METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
X, 2009, str. 185 - 193

NORMOWANIE ZMIENNYCH OPISUJĄCYCH OBIEKTY NIETYPOWE

Kesra Nermend
Instytut Informatyki w Zarządzaniu, US
e-mail: kesra@wneiz.pl

Streszczenie: W wielu badaniach problem stanowią obiekty nietypowe, których cechy opisywane są bardzo dużymi wartościami. Mogą one wpływać w sposób znaczący na wyniki badań z powodu zmniejszania zakresu zmiennych dotyczących obiektów typowych podczas normowania. Niekorzystny wpływ wartości nietypowych można zminimalizować przez wykorzystanie pewnych metod normowania. W artykule zostały przedstawione dwie tego typu metody: standaryzacja z ważonym odchyleniem standardowym oraz unitaryzacja z wartościami progowymi.

Słowa kluczowe: obiekty nietypowe, metody normowania.

WSTĘP

Zmienne biorące udział w tworzeniu miar syntetycznych, czy też wykorzystywane w grupowaniu, często są wyrażone w różnych jednostkach miary, na przykład w osobach na kilometr kwadratowy, czy też złotówkach. Zmienne takie są nieporównywalne. Zmienne mogą mieć też jednakowe jednostki miary, ale być nieporównywalne ze względu na różny zakres wartości. Przykładem może być średni miesięczny dochód na osobę i średnia kwota wydawana miesięcznie na kulturę (kino, teatr itp.). Zmienne te są porównywalne, jednak ze względu na większe wartości średniego miesięcznego dochodu na osobę ta zmienna będzie dominować. Jej ważność będzie znacząco większa niż ważność średniej kwoty wydawanej na kulturę, w związku z tym druga zmienna nieznacznie tylko wpłynie na wartość miary syntetycznej, czy też wyniki grupowania [Kukuła 2000].

W ogólnej formie formułę normowania cech można przedstawić następująco [Grabiński i in. 1989, Kolenda 2006]:

$$x'_{ij} = \left(\frac{x_{ij} - A_i}{B_i} \right)^p \tag{1}$$

gdzie: x_{ij} – wartość i-tej zmiennej dla j-tego obiektu przed normalizacją, x'_{ij} – wartość i-tej zmiennej dla j-tego obiektu po normalizacji, B_i – podstawa normalizacyjna i-tej zmiennej ($B_i \neq 0$), A_i, p – parametry.

Pojawienie się w badanym zbiorze obiektów o nietypowo dużych wartościach zmiennych jest poważnym problemem. Obiekty te powodują podczas standaryzacji „zbicie" pozostałych wartości w pewnym niewielkim zakresie wartości. W przypadku unitaryzacji „zbicie" to jest jeszcze dużo silniejsze. Efekt ten powoduje, zarówno podczas grupowania jak i tworzenia miar syntetycznych, słabą rozróżnialność obiektów ze względu na zmienną, w ramach której występują takie obiekty nietypowe. W skrajnym przypadku może dojść do sytuacji, w której dla tej zmiennej możliwe będzie jedynie rozróżnienie obiektów nietypowych od pozostałych, bez możliwości różnicowania obiektów typowych.

STANDARYZACJA ZMIENNYCH DIAGNOSTYCZNYCH

Po standaryzacji obiekty nietypowe uzyskują bardzo dużą wartość tej zmiennej, dla której posiadają nietypową wartość. W konsekwencji, o wartości tej zmiennej decydować będzie głównie wartość miary syntetycznej dla danego obiektu.

Problem obiektów nietypowych w przypadku standaryzacji można nieco złagodzić stosując zamiast odchylenia standardowego ważone odchylenie standardowe [Kozak i in. 2007]:

$$B_i = \sqrt{\frac{\sum_{j=1}^{N} \left(x_{ij} w_j - \frac{\sum_{k=1}^{N} x_{ik} w_k}{\sum_{k=1}^{N} w_k} \right)^2}{\sum_{j=1}^{N} w_j - 1}} \tag{2}$$

gdzie: w_j, w_k – wartość j-tej (k-tej) wagi.

Odpowiednie zastosowanie wag umożliwia zmniejszenie wpływu obserwacji odstających na wartość odchylenia. Obserwacjom dalekim od średniej nadawane są

wagi mniejsze niż obserwacjom leżącym blisko wartości średniej. W najprostszym przypadku można zastosować system wag zero-jedynkowych:

$$w_j = \begin{cases} 1 \quad dla \quad \left| x_i - \bar{x}_i \right| \le p \\ 0 \quad dla \quad \left| x_i - \bar{x}_i \right| > p \end{cases} \tag{3}$$

gdzie: $\bar{x}_i$ – wartość średnia i-tej zmiennej, p – ustalona wartość progowa (próg).

Ten system wag nadaje wagom wartość jeden jeżeli bezwzględna różnica pomiędzy obserwacją a wartością średnią jest mniejsza lub równa od zadanego progu, a wartość zero jeżeli jest większa.

Obiekty nietypowe występują w populacji dość rzadko - dwa albo trzy razy w zbiorze stuelementowym zawierającym zarówno obiekty typowe, jak i nietypowe. Wartość progu p można zatem uzależnić od przewidywanego prawdopodobieństwa wystąpienia obiektów nietypowych. Można na przykład przyjąć, że prawdopodobieństwo wystąpienia obiektu nietypowego jest mniejsze niż 5%. Wówczas, jeżeli założy się, że rozkład wartości zmiennej jest rozkładem normalnym, to wartość progu p można przyjąć jako równą w przybliżeniu dwóm odchyleniom standardowym 2σ. W przypadku ogólnym, kiedy rozkład jest nieznany, ale istnieje dla niego odchylenie standardowe, aby uzyskać podobne prawdopodobieństwo należałoby przyjąć w przybliżeniu $p = 5\sigma$. Należy jednak pamiętać, że zadane prawdopodobieństwo wystąpienia wartości nietypowej jest prawdopodobieństwem największym z możliwych. Jeżeli nieznany rozkład okazałby się rozkładem normalnym, to faktyczne prawdopodobieństwo wyniosłoby tylko około 0,0000006%. Ze względu na to, że sumowanie rozkładów „przybliża" sumowane rozkłady do rozkładu normalnego, większość rozkładów wartości zmiennych jest zbliżonych do rozkładu normalnego, a więc wartość progu z niewielkim błędem można szacować na podstawie założenia o normalności rozkładu. Wartość progu p można, zatem przyjąć jako 2σ.

UNITARYZACJA ZMIENNYCH DIAGNOSTYCZNYCH

Istnieją również metody zmniejszania wpływu obiektów nietypowych na unitaryzację. Przykładem tego typu metody jest metoda korekcji jasności i kontrastu stosowana w niektórych urządzeniach do automatycznego wykonywania odbitek lub też w niektórych aparatach cyfrowych. Przy przetwarzaniu zdjęć występuje podobny problem jak przy normowaniu zmiennych. Przy zapisie zdjęć jest pewien zakres wartości jasności (lub składowych kolorów) najczęściej $\langle 0;255 \rangle$. Jednak zdjęcie zaraz po zarejestrowaniu ma dużo większy zakres wartości, który musi być

zmniejszony do docelowego. W najprostszym przypadku można zastosować unitaryzację zerowaną (w przetwarzaniu obrazów nazywaną normalizacją), której wartości przemnaża się przez docelową wartość maksymalną, czyli najczęściej 255. Jednak pojawia się tu problem wartości nietypowych, które ze względu na dużą liczbę pikseli obrazu pojawiają się zawsze. Powoduje to, że większość wartości skumulowana jest w pewnej niewielkiej części zakresu docelowego. W konsekwencji zdjęcie wydaje się mało kontrastowe. Aby zwiększyć kontrast zdjęcia, w normalizacji używa się nie wartości skrajnych, a dwie wartości progowe, lewą i prawą, ustalane na podstawie specjalnego algorytmu, najczęściej opartego o analizę histogramu.

Podobne rozwiązanie można przyjąć do normowania zmiennych. Jako podstawę normalizacyjną można przyjąć wartości progowe:

$$B_i = x_{L\,i} - x_{P\,i} \tag{4}$$

gdzie: $x_{L\,i}$ – lewa wartość progowa i-tej zmiennej, $x_{P\,i}$ – prawa wartość progowa i-tej zmiennej.

Jest to pewna odmiana podstawy normalizacyjnej [Kukuła 2000, Nowak 1990]:

$$B_i = \max_j\left(x_{\substack{i\\j}}\right) - \min_j\left(x_{\substack{i\\j}}\right) \tag{5}$$

Wartości progowe podobnie jak dla zdjęć można wyznaczyć na podstawie histogramu częstości. Przy czym, w przypadku zmiennych konieczne jest wykorzystanie histogramu względnych częstości liczonego dla przedziałów wartości [Amir 2000]. Histogram taki liczy się dla zadanej z góry liczby przedziałów, bądź zadanej z góry szerokości przedziałów. Przy czym ten pierwszy przypadek jest o tyle wygodniejszy, że można określić, jaka mniej więcej powinna być minimalna liczba przedziałów. Najlepiej gdyby miała ona taką wartość, aby na jeden przedział nie wypadało średnio mniej niż dziesięć obiektów:

$$N_p \le \frac{N}{10} \tag{6}$$

gdzie: N – liczba obiektów, N_p – liczba przedziałów.

W pierwszym etapie liczenia histogramu częstości wylicza się zakres wartości odejmując od wartości maksymalnej wartość minimalną:

$$zakr = \max_j\left(x_{\substack{i\\j}}\right) - \min_j\left(x_{\substack{i\\j}}\right) \tag{7}$$

Zakres wartości jest podstawą do wyliczenia szerokości przedziałów:

$$szer = \frac{zakr}{N_p} \tag{8}$$

Szerokość przedziałów umożliwia określenie granic poszczególnych przedziałów:

$$\begin{cases} g_{Lk} = szer(k-1) \\ g_{Pk} = szer\ k \end{cases} \tag{9}$$

gdzie: g_{Lk} – lewa granica k-tego przedziału, g_{Pk} – prawa granica k-tego przedziału.

Wyliczone w ten sposób granice definiują przedziały, przy czym jeden przedział jest zawsze domknięty dwustronnie, a pozostałe mogą być domknięte lewostronnie:

$$\langle g_{L1}, g_{P1}), \langle g_{L2}, g_{P2}), \ldots, \langle g_{LN_p}, g_{PN_p} \rangle \tag{10}$$

lub prawostronnie:

$$\langle g_{L1}, g_{P1} \rangle, (g_{L2}, g_{P2} \rangle, \ldots, (g_{LN_p}, g_{PN_p} \rangle \tag{11}$$

Wybór sposobu domykania przedziałów, gdy liczba obiektów jest duża, nie ma znaczącego wpływu na wynik normowania zmiennych. Dla każdego przedziału określa się liczbę wartości zmiennej należących do tego przedziału. Powstaje w ten sposób histogram częstości. Wartości histogramu częstości zależą od liczby wszystkich obiektów oraz liczby przedziałów. Im więcej jest obiektów, tym większe wartości przyjmuje histogram częstości. Natomiast im jest więcej przedziałów, tym histogram przyjmuje mniejsze wartości. Aby uniezależnić wartości histogramu od liczby obiektów i liczby przedziałów dokonuje się ich przeskalowania:

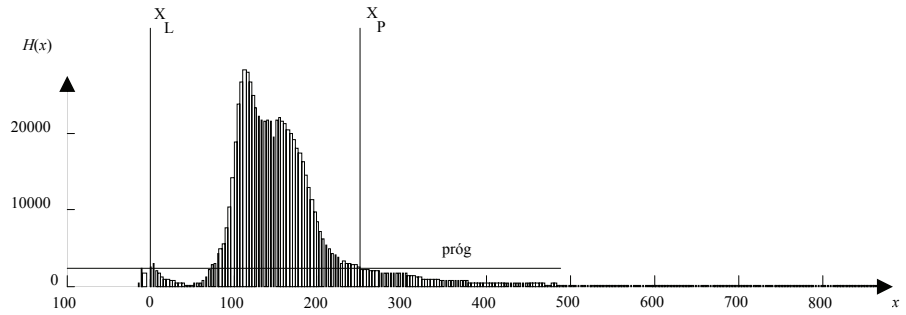
$$h_k = \frac{N_p h_{cz\,k}}{N} \tag{12}$$

gdzie: $h_{cz\,k}$ – k-ty element histogramu częstości, h_k – k-ty element przeskalowanego histogramu częstości.

Przeskalowany histogram częstości jest podstawą do wyliczenia lewej i prawej wartości progowej podstawianej do wzoru (4). Rozłożenie wartości zmiennych charakteryzuje się występowaniem pewnego obszaru skumulowania większości wartości (rys. 1). Poza tym obszarem znajdują się nietypowe wartości zmiennej, występujące dość rzadko. Ze względu na losowy charakter występowania wartości nietypowych można je pominąć przy określaniu lewej i prawej wartości progowej. Lewą i prawą wartość progową można przyjąć za granice skumulowania wartości.

W celu określenia granic można przyjąć lewą i prawą minimalną liczbę elementów znajdujących się w przedziałach należących do skumulowania. Jako lewą granicę przyjmuje się pierwszy z lewej strony przedział, dla którego przekroczona została minimalna liczba elementów lewej strony. Podobnie, jako prawą granicę przyjmuje się pierwszy przedział z prawej strony, dla którego przekroczona została minimalna liczba elementów prawej strony. Ostatecznie wartościami x_{Li} i x_{Pi} będą środki wyznaczonych w ten sposób przedziałów.

Rysunek 1. Wyznaczanie wartości x_{Li} i x_{Pi}



Źródło: obliczenia własne

Wartości x_{Li} i x_{Pi} można także wyznaczać na podstawie odchylenia standardowego:

$$x_{Li} = \bar{x}_i - w_\sigma \sigma_i \tag{13}$$

oraz

$$x_{Pi} = \bar{x}_i + w_\sigma \sigma_i \tag{14}$$

gdzie: σ_i – odchylenie standardowe i-tej zmiennej, w_σ – współczynnik określający krotność odchylenia standardowego.

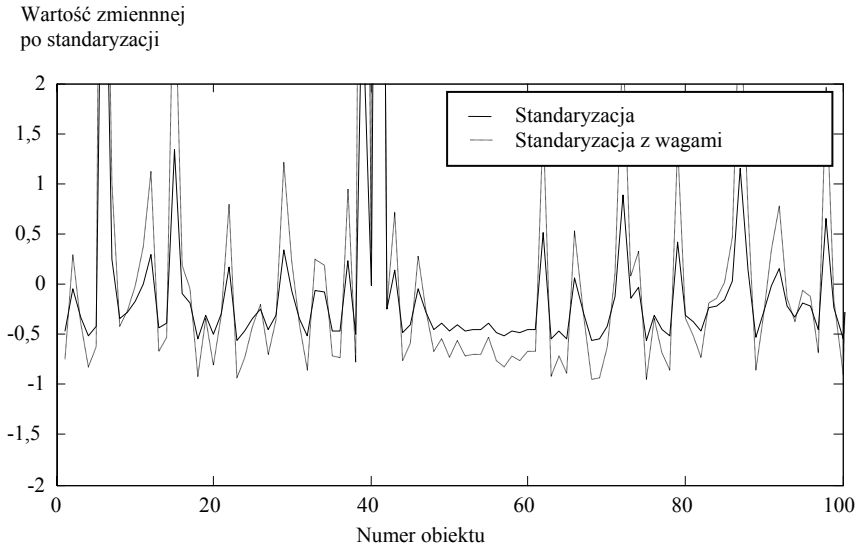
Na ogół wartość współczynnika w_σ przyjmuje się, jako jeden lub dwa. Ta metoda wyznaczania x_{Li} i x_{Pi} jest dużo prostsza od poprzedniej, jednak przy niesymetrycznych histogramach częstości nie gwarantuje, że obie wartości x_{Li} i x_{Pi} będą leżały w zakresie wartości zmiennych.

BADANIA EMPIRYCZNE

Normowaniu poddano wartości emisji zanieczyszczeń pyłowych powietrza z zakładów szczególnie uciążliwych. Badaniu poddano 379 powiatów. Dane zaczerpnięto z GUS-u, dotyczą one 2005 roku. Wskaźnik emisji zanieczyszczeń pyłowych wybrano ze względu na dużą jego rozbieżność w zależności od charakteru zakładów znajdujących się na terenie poszczególnych powiatów. Istnieje wiele powiatów o niewielkiej emisji zanieczyszczeń zbliżonej praktycznie do zera. Są to powiaty z obszarów nieuprzemysłowionych. Istnieje również wiele powiatów znajdujących się w rejonach mocno uprzemysłowionych gdzie emisja zanieczyszczeń jest znaczna. Ponadto istnieje niewielka liczba powiatów, na terenie których znajdują się bardzo duże zakłady przemysłowe, jak na przykład elektrownie. Zakłady te emitują kilkaset razy więcej zanieczyszczeń niż typowe zakłady. Stanowią one obiekty nietypowe zakłócające wartości miar syntetycznych.

Na podstawie wartości wskaźnika emisji zanieczyszczeń pyłowych wyznaczono wartości zmiennej przez podzielenie wartości wskaźnika przez liczbę zarejestrowanych w danym powiecie firm. Otrzymano w ten sposób zmienną: emisja zanieczyszczeń pyłowych na sto firm. Zmienną tą zestandaryzowano z wykorzystaniem zwykłego odchylenia standardowego oraz ważonego odchylenia standardowego. Zestandaryzowane wartości dla pierwszych stu wartości przedstawiono na rys. 2. Wykorzystanie standaryzacji z wagami spowodowało zwiększenie wahań zestandaryzowanej zmiennej.

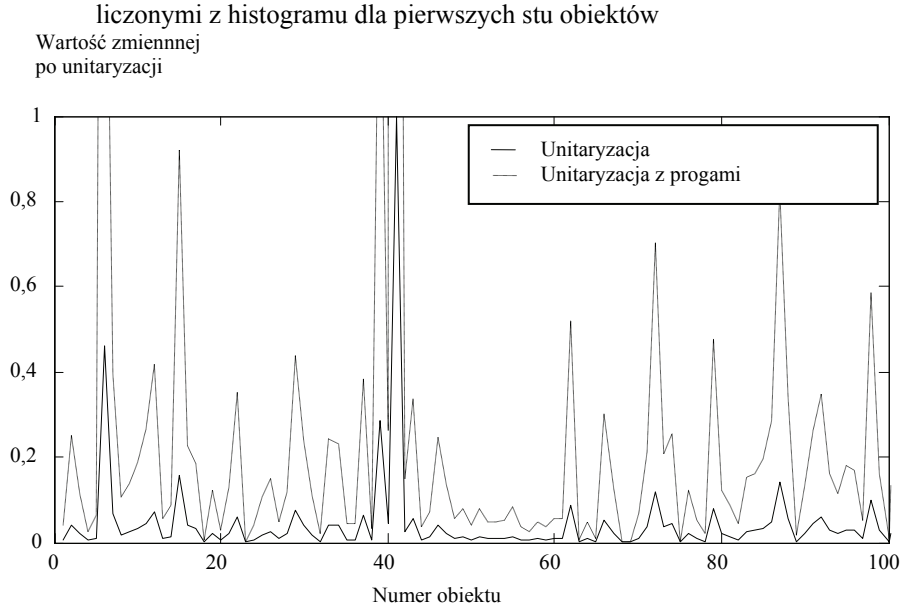
Rysunek 2. Porównanie standaryzacji i standaryzacji z wagami dla pierwszych stu obiektów



Źródło: obliczenia własne

Badaną zmienną podano unitaryzacji zerowanej oraz unitaryzacji z wartościami progowymi (rys. 3). Wartości progowe były wyliczone na podstawie odchyleń standardowych. Przyjęto jako w_σ wartość 1,5. Unitaryzacja z wartościami progowymi spowodowała zwiększenie oscylacji wartości, ale jednocześnie spowodowała znaczne przesunięcie wykresu w górę. To przesunięcie w górę jest spowodowane niesymetrią rozkładu wartości, przy czym w zależności od formy niesymetrii uzyskujemy przesunięcie w górę lub w dół. Te ostatnie jest szczególnie niekorzystne, gdyż pojawiają się wartości ujemne. W przypadku niektórych metod tworzenia miar syntetycznych jest to niedopuszczalne.

Rysunek 3. Porównanie unitaryzacji zerowanej i unitaryzacji z wartościami progowymi liczonymi z histogramu dla pierwszych stu obiektów



Źródło: obliczenia własne

Wady związanej z przesunięciem wykresu nie mają wartości progowe wyznaczane z histogramu. Histogram, który posłużył do wyliczenia wartości progowych miał sto przedziałów. Lewe i prawe progi wyznaczenia wartości progowych były równe i określone zostały na dwadzieścia elementów. W rezultacie uzyskano znaczne zwiększenie wartości wahań zmiennej przy bardzo małym przesunięciu wykresu. Niedogodnością tej metody wyznaczania wartości progowych jest fakt, że konieczna jest dość znaczna liczba obiektów do dokładnego określenia wartości progowych.

PODSUMOWANIE

Przetestowano różne metody eliminacji wpływu obiektów nietypowych na normowanie zmiennych. Wszystkie one powodują zwiększenie wahań wartości, co wpływa na zwiększenie rozróżnialności obiektów typowych ze względu na daną zmienną. W przypadku unitaryzacji, przy małej liczbie obiektów, wartości progowe można wyznaczyć przy pomocy odchylenia standardowego. W tym przypadku dobrze byłoby, gdyby rozkład wartości zmiennej był symetryczny. Przy dużej liczbie obiektów można wyznaczyć wartości progowe z histogramu. W tym przypadku warunek symetrii rozkładu wartości nie jest konieczny.

LITERATURA

Amir D. A. (2000) Statystyka w zarządzanium PWN, Warszawa
Borys T. (1978) Metody normowania cech w statystycznych badaniach porównawczych, Przegląd Statystyczny, nr 2
Grabiński T., Wydymus S., Zeliaś A. (1989) Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych, PWN, Warszawa
Kolenda M. (2006) Taksonomia numeryczna. Klasyfikacja, porządkowanie i analiza obiektów wielocechowych, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu,Wrocław, ISBN 83-7011-805-4
Kozak R., Staudhammer C., Watts S. (2007) Introductory Probability and Statistics: Applications for Forestry and Natural Sciences. CABI, ISBN 1845932757
Kukuła K. (2000) Metoda unitaryzacji zerowanej, PWN, Warszawa, ISBN 83-01-13097-0
Nermend K. (2008) Rachunek wektorowy w analizie rozwoju regionalnego, Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin, ISBN 978-83-7241-660-5
Nowak E. (1990) Metody taksonomiczne w klasyfikacji obiektów społeczno-gospodarczych, PWE, Warszawa, ISBN 83-208-0689-5

Standardization of Variables Describing Untypical Objects

Abstract: In many investigations the problem of untypical objects, whose characteristics are described by very large values, appears. Such objects may affect significantly the investigations results due to the reduction of the scope of variables in the process of standardization. Negative impact of the untypical values can be minimized by the use of certain methods of standardization. The article presents two such methods: standardization with the weighted standard deviation and unitarization with threshold values.

Key words: untypical objects, standardization methods