

Maria Kolenda

ALGORYTM GENEROWANIA MACIERZY DANYCH NA PODSTAWIE MACIERZY ODLEGŁOŚCI

W artykule przedstawiony jest algorytm, który pozwala na generowanie współrzędnych punktów tylko na podstawie odległości metrycznej tych punktów. W statystycznej analizie wielowymiarowej algorytm ten może służyć do generowania (symulacji) macierzy danych na podstawie znanej macierzy odległości $D[n \times n]$, w przypadku gdy odległości zostały utworzone za pomocą metryki euklidesowej.

W algorytmie zadaną macierzą jest macierz odległości $D[n \times n]$, natomiast macierzą konstruowaną jest macierz $W'[n \times m]$, gdzie $m \leq n$. Macierz $W'[n \times m]$ ma taką właściwość, że utworzona na jej podstawie macierz odległości $D'[n \times n]$ będzie taka sama jak zadana macierz odległości $D[n \times n]$. Ponadto wygenerowane wartości cech (wartości współrzędnych punktów) w macierzy W' będą niezależne (nieskorelowane).

Podczas konstruowania (symulacji) macierzy danych na podstawie macierzy odległości zakłada się, że oryginalna macierz danych $X[n \times k]$ będąca podstawą konstruowania macierzy odległości $D[n \times n]$ została utajniona. Obie macierze danych, tzn. utajniona (rzeczywista) macierz $X[n \times k]$ oraz skonstruowana macierz $W'[n \times m]$ będą się różniły, będą miały różne wartości oraz mogą mieć różne rozmiary (różną liczbę kolumn), ale utworzone na ich podstawie macierze odległości będą takie same.

Algorytm może mieć zastosowanie wtedy, gdy ze względu na ochronę danych nie mogą być udostępnione rzeczywiste wartości $X[n \times k]$ albo gdy do obliczeń będzie udostępniona tylko macierz odległości $D[n \times n]$ utworzona metryką euklidesową, a będzie potrzebna macierz ze współrzędnymi punktów (np. w metodzie środków ciężkości).

Podczas konstruowania algorytmu *generowania macierzy danych na podstawie macierzy odległości* autorka korzystała z metody głównych współrzędnych, która

jest opisana w książce K. Jajugi [1]. Jednak poniższy algorytm nie zawiera kopii prezentowanej w książce metody głównych współrzędnych, ale jest odmienną procedurą zaproponowaną przez autorkę.

1. Opis algorytmu

1. Na podstawie macierzy odległości $\mathbf{D}[n \times n]$ o elementach $d_{i,j}$ wyznacza się macierz $\mathbf{C}$ o elementach $c_{i,j}$, gdzie:

$$c_{i,j} = d_{i,j}^2, \quad (1)$$

a następnie normuje się tę macierz i otrzymuje macierz $\mathbf{C}'$ o elementach $c'_{i,j}$, gdzie:

$$c'_{i,j} = c_{i,j} - \bar{c}_i - \bar{c}_j, \text{ gdzie } \bar{c}_i = \frac{1}{n} \sum_{j=1}^n c_{i,j}, \bar{c}_j = \frac{1}{n} \sum_{i=1}^n c_{i,j}, \quad (2)$$

($\bar{c}_i$, $\bar{c}_j$ – średnie arytmetyczne elementów wierszy oraz kolumn macierzy $\mathbf{C}$).

2. Na podstawie macierzy $\mathbf{C}'$ wyznacza się macierz kowariancji $\text{Cov}(\mathbf{C}')$. Dla macierzy kowariancji $\text{Cov}(\mathbf{C}')$ wyznacza się wartości własne $\lambda_1, \dots, \lambda_n$ i wektory własne $\mathbf{w}_j$.

3. Elementy wektorów własnych $\mathbf{w}_j$ przedstawia się w postaci macierzy $\mathbf{W}[n \times m]$ – jedna kolumna macierzy prezentuje wartości jednego wektora własnego. W macierzy $\mathbf{W}[n \times m]$ eliminuje się takie kolumny l , gdzie $\lambda_l \approx 0$. Po eliminacji kolumn macierz będzie miała rozmiary $\mathbf{W}[n \times m]$, gdzie $m < n$.

4. Szukane elementy macierzy $\mathbf{W}'[n \times m]$ wyznacza się na podstawie macierzy $\mathbf{W}[n \times m]$ według wzoru:

$$w'_{i,j} = w_{i,j} \cdot \sqrt[4]{\lambda_j}, \quad (j = 1, 2, \dots, m). \quad (3)$$

5. Na podstawie macierzy $\mathbf{W}'[n \times m]$ konstruuje się macierz odległości $\mathbf{D}'[n \times n]$ i sprawdza, czy jest ona taka sama jak macierz $\mathbf{D}$. Na ogół wartości skonstruowanej macierzy odległości $\mathbf{D}'$ będą się różniły od oryginalnej macierzy odległości $\mathbf{D}$ o pewną stałą a . W takim przypadku należy wszystkie elementy macierzy $\mathbf{W}'[n \times m]$ skorygować (pomnożyć) o stałą a . Wartość parametru a otrzymuje się przez podzielenie dwóch dowolnych elementów w macierzach odległości, czyli:

$$a = \frac{d_{i,j}}{d'_{i,j}}. \quad (4)$$

Algorytm zaprezentowano dalej na przykładach. Pokazano też, że elementy macierzy $W'[n \times k]$ przedstawiają wartości macierzy głównych składowych U utworzonych na podstawie macierzy danych X (zob. przykład 1). Stąd utworzone cechy w macierzy W' są niezależne.

2. Przykłady generowania macierzy danych na podstawie macierzy odległości

W punkcie tym przedstawiono cztery przykłady generowania macierzy danych na podstawie macierzy odległości:

- w przykładzie 1 przedstawiono kolejno etapy algorytmu od punktu 1 do punktu 4,
- w przykładzie 2 przedstawiono punkt 5 algorytmu, w którym pokazano korektę macierzy W' o stałą a ,
- w przykładzie 3 pokazano macierz odległości utworzoną na podstawie empirycznych danych oraz symulacyjną macierz danych W' wygenerowaną na podstawie tylko macierzy odległości,
- w przykładzie 4 przedstawiono macierz odległości utworzoną na podstawie empirycznych danych w przypadku, gdy liczba obserwacji-objektów jest mniejsza od liczby cech.

Przykład 1

W przykładzie najpierw przedstawiono oryginalną macierz danych $X[4 \times 3]$, dla której obliczono macierz odległości metryką euklidesową $D[4 \times 4]$ oraz macierz głównych składowych U (na podstawie macierzy kowariancji $Cov(X)$). Macierze te przedstawione są w tabelach 1, 2 i 3.

Następnie na podstawie macierzy odległości $D[4 \times 4]$ za pomocą przedstawionego algorytmu wyznaczono macierz symulacyjną $W'[4 \times 3]$ (tab. 4-8).

Przedstawiona w tabeli 8 macierz symulacyjna W' zawiera takie same wartości jak macierz głównych składowych U wyznaczona na podstawie macierzy danych X (por. z tab. 3).

Tabela 1. Macierz danych X

Obiekt y	Cecha X	Cecha Y	Cecha Z
o_1	1	7	4
o_2	2	10	2
o_3	4	13	1
o_4	6	15	3

Tabela 2. Macierz odległości D

	o_1	o_2	o_3	o_4
o_1	0,0000	3,7417	7,3485	9,4868
o_2	3,7417	0,0000	3,7417	6,4807
o_3	7,3485	3,7417	0,0000	3,4641
o_4	9,4868	6,4807	3,4641	0,0000

Tabela 3. Macierz głównych składowych U

U_1	U_2	U_3
-4,9632	0,8587	0,0656
-1,6317	-0,8315	-0,1454
2,0785	-1,2420	0,1116
4,5164	1,2148	-0,0318

Tabela 4. Macierz C o elementach $c_{i,j} = d_{i,j}^2$ oraz średnie arytmetyczne elementów wierszy i kolumn

	$c_{i,j}$				$\bar{c}_{\cdot,j}$
	0	14	54	90	39,5
	14	0	14	42	17,5
	54	14	0	12	20
	90	42	12	0	36
$\bar{c}_{i,\cdot}$	39,5	17,5	20	36	

Tabela 5. Macierz C' o elementach $c'_{i,j} = c_{i,j} - \bar{c}_{i,\cdot} - \bar{c}_{\cdot,j}$

-79,00	-43,00	-5,50	14,50
-43,00	-35,00	-23,50	-11,50
-5,50	-23,50	-40,00	-44,00
14,50	-11,50	-44,00	-72,00

Tabela 6. Macierz kowariancji $\text{Cov}(C')$

1284,56	418,06	-541,31	-1161,31
418,06	141,56	-171,81	-387,81
-541,31	-171,81	231,56	481,56
-1161,31	-387,81	481,56	1067,56

Tabela 7. Macierz wektorów własnych w_j wartości własnych λ_i wyznaczonych dla $\text{Cov}(C')$ (macierzy z tab. 6)

	w_1	w_2	w_3	w_4
	0,6882	0,4072	0,3326	0,5
	0,2262	-0,3943	-0,7371	0,5
	-0,2882	-0,5890	0,5657	0,5
	-0,6262	0,5761	-0,1612	0,5
λ_i	2705,4733	19,7751	0,0015	0
$\sqrt[4]{\lambda_i}$	7,212085	2,10877	0,19726	

Macierz odległości wyznaczona na podstawie macierzy W' jest taka sama jak wyznaczona na podstawie macierzy danych X . Przykładowo: odległość obiektu o_2 do o_3 , czyli $d(o_2, o_3)$ skonstruowana dla drugiego i trzeciego wiersza macierzy:

– danych X wynosi: $\sqrt{(2-4)^2 + (10-13)^2 + (2-1)^2} = \sqrt{14} = 3,7417$ (zob. $d(o_2, o_3)$ tab. 2.)

– W' wynosi: $\sqrt{(1,6317+2,0785)^2 + (-0,8315+1,2420)^2 + (-0,1454-0,1116)^2} \approx \sqrt{14}$.

Różnice w odległościach metrycznych wynikają tylko z dokładności prezentowanych wartości macierzy W' (liczby miejsc po przecinku). W programie TAKSONOMIA, gdzie przedstawiony jest ten algorytm, obie macierze odległości (tj. macierz odległości utworzona na podstawie danych X i na podstawie macierzy W') będą takie same.

Przykład 2

Według powyższego algorytmu otrzymuje się wartości macierzy W' , dla których macierz odległości będzie się na ogół różniła od pierwotnej macierzy odległości o pewną stałą a . W takim przypadku wszystkie wygenerowane wartości macie-

Tabela 8. Macierz symulacyjna W' (por. z tab. 3)

$w_1 \cdot 7,212$	$w_2 \cdot 2,108$	$w_3 \cdot 0,197$
4,9632	0,8587	0,0656
1,6317	-0,8315	-0,1454
-2,0785	-1,2420	0,1116
-4,5164	1,2148	-0,0318

rzy W' należy pomnożyć (skorygować) przez stałą α . Przypadek taki przedstawiony jest na przykładzie opisanym w tab. 9-14.

Tabela 9. Początkowa macierz odległości

0,00000	1,86011	2,31948
1,86011	0,00000	1,61245
2,31948	1,61245	0,00000

Tabela 10. Symulacyjna macierz danych W'

w'_1	w'_2
1,35110	-0,33059
-0,22153	0,90314
-1,12957	-0,57255

Tabela 11. Macierz odległości wyznaczona na podstawie danych z tab. 10

0,00000	1,99882	2,49245
1,99882	0,00000	1,73269
2,49245	1,73269	0,00000

Tabela 12. Iloraz elementów macierzy odległości z tab. 9 oraz z tab. 11. Wyznaczanie wartości stałej α

0,00000	0,93060	0,93060
0,93060	0,00000	0,93060
0,93060	0,93060	0,00000

Tabela 13. Korekta konstruowanej macierzy danych o stałą $\alpha = 0,9306$

$w'_1 \cdot 0,9306$	$w'_2 \cdot 0,9306$
1,25734	-0,30765
-0,20615	0,84047
-1,05119	-0,53282

Tabela 14. Macierz odległości wyznaczona na podstawie danych z tab. 13 (po korekcie)

0,00000	1,86011	2,31948
1,86011	0,00000	1,61245
2,31948	1,61245	0,00000

Przykład 3

Przykład symulacji macierzy danych W' na podstawie macierzy odległości dla większego zbioru obiektów.

Tabela 15. Macierz odległości D

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8
o_1	0,0000	58,1073	58,1238	300,5458	132,7439	114,0740	152,7015	218,1622
o_2	58,1073	0,0000	1,6125	248,6166	177,9467	160,6440	197,6440	264,6680
o_3	58,1238	1,6125	0,0000	248,6232	177,9860	160,6707	197,7145	264,7116
o_4	300,5458	248,6166	248,6232	0,0000	426,3936	409,2501	445,8967	513,0268
o_5	132,7439	177,9467	177,9860	426,3936	0,0000	18,9868	20,3066	86,7431
o_6	114,0740	160,6440	160,6707	409,2501	18,9868	0,0000	38,9710	104,5779
o_7	152,7015	197,6440	197,7145	445,8967	20,3066	38,9710	0,0000	67,1863
o_8	218,1622	264,6680	264,7116	513,0268	86,7431	104,5779	67,1863	0,0000

Tabela 16. Symulacyjna macierz danych W' wygenerowana na podstawie macierzy odległości D

	w_1	w_2	w_3	w_4
o_1	25,66964	29,22258	-1,39870	0,11492
o_2	75,20089	-1,04220	1,27423	-0,01433
o_3	75,22473	-0,85223	2,86368	0,17394
o_4	323,63738	-10,05801	-1,59591	0,08097
o_5	-102,71214	-4,44945	0,82447	-0,44565
o_6	-85,38408	3,29503	1,33040	-0,39252
o_7	-122,25526	-8,64325	-2,75863	-0,43267
o_8	-189,38114	-7,47247	-0,53954	0,91533

Tabela 17. Rzeczywista macierz danych **X**, na podstawie której skonstruowano macierz odległości **D**

	Moc maksymalna [KM]	Maksymalny moment obrotowy [Nm]	Przyspieszenie 0-100 km/h [s]	Średnie zużycie paliwa [l]
ALFA 156	190	222	7.3	11.4
AUDI A8	193	280	8.4	9.9
BMW 3 LIM	193	280	7	9.1
BMW 8	326	490	6.3	14
XSARA COP	90	135	13.6	7.9
LANOS SED	106	145	11.5	8.2
POLONEZ	76	121	18	8.9
SEICENTO 899	39	65	18	5.9

Przykład 4

Przykład generowania macierzy danych **W'** do macierzy odległości, która jest skonstruowana na podstawie macierzy danych **X**, gdzie liczba obiektów była mniejsza od liczby cech.

Tabela 18. Rzeczywista macierz danych **X**

	X_1	X_2	X_3	X_4	X_5
α_1	10	2.22	7.3	11.4	1.32
α_2	19.3	2.8	8.4	9.9	1.51
α_3	16	3	7	6.1	1.445

Tabela 19. Macierz odległości **D** skonstruowana na podstawie rzeczywistej macierzy danych **X**

$D =$	0	9.503815	8.050095
	9.503815	0	5.228214
	8.050095	5.228214	0

Tabela 20. Symulacyjna macierz danych **W'**

$W' =$	w_1	w_2
	5.576144	-0.58485
	-3.80345	-2.11643
	-1.77269	2.701275

Tabela 21. Macierz odległości skonstruowana na podstawie macierzy **W'**

$D' =$	0	9.503815	8.050095
	9.503815	0	5.228214
	8.050095	5.228214	0

3. Zakończenie

1. Inspiracją do napisania tego artykułu było ukazanie się w Internecie pytania o algorytm konstruowania (generowania) macierzy danych (macierzy współrzędnych punktów) tylko na podstawie macierzy odległości.

2. *Algorytm generowania macierzy danych na podstawie macierzy odległości* może być przydatny w metodologii taksonomicznej.

Literatura

ALGORITHM OF GENERATING DATA MATRICES ON THE BASIS OF DISTANCE MATRICES

Summary

The paper describes techniques of classification scoring. The basis of a score is determined by a heterogeneity of every cluster and every isolated object in a given classification. A few coefficients has been proposed, which determine various degrees of heterogeneity of clusters and isolated objects. On the basis of these coefficients, four types of classifications have been defined: very good, good, weak but sufficient, and bad classification.

The paper also describes how to easily find „the best” classification for hierarchical methods. The proposed coefficients have been commented and compared with different coefficients, such as SC (Silhouette Coefficient) and average innercluster variation.

Dr Maria Kolenda jest pracownikiem Katedry Statystyki i Cybertyki Ekonomicznej w Akademii Ekonomicznej we Wrocławiu.