Dobór zmiennych diagnostycznych:

Metoda Hellwiga

Punktem wyjścia tej metody jest ustalenie wartości progowej współczynnika korelacji *r**, powyżej której przyjmujemy, że zmienne są ze sobą istotnie skorelowane. Wartość tę można ustalić w sposób arbitralny lub też wyznaczyć w sposób formalny. Przy podejściu arbitralnym jest ona najczęściej ustalana na poziomie 0,5. W podejściu formalnym najczęściej stosuje się metodę minimaksową lub procedurę weryfikacji istotności skorelowania zmiennych diagnostycznych.

W pierwszej z metod wartość progową współczynnika korelacji wyznaczamy na podstawie wzoru:

$$r^* = \min_{j} \max_{j'} \left| r_{jj'} \right|, \qquad j,j'=1,2,...,m. \qquad (1)$$

W drugiej z metod wzór na wartość progową współczynnika korelacji ma postać:

$$\mathrm{r}^* = \left[\frac{t_{\alpha,s}^2}{t_{\alpha,s}^2 + n - 2} \right]^{\frac{1}{2}}, \qquad (2)$$

gdzie:

$\mathrm{t}_{\alpha,s}$ - wartość odczytana z tablic dystrybuanty rozkładu *t*-Studenta dla *s* =*n*-2 stopni swobody oraz przyjętego poziomu istotności α.

Procedura metody parametrycznej przebiega zgodnie z następującymi krokami:

1. Wyznaczamy sumy bezwzględnych wartości każdej kolumny macierzy korelacji ***R***:

$$R_{j'} = \sum_{j=1}^{m} \left| r_{jj'} \right|, \qquad j'=1,2,...,m. \qquad (3)$$

2. Wyszukujemy kolumnę, dla której suma $R_{j'}$ jest największa:

$$R_{j_0'} = \max_{j'} \left\{ R_{j'} \right\}. \qquad (4)$$

3. W kolumnie $R_{j_0'}$ wybieramy elementy o wartościach bezwzględnych większych od wartości progowej *r**:

$$\left| r_{jj'_0} \right| > r^*, \qquad j'=1,2,...,m, \qquad (5)$$

oraz identyfikujemy odpowiadające tym elementom wiersze. Potencjalną zmienną dopuszczalną odpowiadającą wyróżnionej kolumnie nazywamy zmienną centralną, a zmienne odpowiadające wyróżnionym wierszom za jej zmienne satelitarne. Tworzą one podzbiór potencjalnych zmiennych dopuszczalnych (tzw. skupienie). Zmienne należące do danego skupienia są ze sobą istotnie skorelowane, co oznacza znaczące powielanie informacji przez te zmienne.

4. Redukujemy macierz korelacji ***R,*** wykreślając z niej kolumnę oraz wiersze odpowiadające wyróżnionej zmiennej centralnej oraz jej zmiennym satelitarnym.

5. Powtarzamy kroki 1-4, operując kolejnymi zredukowanymi macierzami korelacji, otrzymując kolejne skupienia, aż do wyczerpania zbioru dopuszczalnych zmiennych diagnostycznych.

Do ostatecznego zbioru zmiennych diagnostycznych, będącego podstawą analiz porównawczych, wchodzą wszystkie zmienne centralne oraz zmienne izolowane

Porządkowanie liniowe:

Metoda rang

Na wstępie dokonujemy stymulacji zmiennych. W kolejnym kroku dla każdego obiektu wyznacza się sumę przyporządkowanych mu rang, ze względu na wszystkie zmienne. Gdy dana wartość zmiennej występuje w więcej niż jednym obiekcie, przyporządkowujemy im jednakową rangę będącą średnią arytmetyczną z przysługujących im rang. Następnie oblicza się wartości zmiennej syntetycznej jako średnią wartość rang:

$$s_i = \frac{1}{m}\sum_{j=1}^{m} z_{ij}\,, \qquad i=1,2,...,n, \qquad (6)$$

gdzie z_{ij} została wystandaryzowana.

Metoda sum

W pierwszym etapie metody dokonujemy stymulacji zmiennych. Następnie obliczamy wartości zmiennej syntetycznej dla każdego obiektu jako średnią arytmetyczną z wartości zmiennych.

W kolejnym kroku eliminujemy wartości ujemne zmiennej syntetycznej przesuwając jej skalę do punktu zerowego poprzez przekształcenie:

$$s_i^{'} = s_i - \min_i \{s_i\}, \qquad i=1,2,...,n. \qquad (7)$$

Ostateczną postać zmiennej syntetycznej otrzymujemy przeprowadzając jej normalizację według następującego wzoru:

$$s_i^{''} = \frac{s_i^{'}}{\max_i \{s_i^{'}\}}, \qquad i=1,2,...,n. \qquad (8)$$

Dokonane przekształcenia powodują unormowanie miary syntetycznej w przedziale [0,1].

Metoda Hellwiga

Na podstawie macierzy wystandaryzowanych zmiennych wejściowych wyznacza się obiekt wzorcowy o współrzędnych:

$$\boldsymbol{O}_0 = [z_{oj}], \qquad j=1,2,...,m. \qquad (9)$$

Współrzędne obiektu wzorcowego obliczamy na podstawie następującego wzoru:

$$z_{oj} = \begin{cases} \max_i \{z_{ij}\} & \text{dla} \quad z_j^S \\ \min_i \{z_{ij}\} & \text{dla} \quad z_j^D \end{cases}, \qquad j=1,2,...,m;\ i=1,2,...,n. \qquad (10)$$

W przypadku zmiennych nominat powinny zostać one wstępnie przekształcone w stymulanty. Następnie obliczamy dla każdego obiektu jego odległość od obiektu wzorcowego, stosując najczęściej metrykę euklidesową o postaci:

$$d_{i0} = \left[\sum_{j=1}^{m} \left(z_{ij} - z_{0j} \right)^2 \right]^{\frac{1}{2}}, \qquad i=1,2,...,m. \qquad (11)$$

Miara syntetyczna jest ostatecznie definiowana jako:

$$s_i = 1 - \frac{d_{i0}}{d_0}, \qquad i=1,2,...,m, \qquad (12)$$

gdzie:

$$d_0 = \bar{d}_0 + 2S(d_0), \qquad (13)$$

przy czym:

$$\bar{d}_0 = \frac{1}{n} \sum_{i=1}^{n} d_{i0}; \qquad S(d_0) = \left[\frac{1}{n} \sum_{i=1}^{n} \left(d_{i0} - \bar{d}_0 \right)^2 \right]^{\frac{1}{2}}. \qquad (14)$$

Wartości te są tym wyższe, im mniej oddalony od wzorca jest dany obiekt.

Metoda dystansowa

Punktem wyjścia wyznaczania zmiennej syntetycznej jest obliczenie odległości (dystansu) od obiektu wzorca, dla każdego z porównywanych obiektów, np. według wzoru (12). Konstrukcja miary syntetycznej, wykorzystująca przekształcenie unitaryzacyjne, przyjmuje postać:

$$s_i = \left(\frac{d_{i0} - \min\limits_i \{d_{i0}\}}{\max\limits_i \{d_{i0}\} - \min\limits_i \{d_{i0}\}} \right)^p, \qquad i=1,2,...,m, \qquad (15)$$

gdzie *p* jest dodatnim wykładnikiem potęgowym.

Miara syntetyczna uzyskana za pomocą metody dystansowej jest unormowana i przyjmuje wartości z przedziału [0; 1]. Czym wyższa wartość miary, tym bliżej obiektu wzorcowego leży dany obiekt.

Taksonomia wrocławska

Zasady metody taksonomii wrocławskiej zostały opracowane przez członków Grupy Zastosowań Państwowego Instytutu Matematycznego we Wrocławiu (Florek i in., 1951). Sam proces budowy dendrytu wrocławskiego, w ramach powyższej metody, jest procesem wieloetapowym.

W pierwszym etapie szukamy dla każdego obiektu O_i obiektu $O_{i'}$ najbardziej do niego podobnego. W tym celu w każdym wierszu (kolumnie) macierzy odległości ***D*** wyznaczamy najmniejszy element:

$d_{ii'} = \min_{i'} \{d_{ii'}\}$, $i,i'=1,2,...,n;\ i \neq i'$.

Otrzymane pary najbardziej podobnych do siebie obiektów przedstawiamy w postaci grafu niezorientowanego, tzn. grafu, w którym wierzchołki odpowiadające tym obiektom są połączone wiązadłami bez zaznaczania kierunku połączenia. Długości tych krawędzi są proporcjonalne do odległości między obiektami. Wśród wyznaczonych par połączeń mogą znajdować się połączenia występujące dwukrotnie. Ponieważ kolejność połączeń w dendrycie nie odgrywa roli, jedno z podwójnych połączeń jest eliminowane. Ponadto w łączeniu mogą występować wielokrotnie te same obiekty, a w dendrycie dany obiekt może występować tylko jeden raz. Dla zapewnienia powyższego warunku połączenia te łączone są w zespoły, nazywane skupieniami.

Po utworzeniu w powyższy sposób grafu sprawdzamy czy jest on spójny. Jeżeli uzyskaliśmy spójny graf, budowa dendrytu została zakończona. Jeżeli natomiast otrzymany graf nie jest spójny, to jego poszczególne składowe (skupienia) łączy się w większe zespoły. Poszczególne skupienia łączymy ze sobą w miejscach określonych przez minimalną odległość między nimi. Tworzymy w ten sposób skupienia 2-giego rzędu. Znajdujemy w tym celu najmniejszą odległość każdego obiektu jednego skupienia od obiektów należących do pozostałych skupień. Z odległości tych wybieramy odległość najmniejszą, która zostaje wiązadłem łączącym skupienia.

Jeżeli graf w dalszym ciągu nie jest spójny, proces ten jest kontynuowany poprzez tworzenie skupień wyższego rzędu. Otrzymanie spójnego grafu kończy proces tworzenia dendrytu. Uzyskane w ten sposób uporządkowanie dendrytowe jest najkrótsze (suma długości wiązadeł dendrytu jest najmniejsza) ze wszystkich możliwych uporządkowań dendrytowych.

Grupowanie:

Metoda maksymalnego gradientu

Podział liniowo uporządkowanych obiektów na grupy według metody maksymalnego gradientu przebiega na podstawie wartości syntetycznych miar odległości w badanych obiektach. Na wstępie ustalamy liczbę grup obiektów (*z*), którą chcielibyśmy otrzymać w wyniku grupowania. Dla uporządkowanego liniowo ciągu obiektów, ze względu na niemalejące wartości miar syntetycznych, liczymy różnice pomiędzy tymi wartościami dla kolejnych par obiektów:

$\Delta_i = s_{i+1} - s_i$, $i=1,...,n-1$. (16)

Ciąg obiektów dzielimy na ustaloną grupę podciągów (grup obiektów) przerywając go w *z-1* miejscach odpowiadających *z-1* najwyższym wartościom bezwzględnym miary (16).

Metoda odchyleń standardowych

W metodzie tej, przy grupowaniu obiektów uporządkowanych liniowo, dzielimy obiekty na grupy na podstawie badania odchyleń wartości zmiennej syntetycznej obiektów od średniej wartości tej zmiennej syntetycznej o wartości odchyleń standardowych (Nowak, 1990, s. 93). Zbiór badanych obiektów jest dzielony na cztery grupy, zawierające obiekty o wartościach zmiennej syntetycznej należącej do następujących czterech przedziałów klasowych:

- $\mathrm{G}_1 : s_i < \bar{s} - S(s),$
- $\mathrm{G}_2 : \bar{s} > s_i \geq s_i - S(s),$
- $\mathrm{G}_3 : \bar{s} + S(s) > s_i \geq \bar{s},$ (17)
- $\mathrm{G}_4 : s_i \geq \bar{s} + S(s),$

gdzie:

$\bar{s}, S(s)$ - odpowiednio wartość średniej arytmetycznej i odchylenia standardowego zmiennej syntetycznej.

Metoda eliminacji wektorow

Zasadą grupowania obiektów w metodzie eliminacji wektorów jest dążenie do takiego grupowania, aby w poszczególnych grupach znajdowały się pary obiektów o podobnych strukturach. Za parę obiektów podobnych uważa się obiekty, dla których miara odległości struktur:

$$d_{ii'} = 1 - \sum_{j=1} \min \left\{ x_{ij}, x_{i'j} \right\}. \quad (18)$$

jest mniejsza od ustalonej wartości progowej *d**.

Punktem wyjścia metody eliminacji wektorów jest przekształcenie macierzy odległości ***D*** w macierz binarną podobieństwa obiektów, przy przyjęciu wartości progowej odległości *d**, o postaci:

$$\boldsymbol{P} = [p_{ii'}], \qquad i,i'=1,2,...,n, \quad (19)$$

przy czym:

$$p_{ii'} = \begin{cases} 0 & \text{gdy} \quad d_{ii'} < d^* \\ 1 & \text{gdy} \quad d_{ii'} \geq d^* \end{cases}, \quad (20)$$

gdzie:

$p_{ii'}$ – miara podobieństwa *i*-tego i *i'*-tego obiektu.

Następnie wyznaczamy wektor eliminacji ***p*** zdefiniowany jako:

$$\boldsymbol{p} = \boldsymbol{P} \cdot \boldsymbol{1}, \quad (21)$$

gdzie:

1 – wektor (*n* x 1) składający się z jedynek.

Maksymalna wartość w wektorze eliminacji wskazuje obiekt, który przy danej wartości progowej *d** jest niepodobny do największej liczby pozostałych obiektów. Obiekt ten zostaje wyeliminowany ze zbioru obiektów. W sytuacji, gdy więcej niż jedna składowa wektora eliminacji jest równa wartości maksymalnej, należy zastosować dodatkowe kryterium eliminacji. Autorzy metody sugerują, aby ze zbioru obiektów o maksymalnych wartościach składowej wektora eliminacji eliminować ostatecznie obiekt, któremu odpowiada maksymalna wartość lub suma wartości w odpowiadających mu wierszu macierzy odległości.

Po eliminacji obiektu ze zbioru obiektów tworzymy zredukowaną macierz binarną $\boldsymbol{P}_1$ wykreślając z macierzy $\boldsymbol{P}$ wiersz i kolumnę odpowiadające wyeliminowanemu obiektowi. Następnie wyznaczamy wektor eliminacji $\boldsymbol{p}_1$. Procedurę kontynuujemy, aż wszystkie składowe wektora eliminacji będą zerami. Obiekty, które nie zostały wyeliminowane, tworzą pierwszą grupę obiektów. Przedstawioną procedurę powtarzamy do grupowania obiektów nie należących do już wyodrębnionej grupy obiektów, uzyskując kolejne grupy obiektów o podobnej strukturze.

Naturalny podział dendrytu

Na wstępie porządkujemy nierosnąco wiązadła dendrytu:

$$d_1 \geq d_2 \geq d_3 \geq \ldots \geq d_{n-1},$$

gdzie:

$d_1, d_2, \ldots, d_{n-1}$ – uporządkowane długości wiązadeł.

Następnie obliczamy ilorazy długości sąsiednich wiązadeł:

$$i_k = \frac{d_k}{d_{k+1}}, \qquad k=1,2,\ldots,n-1, \qquad (22)$$

gdzie:

d_k – długość k-tego wiązadła w uporządkowanym nierosnąco szeregu długości wiązadeł.

W kolejnym kroku dla każdej pary sąsiednich ilorazów sprawdzamy, czy zachodzi relacja:

$$i_k < i_{k+1}, \qquad k=1,2,\ldots,n-1. \qquad (23)$$

Jeżeli relacja (23) spełniona jest tylko dla jednej pary sąsiednich wiązadeł, to zbiór obiektów należy podzielić na $z=k$ grup, usuwając z grafu k-1 najdłuższych wiązadeł. Natomiast w sytuacji, gdy

kryterium (23) spełnione jest więcej niż jeden raz, wprowadzamy dodatkowe kryterium, pozwalające wybrać lepszy spośród dwóch podziałów dendrytu, o postaci:

$$i_k < i_{k'}, \qquad k,k'=1,2,...,n-1;\ k \neq k'. \tag{24}$$

Lepszym podziałem dendrytu jest, według powyższego kryterium, podział na *z=k* grup niż podział na *z'=k'* grup. Oznacza to, że dla ustalenia liczby wiązadeł, które usuwamy z dendrytu bierzemy pod uwagę najmniejszy iloraz sąsiednich wiązadeł z ilorazów spełniających warunek (22). Uzyskany podział na grupy obiektów nazywany jest podziałem naturalnym, gdyż dendryt rozpada się w sposób niejako naturalny.

Metody przecięcia gałęzi dendrogramu

W wyniku porządkowania nieliniowego na podstawie metod drzewkowych uzyskujemy hierarchiczny system grup obiektów, w postaci drzewka połączeń (dendrogramu), rozłączny na każdym z jego poziomów. W celu wyodrębnienia grup obiektów podobnych, ze względu na opisujące je zmienne, można dokonać podziału drzewka przez przecięcie wszystkich połączeń o długości wyższej niż przyjęty krytyczny poziom d^*_{h-1}. Najpopularniejsze zasady podziału drzewka to:

1. **metoda Moyeny**

   $d^*_{h-1} > \overline{d} + kS(d)$,

   $\overline{d}, S(d)$ - odpowiednio średnia arytmetyczna i odchylenie standardowe długości gałęzi drzewka,

   k – parametr, którego wartości według R. Mojeny powinny zawierać się w przedziale <2,5;3,5>. W innych pracach sugerowana jest optymalna wartość parametru równa 1,25

2. **metoda najwyższej różnicy długości wiązań**

   $d^*_{h-1} > \max_h \{d_h - d_{h-1}\}$, $\qquad h=2,3,...,n-1,$

   gdzie:

   d_h – długość *h*-tej gałęzi drzewka,

   d^*_{h-1} - wartość krytyczna odległości odpowiadająca *h*-1 długości gałęzi drzewka.

3. **metoda najwyższego ilorazu długości wiązań**

   $$d^*_{h-1} > \max_h \left\{ \frac{d_h}{d_{h-1}} \right\}, \qquad h=2,3,..,n-1.$$

Operacje na danych

Stymulacja destymulant

1. **Przekształcenie ilorazowe**

$$x_{ij}^S = b\left[x_{ij}^D\right]^{-1}, \qquad i=1,2,...,n;\ j=1,2,...m;\ b>0, \qquad (25)$$

gdzie:

x_{ij}^D - wartość *j*-tej zmiennej destymulanty w *i*-tym obiekcie,

x_{ij}^S - wartość *j*-tej zmiennej po przekształceniu w stymulantę w *i*-tym obiekcie,

b – stała przyjmowana w sposób arbitralny, najczęściej *b*=1.

2. **Przekształcenie różnicowe**

Przekształcenie różnicowe ma następującą postać:

$$x_{ij}^S = a - bx_{ij}^D, \qquad i=1,2,...,n;\ j=1,2,...m; \quad b>0, \qquad (26)$$

gdzie:

a, *b* – stałe przyjmowane w sposób arbitralny, najczęściej przyjmuje się *b*=1, *a*=0 lub $a = \max_i \left\{x_{ij}^D\right\}$.

Normalizacja zmiennych

Ogólny wzór na przekształcenie normalizacyjne zmiennych diagnostycznych można przedstawić w następującej postaci:

$$z_{ij} = \left(\frac{x_{ij} - a}{b}\right)^p, \qquad i=1,2,..,n;\ j=1,2,...,m;\ b\neq 0, \qquad (27)$$

gdzie:

z_{ij} – znormalizowana wartość *j*-tej zmiennej w *i*-tym obiekcie,

a,b,p – parametry normalizacyjne.

Normalizacja może przebiegać zarówno przy zastosowaniu jako parametrów normalizacyjnych miar klasycznych, jak i miar pozycyjnych. Najpopularniejsze warianty normalizacji to:

1. Standaryzacja klasyczna ($p=1; a=\bar{x}_j; b=S\left(x_j\right)$),

2. Standaryzacja pozycyjna ($p=1; a=M(x_j); b=MOB(x_j)$),
3. Unitaryzacja ($p=1; \quad a=\min_i\{x_{ij}\}; b=\max_i\{x_{ij}\}-\min_i\{x_{ij}\};$),
4. Przekształcenie ilorazowe ($p=1; \quad a=0; \quad b\in(\bar{x}_j, \min_i\{x_{ij}\}, \max_i\{x_{ij}\})$),

Wyeliminowanie z obliczeń ujemnych wartości zmiennych

Zapewnienie postulatu dodatniości wartości zmiennych można uzyskać dokonując ich stymulacji za pomocą transformacji różnicowej (26), przyjmując wartość parametru;

$$a=\max_i\{x_{ij}\}+\varepsilon$$

gdzie ε jest dowolnie małą liczbą dodatnią.

.