

Detection of Atypical Elements by Transforming Task to Supervised Form

Piotr Kulczycki^{1,2}  and Damian Kruszewski¹ 

¹ Centre of Information Technology for Data Analysis Methods,
Polish Academy of Sciences, Systems Research Institute, Warsaw, Poland
kulczycki@ibspan.waw.pl, kulczycki@agh.edu.pl

² Division for Information Technology and Systems Research,
AGH University of Science and Technology,
Faculty of Physics and Applied Computer Science, Cracow, Poland

Abstract. The problem of identifying atypical elements in a data set presents many difficulties at every stage of analysis. For instance, it is not clear which traits should distinguish such elements, and what more we cannot know in advance of their natural pattern, which even if it did exist, would in its nature be significantly limited. The subject of the presented research is the procedure for transforming the problem of detection of atypical elements from an unsupervised task to a supervised one with equal-sized patterns. This allows a suitable analysis, in particular the use of diverse well-developed methods of classification. Elements are considered atypical by their rare occurrence, which when coupled with the application of nonparametric methodology enables their detection not only on the peripheries of the distribution, but also – in the multimodal case – potentially located inside.

Keywords: Atypical element · Rare element · Outlier · Atypical elements detection · Classification · Distribution-free methods

1 Introduction

Atypical elements (often rashly referred to as outliers) can intuitively be considered as significantly differing from the rest of a data set (Aggarwal 2013; Barnett and Lewis 1994). The immense diversity of interpretations of such an intuitive definition means that even the concept of an atypical element itself is ambiguous, from the trivial, where they are elements furthest away from the remaining population (outliers), to the functional, when they have the greatest – or rather excessive – influence on a system operation. This paper will apply the most universal frequency approach, whereby atypical elements are rare, i.e. the probability of their appearance is faint. Thanks to the application of distribution-free nonparametric methodology, we can identify atypical observations not only on the peripheries of the population, but in the case of multimodal distributions with wide-spreading segments, also those lying in between such segments, even if close to the center of the set.

A different problem results from the unsupervised nature of the task, which manifests in the lack of a priori natural pattern of atypical elements. It is worth noting that

even if it existed, it would obviously occur significantly less than the pattern of typical ones. The subject of this paper is the transformation of an unsupervised task to a supervised one with equal-sized patterns, which in consequence enables the use of a well-developed valuable and distinctive classification apparatus.

The procedure investigated and presented here is ready-to-use without laborious research. Its easy and illustrative interpretation is particularly valuable.

Section 2 presents the distribution-free statistical kernel estimators methodology. Then, the basic formula of the procedure for identifying atypical, i.e. rarely occurring, elements is described in Sect. 3. Due to difficult conditioning, mainly stemming from a naturally very low number of elements considered atypical, the quality of the procedure is considerably improved in Sect. 4 by significantly increasing the set of elements representative for the population. Next, in Sect. 5, patterns of atypical and typical elements, equal in size, will be generated, which among others form the basis for the convenient application of classification methods, according to the researcher's preferences and specifics of the task under consideration. Final comments are shortly presented in Sect. 6.

A broader description of the concept worked out here and detailed results of empirical verification can be found in the paper (Kulczycki and Kruszewski 2017a). In the publication (Kulczycki and Kruszewski 2017b) the procedure design to submit the result in fuzzy and intuitionistic fuzzy forms is investigated.

2 Nonparametric Kernel Estimators

In the presented method, the characteristics of a data set will be defined using the nonparametric methodology of kernel estimators. It is distribution-free, i.e. the preliminary assumptions concerning the types of appearing distributions are not required. A broad description can be found in the monographs (Kulczycki 2005; Wand and Jones 1995). Exemplary applications for data analysis tasks are described in the publications (Kulczycki et al. 2012; Kulczycki and Charytanowicz 2016; Kulczycki and Kowalski 2016); see also (Kulczycki and Lukasik 2014).

Let the n -dimensional continuous random variable X be given, with a distribution characterized by the density X . Its kernel estimator $\hat{f} : \mathbb{R}^n \rightarrow [0, \infty)$, calculated using the experimentally obtained m -element random sample x_i for $i = 1, 2, \dots, m$, in its basic form is defined as

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

where $m \in \mathbb{N} \setminus \{0\}$, the coefficient $h > 0$ is called a smoothing parameter, while the measurable function $K : \mathbb{R}^n \rightarrow [0, \infty)$ of unit integral $\int_{\mathbb{R}^n} \hat{f}(x) dx = 1$, symmetrical with respect to zero and having a weak global maximum in this place, takes the name of a kernel.

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 into account primarily

properties of the estimator obtained or computational aspects, advantageous from the point of view of the applicational problem under investigation; for broader discussion see the books (Kulczycki 2005 – Sect. 3.1.3; Wand and Jones 1995 – Sects. 2.7 and 4.5). In the one-dimensional case (i.e. when $n = 1$) the normal (Gauss) kernel

$$K_j(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{2}$$

and the uniform kernel

$$K_j(x) = \begin{cases} \frac{1}{2} & \text{dla } x \in [-1, 1] \\ 0 & \text{dla } x \notin [-1, 1] \end{cases} \tag{3}$$

will be used in the following. The normal kernel is generally held as basic. The uniform kernel has bounded support and assumes a finite number of values, which will be taken advantage of later in this paper. In the multidimensional case, a so-called product kernel will be applied in the following. The main idea here is the division of particular variables with the multidimensional kernel then becoming a product of n one-dimensional kernels for particular coordinates. Thus the kernel estimator (1) is then given as

$$\hat{f}(x) = \frac{1}{mh_1h_2\dots h_n} \sum_{i=1}^m K_1\left(\frac{x_1 - x_{i,1}}{h_1}\right) K_2\left(\frac{x_2 - x_{i,2}}{h_2}\right) \dots K_n\left(\frac{x_n - x_{i,n}}{h_n}\right), \tag{4}$$

where K_j ($j = 1, 2, \dots, n$) denote one-dimensional kernels, e.g. (2) or (3), h_j ($j = 1, 2, \dots, n$) are smoothing parameters individualized for particular coordinates, while assigning to coordinates

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad x_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,n} \end{bmatrix} \quad \text{for } i = 1, 2, \dots, m. \tag{5}$$

The above kernels fulfill the additional requirements of the particular procedures used in the following.

The fixing of the smoothing parameter has significant meaning for quality of estimation. Fortunately many suitable procedures for calculating its value on the basis of a random sample have been worked out. For the purposes of the research investigated here, the simplified method (Kulczycki 2005 – Sect. 3.1.5; Wand and Jones 1995 – Sect. 3.2.1) will be applied, according to which

$$h_j = \left(\frac{8\sqrt{\pi} W(K_j)}{3 U(K_j)^2 m} \right)^{1/5} \hat{\sigma}_j \quad \text{for } j = 1, 2, \dots, n, \tag{6}$$

where $W(K_j) = \int_{-\infty}^{\infty} K_j(x)^2 dx$ and $U(K_j) = \int_{-\infty}^{\infty} x^2 K_j(x) dx$, while $\hat{\sigma}_j$ denotes the estimator of a standard deviation for the j -th coordinate:

$$\hat{\sigma}_j = \sqrt{\frac{1}{m-1} \sum_{i=1}^m x_{i,j}^2 - \frac{1}{m(m-1)} \left(\sum_{i=1}^m x_{i,j} \right)^2} \quad \text{for } j = 1, 2, \dots, n. \quad (7)$$

As shown in verification testing the presented procedure, this method seems to be sufficiently precise, and furthermore it is simple and fast. The functional values occurring in formula (6) are, respectively, for normal kernel (2)

$$W(K_j) = \frac{1}{2\sqrt{\pi}}, \quad U(K_j) = 1 \quad (8)$$

and for uniform (3)

$$W(K_j) = \frac{1}{2}, \quad U(K_j) = \frac{1}{3}. \quad (9)$$

For specific cases the more sophisticated yet effective plug-in method (Kulczycki 2005 – Sect. 3.1.5; Wand and Jones 1995 – Sect. 3.6.1) can be also proposed. It is provided for one-dimensional tasks but, of course, this method can be also applied in the n -dimensional case when a product kernel is used, sequentially n times for each coordinate.

In practice, various modifications and generalizations of the standard form of the kernel estimator presented above are possible, fitting its properties to specific realities. It is worth remembering however, that they increase complexity of formulas, their interpretation becomes more difficult and in consequence the problem is less convenient for potential users to solve. For many aspects concerning the kernel estimators method, see the classic monographs (Kulczycki 2005; Wand and Jones 1995).

3 Basic Version of Procedure

The basic idea of the presented procedure for identification of atypical elements stems from the significance test proposed in the work (Kulczycki and Prochot 2002). Thanks to the application of nonparametric methods it is unnecessary to introduce assumptions concerning distribution type for an examined population.

Let the set be given, with elements representative for the population

$$x_1, x_2, \dots, x_m. \quad (10)$$

Treat these elements as realizations of the n -dimensional continuous random variable X with distribution having density f and calculate – in accordance with Sect. 2 (using a normal kernel) – the kernel estimator \hat{f} . Next consider the set of its value for elements of set (10), so

$$\hat{f}(x_1), \hat{f}(x_2), \dots, \hat{f}(x_m). \tag{11}$$

It is worth noticing that, regardless of the dimension of the random variable X , the values of set (11) are real (one-dimensional).

Define now the number

$$r \in (0, 1) \tag{12}$$

establishing sensitivity of the procedure for identifying atypical elements. This number will determine the assumed proportion of atypical elements in relation to the total population, therefore the ratio of the number of atypical to the sum of atypical and typical elements. In practice

$$r = 0.01, 0.05, 0.1 \tag{13}$$

is the most often used, with particular attention paid to the second option. In certain applications it is possible to use other, approximate values of the above parameter.

Let us treat set (11) as realizations of a real (one-dimensional) random variable and calculate the estimator for the quantile of the order r . The positional estimator of the second order (Parrish 1990) will be applied in the following, given by the formula

$$\hat{q}_r = \begin{cases} z_1 & \text{for } mr < 0.5 \\ (0.5 + i - mr)z_i + (0.5 - i + mr)z_{i+1} & \text{for } mr \geq 0.5 \end{cases}, \tag{14}$$

where

$$i = [mr + 0.5], \tag{15}$$

while $[d]$ denotes an integral part of the number $d \in \mathbb{R}$, and z_i is the i -th value in size of set (11) after its sorting, thus

$$\{z_1, z_2, \dots, z_m\} = \{\hat{f}(x_1), \hat{f}(x_2), \dots, \hat{f}(x_m)\} \tag{16}$$

with $z_1 \leq z_2 \leq \dots \leq z_m$. Application of the positional quantile estimator guarantees its value does not exceed beyond support of the random variable under investigation, or rather to be more precise, thanks to the use of kernel (2) with positive values, the condition $\hat{q}_r > 0$ is fulfilled.

Generally there are no special recommendations concerning choice of sorting algorithm (Canaan et al. 2011) used for specifying set (16). However, let us interpret definition (14) and (15), taking into account condition (13). So, it is enough to sort only the $i + 1$ smallest values in the set $\{z_1, z_2, \dots, z_m\}$, therefore about 1-10% of its size. One can apply a simple algorithm that subsequently finds the $i + 1$ smallest elements of the set $\{z_1, z_2, \dots, z_m\}$.

Finally, if for a given tested element $\hat{x} \in \mathbb{R}^n$, the condition $\hat{f}(\hat{x}) \leq \hat{q}_r$ is fulfilled, then this element should be considered atypical; for the opposite $\hat{f}(\hat{x}) > \hat{q}_r$ it is typical. What is noteworthy is that for the correctly estimated quantities \hat{f} and \hat{q}_r , the above

guarantees obtaining the proportion of the number of atypical elements to total population at the assumed level r .

The above procedure for identifying atypical elements, combined with the properties of kernel estimators, allows in the multidimensional case for inferences based not only on values for specific coordinates of a tested element, but above all on the relations between them.

4 Extended Pattern of Population

Although, from a theoretical point of view, the procedure presented in the previous section seems complete, when the values r are applied in practice – see condition (13) – and the size m is not big, the estimator of the quantile \hat{q}_r is encumbered with a large error, due to the low number of elements z_i smaller than the estimated value. To counteract this, a data set will be extended by generating additional elements with distribution identical to that characterizing the subject population, based on set (10).

The methodology for enlarging a set representative for the investigated population is suggested using von Neumann’s elimination concept (Gentle 2003). This allows the generation of a sequence of random numbers of distribution with support bounded to the interval $[a, b]$, while $a < b$, characterized by the density f of values limited by the positive number c , i.e.

$$f(x) \leq c \text{ for every } x \in [a, b]. \tag{17}$$

In the multidimensional case, the interval $[a, b]$ generalizes to the n -dimensional cuboid $[a_1, b_1] \times [a_2, b_2] \times \dots \times [a_n, b_n]$, while $a_j < b_j$ for $j = 1, 2, \dots, n$.

First the one-dimensional case is considered. Let us generate two pseudorandom numbers u and v of distribution uniform to the intervals $[a, b]$ and $[0, c]$, respectively. Next one should check that

$$v \leq f(u). \tag{18}$$

If the above condition is fulfilled, then the value u ought to be assumed as the desired realization of a random variable with distribution characterized by the density f , that is

$$x = u. \tag{19}$$

In the opposite case the numbers u and v need to be removed and steps (18) and (19) repeated, until the desired number of pseudorandom numbers x with density f is obtained.

In the presented procedure the density f is established by the kernel estimators methodology, described in Sect. 2. Denote its estimator as \hat{f} . The uniform kernel will be employed, allowing easy calculation of the support boundaries a and b , as well as the parameter c appearing in condition (17). Namely:

$$a = \min_{i=1,2,\dots,m} x_i - h \tag{20}$$

$$b = \max_{i=1,2,\dots,m} x_i + h \tag{21}$$

$$c = \max_{i=1,2,\dots,m} \{ \hat{f}(x_i - h), \hat{f}(x_i + h) \}. \tag{22}$$

The last formula results from the fact that the maximum for a kernel estimator with the uniform kernel must occur on the edge of one of the kernels. It is also worth noting that calculations of parameters (20)–(22) do not require much effort. This is thanks to the appropriate choice of kernel form.

In the multidimensional case, von Neumann’s elimination algorithm is similar to the previously discussed one-dimensional version. The edges of the n -dimensional cuboid $[a_1, b_1] \times [a_2, b_2] \times \dots \times [a_n, b_n]$ are calculated from formulas comparable to (20)–(22) separately for particular coordinates. The kernel estimator maximum is thus located in one of the corners of one of the kernels; therefore

$$c = \max_{i=1,2,\dots,m} \left\{ \hat{f} \left(\begin{bmatrix} x_{i,1} \pm h_1 \\ x_{i,2} \pm h_2 \\ \vdots \\ x_{i,n} \pm h_n \end{bmatrix} \right) \right\} \text{ following all combinations of } \pm . \tag{23}$$

The number of these combinations is finite and equal to 2^n . Using the formula presented, n particular coordinates of pseudorandom vector u and the subsequent number v are generated, after which condition (18) is checked.

The results of verification presented in Sect. 6 show that for the properly extended set (10), the procedure investigated here for identifying atypical elements allows us to obtain a proportion of this type of element throughout the whole population, with great accuracy, sufficient from an applicational point of view.

5 Equal-Sized Patterns of Atypical and Typical Elements

Let us consider set (10) introduced in Sect. 3, consisting of elements representative for an investigated population, and extended as described in accordance with Sect. 4. In taking its subset comprising these observations x_i for which $\hat{f}(x_i) \leq \hat{q}_r$, one can treat it as a pattern of atypical elements. Denote it thus:

$$x_1^{at}, x_2^{at}, \dots, x_{m_{at}}^{at}. \tag{24}$$

Similarly, the set of observations for which $\hat{f}(x_i) > \hat{q}_r$ may be considered as a pattern of typical elements:

$$x_1^t, x_2^t, \dots, x_{m_t}^t. \tag{25}$$

Sizes of the above patterns equal respectively m_{at} and m_t . Of course $m_{at} + m_t = m$; we also have

$$\frac{m_{at}}{(m_{at} + m_t)} \cong r \tag{26}$$

In this way, unsupervised in its nature, the problem of identifying atypical elements has been reduced to a supervised classification task, although with strongly unbalanced patterns – taking into account relation (26) with (13), set (24) is in practice around 10-100 times smaller than (25). Classification is relatively conveniently conditioned and can use many different well developed methods. However most procedures work much better if patterns are of similar or even equal sizes (Kaufman and Rousseeuw 1990). Using once again the algorithm presented in Sect. 5, the size of the set can be increased to m_t , so that $m_{at} = m_t$, thus equaling patterns of atypical (24) and typical (25) elements.

Finally, a method for the unsupervised identification of atypical elements, has been thus brought to supervised classification with two patterns of equal, relatively large size, thereby creating the conveniently conditioned task with rich and diverse methodology, allowing for the selection of the best procedure regarding the character of the problem and user preferences.

6 Final Comments

The operation of the procedure was tested in details. First with the use of generated data the quantitative aspects were verified, in particular suggestions for fixing parameters. Figure 1 presents an exemplary decision tree attained for the bimodal distribution:

$$N(-3, 1) \quad 40\%, \quad N(3, 1) \quad 60\%. \tag{27}$$

This classification method offers an illustrative interpretation of a problem. Thanks to the equal-sized patterns, the results obtained in this way are close to those obtained with an unsupervised procedure of identification for atypical elements, however the potential fundamental analysis of decision trees brings great additional possibilities to enhance a model as information concerning its correctness is obtained, and flexibly adapt it to a changing environment.

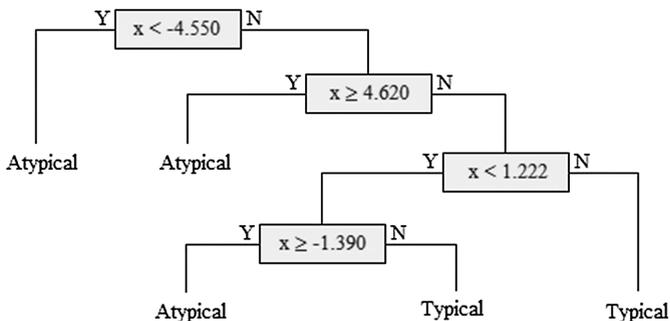


Fig. 1. Decision tree for bimodal distribution (27); $r = 0.1$, $m = 1,000$, $m^* = 10,000$.

Next, real experimental data taken from medicine (National Health and Nutrition Examination Survey 2016; National Cancer Institute 2016) were applied to demonstrate the comprehensive application of the presented procedure. The obtained results fully positively confirmed the correct functioning of the procedure presented in this paper. A detailed description of the experimental verification is presented in the paper (Kulczycki and Kruszewski 2017). The concept's independence from a distribution characterizing an analyzed set (in particular multimodality) as well as the dimensionality of a problem (within a reasonable range) should be underlined.

References

- Aggarwal, C.C.: *Outlier Analysis*. Springer, New York (2013)
- Barnett, V., Lewis, T.: *Outliers in Statistical Data*. Wiley, New York (1994)
- Canaan, C., Garai, M.S., Daya, M.: Popular sorting algorithms. *World Applied Programming* **1**, 62–71 (2011)
- Gentle, J.E.: *Random Number Generation and Monte Carlo Methods*. Springer, New York (2003)
- Kaufman, L., Rousseeuw, P.J.: *Finding groups in data: An introduction to cluster analysis*. Wiley, New York (1990)
- Kulczycki, P.: *Estymatory jądrowe w analizie systemowej*. WNT, Warsaw (2005)
- Kulczycki, P., Charytanowicz, M.: An algorithm for conditional multidimensional parameter identification with asymmetric and correlated losses of under- and overestimations. *J. Stat. Comput. Simul.* **86**, 1032–1055 (2016)
- Kulczycki, P., Charytanowicz, M., Kowalski, P.A., Lukasik, S.: The complete gradient clustering algorithm: properties in practical applications. *J. Appl. Stat.* **39**, 1211–1224 (2012)
- Kulczycki, P., Kowalski, P.A.: A complete algorithm for the reduction of pattern data in the classification of interval information. *Int. J. Comput. Methods* **13**, 1650018 (2016)
- Kulczycki, P., Kruszewski, D.: Identification of atypical elements by transforming task to supervised form with fuzzy and intuitionistic fuzzy evaluations. *Appl. Soft Comput.* **60**, 623–633 (2017a)
- Kulczycki, P., Kruszewski, D.: Detection of atypical elements with fuzzy and intuitionistic fuzzy evaluation. In: Mitkowski, W., Kacprzyk, J., Oprzedkiewicz, K., Skrush, P. (eds.) *Trends in Advanced Intelligent Control, Optimization and Automation*, pp. 774–786. Springer, Cham (2017b)
- 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)
- Kulczycki, P., Prochot C.: Identyfikacja stanów nietypowych za pomocą estymatorów jądrowych. In: Bubnicki, Z., Hryniewicz, O., Kulikowski, R. (eds.) *Metody i techniki analizy informacji i wspomaganie decyzji*. EXIT, Warsaw, pp. 57–62 (2002)
- National Health and Nutrition Examination Survey, <http://www.cdc.gov/nchs/nhanes.htm/>. Accessed 10 May 2016
- National Cancer Institute, <http://ctep.cancer.gov/>. Accessed 10 May 2016
- Parrish, R.: Comparison of quantile estimators in normal sampling. *Biometrics* **46**, 247–257 (1990)
- Wand, M., Jones, M.: *Kernel Smoothing*. Chapman and Hall, London (1995)