PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1076 2005
TAKSONOMIA 12
Klasyfikacja i analiza danych – teoria i zastosowania

Krzysztof Jajuga
Akademia Ekonomiczna we Wrocławiu

ZALEŻNOŚĆ W ANALIZIE DANYCH – DYSKUSJA NAD NIEKTÓRYMI PODEJŚCIAMI

1. Wstęp – systematyzacja niektórych metod analizy zależności

Jednym z podstawowych zadań realizowanych za pomocą metod statystycznej analizy wielowymiarowej jest analiza zależności między zmiennymi wchodzącymi w skład pewnego wektora. W tym artykule przedstawimy rozważania na temat metod analizy zależności. Rozważania te mają charakter systematyzacji niektórych zagadnień związanych z analizą zależności.

Na wstępie zajmiemy się problemem podziału różnych podejść w analizie zależności. W tym zakresie rozważymy kilka kryteriów podziału. Dwoma podstawowymi kryteriami są:

Kryterium zależności wewnątrzgrupowej i międzygrupowej

Zależność wewnątrzgrupowa to taka zależność, w której rozpatrywane zmienne są traktowane jako jeden zbiór i analizowana jest „łączna" zależność między zmiennymi należącymi do tego zbioru. W szczególnym przypadku rozpatrywany zbiór liczy dwie zmienne i przedmiotem zainteresowań jest zależność między tymi zmiennymi (np. mierzona współczynnikiem korelacji). Gdy zaś tych zmiennych jest więcej, przykładową metodą analizy zależności jest np. analiza głównych składowych.

Zależność międzygrupowa to taka zależność, w której zmienne są pogrupowane w dwa zbiory lub więcej zbiorów, analiza jest zaś przeprowadzona między zbiorami. W dwóch zbiorach zmiennych klasyczną metodą jest analiza korelacji kanonicznej, a miarą zależności – współczynnik korelacji kanonicznej.

Kryterium liczby zmiennych

Tutaj wyróżnia się dwie sytuacje: kiedy mamy dwie zmienne lub więcej zmiennych. Przypadek dwóch zmiennych nie wymaga komentarza, jest klasyczny. Jeśli zaś dysponujemy większą liczbą zmiennych, to często postępuje się tak, że część zmiennych zastępuje się pewną zmienną syntetyczną, będącą kombinacją liniową tych zmiennych, i wtedy bada się zależność między jedną zmienną i zmienną syntetyczną.

Łącząc te dwa kryteria, możemy wyróżnić kolejne możliwe rodzaje analizy zależności. Przedstawię je poniżej i wskaże typowe metody ich stosowania.

Przypadek 1 – łączna zależność, dwie zmienne.

Typowe metody: korelacja, regresja z jedną zmienną objaśniającą.

Przypadek 2 – łączna zależność, więcej zmiennych.

Typowe metody: analiza głównych składowych.

Przypadek 3 – zależność między dwiema grupami, obie zawierają po jednej zmiennej.

Typowe metody: korelacja, regresja z jedną zmienną objaśniającą.

Przypadek 4 – zależność między dwiema grupami, jedna zawiera jedną zmienną, zaś druga – więcej zmiennych.

Typowe metody: korelacja wieloraka, regresja z wieloma zmiennymi objaśniającymi.

Przypadek 5 – zależność między dwiema grupami, obie zawierają więcej zmiennych.

Typowe metody: korelacja kanoniczna, regresja wielowymiarowa.

Zauważmy, że przypadki 1 i 3 są formalnie takie same, jakkolwiek rozpatrywane z różnych punktów widzenia. Dodajmy również, iż duża część wymienionych metod polega na tworzeniu syntetycznej zmiennej, będącej kombinacją liniową pewnej grupy zmiennych. Taka zmienną syntetyczną jest np. główna składowa, zmienna kanoniczna czy też zmienna objaśniana w funkcji regresji.

Trzecim kryterium podziału metod analizy zależności jest podział ze względu na zależność „jednostronną" i „dwustronną". Zależność zwana przez nas „jednostronną" to taka zależność, w której jedna zmienna traktowana jest jako objaśniana, inne zmienne zaś (jedna lub więcej) traktowane są jako objaśniające i badacza interesuje wpływ zmiennych objaśniających na zmienną objaśnianą, czyli coś, co przypomina zależność „przyczynowo-skutkową". Z kolei w zależności zwanej przez nas „dwustronną" badacza interesuje jedynie wskazanie powiązań między dwiema (lub więcej) zmiennymi, bez wnikania w to, co jest ewentualną przyczyną, a co jest skutkiem – czyli występuje dwustronne powiązanie.

Zauważmy, że gdy przyjmiemy stochastyczne ujęcie w analizie danych, wówczas najpełniejsza informacja o strukturze zależności znajduje się w rozkładzie wielowymiarowym, danym (w przypadku zmiennych ciągłych, a takie tu rozpatrujemy) poprzez funkcję gęstości:

$$f(x_1, x_2, ..., x_m). \tag{1}$$

W takiej sytuacji analiza „dwustronnej" zależności polega na rozpatrywaniu rozkładu łącznego, a analiza „jednostronnej" zależności polega na rozpatrywaniu rozkładu warunkowego – zmiennej objaśnianej pod warunkiem zmiennych objaśniających.

Kolejne kryterium podziału metod analizy zależności wywodzi się z tego, czy zależność jest analizowana na podstawie pełnej informacji zawartej w rozkładzie wielowymiarowym, czy też na podstawie syntetycznej informacji, którą niosą parametry tego rozkładu. Gdy w celu uproszczenia pod uwagę weźmiemy jedynie dwie zmienne: X oraz Y, wówczas możemy wyróżnić modele teoretyczne analizy zależności.

Pełny model wykorzystuje informacje w postaci rozkładu dwuwymiarowego, rozpatrywanego poprzez funkcję dystrybuanty. W ten sposób otrzymujemy:

– w przypadku analizy rozkładu łącznego:

$$F(x, y) = P(X \leq x, Y \leq y), \tag{2}$$

– w przypadku analizy rozkładu warunkowego:

$$F(y|x) = P(Y \leq y|X = x). \tag{3}$$

Uproszczony – lecz częściej stosowany – model wykorzystuje jedynie informacje o parametrach rozkładu. W ten sposób – przy założeniu, że rozpatrywanym parametrem jest wartość oczekiwana – otrzymujemy:

– w przypadku analizy rozkładu łącznego:

$$E(XY), \tag{4}$$

– w przypadku analizy rozkładu warunkowego:

$$E(Y|X). \tag{5}$$

Zauważmy, iż dystrybuanta rozkładu wielowymiarowego zawiera pełną informację o zależności, jednak oprócz tego zawiera również informacje o innych charakterystykach, mianowicie o położeniu, skali itp. W tym sensie można powiedzieć, że ta informacja jest „zanieczyszczona". Dokonując analizy zależności, należy „uwolnić" rozkład wielowymiarowy od tych innych charakterystyk. Jak się okazuje, są możliwe co najmniej dwa ogólne sposoby postępowania.

Pierwszy sposób postępowania polega na standaryzacji każdej zmiennej, „uwalniając" te zmienne w ten sposób od parametrów położenia i parametrów rozrzutu, czyli rozpatrując rozkład zmiennych standaryzowanych. Standaryzacja polega na odjęciu parametru położenia i podzieleniu przez parametr skali. Zastosowanie tu ma następujący wzór:

$$Z = \frac{X - \mu}{\sigma}. \tag{6}$$

W przypadku dwóch zmiennych, gdy założymy, że parametrem położenia jest wartość oczekiwana, a parametrem rozrzutu (skali) jest odchylenie standardowe, wzór (4), określający wartość oczekiwaną iloczynu (analiza łącznej zależności), staje się wzorem na zwykły współczynnik korelacji.

Oczywiście, zakładając inne parametry położenia i parametry skali, otrzymujemy inne miary zależności. Można tu na przykład wykorzystać podejście, którego idea została przedstawiona przez Jajugę i Walesiaka [2004], z zastosowaniem L-normy. Gdy np. przyjmiemy normę wynikającą z zastosowania odległości miejskiej, wówczas parametrem położenia jest mediana, a naturalnym parametrem skali jest wartość oczekiwana bezwzględnych odchyleń danej zmiennej od jej mediany. Oczywiście dalszych badań wymaga określenie właściwości takich miar zależności. Jeśli zaś chodzi o drugi możliwy sposób, to polega on na zastosowaniu tzw. analizy połączeń (*copula analysis*). Sposób ten jest opisany w następnej części artykułu.

2. Analiza zależności za pomocą funkcji połączeń

Główna idea analizy połączeń polega na oddzieleniu analizy rozkładów brzegowych od analizy zależności. Dzięki temu parametry zależności są „oddzielone" od parametrów położenia i parametrów skali.

Formalne podstawy analizy połączeń najlepiej ilustruje twierdzenie Sklara (por. [Sklar 1959]), mówiące o dekompozycji ciągłej dystrybuanty rozkładu wielowymiarowego jako funkcji dystrybuant rozkładów jednowymiarowych. Twierdzenie to zapisane jest następująco:

$$F(x_1,...,x_m) = C(F_1(x_1),...,F_m(x_m)). \tag{7}$$

Funkcja C, występująca w tej dekompozycji, wiążąca dystrybuantę rozkładu wielowymiarowego z dystrybuantami rozkładów jednowymiarowych, jest nazywana funkcją połączenia, inaczej funkcją *copula*. Funkcja ta jest funkcją dystrybuanty wielowymiarowego rozkładu jednostajnego.

Relację między rozkładem wielowymiarowym i rozkładami jednowymiarowymi można również przedstawić za pomocą tzw. gęstości *copula*, oznaczonej przez c:

$$f(x_1,...,x_m) = c(F_1(x_1),...,F_m(x_m)) \cdot f_1(x_1) \cdot ... \cdot f_m(x_m), \tag{8}$$

$$c(u_1,...,u_m) = \partial^m C(u_1,...,u_m). \tag{9}$$

Szczególnym przypadkiem funkcji połączenia jest tzw. normalna (gaussowska) *copula*. Wiąże ona dystrybuantę wielowymiarowego rozkładu normalnego z dystrybuantami jednowymiarowych rozkładów normalnych według wzoru:

$$C(u_1,...,u_m) = \Phi^m(\Phi^{-1}(u_1),...,\Phi^{-1}(u_m)). \tag{10}$$

W szczególnym, dwuwymiarowym przypadku otrzymujemy:

$$C(u_1,u_2) = \Phi^2(\Phi^{-1}(u_1),\Phi^{-1}(u_2)), \tag{11}$$

$$C(u_1,u_2) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right) dxdy. \tag{12}$$

Jak widać, funkcja normalna *copula* to jedyna funkcja, która transformuje jednowymiarowe rozkłady normalne w wielowymiarowy rozkład normalny. Innymi

słowy: jeśli jednowymiarowe rozkłady brzegowe są rozkładami normalnymi, to wielowymiarowy rozkład jest normalny tylko wtedy, gdy odpowiednia funkcja połączenia (funkcja *copula*) jest funkcją normalną (gaussowską). Widać również, iż w przypadku dwuwymiarowym jedynym parametrem, który występuje w normalnej funkcji copula, jest współczynnik korelacji, będący naturalną miarą zależności w przypadku dwuwymiarowego rozkładu normalnego.

Innymi często rozważanymi funkcjami połączenia (funkcjami *copula*) są funkcje należące do rodziny *copula* Archimedesa. Te funkcje, zdefiniowane za pomocą ściśle malejących i wypukłych funkcji, zwanych generatorami, spełniają następujący warunek:

$$\psi : [0;1] \to [0;\infty), \tag{13}$$

$$\psi(1) = 0. \tag{14}$$

Copula Archimedesa w przypadku dwuwymiarowym jest określona następująco:

$$C(u_1,u_2) = \psi^{-1}(\psi(u_1) + \psi(u_2)). \tag{15}$$

Oczywiście w zależności od postaci funkcji generatora można wyróżnić różne funkcje *copula* Archimedesa. Najbardziej znaną jest *copula* Gumbela, w której generator jest przedstawiony następującym wzorem:

$$\begin{gathered} \psi(t) = -(\ln(t))^{\theta}, \\ \theta \geq 1. \end{gathered} \tag{16}$$

Copula Gumbela jest zależna od jednego parametru. Podobnie jest w przypadku wielu innych *copula* Archimedesa. Okazuje się, że ten parametr może być interpretowany jako parametr zależności. Jak widać, ze względu na to, że argumentami funkcji *copula* są wartości dystrybuant jednowymiarowych, ten parametr zależności jest „oddzielony" od jednowymiarowych parametrów położenia i skali.

Istotne właściwości funkcji połączenia wiążą się właśnie z interpretacją tej funkcji w kategoriach zależności. Podstawowe właściwości są następujące:

– w odniesieniu do zmiennych niezależnych otrzymujemy:

$$C(u_1,...,u_n) = C^{-}(u_1,...,u_n) = u_1u_2...u_n, \tag{17}$$

– dolna granica funkcji *copula* osiągnięta jest, gdy:

$$C^{-}(u_1,...,u_n) = \max(u_1 + ... + u_n - n + 1;0), \tag{18}$$

– górna granica funkcji *copula* jest osiągnięta, gdy:

$$C^{+}(u_1,...,u_n) = \min(u_1,...,u_n). \tag{19}$$

Kluczowa właściwość funkcji *copula* wiąże się z pojęciem całkowitej dodatniej zależności i całkowitej ujemnej zależności. Jest to uogólnienie klasycznej zależności liniowej. Pojęcia te – w przypadku dwuwymiarowym i zmiennych X i Y – są określone następująco:

– całkowita dodatnia zależność występuje wtedy, gdy $Y = T(X)$, gdzie: T – funkcja rosnąca,

– całkowita ujemna zależność występuje wtedy, gdy $Y = T(X)$, gdzie: T – funkcja malejąca.

Okazuje się, że zachodzą następujące relacje:

– w przypadku całkowitej dodatniej zależności:

$$C(u_1,u_2) = C^+(u_1,u_2) = \min(u_1,u_2),$$

– w przypadku całkowitej ujemnej zależności:

$$C(u_1,u_2) = C^-(u_1,u_2) = \max(u_1+u_2-1;0),$$

Estymacja funkcji *copula* jest przeprowadzana zazwyczaj w dwóch etapach z zastosowaniem metody największej wiarygodności. W pierwszym etapie metoda największej wiarygodności jest stosowana do oszacowania parametrów rozkładów brzegowych. W drugim etapie metoda ta jest stosowana do wyznaczenia parametrów funkcji *copula* dla otrzymanych parametrów rozkładów brzegowych.

3. Analiza zależności ekstremalnej

Przedstawione do tej pory rozważania dotyczyły sytuacji, w której dokonuje się pomiaru zależności dla całego zbioru danych traktowanego jako jednorodny. W praktyce często zdarza się jednak, iż w różnych klasach rozpatrywanego zbioru danych jest różna zależność. Przedstawimy teraz dwa rodzaje analiz, w których mamy do czynienia z tym problemem.

Pierwsza sytuacja dotyczy analizy tzw. zależności ekstremalnej. Jest to zależność, którą charakteryzują się „ogony" rozkładu. Innymi słowy, chodzi o określenie, w jakim stopniu bardzo wysokie (bardzo niskie) wartości jednej zmiennej są powiązane z bardzo wysokimi (bardzo niskimi) wartościami drugiej zmiennej. Występują dwa podstawowe sposoby analizy:

– modyfikacja klasycznych miar zależności,

– propozycja specyficznych miar zależności ekstremalnej.

Jeśli chodzi o zmodyfikowane miary zależności, to najprostszym przykładem jest tzw. warunkowy współczynnik korelacji. Dany jest on następującym wzorem:

$$\rho_u = \frac{COV(X,Y|Y>u)}{\sqrt{V(X|Y>u)V(Y|Y>u)}}. \tag{20}$$

Jak wynika ze wzoru (20), warunkowy współczynnik korelacji jest określony ze względu na warunek, że jedna ze zmiennych przekroczy pewien próg, oznaczony tutaj przez u. Współczynnik ten jest oczywiście funkcją tego progu.

Warunkowy współczynnik korelacji jest powiązany z klasycznym (bezwarunkowym) współczynnikiem korelacji. Można dowieść, iż w przypadku dwuwymiarowego rozkładu normalnego graniczna (dla dużej próby) wartość warunkowego współczynnika korelacji określona jest następująco:

$$\rho_u \to \frac{\rho}{\sqrt{1-\rho^2}} \frac{1}{u}. \tag{21}$$

Widać, że przy zwiększającym się progu warunkowy współczynnik korelacji dąży do zera.

Drugim podejściem w analizie zależności ekstremalnej jest wyznaczenie specyficznych miar zależności ekstremalnej. Najbardziej popularne są współczynniki zależności w ogonie. Dla rozkładu dwuwymiarowego, w którym dystrybuanty dane są jako F i G, współczynniki te określone są następująco:

– współczynnik zależności w dolnym ogonie:

$$\lambda_L = \lim_{u \to 0} P(Y \leq G^{-1}(u) \Big| X \leq F^{-1}(u)), \tag{22}$$

– współczynnik zależności w górnym ogonie:

$$\lambda_U = \lim_{u \to 1} P(Y > G^{-1}(u) \Big| X > F^{-1}(u)). \tag{23}$$

Oba współczynniki określają asymptotyczną zależność. Na przykład współczynnik zależności w górnym ogonie jest określony jako graniczne prawdopodobieństwo, że jedna ze zmiennych przyjmie wartość z górnego ogona pod warunkiem, że druga zmienna przyjmuje wartość z górnego ogona. Granica jest tu rozumiana ze względu na to, jaki wysoki (bądź niski) kwantyl jest rozpatrywany.

Okazuje się, że przedstawione współczynniki zależności w ogonie formalnie zależą od omówionej w poprzednim punkcie funkcji połączenia. Wyrażają to następujące wzory:

$$\lambda_L = \lim_{u \to 0} [C(u,u)/u], \tag{24}$$

$$\lambda_U = \lim_{u \to 1} [(1 - 2u + C(u,u))/(1-u)]. \tag{25}$$

Fakt ten potwierdza przydatność analizy połączeń (*copula analysis*) w badaniach zależności.

4. Analiza zależności w przypadku niejednorodnego zbioru danych

Jak wskazywaliśmy, w praktyce często zdarza się, że zbiór danych jest niejednorodny i interesuje nas określenie zależności dla jednorodnych klas w tym zbiorze. Mamy wówczas do czynienia z zagadnieniem klasyfikacji, która jest dokonywana właśnie ze względu na zależność. Podamy teraz propozycję analizy zależności w takiej sytuacji.

Zostanie zastosowane tutaj tzw. podejście klasyfikacyjne, opisane w pracy Jajugi [1993]. Rozważamy podejście stochastyczne, zbiór liczący n obserwacji pochodzi zaś z K populacji o rozkładach tej samej postaci, różniących się jedynie parametrami. Wtedy funkcja wiarygodności zbioru obserwacji określona jest następująco:

$$L(\theta|x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} f_{\gamma_i}(x_i|\theta_{\gamma_i}), \quad (26)$$

przy czym:

$$\gamma_i = j \Leftrightarrow x_i \in \Pi_j.$$

W praktyce minimalizacja funkcji wiarygodności danej wzorem (26) odbywa się za pomocą algorytmu iteracyjnego, przebiegającego w następujących etapach:

1. Startuje się od zadanej klasyfikacji początkowej zbioru obserwacji.
2. W każdej iteracji algorytmu:
   - dokonuje się estymacji parametrów rozkładu w każdej klasie,
   - wyznacza się wartość funkcji gęstości dla każdej obserwacji i każdej klasy (w sumie *n* razy *K* wartości),
   - wyznacza się nową klasyfikację poprzez przydzielenie każdej obserwacji do klasy o największej wartości funkcji gęstości.
3. Postępowanie iteracyjne jest kontynuowane do momentu, gdy klasyfikacja otrzymywana w kolejnych iteracjach przestanie się zmieniać.

W przedstawionym podejściu klasyfikacyjnym i algorytmie pojawiają się dwa problemy, które muszą być rozwiązane. Pierwszy z nich to konieczność ustalenia liczby klas, o której w podejściu klasyfikacyjnym zakłada się, że jest znana. Drugi problem to możliwość otrzymania rozwiązania odpowiadającego ekstremum lokalnemu, co z reguły oznacza konieczność wykorzystania metod globalnej optymalizacji bądź stosowania przedstawionego algorytmu kilkakrotnie przy różnych zadanych klasyfikacjach początkowych i spośród otrzymanych wybór klasyfikacji odpowiadającej największej wartości funkcji wiarygodności. Przedstawione podejście jest ogólne i dopuszcza możliwość założenia różnych postaci rozkładów wielowymiarowych w poszczególnych klasach. Podejście to może być traktowane jako odpowiednik idei *model-based clustering*, czyli metod klasyfikacji wynikających z pewnych modeli opisujących każdą klasę. Idea ta została dobrze przedstawiona przez Banfielda i Raftery'ego [Banfield, Raftery 1993]. Ogólny model dla każdej klasy, przedstawiony przez nich, to wielowymiarowy rozkład normalny, którego parametry odzwierciedlają orientację, objętość i kształt klasy, dzięki temu, że macierz kowariancji może być przedstawiona za pomocą dekompozycji spektralnej (według wartości własnych).

Oczywiście model zdefiniowany dla każdej klasy może mieć bardzo różny charakter. Może być nim właściwie dowolny rozkład wielowymiarowy, może to być syntetyczna charakterystyka, taka jak główna składowa czy też funkcja regresji.

Przedstawimy autorską propozycję modelu wynikającego z tego podejścia. Model ten może być przydatny właśnie w analizie zależności w przypadku niejednorodnego zbioru danych.

Pod uwagę bierzemy funkcję wiarygodności daną wzorem (26). Model w każdej klasie to rozkład wielowymiarowy, jednak jest on opisany za pomocą rozkła-

dów brzegowych i funkcji połączenia (funkcji *copula*). Otrzymujemy zatem następujące przedstawienie funkcji dystrybuanty i funkcji gęstości dla danej klasy:

$$F_j(x_1,...,x_m) = C_j(F_{j1}(x_1),...,F_{jm}(x_m)), \tag{27}$$

$$f_j(x_1,...,x_m) = c_j(F_{j1}(x_1),...,F_{jm}(x_m)) \cdot f_{j1}(x_1) \cdot ... \cdot f_{jn}(x_m). \tag{28}$$

W praktycznym zastosowaniu tego podejścia wykorzystuje się taki sam algorytm iteracyjny, jak w podejściu klasycznym, z tym że w każdej iteracji estymacja jest przeprowadzana odrębnie dla parametrów rozkładów brzegowych i parametrów funkcji połączenia.

Proponowane podejście umożliwia analizę zależności w niejednorodnym zbiorze danych, przy czym jest to zależność rozumiana ogólniej niż klasyczna zależność liniowa, dzięki wprowadzeniu analizy połączeń.

Literatura

Banfield J.D., Raftery A.E., *Model-Based Gaussian and Non-Gaussian Clustering*, „Biometrics" 1993, 49, s. 803-821.

Jajuga K., *Statystyczna analiza wielowymiarowa*, Wydawnictwo Naukowe PWN, Warszawa 1993.

Jajuga K., Walesiak M., *Remarks on the Dependence Measures and the Distance Measures*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1022, AE, Wrocław 2004, s. 348-354.

Sklar A., *Fonctions de repartition à n dimensions et leurs marges*, Publications de l'Institut de Statistique de l'Université de Paris, 8, Paris 1959, s. 229-231.

DEPENDENCE IN DATA ANALYSIS – DISCUSSION ON SOME APPROACHES

Summary

The paper presents the discussion on the methods of dependence analysis, where one considers continuous variables. Different criteria to systematize dependence analysis methods are proposed. The particular attention is paid to copula analysis, which is the generalization of the traditional approach to analyze linear dependence. In the paper the methods to analyze extreme dependence are also given. Finally author proposes the approach to analyze the dependence in the case of heterogeneous data set.