PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 327

Taksonomia 22
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga, Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Beata Mazur
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 327 • 2014
Taksonomia 22 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

Michał Trzęsiok
Uniwersytet Ekonomiczny w Katowicach

WYBRANE METODY IDENTYFIKACJI OBSERWACJI ODDALONYCH

Streszczenie: W artykule podjęto próbę uporządkowania definicji obserwacji oddalonych. Zestawiono kilka, znacząco różniących się podejściem do problemu, metod identyfikacji obserwacji oddalonych: jednowymiarową metodę kwartylową, krzywe Andrewsa, metodę bazującą na odległości Mahalonobisa, metodę uwzględniającą lokalne zagęszczenie obserwacji, metodę wektorów nośnych. Ponadto sprawdzono empirycznie, w jakim stopniu wybrane metody pokrywają się we wskazaniach obserwacji oddalonych. Część empiryczną przeprowadzono na zbiorze danych [*Diagnoza społeczna* 2011].

Słowa kluczowe: obserwacje oddalone, wielowymiarowa analiza statystyczna.

1. Wstęp

Niezależnie od tego, jak bardzo wyrafinowana metoda zostanie użyta do zbudowania modelu statystycznego, jakość tego modelu zależy wprost od jakości czy też nieco szerzej – od specyfiki danych wykorzystanych do jego wyznaczenia. W rzeczywistych zbiorach danych występują niekiedy pewne obserwacje nietypowe. Ponieważ obserwacje te mogą mieć istotny wpływ na wyniki analizy, wymagają szczególnej uwagi.

W artykule podjęto próbę uporządkowania definicji obserwacji oddalonych. Ponadto zestawiono kilka, znacząco różniących się podejściem do problemu, metod identyfikacji obserwacji oddalonych oraz zweryfikowano empirycznie przydatność wybranych metod na wymagającym i różnorodnym pod względem skal pomiaru zbiorze danych [*Diagnoza społeczna* 2011]. W szczególności na przykładzie tym sprawdzono, w jakim stopniu tak różnorodne metody pokrywają się w kwestii wskazań obserwacji oddalonych.

2. Definicje oraz charakterystyka wartości oddalonych

Pojęcie obserwacji oddalonej nie jest w literaturze zdefiniowane jednoznacznie. W niniejszej pracy posłużono się dosyć ogólną definicją zaczerpniętą z pracy Hawkinsa [Hawkins 1980], który przez obserwację oddaloną rozumie taką obserwację,

która odchyla się tak bardzo od innych obserwacji, że rodzi to przypuszczenie, że powstała w wyniku działania innego mechanizmu, tj. że pochodzi z innego rozkładu niż pozostałe obserwacje w zbiorze danych. Należy tu podkreślić, że jeśli owo przypuszczenie zawarte w definicji okazałoby się słuszne względem obserwacji z analizowanego zbioru danych, to oznaczałoby to brak spełnienia jednego z najważniejszych założeń metod wielowymiarowej analizy statystycznej. W grupie tych metod przyjmuje się wprawdzie różne założenia, lecz na ogół minimalny poziom założeń dotyczących zbioru danych to i.i.d. (*independent and identically distributed*), czyli założenie, że zbiór danych tworzą obserwacje wylosowane w sposób niezależny, o jednakowym, wielowymiarowym rozkładzie określonym przez (nieznaną, ale wspólną) funkcję gęstości.

W literaturze spotkać można wiele innych definicji obserwacji oddalonych. Często są to definicje odnoszące się do pojęcia obserwacji oddalonej przez pewien szczególny kontekst. Wyróżnić tu można trzy rodzaje obserwacji oddalonych [Rousseeuw, Leroy 2003]:

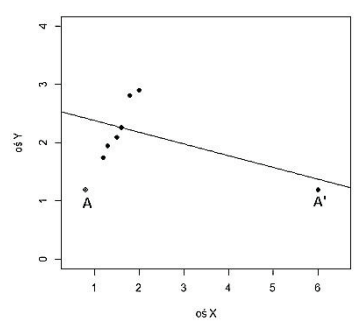
1) obserwacje *nietypowe* (lub *odstające*, ang. *outliers*), w których wyróżniona jest zmienna objaśniana Y i właśnie wartość tej zmiennej znacząco odchyla się od wartości dla innych obserwacji;

2) obserwacje *wysokiej dźwigni* (lub *dźwigniowe*, ang. *leverage*), w których wartość przynajmniej jednej ze zmiennych objaśniających (**X**) znacząco odchyla się od wartości tej zmiennej dla innych obserwacji (por. rys. 1);

3) obserwacje *wpływowe* (ang. *influential observations*), których wyłączenie ze zbioru danych powoduje istotną zmianę modelu (por. rys. 1).

Przedstawiona klasyfikacja nie jest rozłączna, np. obserwacja może być jednocześnie odstająca i wpływowa bądź odstająca i dźwigniowa.

Na rys. 1 widać konsekwencje wprowadzenia do zbioru danych obserwacji A', której poprawne położenie oznaczono literą A, lecz wartość zmiennej objaśniającej została błędnie wprowadzona. Wystąpienie A' bardzo istotnie wpłynęło na model regresji liniowej (A' jest obserwacją wpływową oraz dźwigniową).



Rys. 1. Ilustracja konsekwencji wystąpienia w zbiorze danych obserwacji o wysokiej dźwigni, będącej jednocześnie obserwacją wpływową

Źródło: opracowanie własne na podstawie [Rousseeuw, Leroy 2003, rys. 2, s. 5] .

Zagadnienie identyfikacji obserwacji oddalonych zawiera w sobie kilka poważnych trudności. Po pierwsze, nie zawsze występowanie obserwacji oddalonych jest zjawiskiem negatywnym. Owszem, niekiedy są one rezultatem błędów pomiaru zmiennych, jednak czasem są wynikiem poprawnych pomiarów i obrazują prawdziwe, choć rzadkie i nietypowe zachowanie badanego zjawiska. W tym drugim przypadku zdecydowanie nie należy usuwać tych obserwacji, gdyż na ogół ich zawartość informacyjna jest bardzo duża [Webb 2002]. W obu przypadkach ważne jest, by zidentyfikować obserwacje oddalone i w odpowiedni sposób je potraktować. Po drugie, wiele klasycznych metod identyfikacji obserwacji nietypowych nie potrafi wykrywać mnogich wartości oddalonych (efekt wzajemnego maskowania się dwóch lub więcej obserwacji oddalonych leżących blisko siebie) [por. Huber, Ronchetti 2009]. Po trzecie, niektóre metody skupione są na identyfikowaniu obserwacji oddalonych, wykorzystując tylko jedną z wielu możliwych konsekwencji ich występowania, np. badając reszty modelu. Tymczasem nie zawsze duża reszta modelu dla danej obserwacji oznacza, że jest to obserwacja oddalona [Maddala 2006], gdyż model może być źle dopasowany na przykład z powodu wystąpienia obserwacji oddalonych.

Podkreślić należy, że celem stosowania metod identyfikacji obserwacji oddalonych nie jest późniejsze usunięcie tych obserwacji (chyba że przyczyną ich powstania były błędy). Badania empiryczne wskazują, że na ogół znacznie lepsze wyniki niż usuwanie obserwacji oddalonych dają metody odporne (*robust methods*) [Huber, Ronchetti 2009].

3. Krótki opis wybranych metod identyfikacji obserwacji oddalonych

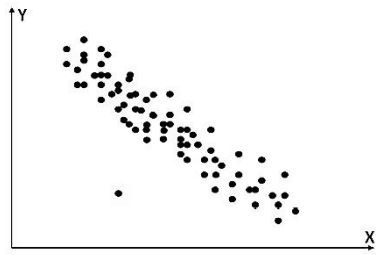
3.1. Metody jednowymiarowe – kryterium kwartylowe

Niech $\mathbf{X} = (X_1, \ldots, X_k)$ będzie wektorem zmiennych objaśniających w n elementowym zbiorze danych. Najprostsze i najstarsze metody identyfikowania obserwacji oddalonych to metody jednowymiarowe, na ogół połączone z prezentacją graficzną wartości zmiennej. Do takich metod zaliczyć należy kryterium kwartylowe wykorzystywane w budowie wykresów pudełkowych wprowadzonych przez Tukeya [Tukey 1977]. Wartość pojedynczej zmiennej jest uznana za oddaloną, jeśli znajduje się poza przedziałem:

$$\langle Q_1 - 1{,}5 \cdot IQR, \quad Q_3 + 1{,}5 \cdot IQR \rangle, \tag{1}$$

gdzie Q_1, Q_3 to odpowiednio pierwszy i trzeci kwartyl, a IQR to odchylenie ćwiartkowe. Niektórzy autorzy przyjmują nawet dopełnienie przedziału danego wzorem (1) jako definicję obserwacji oddalonej [por. Giudici 2003, s. 42]. Wykresy pudełkowe są bardzo cennym narzędziem do wstępnego zapoznania się z anali-

zowanym zbiorem danych, lecz jednowymiarowe podejście do zagadnienia identyfikacji obserwacji oddalonych jest niewystarczające. Na rysunku 2 przedstawiono prosty dwuwymiarowy przykład występowania obserwacji oddalonej, która ze względu zarówno na zmienną objaśniającą, jak i objaśnianą nie odbiega znacząco od mediany. Kryterium kwartylowe nie jest skutecznym narzędziem identyfikowania obserwacji oddalonych w przypadku danych wielowymiarowych.



Rys. 2. Przykład zbioru z jedną obserwacją oddaloną, której nie można zidentyfikować jednowymiarowymi metodami kwartylowymi

Źródło: opracowanie własne na podstawie [Rousseeuw, Leroy 2003, rys. 4, s. 7].

3.2. Graficzna metoda wielowymiarowa – krzywe Andrewsa

Do identyfikacji wielowymiarowych obserwacji oddalonych można wykorzystać metody redukcji wymiaru, np. metodę Andrewsa, która każdą obserwację sprowadza do pewnej krzywej na płaszczyźnie [Andrews 1972]. Andrews zaproponował kilka typów przekształceń wielowymiarowych obserwacji do krzywych. W niniejszej pracy wykorzystano przekształcenie:

$$f(t) = x_1 \cdot \sin(t) + x_2 \cdot \cos(t) + x_3 \cdot \sin(2 \cdot t) + x_4 \cdot \cos(2 \cdot t) + \ldots \tag{2}$$

Metoda Andrewsa wykorzystuje ideę rozwinięcia funkcji w szereg Fouriera i choć jest elegancka w swojej matematycznej warstwie, to jednak ma ograniczone zastosowanie w przypadku zbiorów danych o dużej liczebności, gdyż otrzymywany rysunek jest nieczytelny (zbyt wiele nakładających się krzywych).

3.3. Metody bazujące na odległości Mahalanobisa

Szczególnie w ekonometrii stosuje się metody identyfikacji obserwacji oddalonych, wykorzystujące kryterium bazujące na odległości Mahalonobisa [Healy 1968]:

$$MD^2(\mathbf{x}) = (\mathbf{x} - \hat{\boldsymbol{\mu}})\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}), \tag{3}$$

gdzie $\hat{\boldsymbol{\mu}} = \overline{\mathbf{x}}$ jest wartością przeciętną, a $\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^{\mathrm{T}}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$ – macierzą wariancji i kowariancji. Punkty o dużych (w porównaniu z wartościami krytycz-

nymi rozkładu χ^2) wartościach kwadratu odległości Mahalanobisa traktowane są jako obserwacje oddalone. To podejście ma jednak tę podstawową wadę, że wartość samego kryterium (3) w bezpośredni sposób zależy od statystyk (klasycznych), które są bardzo wrażliwe na występowanie wartości oddalonych. W celu wyeliminowania tej wady zaproponowano modyfikacje obliczania wartości miernika (3) poprzez zastąpienie średniej $\hat{\boldsymbol{\mu}}$ przez odporny parametr położenia. Jedna z propozycji to wykorzystanie estymatora *MVE* (*Minimum Volume Ellipsoid Estimator*), tj. estymatora o minimalnej objętości elipsoidy [Rousseeuw 1984]:

$$\hat{\boldsymbol{\mu}} = \text{środek ciężkości elipsoidy o minimalnej objętości zawierającej co najmniej } h \text{ obserwacji danego zbioru,} \tag{4}$$

gdzie $h = \lfloor n/2 \rfloor + 1$. Druga z propozycji [także w: Rousseeuw 1984] to wyznaczenie parametru położenia $\hat{\boldsymbol{\mu}}$ we wzorze (3) wg formuły:

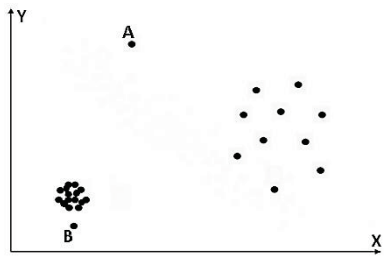
$$\hat{\boldsymbol{\mu}} = \text{średnia z tych } h \text{ obserwacji danego zbioru, dla których wyznacznik macierzy kowariancji jest najmniejszy.} \tag{5}$$

Odporny estymator położenia (5) jest nazywany estymatorem *MCD* (*Minimum Covariance Determinant Estimator*), tj. estymatorem o minimalnym wyznaczniku macierzy kowariancji. Trzecie podejście, zasugerowane w pracy [Filzmoser i in. 2008], wykorzystuje analizę głównych składowych i identyfikuje obserwacje oddalone właśnie po przekształceniu wszystkich obserwacji w przestrzeni głównych składowych przez wyznaczenie w tej przestrzeni wartości kwadratu odległości Mahalanobisa. Autorzy tego podejścia sugerują zastosowanie na etapie przygotowania danych do analizy, standaryzacji zmiennych z wykorzystaniem mediany jako parametru położenia oraz MAD, czyli medianowego odchylenia bezwzględnego, jako parametru rozproszenia. Po zastosowaniu takiej standaryzacji obliczanie odległości euklidesowej w przestrzeni głównych składowych jest równoważne z obliczeniem odpornego wariantu odległości Mahalanobisa.

3.4. Metoda uwzględniająca lokalne zagęszczenie obserwacji

Przedstawione w punktach 3.1–3.3 metody identyfikacji obserwacji oddalonych traktują to zagadnienie zero-jedynkowo, czyli albo obserwacja jest oddalona, albo nie. Odmienne podejście do tego problemu prezentują Breunig, Kriegel, Ng i Sander [Breunig i in. 2000], którzy proponują miernik wskazujący stopień oddalenia danego obiektu od pozostałych obserwacji ze zbioru danych. Miernik ten nazywają *LOF* (*Local Outlier Factor*) – lokalnym miernikiem stopnia oddalenia obserwacji. Definicja tego miernika ma złożoną postać analityczną oraz zagnieżdżoną strukturę i wymaga zdefiniowania trzech innych pojęć. W tym miejscu podana zostanie jedynie główna idea jego konstrukcji. Miernik *LOF* jest zainspirowany metodą *k*

najbliższych sąsiadów i wskazuje stopień oddalenia danej obserwacji od pozostałych, uwzględniając zagęszczenie obiektów z *k*-elementowego sąsiedztwa. Takie podejście pozwala identyfikować obserwacje oddalone również w przypadku, gdy zbiór danych tworzą skupienia o różnym stopniu zagęszczenia, czyli różnym poziomie koncentracji wokół środka ciężkości (por. rys. 3). Na rys. 3 przedstawiono przykład zbioru, w którym są dwie klasy o różnym stopniu zagęszczenia oraz dwie obserwacje oddalone (oznaczone: A i B). Większość metod zidentyfikuje poprawnie obserwację A jako oddaloną. Zidentyfikowanie obserwacji B jako oddalonej wymaga uwzględnienia stopnia lokalnego zagęszczenia obiektów.



Rys. 3. Przykład zbioru, w którym są dwie klasy o różnym stopniu zagęszczenia oraz dwie obserwacje oddalone (oznaczone: A i B)

Źródło: opracowanie własne.

To, czy pewna odległość punktu od pozostałych jest wystarczająco duża, by uznać punkt za oddalony, jest wszak zależne od stopnia zróżnicowania odległości punktów w danym fragmencie przestrzeni.

3.5. Metoda wyznaczania uogólnionego wielowymiarowego kwantyla rozkładu

Jeden z wariantów metody wektorów nośnych *SVM* (*Supprot Vector Machines*) pozwala na wyznaczenie uogólnionego wielowymiarowego kwantyla rozkładu generującego dane z analizowanego zbioru. Przez uogólniony kwantyl rozkładu rozumieć należy taki obszar $Q \subset \mathbf{R}^k$ przestrzeni danych, który spełnia warunek, że niemal wszystkie obserwacje wygenerowane z rozkładu należą do Q, z drugiej strony obiekty niepochodzące z tego rozkładu, należą do dopełnienia zbioru Q. Wykorzystując funkcje jądrowe, określające pewne nieliniowe przekształcenie przestrzeni danych, poszukiwanie rozwiązania problemu przeniesione zostaje w przestrzeń $\mathbf{Z}$ o znacznie większym wymiarze i w tej nowej przestrzeni wyznaczana jest optymalna hiperkula (o najmniejszym możliwym promieniu, tzw. hiperkula Czebyszewa), zawierająca obrazy obserwacji ze zbioru danych. Tej hiperkuli w przestrzeni $\mathbf{Z}$ odpowiada (jako przeciwobraz) pewien zbiór w pierwotnej przestrzeni danych. Jest nim poszukiwany uogólniony kwantyl Q. Ze względu na uela-

stycznienie metody na występowanie potencjalnych błędów pomiaru, wyznaczona hiperkula nie musi zawierać obrazów wszystkich obserwacji ze zbioru danych. Obiekty, które znalazły się poza tą hiperkulą można łatwo zidentyfikować. Są to obserwacje, które znajdują się poza uogólnionym kwantylem rozkładu i potencjalnie pochodzą z innego rozkładu, czyli obserwacje oddalone. Szczegóły metody znaleźć można w pracy [Ben-Hur i in. 2001].

3.6. Inne możliwe podejścia do identyfikacji obserwacji oddalonych

Do identyfikacji obserwacji oddalonych można również posłużyć się metodami taksonomicznymi, licząc, że w wyniku grupowania zostaną one wyodrębnione tworząc jednoelementowe klasy. Takie podejście jest jednak krytykowane [Breunig i in. 2000], gdyż celem metod taksonomicznych jest wyznaczenie skupień i temu podporządkowany jest ich mechanizm (optymalizacyjny), a nie rozpoznaniu obserwacji oddalonych.

W literaturze przedmiotu znaleźć można bardzo wiele propozycji testów statystycznych do weryfikacji hipotezy, czy dana obserwacja jest obserwacją oddaloną. Obszerny zestaw takich testów zawiera praca Barnetta i Lewisa [Barnett, Lewis 1998].

Inne podejście wykorzystuje pojęcie głębi [Tukey 1977], lecz w praktyce metoda ta okazuje się niewydajna dla danych *k*-wymiarowych dla $k \geq 4$, gdyż wiąże się z wyznaczaniem otoczek wypukłych, co jest wymagające obliczeniowo [Breunig i in. 2000].

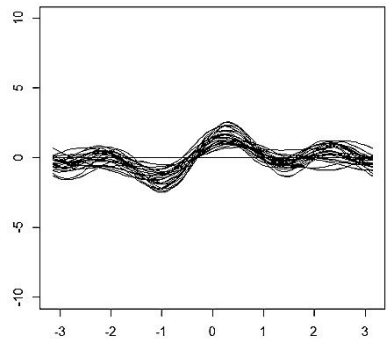
4. Empiryczne porównanie wyników działania wybranych metod

Analiza empiryczna została przeprowadzona na zbiorze danych [*Diagnoza społeczna* 2011]. Do analizy wykorzystano tylko wybrane zmienne mierzone na mocnych skalach: wiek respondenta, liczba osób w gospodarstwie domowym, liczba lat nauki, liczba osób uznawanych za przyjaciół, dochód netto i staż pracy.

Zbiór danych [*Diagnoza społeczna* 2011] poddano najpierw filtrowaniu, usuwając obserwacje, w których dla wybranych sześciu zmiennych występowały braki wartości. Liczebność zbioru poddanego dalszej analizie to 1070. Wszystkie obliczenia przeprowadzone zostały z wykorzystaniem programu statystycznego **R**.

Podjęto próbę identyfikacji obserwacji oddalonych metodą krzywych Andrewsa, ale umieszczenie 1070 krzywych na jednym wykresie daje nieczytelny obraz. W celu zilustrowania metody na rys. 4 przedstawiono 30 krzywych Andrewsa dla pierwszych 30 obserwacji.

W dalszej części przeprowadzono identyfikację obserwacji oddalonych czterema metodami: jednowymiarową metodą kwartylową, metodą *MD** bazującą na odległości Mahalanobisa z poprawkami zaproponowanymi przez Filzmosera i in.,



Rys. 4. Krzywe Andrewsa dla pierwszych 30 obserwacji ze zbioru [*Diagnoza społeczna* 2011]

Źródło: opracowanie własne.

metodą *LOF*, uwzględniającą lokalne zagęszczenie obserwacji, oraz metodą wektorów nośnych *SVM*. Liczba obserwacji oddalonych zidentyfikowanych przez każdą z metod nie pozwala na ich prezentację tabelaryczną, ale fragment wyników dla metody MD^* przedstawiono na rys. 5.

L.p.	wiek	l. osób	lata nauki	l. przyjaciół	dochód netto	staż pracy
62	80	3	4	15	0	0
77	62	1	13	30	1 000	30
78	62	2	10	20	600	35
95	56	2	17	14	2 800	28
128	29	3	15	30	0	0
243	49	4	13	20	1 000	25
283	58	2	13	1	6 000	30
314	84	2	16	21	840	24
328	66	2	17	50	2 500	35
1052	54	4	13	20	1 500	27
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Rys. 5. Fragment wyników identyfikacji obserwacji oddalonych dla metody bazującej na odległości Mahalanobisa MD^* wraz z wykresami pudełkowymi dla każdej ze zmiennych (wynikami zastosowania metody kwartylowej)

Źródło: opracowanie własne.

Do rys. 5 dołączono wykresy pudełkowe dla każdej zmiennej diagnostycznej. W tabeli 1 przedstawiono zgodność klasyfikacji trzech zastosowanych metod parami, tj. liczbę obserwacji, które zgodnie zostały zidentyfikowane przez dwie metody jako oddalone.

Elementy występujące poza przekątną w tabeli 1 wskazują, że wyniki działania przedstawionych metod znacząco różnią się – zbiory obserwacji oddalonych

Tabela 1. Zgodność klasyfikacji zastosowanych metod parami, tj. liczba obserwacji, które zgodnie zostały zidentyfikowane przez dwie metody jako oddalone

Metoda	*MD**	*LOF*	*SVM*
*MD**	70	12	25
LOF		51	12
SVM			54

Źródło: opracowanie własne.

w niewielkim stopniu się pokrywają. Nadmienić należy, że liczba zidentyfikowanych obserwacji oddalonych jest zależna od parametrów metody, które ustalane były symulacyjnie. Oznacza to, że przedstawione wyniki są tylko jednym z możliwych wariantów. Brak jednoznaczności wyników oraz rozbieżności między metodami wynikają z natury zagadnień klasyfikacji bezwzorcowej. W celu zredukowania subiektywizmu w doborze wartości parametrów wykorzystanych metod można zbudować wiele modeli dla różnych kombinacji parametrów i na przykład zastosować regułę majoryzacyjną.

5. Podsumowanie

Zaprezentowano wybrane metody identyfikacji obserwacji oddalonych. Te metody na różne sposoby realizują cel identyfikacji obserwacji oddalonych, co przekłada się również na odmienne rezultaty ich działania (zbiory zidentyfikowanych obserwacji oddalonych dla różnych metod w niewielkim stopniu się pokrywają). Nie oznacza to jednak, że niektóre metody są gorsze, tylko że metody te można traktować jako komplementarne.

Problem identyfikacji obserwacji oddalonych ma być jedynie narzędziem wstępnej poprawy jakości danych – zwróceniem uwagi na występujące w zbiorze anomalie. Wszystkie przedstawione metody spełniają ten postulat, choć każda w nieco inny sposób.

Literatura

Andrews D.F. (1972), *Plots of High-Dimensional Data*, Biometrics, vol. 28, 1, s. 125-136.

Barnett V., Lewis T. (1998), *Outliers in Statistical Data*, 3rd Edition, John Wiley & Sons, New York.

Ben-Hur A., Horn D., Siegelman H. T., Vapnik V. (2001), *Support Vector Clustering*, Journal of Machine Learning Research, 2, s. 125-137.

Breunig M.M., Kriegel H.-P., Ng R.T., Sander J. (2000), *LOF: Identifying Density-Based Outliers*, Proc. 29th ACM SIDMOD Int. Conf. on Management of Data, s. 93-104.

Diagnoza społeczna (2011), *Diagnoza społeczna: zintegrowana baza danych*, Rada Monitoringu Społecznego, www.diagnoza.com (24.02.2013).

Filzmoser P., Maronna R.A., Werner M. (2008), *Outlier Identification in High Dimensions*, Computational Statistics & Data Analysis, 52, s. 1694-1711.

Giudici P. (2003), *Applied Data Mining: Statistical Methods for Business and Industry*, John Wiley & Sons, Southern Gate, Chichester.
Hawkins D. (1980), *Identification of Outliers*, Chapman and Hall, London – New York.
Healy M.J.R. (1968), *Multivariate Normal Plotting*, Applied Statistics, 17, s. 157-161.
Huber P.J., Ronchetti E.M. (2009), *Robust Statistics*, Second Edition, John Wiley & Sons, Hoboken, New Jersey.
Maddala G.S. (2006), *Ekonometria*, Wydawnictwo Naukowe PWN, Warszawa.
Rousseeuw P.J. (1984), *Least Median of Squares Regression*, Journal of the American Statistical Association, 79, s. 871–880.
Rousseeuw P.J., Leroy A.M. (2003), *Robust Regression and Outlier Detection*, John Wiley & Sons, New York.
Tukey J.W. (1977), *Exploratory Data Analysis*, Addison-Wesley, Southern Gate, Chichester.
Webb A.R. (2002), *Statistical Pattern Recognition*, Second Edition, John Wiley & Sons, Chichester.

SELECTED METHODS FOR OUTLIER DETECTION

Summary: In the paper we try to set in order different definitions of outliers. We also collate a few selected outlier detection techniques, which represent very different approaches to outliers identification: classical univariate method embodied in boxplots, Andrews' curves, methods based on Mahalonobis distance, local outlier factor method, support vector machines. Moreover, we empirically examine the agreement between the results of outlier detection methods on the demanding dataset *Social Diagnosis 2011*.

Keywords: outlier detection, multivariate statistical analysis.