.1], which for Cauchy kernel (4) takes the plain form

$$h = \left(\frac{10}{3\sqrt{\pi}m}\right)^{1/5} \hat{\sigma}, \tag{5}$$

where $\hat{\sigma}$ denotes the estimator of a standard deviation for a given coordinate. The above value may be sufficiently precise for many practical applications, whereas – thanks to its simplicity – this method significantly increases the calculation velocity.

The above concept will now be generalized for the conditional case. Here, besides the basic (sometimes termed the describing) n_Y-dimensional random variable Y, let also be given the n_W-dimensional random variable W, called hereinafter the conditioning random variable. Their composition $X = \begin{bmatrix} Y \\ W \end{bmatrix}$ is a random variable of the dimension $n_Y + n_W$. Assume that distributions of the variables X and, in consequence, W have densities, denoted below as $f_X : \mathrm{R}^{n_Y + n_W} \to [0, \infty)$ and $f_W : \mathrm{R}^{n_W} \to [0, \infty)$, respectively. Let also be given the so-called conditioning value, i.e., the fixed value of conditioning random variable $w^* \in \mathrm{R}^{n_W}$, such that

$$f_W(w^*) > 0. \tag{6}$$

Then the function $f_{Y|W=w^*} : \mathrm{R}^{n_Y} \to [0, \infty)$ given by

$$f_{Y|W=w^*}(y) = \frac{f_X(y, w^*)}{f_W(w^*)} \quad \text{for every } y \in \mathrm{R}^{n_Y} \tag{7}$$

constitutes a conditional density of probability distribution of the random variable Y for the conditioning value w^*. The conditional density $f_{Y|W=w^*}$ can, therefore, be treated as a "classic" density, whose form has been made more accurate in practical applications with w^* – a concrete value taken by the conditioning variable W in a given situation.

Let, therefore, the random sample

$$\begin{bmatrix} y_1 \\ w_1 \end{bmatrix}, \begin{bmatrix} y_2 \\ w_2 \end{bmatrix}, \ldots, \begin{bmatrix} y_m \\ w_m \end{bmatrix}, \tag{8}$$

obtained from the variable $X = \begin{bmatrix} Y \\ W \end{bmatrix}$, be given. The particular elements of this sample are interpreted as the values y_i taken in measurements from the random variable Y, when the conditioning variable W assumes the respective values w_i. On the basis of sample (6) one can calculate $\hat{f}_X$, i.e. the kernel estimator of density of the random variable X probability distribution, while the sample

$$w_1,\ w_2,\ \ldots,\ w_m \tag{9}$$

gives $\hat{f}_W$ – the kernel density estimator for the conditioning variable W. The kernel estimator of conditional density of the random variable Y distribution for the conditioning value w^*, is defined then – in natural consequence of formula (5) – as the function $\hat{f}_{Y|W=w^*} : \mathsf{R}^{n_Y} \to [0, \infty)$ given by

$$\hat{f}_{Y|W=w^*}(y) = \frac{\hat{f}_X(y, w^*)}{\hat{f}_W(w^*)}. \tag{10}$$

If for the estimator $\hat{f}_W$ one uses a kernel with positive values, then the inequality $\hat{f}_W(w^*) > 0$ implied by condition (4) is fulfilled for any $w^* \in \mathsf{R}^{n_W}$.

In the case when for the estimators $\hat{f}_X$ and $\hat{f}_W$ the product kernel (3) is used, applying in pairs the same kernels to the estimator $\hat{f}_X$ for coordinates which correspond to the vector W and to the estimator $\hat{f}_W$, then the expression for the kernel estimator of conditional density becomes particularly helpful for practical applications. Formula (8) can then be specified to the form

$$\hat{f}_{Y|W=w^*}(y) = \hat{f}_{Y|W=w^*}\left(\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_Y} \end{bmatrix}\right)$$

$$= \frac{\frac{1}{h_1 h_2 \ldots h_{n_Y}} \sum_{i=1}^{m} K_1\left(\frac{y_1 - y_{i,1}}{h_1}\right) K_2\left(\frac{y_2 - y_{i,2}}{h_2}\right) \ldots K_{n_Y}\left(\frac{y_{n_Y} - y_{i,n_Y}}{h_{n_Y}}\right) K_{n_Y+1}\left(\frac{w_1^* - w_{i,1}}{h_{n_Y+1}}\right) K_{n_Y+2}\left(\frac{w_2^* - w_{i,2}}{h_{n_Y+2}}\right) \ldots K_{n_Y+n_W}\left(\frac{w_{n_W}^* - w_{i,n_W}}{h_{n_Y+n_W}}\right)}{\sum_{i=1}^{m} K_{n_Y+1}\left(\frac{w_1^* - w_{i,1}}{h_{n_Y+1}}\right) K_{n_Y+2}\left(\frac{w_2^* - w_{i,2}}{h_{n_Y+2}}\right) \ldots K_{n_Y+n_W}\left(\frac{w_{n_W}^* - w_{i,n_W}}{h_{n_Y+n_W}}\right)}, \tag{11}$$

where $h_1, h_2, \ldots, h_{n_Y+n_W}$ represent – respectively – smoothing parameters mapped to particular coordinates of the random variable X, while the coordinates of the vectors w^*, y_i and w_i are denoted as

$$w^* = \begin{bmatrix} w_1^* \\ w_2^* \\ \vdots \\ w_{n_W}^* \end{bmatrix} \text{ and } y_i = \begin{bmatrix} y_{i,1} \\ y_{i,2} \\ \vdots \\ y_{i,n_Y} \end{bmatrix}, \quad w_i = \begin{bmatrix} w_{i,1} \\ w_{i,2} \\ \vdots \\ w_{i,n_W} \end{bmatrix} \quad \text{for} \quad i = 1, 2, \ldots, m. \tag{12}$$

Define the so-called conditioning parameters d_i for $i = 1, 2, \ldots, m$ by the following formula:

$$d_i = K_{n_Y+1}\left(\frac{w_1^* - w_{i,1}}{h_{n_Y+1}}\right) K_{n_Y+2}\left(\frac{w_2^* - w_{i,2}}{h_{n_Y+2}}\right) \ldots K_{n_Y+n_W}\left(\frac{w_{n_W}^* - w_{i,n_W}}{h_{n_Y+n_W}}\right). \tag{13}$$

If the values of the kernels K_{n_Y+1}, $K_{n_Y+2}, \ldots, K_{n_Y+n_W}$ are positive, then these parameters are also positive. So the kernel estimator of conditional density (9) can be finally presented in the form

$$\hat{f}_{Y|W=w^*}(y) = \hat{f}_{Y|W=w^*}\left(\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_Y} \end{bmatrix}\right) = \frac{1}{h_1 h_2 \ldots h_{n_Y} \sum_{i=1}^{m} d_i} \sum_{i=1}^{m} d_i K_1\left(\frac{y_1 - y_{i,1}}{h_1}\right) K_2\left(\frac{y_2 - y_{i,2}}{h_2}\right) \ldots K_{n_Y}\left(\frac{y_{n_Y} - y_{i,n_Y}}{h_{n_Y}}\right). \tag{14}$$

The value of the parameter d_i – the "distance" of the given conditioning value w^* from w_i – that of the conditioning variable for which the i-th element of the random sample was obtained. Then estimator (12) can be interpreted as the linear combination of kernels mapped to particular elements of a random sample obtained for the variable Y, when the coefficients of this combination characterize how representative these elements are for the given value w^*.

More details concerning kernel estimators can be found in the classic monographs [10, 21, 22]. Sample applications for data analysis tasks are described in the publications [11–13, 16, 18]. See also [19].

3 Conditional Atypical Elements Detection

Consider – in relation to notations introduced in the previous section – the data-set comprised of elements representative for a population $y_1, y_2, \ldots, y_m$, obtained for the conditioning values $w_1, w_2, \ldots, w_m$, respectively. The aim of the developed procedure is the isolation from the set $y_1, y_2, \ldots, y_m$ of elements which are atypical in the sense that they occur most rarely, in conditions when the specific conditioning value w^* appears.

First, fix the number

$$r \in (0, 1) \tag{15}$$

defining a desired proportion of atypical to typical elements, more accurately the share of atypical elements in a population. In practice, the values $r = 0.01, 0.05, 0.1$ are commonly used. In reference to the notations in the previous section, let us treat the set $y_1, y_2, \ldots, y_m$ as the realization of the n_Y-dimensional continuous random variable Y, and the set $w_1, w_2, \ldots, w_m$ as their respective set of realizations of the conditioning random variable W, and then calculate the conditional density $\hat{f}_{Y|W=w^*}$. Next, let us consider the set of its values for the elements of the set $y_1, y_2, \ldots, y_m$, therefore

$$\hat{f}_{Y|W=w^*}(y_1), \hat{f}_{Y|W=w^*}(y_2), \ldots, \hat{f}_{Y|W=w^*}(y_m) \in \mathsf{R}. \tag{16}$$

The value $\hat{f}_{Y|W=w^*}(y_i)$ refers to the probability of occurrence of the element y_i with the assumption that the value of the conditioning variable is w^*. So, the greater the value $\hat{f}_{Y|W=w^*}(y_i)$, the more typical element y_i can be interpreted to be for the given w^*. Let's treat as typical these elements y_i for which $\hat{f}_{Y|W=w^*}(y_i)$ is bigger than a given limit value, while atypical – those y_i for which $\hat{f}_{Y|W=w^*}(y_i)$ is smaller. In accordance with the assumptions made, a quantile of the order r for the condition w^* should be accepted as the above limit value. Thus, the set of elements y_i was hereby divided into $[100 \cdot r]$-percent of elements of lower probability and $[100 \cdot (1-r)]$-percent of those of higher probability.

There remains, however, to calculate the above mentioned value of the quantile. To this end, the kernel estimator scheme presented in the paper [14], fitted to the task investigated here, will be applied. The values (14) will be treated as realizations of the one-dimensional describing random variable Z, obtained, as before, for the realizations $w_1, w_2, \ldots, w_m$ of the n_W-dimensional conditioning random variable W. The kernel estimator of a quantile of the order r for the condition w^* can be effectively calculated on the basis of Newton's algorithm [7] as the limit of the sequence $\{\hat{q}_{r|w^*,j}\}_{j=0}^{\infty}$ defined by

$$\hat{q}_{r|w^*,0} = \frac{\sum_{i=1}^{m} d_i \hat{f}_{Y|W=w^*}(y_i)}{\sum_{i=1}^{m} d_i} \tag{17}$$

$$\hat{q}_{r|w^*,j+1} = \hat{q}_{r|w^*,j} - \frac{L(\hat{q}_{r|w^*,j})}{L'(\hat{q}_{r|w^*,j})} \quad \text{for} \quad j = 0, 1, \ldots, \tag{18}$$

with the functions L and L' being given by dependencies

$$L(\hat{q}_{r|w^*}) = \sum_{i=1}^{m} d_i I\left(\frac{\hat{q}_{r|w^*} - \hat{f}_{Y|W=w^*}(y_i)}{h}\right) - r\sum_{i=1}^{m} d_i \tag{19}$$

$$L'(\hat{q}_{r|w^*}) = \frac{1}{h}\sum_{i=1}^{m} d_i K\left(\frac{\hat{q}_{r|w^*} - \hat{f}_{Y|W=w^*}(y_i)}{h}\right), \tag{20}$$

where $I : \mathsf{R} \to [0,1]$ means a primitive of the kernel K, i.e., $I(x) = \int_{-\infty}^{x} K(y)\,\mathrm{d}y$, whereas a stop criterion takes on the form

$$|\hat{q}_{r|w^*,j} - \hat{q}_{r|w^*,j-1}| \leq 0.01\ \hat{\sigma}_Z, \tag{21}$$

while $\hat{\sigma}_Z$ denotes the estimator of the standard deviation of the random variable Z, found on the basis of set (16). For Cauchy kernel (4), its primitive is given as

$$I(x) = \frac{1}{\pi}\left[\text{arctg}(x) + \frac{x}{(1+x^2)}\right] + \frac{1}{2}. \tag{22}$$

In formulas (19)–(20), the smoothing parameter h should be calculated for set (16).

Thanks to the use of kernel estimators with strong averaging properties, inference takes place not only for data obtained exactly for w^* (among the values w_i there may be some too small for reliable consideration or even not at all), but also for neighboring values proportional to their "closeness" with respect to w^*.

4 Final Remarks and Conclusion

The procedure presented in this paper has been numerically verified in detail. The obtained results confirmed its correct functioning and full completion of the objectives and goals set out in the Introduction. Particularly, in the case of a positive correlation between the describing and conditioning factors, the greater (or smaller) the value of the conditioning attributes, the greater (or smaller) the values of the describing elements detected to be atypical. For the negative correlation, the above relation is inverse.

The procedure also successfully underwent verification in solving a practical problem in control engineering. Based on the current state of the system, the atypical elements discovered, provided evidence of arising failures of an observed device [8, 9]. The conditioning factors allowed the model used to be significantly refined.

A detailed description of the numerical and empirical verifications can be found in the article [15], currently in press.

Finally, this paper presents the algorithm for atypical (rare) elements also for a multivalued case, with continuous coordinates of describing and conditioning variables. The conditional approach allows in practice for refinement of the model by including the current value of the conditioning factors. Use of the nonparametric concepts frees the worked out procedure from distributions of describing and conditioning attributes. The investigated algorithm is ready for direct use without any additional laborious research or calculations. A full version of the material described here is presented in the paper [15].

Acknowledgments. Our heartfelt thanks go to our colleagues Damian Kruszewski and Cyprian Prochot, with whom we collaborated on the subject presented here.

References

1. Aggarwall, C.C.: Outlier Analysis. Springer, Heidelberg (2013)
2. Aggarwal, C.C.: Data Mining. Springer, Heidelberg (2015)
3. Barnett, V., Lewis, T.: Outliers in Statistical Data. Wiley, Hoboken (1994)
4. Dawid, A.P.: Conditional independence in statistical theory. J. Roy. Stat. Soc. Ser. B **41**, 1–31 (1979)
5. Hawkins, D.M.: Identification of Outliers. Chapman and Hall, London (1980)

6. Hodge, V., Austin, J.: A survey of outlier detection methodologies. Artif. Intell. Rev. **22**, 85–126 (2004)
7. Kincaid, D., Cheney, W.: Numerical Analysis. Brooks/Cole, Pacific Grove (2002)
8. Korbicz, J., Kościelny, J.M., Kowalczuk, Z., Cholewa, W. (eds.): Fault Diagnosis: Models, Artificial Intelligence, Applications. Springer, Heidelberg (2004)
9. Kulczycki, P.: Wykrywanie uszkodzen w systemach zautomatyzowanych metodami statystycznymi. Alfa, Warsaw (1998)
10. Kulczycki, P.: Estymatory jadrowe w analizie systemowej. WNT, Warszawa (2005)
11. Kulczycki, P.: Kernel estimators in industrial applications. In: Prasad, B. (ed.) Soft Computing Applications in Industry, pp. 69–91. Springer, Heidelberg (2008).
12. Kulczycki, P., Charytanowicz, M.: A Complete gradient clustering algorithm formed with kernel estimators. Int. J. Appl. Math. Comput. Sci. **20**, 123–134 (2010)
13. Kulczycki, P., Charytanowicz, M.: Conditional parameter identification with different losses of under- and overestimation. Appl. Math. Model. **37**, 2166–2177 (2013)
14. Kulczycki, P., Charytanowicz, M., Dawidowicz, A.: A Convenient ready-to-use algorithm for a conditional quantile estimator. Appl. Math. Inf. Sci. **9**, 841–850 (2015)
15. Kulczycki P., Charytanowicz M., Kowalski P.A., Lukasik S.: Identification of Atypical (Rare) Elements – A Conditional, Distribution-Free Approach. IMA J. Math. Control I. (2017, in press)
16. Kulczycki, P., Daniel, K.: Metoda wspomagania strategii marketingowej operatora telefonii komorkowej. Przeglad Statystyczny **56**(2), 116–134 (2009). Errata: **56**(3-4), 3
17. Kulczycki, P., Hryniewicz, O., Kacprzyk, J. (eds.): Techniki informacyjne w badaniach systemowych. WNT, Warszawa (2007)
18. Kulczycki P., Kowalski P.A.: Bayes classification for nonstationary patterns. Int. J. Comput. Methods **12** (2015). Article ID: 1550008
19. Kulczycki, P., Lukasik, S.: An algorithm for reducing dimension and size of sample for data exploration procedures. Int. J. Appl. Math. Comput. Sci. **24**, 133–149 (2014)
20. Larose, D.T.: Discovering Knowledge in Data. An Introduction to Data Mining. Wiley, Hoboken (2005)
21. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (1986)
22. Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman and Hall, London (1995)