PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1126 2006
TAKSONOMIA 13
Klasyfikacja i analiza danych – teoria i zastosowania

Jan W. Owsiński
Systems Research Institute, Polish Academy of Sciences

OUTLIER DETECTION: NOTIONS, PROBLEMS, AND METHODOLOGICAL PROPOSALS

1. Introduction

Outlier detection has become during the recent years one of the favourite subjects in data analysis and related fields of scientific endeavour. This fact is linked with the appearance or increase of importance of some areas of application, mostly associated with large quantities of observations (like fraud detection in banking operations, abnormal communication behaviour, internet searches which may be indicative of terrorist activity, highly irregular spatial behaviour, etc.). Detection of outliers and identification of their characteristics may lead to their rejection from the sample or population, but also, more importantly, to knowledge related to uncovering of important singular or systematic phenomena.

A number of methodologies have been developed to identify outliers in data sets. They are shortly characterised in Section 2. We can very roughly classify them into the ones based on statistical models (like the GLM or the +/-3σ charts) and on clustering (like the most popular LOF, COF, or two-phase clustering, but also convex-hull determination based).

Additionally, the respective methodologies are quite often oriented at definite categories of problems (e.g. low or even very low dimensionality of the object space, like in spatial applications, but also in banking operations), and at definite problem formulations. This is associated with both the pragmatic considerations of computational effectiveness and the fundamental feature of outlier detection, namely dependence of technical outlier definition upon the model of a given process or situation.

The methods, which operate on large data sets, like, e.g., those associated with spatial analysis, strive mainly at the possibly rapid operation (e.g. linear in the number of objects). They usually refer to data sets of low dimensionality, and they can make use of very special characteristic features of these sets for outlier detection. Low dimensionality is often also the characteristic feature of the statistically based methods, not to mention the fact that they depend upon the distributions assumed.

In all such cases we deal with a very limited capacity of transferring methods across various applications. It is primarily because most of the methods suffer from the original problem of insufficiently precise definition of an *outlier.*

The paper starts with the standard understanding of *what an outlier is*, to then classify the approaches proposed and used in a bit more of detail. Then, some representative methods and techniques are shortly characterised, along with their implicit outlier definitions. In conclusion it is shown that the existing approaches and their implicit outlier definitions are not constructive, and generally lead to circular reasoning, even if it is claimed that the respective methods do without arbitrary assumptions (like in LOCI).

Against this background an understanding of what an outlier is is proposed. This intuitive quasi-definition distinguishes the situations in which we dispose of a "model", and the second, when no model is available, like in exploratory analysis. The consequences of these two different situations for the outlier identification are shown.

2. The standard problem formulation and the existing approaches

2.1. What is an outlier?

In the standard, and definitely quite intuitive formulation, outliers are "observations which seem to deviate strongly from the main part of the data" [Kuhnt, Pawlitschko 2005], or which "appear to be inconsistent with the remainder of [...] set of data" [Barnett, Lewis 1994]. Thus, an outlier is an observation that does not fit the "general pattern" of the data set, a population or a sample. This quasi-definition is very vague, and so there are many ways of making it operational, whether within the statistical methods or other approaches, including the cluster analysis based algorithms ("outliers [...] may be considered [as] the smallest clusters which are far away enough from the other objects" – [Zakrzewska, Osada 2004]).

There are some key ingredients to the concrete understanding of the above quasi-definitions. *First* outliers **seem** or **appear** to be outliers, this being a concession that we can only use relative measures. *Second*, they seem or appear to **deviate** or be **inconsistent** with (or **far away enough from**) the general body of observations. This suggests that we must be able to measure "distances" in the space of observations. Then, *third*, outliers are distinguished from the **main part** or **remainder** of the data set. In fact, when determining the outliers, we determine thereby also the non-outliers, or at least assume something about them.

It is mainly with respect to the latter issue that the notion of *model* is brought into the outlier detection problem. Thus, a model, understood in a variety of manners, stands for the definition of the main body of the data or the regular pattern of it. Thereby, however, also the other notions get more easily defined ("the problem is that you can't catch an outlier without a model (at least a mild one) for your data" – see [*Outlier detection...* 2004]).

2.2. An abbreviated classification of approaches

Identification and treatment of outliers has long been a subject of research, mainly in connection with their consideration in various statistical analyses (e.g. regression). Here, let us mention the seminal book by Hawkins [1980], but also, e.g., Barnett and Lewis [1994], while of the Polish literature of the subject, e.g. the paper by Zeliaś [1986]. The recent developments made new approaches appear, less related to the classical statistical paradigm. Following roughly Papadimitriou et al. [2002] we will now give short characterisations of the existing approaches as classified into:

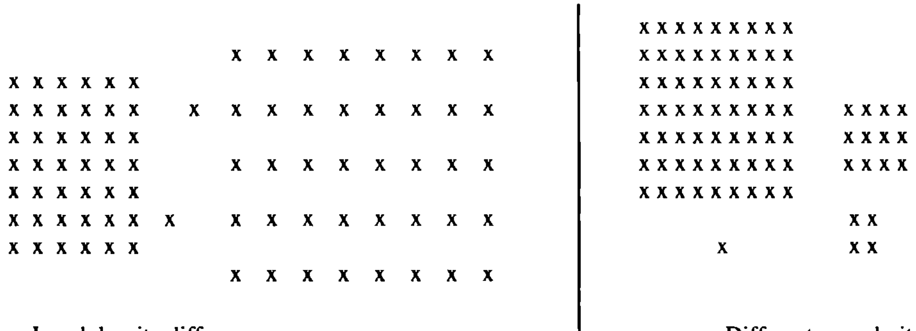
- *distribution based*, in which some distribution and/or its properties, are assumed, and when an object or objects deviate ("sufficiently") from the conditions implied by the assumption, they are labelled as (potential) outliers; the approach by Filzmoser [2004], is a classical example of such an approach; in fact, the vast majority of the traditional approaches to outliers relies on distributions;
- *geometrical* or *depth-based*, consisting in the use of geometric properties of respective sets, particularly determination of convex hulls corresponding to the subsets of the data set; although this approach has very attractive theoretical properties (e.g. possibility of evaluating global degree of belongingness), it is computationally most cumbersome of all;
- *classical clustering*, in which clusters "small enough" and "far away enough" are considered outliers; this approach has two important advantages: first, it does not require much of initial information, and second (contrary to what Papadimitriou et al. [2004], claim) – it can be used for high-dimensional and large data sets, if the clustering algorithm used is appropriately selected;
- *distance based*, in which an object is assigned a measure based on the distance between this object and the other objects in the set, and, when this measure exceeds a threshold, the object is labelled outlier;
- *density based*, in theoretical terms not far from the previous methods, these ones gained wide popularity, as capable of finding outliers notwithstanding the local differences in the density of the data set; Breunig et al. [2000] provided the seminal technique of this kind, to which most of the other related papers refer.

On the top of this, there are also, as mentioned, more specialised methods, meant for narrower applications. Special approaches and definitions are used for the time series data (see, e.g. [Suchecka, Kowalik 2005]), where outliers are a priori categorised according to their "nature". The reference quoted makes appeal to geostatistical methods. Indeed, spatial data analysis is another domain, in which outlier identification is of high importance (see [Mehra, Stello 2005; Janeja et al. 2004]). The specialised approaches are often (see, e.g. [Janeja et al. 2004; Zakrzewska, Osada 2004]) hybrid techniques, referring to various principles, or using simultaneously different indicators. The latter kind of composite approach is insofar justified as most methods determine outliers-to-a-degree, the degree being defined by the coefficients involved.

Finally, let us note that despite the separate classification provided above, the density and distance based methods are in fact very closely linked with the clustering ones, and that there is a very clear connection between the outlier identification and the cluster number problems.

2.3. Global vs. local criteria of "outlierness"

There are two essential geometric features of the data sets, which constitute problems for the outlier detection methods. These two are usually referred to as "local density" (different densities of various actually existing clusters, parts of the sample or of the population) and "different granularity" (different magnitude of these parts). They are illustrated below (Figure 1).



Local density differences

Different granularity

Figure 1.

It is these potential features of the data sets that make effective design and functioning of the outlier detection methods difficult, in view of the fact that global – or rather universal – criteria may fail. The obvious consequence is the design and use of the techniques having local properties, especially related to local density.

2.4. Some illustrative examples of the existing approaches

Let us first note that the standard trick of the trade of 3σ("objects located at the distance of more than 3σ from the centre of the distribution of the main body of data are labelled as outliers") belongs to the distribution-based approaches. More elaborate techniques involve the use of more refined tools than just the distance of 3σ from the mean (or mode), even though they remain within the same distribution-based paradigm. Thus, the approach proposed by Filzmoser [2004] refers also to the normal distributions and the distance from the "centre". A Mahalanobis-distance-based χ^2 statistic is proposed for comparing tails of theoretical and empirical distance distributions. This statistic is then used to assess the "degree of outlierness" of an object. Similarly, Kuhnt and Pawlitschko [2005] refer to the same paradigm in developing the rules for identification of ("candidate") outliers in the cases of loglinear Poisson and logistic regression models.

Now, as we turn towards the density-based methods, it is perhaps most appropriate to refer to the seminal paper by Breunig et al. [2000]. They introduced the Local Outlier Factor (LOF), measuring the degree of "outlierness" of an object against the background of its neighbours. For a better assessment of when an object is actually an outlier Breunig et al. [2000] prove several properties of LOF values for geometrically different positions of objects in a data set.

The development of LOCI (a method for fast outlier detection using the Local Correlation Integral) by Papadimitriou et al. [2002] resulted from the criticism of the previous methods, including LOF. In place of LOF they introduce MDEF (multi-granularity deviation factor), supposed also to measure the degree of outlierness of an object against its neighbourhood. In distinction, however, from the LOF methodology, they treat MDEF of an object as a kind of statistic and compare it with the corresponding empirical standard deviation calculated for the set of all objects. At this point, Papadimitriou et al. [2000] use exactly the rule of 3σ.

2.5. Some doubts

The doubts are obvious, and can be summarised as follows:

- in case of the distribution based methods – how do we ascertain that our result holds for the entire range of theoretical distributions, which a given empirical distribution (sufficiently) matches? what level of "outlierness" makes an object a true outlier?
- in case of the density-based methods – what neighbourhoods (distances, numbers of nearest neighbours, proportions of the entire set) should we use for assessment? and, same as before: what level of "outlierness" makes an object a true outlier?

3. Can we bring more objectivity into the outlier definition?

We will now show some situations, in which one may bring into the identification of outliers some justified objectivity. In doing so, we will refer to the presence of a model, and to a more "concrete" understanding of the notion of an outlier.

3.1. Case 1 of justified "sufficiency of outlierness": the decision-analytic framework

The distribution-based methods are, as mentioned, actually using the classical Chebyshev inequality, i.e. $P(|X\text{-}m|\geq k\sigma) \leq 1/k^2$. Assuming any value of k may therefore be equivalent to an arbitrary definition of an outlier. Yet, there is a class of tasks, for which there exists an "objective" definition of an outlier. Namely, when we look for objects X that are "suspect", and we can "verify", whether the suspicion is justified. Possible cases are illustrated below:

Costs involved in particular cases	Real outlier	Non outlier
Classified for checking	c-v<0	c
Not classified for checking	V	0

c – cost of verification, v – loss from accepting an outlier as a "normal" object, c, v > 0

Without any further details, as this is not the proper subject of this paper: depending upon the values of c and v, and the parameters of the Chebyshev equation, we obtain the "outlier definition" in terms of k^* meaning that for $k\geq k^*$ an observation X ought to be positively "verified for outlierness". Here, k^* is the obtained threshold of "outlierness". Otherwise, we can only speak of different probabilities, not of identification of outliers (like, for instance, when $c>v$, when it is does not pay to check at all). (A more realistic model would involve $v = v(|X\text{-}m|)$, an increasing function, and then the trivial case mentioned is altogether avoided.)

3.2. The return to the intuitive meaning

Aside from the situation of the preceding section, where a model referred to had very precise (general) shape, it is assumed here that an outlier is an observation that cannot be classified as "(measurement) error" or "noise". Thus, say, if we subtract the "systematic" and "random" error from an observation, and it is still far away from the "general pattern", then we can speak of an outlier. Then, the observation can be called a "mistake" rather than an "error".

3.3. Case 2: the frequency argument

To strengthen the above meaning, there is a "*frequency*" or "*possibility*" argument for identification of outliers. The argument applies, anyway, also to other paradigms here quoted.

Assume we deal with a set of observations fitting (sufficiently) well a certain theoretical distribution function. The potential or actual population, to which this model (distribution) applies is composed (at most) of some $O(10^9)$ objects (individuals). Assume an observation in the area of the distribution function, for which the calculated probability of occurrence is in the order of $O(10^{-25})$. Apparently, such an observation is an outlier, a mistake. The frequency of appearance of such an observation is $O(10^{-16})$, which, if we speak of the time scale of, say, at least days or weeks, is well below any reasonable level (the average time between the occurrences being far above the age of the Universe).

This frequency reasoning can be illustrated by the following example. Weights, w [kg] and heights, h [m] of adult people are measured. Most measurements will concentrate around the BMI (body mass or Quetelet index) relation, according to some distribution. Conform to the US National Institute of Health's postulates, *normal* BMI values for the adults (> 20 years of age), BMI = $w/h^2 \in [18.5, 24.9]$. The intervals are also defined for "skinny", "overweight" ($\in [25, 29.9]$) and "obese" (≥ 30) BMI values. "Strange" observations, like [238, 1.57], far away from the "normal" interval (BMI = 96.56) may occur. Yet, they would not be regarded as "mistakes", at least not in terms of our "frequency" reasoning. Observations like [17, 1.98] (BMI = 4.33) or [238, 0.78] (BMI = 391.19) would, however, be classified "mistakes", either because the variables (weight↔height) and/or their units (m↔kg) were switched, or there was a grave error (e.g. one digit missing), or the measurements are not those of a human.

Thus, we have outlined two situations, in which outliers can be identified with a justified "objectivity", both referring to a "model". Indeed, if many methods of outlier identification fail, it is exactly because they cannot refer to any sort of model in the sense depicted here. In particular, Zakrzewska and Osada [2004] mention that the methods they used *determined successfully outliers*, but not the observations they actually looked for, that is – the ones associated with a fraud or other criminal activity. This is the effect of lack of a model.

3.4. More than one regular pattern and the significance of a model

All the above considerations, together with the examples, apply also to the situations, in which we deal with more than one "regular pattern" or "main population". In the statistical setting this would amount, in particular, to the mixture model, to which the "frequency" reasoning fully applies. At the same time, this is a straight case for cluster analysis, also in the vein of Zhuk [1996], who analyses the k-means, very sensitive to outliers.

Yet, in the latter case the question arises as to when a separate observation or observations are referred to as outliers or "other populations", "other clusters". This question, though, is valid only for groups of very low cardinality or marginal share in the data set. We can also ask it in an opposite manner: given that a singleton outlier is always classified as "outlier", what decision we take when it is clustered with another one? *Note, though, that the answer to this question may be of no importance, provided we have correctly separated the observations.*

Thus, while looking for candidate outliers, we in fact classify the data set, and it is beyond the knowledge contained in the data set itself that we might be able to label the groups obtained as the "main population", "other population", "outlier 1", "outlier 2" etc. An example of the basis for such labelling is that the (expected) variable values are inverted, which, though, can hardly be done automatically. Note that it is only a definite model, e.g. in the form of a probability density function, or of the rules that the object system obeys, that can tell us whether, like in the BMI example, the value 4.33 is a "mistake", while the value 96.56 is just "strange".

Looking from another angle, which is, anyway, very often the purpose of outlier detection, we can say that an "outlier is an object of which we *know* we should *exclude* it from the considerations". *Knowing* means here disposing of a model including the aspects mentioned before (e.g. probability distribution, verification cost and loss, frequency, time scale, distances etc.).

3.5. What if there is no model?

It is proposed to then refer to the basic clustering paradigm and look for the "robust" structures. If, namely, we are less interested in the overall structure of the data set, and more in the observations that "deviate" from this overall structure, then "robustness" with respect to various assumptions that may lead to differing structures may be the proper approach. There are two such obvious dimensions in clustering, against which a parameterisation may provide indication of robustness: (i) the definitions of distance, especially the Minkowski exponents, and (ii) the Lance-Williams-Jambu coefficients of the hierarchical merger procedures. If an observation remains a singleton under the entire range of parameterisation, the suspicion of outlierness gains a deeper foundation. A rough illustration is provided in the following figures, where two similar academic cases are compared for the single linkage and complete linkage algorithms, constituting in a way the extremes of the Lance-Williams-Jambu formula parameterisation (Figure 2).

It can easily be seen (in qualitative terms) that in Case B observation 3' might be labelled "outlier" in a definitely more justified manner than observation 3 in Case A, as a "robust" singleton. Such a procedure could be proposed also for the statistical approaches with checking against an appropriately broad class of distributions.

4. Some conclusions

It was indicated that there are classes of problems, for which the notion of outlier can be made more "objective" than usually to fate. In some such kinds of problems definite procedures can be proposed. In general, however, only the verification of robustness can be proposed.

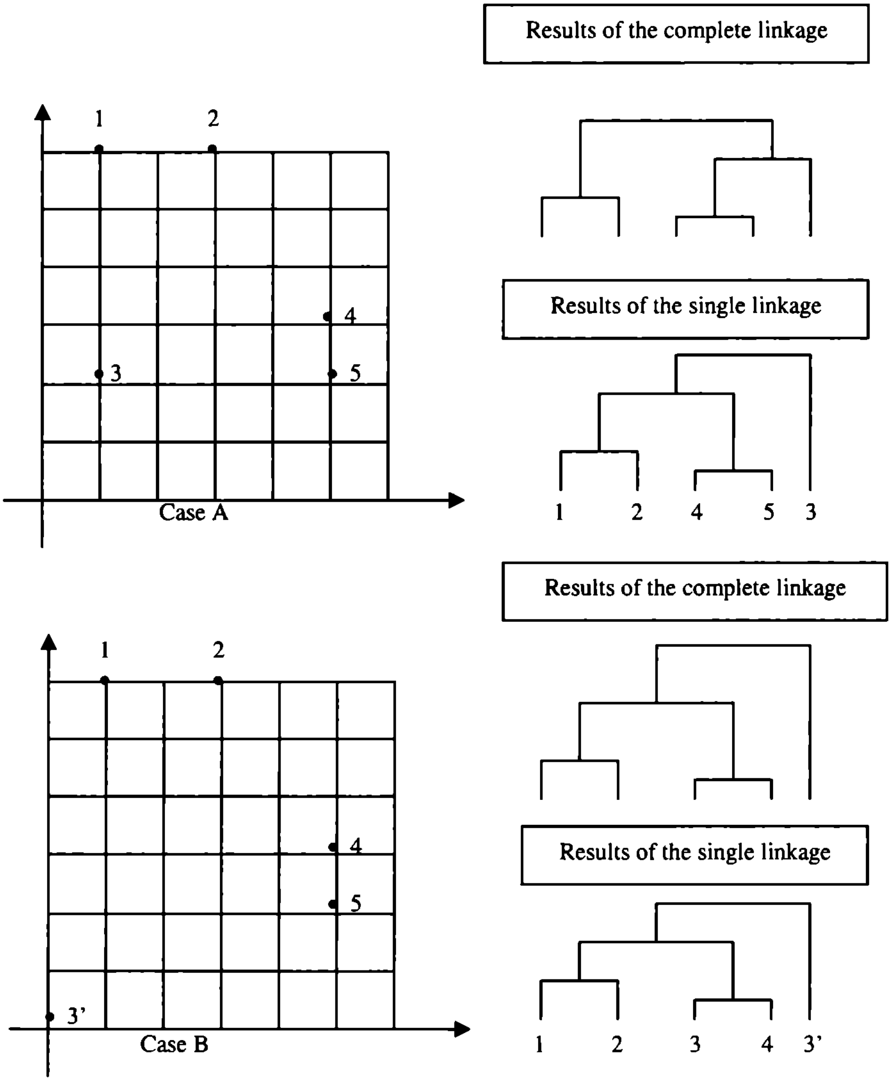

Results of the complete linkage
1
2
4
3
5
Case A
Results of the single linkage
1 2 4 5 3
Results of the complete linkage
1
2
4
5
3'
Case B
Results of the single linkage
1 2 3 4 3'

Figure 2.

This leaves aside the technical issue of computational complexity, which, however, is mainly associated with the selection of candidate outliers. Then, they can be checked against the yardsticks similar to those proposed here.

References

Barnett V., Lewis T. (1994), *Outliers in Statistical Data*, Wiley, New York.

Breunig M.M., Kriegel H.-P., Ng R.T., Sander J. (2000), *LOF: Identifying Density-Based Local Outliers*, Proc. ACM SIGMOD 2000: Int. Conf. On Management of Data, Dallas, TX.

Filzmoser P. (2004), *A Multivariate Outlier Detection Method*, [in:] S. Aivazian, P. Filzmoser, Yu. Kharin, (eds.), *Computer Data Analysis and Modeling: Robustness and Computer Intensive Methods*, Proc. of the 7th International Conference, Minsk, September 6-10. Minsk, p. 18-22.

Hawkins D. M. (1980), *Identification of Outliers*, Chapman and Hall.

Janeja V., Atluri V., Adam N. R. (2004), *OUTLAW: Using Geo-Spatial Associations for Outlier Detection and Visual Analysis of Cargo Routes*, www.rutgers.edu.

Kuhnt S., Pawlitschko J. (2005), *Outlier Identification Rules for Generalized Linear Models*, [in:] D. Baier, K.-D. Wernecke, (eds.), *Innovations in Classification, Data Science, and Information Systems*, Springer, Berlin-Heildelberg-New York, p. 165-172.

Mehra K., Stello R. (2005), *The Evolution of Spatial Outlier Detection Algorithms. An Analysis of Design*, typescript on www.cs.umn.edu.

Outlier Detection and +/-3σCharts (2004), http://www.autobox.com/outlier.html.

Papadimitriou S., Kitagawa H., Gibbons Ph.B., Faloutsos Ch. (2002), *LOCI: Fast Outlier Detection Using the Local Correlation Integral*, CMU-CS-02-188, November 2002, School of Computer Science, Carnegie Mellon University.

Suchecka J., Kowalik J. (2005), *Analiza danych czasowych z obserwacjami nietypowymi z wykorzystaniem metod geostatystyki*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1076, Taksonomia 12, p. 96-105.

Zakrzewska D., Osada M. (2004), *Outlier Analysis for Financial Fraud Detection*, typescript.

Zeliaś A. (1986), *Metody wykrywania obserwacji nietypowych w badaniach ekonomicznych*, „Wiadomości Statystyczne" nr 8, p. 16-27.

Zhuk E.E. (1995), *Robust L-Means Decision Rule Based on Observation-"Outlier" Rejection*, [in:] Yu. S. Kharin (ed.), *Computer Data Analysis and Modeling*, Proc. of International Conference, September 4-8, 1995, Minsk, p. 159-164.

WYKRYWANIE OBIEKTÓW NIETYPOWYCH: POJĘCIA, ZAGADNIENIA I PROPOZYCJE METODYCZNE

Streszczenie

Wykrywanie obiektów (obserwacji) nietypowych stało się ostatnio niezwykle popularne. Wynika to z bardzo szybkiego rozwoju dziedzin, w których istnieją ogromne zbiory danych i w związku z tym pojawia się możliwość wykrywania zachowań nieprawidłowych i kryminalnych. Artykuł pokazuje kilka reprezentatywnych przykładów podejść do wykrywania obserwacji nietypowych. Proponuje także pewien sposób rozumienia pojęcia obserwacji nietypowej. Na tle tego rozumie-

nia obserwacji nietypowej zarysowano dwie podstawowe sytuacje, a mianowicie wtedy, gdy dysponujemy pewnym „modelem" procesu oraz gdy mamy do czynienia ze wstępna analizą danych, nie dysponujemy modelem i nasza wiedza jest bardzo ograniczona. Pokazano, że w sytuacji, gdy dysponujemy pewnym modelem, możliwe jest bliskie obiektywnemu wykrywanie obserwacji nietypowych. Dla przypadku ogólnego zaproponowano badanie odporności obserwacji klasyfikowanych jako pojedyncze skupienia.