Jarosław GRAMACKI*, Artur GRAMACKI*

Statystyczne odkrywanie zależności w danych

Powszechny dostęp do wydajnych i względnie tanich systemów baz danych powoduje, iż ilość przechowywanych danych jest ogromna i ciągle rośnie. Często charakteryzują się one dużą wymiarowością (*high-dimensional data sets*), która stanowi istotny problem podczas ich analizowania. Mówiąc o dużej wymiarowości rozumiemy, że poszczególne porcje danych (np. rekordy w tabelach relacyjnych) mają dużą liczbę atrybutów (zmiennych)[1)]. W praktyce często okazuje się, że wiele z tych atrybutów jest ze sobą dosyć mocno powiązanych (skorelowanych) i do otrzymania pełnego obrazu opisywanego zjawiska czy zauważenia pewnych prawidłowości w danych wystarczy uwzględnić jedynie niewielki ich podzbiór. Wychwycenie takich prawidłowości przy analizowaniu wszystkich (jak wspomniano wyżej, często skorelowanych ze sobą) atrybutów jest zwykle albo niemożliwe, albo bardzo trudne do uzyskania. Spośród wielu zadań, które definiuje się w dziedzinie eksploracji danych (*data mining*) można wyróżnić metody zmierzające do redukcji wspomnianej dużej wymiarowości zbioru danych oraz metody grupowania danych. Do pierwszej grupy można zaliczyć m. in. metodę analizy składowych głównych (*Principal Component Analysis* – **PCA**), do drugiej np. grupowanie hierarchiczne (*Hierarchical Clustering*) lub metodą *k*-średnich (*k-Means Clustering*).

REDUKCJA WYMIAROWOŚCI

Niech analizowany zbiór danych zawiera *n* niezależnych obserwacji *p*-elementowych wektorów. O takim zbiorze będziemy mówić, że jest on *p*-wymiarowy lub że jest opisany *p* atrybutami (zmiennymi). Z reguły analizowane dane są zgromadzone w jakiejś bazie danych i wtedy mówimy o tabeli składającej się z *n* wierszy i *p* kolumn (atrybutów). Głównym celem metody PCA jest redukcja tego wymiaru do $m<<p$, przeprowadzona w taki sposób, aby w *m*-wymiarowej przestrzeni pozostała zachowana większość informacji z pierwotnego zbioru danych. Innymi słowy dokonujemy transformacji pierwotnych danych z układu o *p* składowych do nowego zbioru (nieskorelowanych ze sobą) *m* składowych, zwanych *składowymi głównymi*. Zakłada się dalej, że *m* pierwszych takich składowych „gromadzi" w sobie większość (np. 80%) zmienności (wariancji) pierwotnego zbioru danych [1–5]. Okazuje się, że dla wielu rzeczywistych zbiorów danych liczba składowych głównych, pokrywająca duży procent zmienności pierwotnych danych, jest rzeczywiście dużo mniejsza, niż pierwotna wymiarowość danych. Często będzie wystarczające uwzględnienie jedynie dwóch pierwszych składowych, co umożliwi wygodne analizowanie danych na wykresie 2D. Dodatkowo na danych o zmniejszonym wymiarze łatwiej przeprowadzać kolejne analizy, np. polegające na grupowaniu danych. Czasami jest również możliwe fizyczne zinterpretowanie nowego zbioru zmiennych, choć często nie jest to konieczne.

* Uniwersytet Zielonogórski, Instytut Informatyki I Elektroniki,
e-mail: J.Gramacki@iie.uz.zgora.pl, A.Gramacki@iie.uz.zgora.pl

ANALIZA SKŁADOWYCH GŁÓWNYCH – PCA

Wyliczenie wspomnianych składowych głównych jest oparte na wartościach i wektorach własnych tzw. *macierzy kowariancji* pierwotnego zbioru danych[2)]. Macierz kowariancji jest miarą stopnia liniowej zależności pomiędzy zmiennymi[3)] lub, gdy interpretować ją inaczej, miarą rozproszenia danych w zbiorze danych [2]. Wektory własne odpowiadające *m* największym wartościom własnym macierzy kowariancji okazują się wyznaczać szukane składowe główne – określają one kierunki szukanych nowych osi układu. W sytuacjach, gdy atrybuty opisujące dane są nieporównywalne ze sobą (np. dochód osoby i wielkość miasta) lub charakteryzują się dużą zmiennością wariancji pomiędzy sobą, zamiast – jak poprzednio, pracować z macierzą *kowariancji* należy zbudować macierz *korelacji* (po wcześniejszym odpowiednim przeskalowaniu danych wejściowych) [1, 3].

Przykład pierwszy

Metoda PCA zostanie zilustrowana na przykładzie danych opisujących strukturę pozyskiwania energii ze źródeł odnawialnych w wybranych krajach UE w 2005 roku (tabela 1 oraz [6]). Interesuje nas określenie, czy istnieją jakieś podobieństwa wśród uwzględnianych w zestawieniu krajów w zakresie produkcji energii ze źródeł odnawialnych. Analizując oryginalne dane opisane (aż) 8 atrybutami, trudno dostrzec zależności pomiędzy nimi. Wykonamy więc analizę PCA, która pokaże, że jest możliwe istotne zmniejszenie liczby atrybutów, co z kolei umożliwi zauważenie niewidocznych w pierwotnej postaci zależności pomiędzy danymi. W tabeli poszczególne źródła energii to: X1 – biomasa stała, X2 – energia promieniowania słonecznego, X3 – energia wody, X4 – energia wia-

Tabela 1. Struktura pozyskania energii według wybranych źródeł w wybranych krajach UE w 2005 roku [%]

	X1	X2	X3	X4	X5	X6	X7	X8
UE-25	51,3	0,7	21,4	5,4	3,8	4,0	4,7	8,7
Austria	49,5	1,3	43,5	1,6	0,4	0,8	0,5	2,4
Czechy	76,4	0,1	10,2	0,1	2,8	5,6	0	4,8
Estonia	99,0	0	0,3	0,7	0	0	0	0
Finlandia	82,7	0	14,7	0,2	0,5	0	0	1,9
Litwa	92,9	0	5,0	0	0,3	1,4	0,4	0
Łotwa	86,9	0	12,5	0,2	0,3	0,1	0	0
Niemcy	41,3	2,2	10,1	14,0	8,6	13,1	0,8	9,9
Polska	91,2	0	4,1	0,3	1,2	2,6	0,2	0,4
Słowacja	45,1	0	45,2	0,1	0,6	4,1	0,9	4,0
Szwecja	51,7	0	40,7	0,5	0,2	2,1	0	4,8
Wariancja	485,0	0,5	274,0	18,0	6,7	14,5	1,9	12,2

[1)] Opisywane metody wywodzą się ze statystyki i dlatego często mówiąc o zbiorze danych, używa się pojęcia *obserwacja*

[2)] Obliczenia są również możliwe na podstawie rozkładu SVD macierzy. Jest to jednak w pewnym sensie analogiczne podejście.

[3)] Założenie o liniowości jest tu bardzo ważne. Skutkuje to tym, że nie wszystkie zbiory danych można analizować metodą PCA. Rozszerzeniem jest tu metoda zwana *kernel* PCA.

tru, X5 – biogaz, X6 – biopaliwa, X7 – energia geotermalna, X8 – odpady komunalne. W ostatnim wierszu podano wartości wariancji dla każdego atrybutu. Duża zmienność ich wartości uzasadnia korzystanie z macierzy korelacji zamiast macierzy kowariancji.

W pierwszym kroku [3] oblicza się macierz korelacji opisującą stopień zależności poszczególnych atrybutów od siebie (tabela 2). Im większa bezwzględna wartość, tym większa zależność (korelacja) pomiędzy poszczególnymi atrybutami. Macierz jest symetryczna, dlatego dla czytelności pominięto elementy nad główną przekątną. Przykładowo pomiędzy atrybutami X5 oraz X6 istnieje duża zależność. Pomiędzy atrybutami X3 oraz X6 takiej zależności prawie nie ma. Zauważmy również istnienie dodatnich i ujemnych korelacji pomiędzy zmiennymi.

Tabela 2. Macierz korelacji dla zbioru z tabeli 1

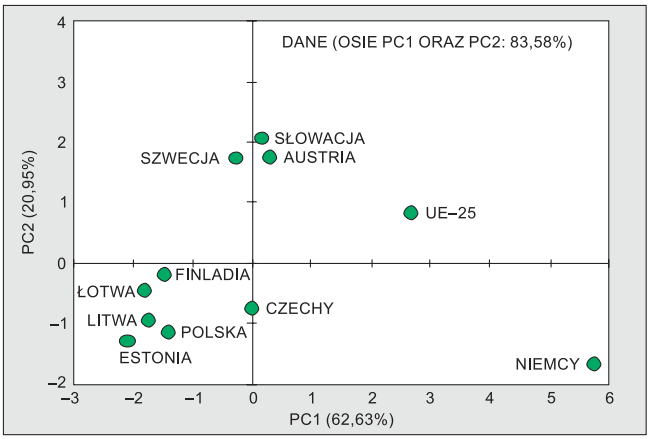
	X1	X2	X3	X4	X5	X6	X7	X8
X1	1,00							
X2	–0,63	1,00						
X3	–0,75	0,11	1,00					
X4	–0,55	0,89	-0,11	1,00				
X5	–0,49	0,80	-0,20	0,94	1,00			
X6	–0,57	0,72	**-0,06**	0,85	**0,94**	1,00		
X7	–0,43	0,28	0,14	0,37	0,37	0,22	1,00	
X8	–0,80	0,67	0,26	0,78	0,83	0,81	0,60	1,00

Tabela 3. Wartości własne macierzy korelacji z tabeli 2 uszeregowane malejąco

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Wart. własne	5,01	1,68	0,85	0,36	0,06	0,04	0,01
Zmienność (%)	62,6	20,95	10,58	4,51	0,75	0,51	0,08
Zm. skum. (%)	**62,6**	**83,58**	94,16	98,67	99,42	99,92	100

Następnie wyznaczmy wartości własne macierzy korelacji (tabela 3). Są one miarą zmienności pierwotnych danych przedstawionych we współrzędnych składowych głównych. Konkretne wartości tych zmienności pokazano w drugim wierszu tej tabeli.

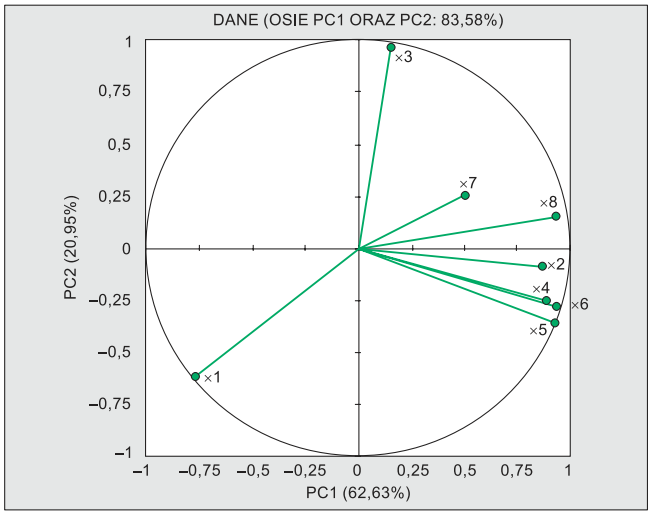
W kolejnym kroku szereguje się wartości własne od największej do najmniejszej. Wektory własne (nie pokazane) określają poszczególne składowe główne PC1 – PC7. Widać, że dwie pierwsze z nich „przenoszą" ponad 80% zmienności pierwotnych danych. Można więc z dobrym przybliżeniem analizować pierwotny zbiór danych jedynie w dwóch wymiarach. Po odpowiednim zrzutowaniu danych z tabeli 1 na nowe osie (tu: dwie pierwsze składowe główne), otrzymuje się wykres jak na rys. 1.



Rys. 1. Wykres obserwacji. Dane o produkcji energii ze źródeł odnawialnych z tabeli 1, zrzutowane na dwie pierwsze składowe główne

Wyraźnie widać na nim, które kraje w rozważanej kategorii są do siebie podobne, które wyraźnie odróżniają się od siebie oraz jak mają się one do średnich wartości całej Unii Europejskiej (UE-25).

Możliwe jest również uwidocznienie wpływu zmiennych (atrybutów) X1 – X8 na poszczególne składowe główne. Każda z tych zmiennych jest reprezentowana na rys. 2 w postaci wektora. Kierunek i długość tego wektora określa, w jakim stopniu każda ze zmiennych wpływa na poszczególne składowe główne. Wektory takie są nazywane wektorami ładunków (*loadings*).

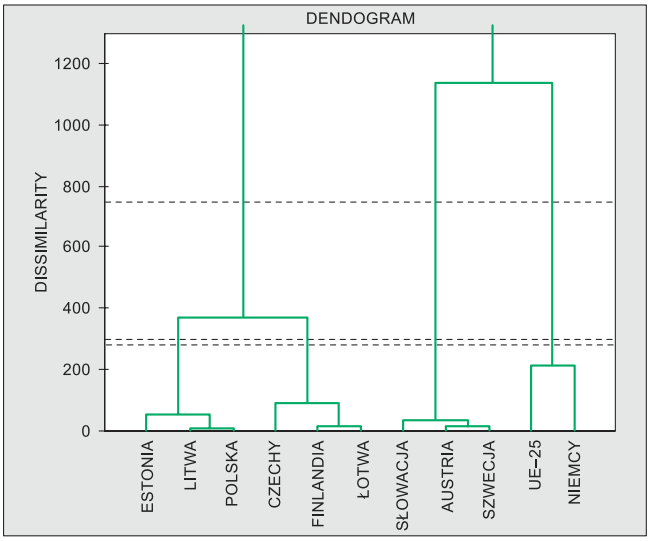


Rys. 2. Wykres zmiennych. Położenie wektorów ładunków względem dwóch pierwszych składowych głównych

Gdy dwie zmienne leżą na wykresie blisko siebie, oznacza to, że są ze sobą silnie dodatnio skorelowane (np. zmienne X4 i X5). Gdy są do siebie prostopadłe, oznacza to, że nie są ze sobą skorelowane (np. X2 i X3). Gdy znajdują się po przeciwnych stronach, oznacza to, że są silnie ujemnie skorelowane (np. X1 i X3).

Często dane z rys. 1 i 2 przedstawia się na jednym, odpowiednio przeskalowanym wykresie (w pracy nie pokazano tego), co ułatwia analizę współzależności obserwacji i zmiennych. Im *n*-ta obserwacja leży bliżej od *p*-tej zmiennej, tym większy jest wpływ tej zmiennej na daną obserwację. Wykres taki nazywa się biwykresem (*biplot*).

Na rysunku 1 widać w jaki sposób grupują się analizowane kraje. Jest tak oczywiście dzięki wcześniej wykonanej analizie PCA i zrzutowaniu pierwotnych danych na odpowiedni wykres 2D. Bez



Rys. 3. Dendrogram grupowania metodą hierarchiczną danych z tabeli 1. Dla czytelności ograniczono oś Y do wartości 1200 (było 7000)

niej grupowanie wykonane na pierwotnych danych może dać niepoprawne wyniki, gdyż należy pamiętać, że wtedy grupujemy dane potencjalnie silnie ze sobą skorelowane. Jako przykład pokazano wyniki przeprowadzonego grupowania pierwotnych danych metodą aglomeracyjną, która nie zakłada wstępnego podania liczby grup, na które chce się podzielić dane. Wyniki grupowania typowo w takich sytuacjach przedstawia się w postaci tzw. dendrogramu (rys. 3). Widać, że otrzymane grupowanie odbiega nieco od tego, które uzyskano po zastosowaniu PCA (rys. 1). Poziomą linią przerywaną zaznaczono kryterium podziału na 3 grupy. Poziomą linią przerywaną podwójną zaznaczono kryterium podziału na 4 grupy.

Przykład drugi

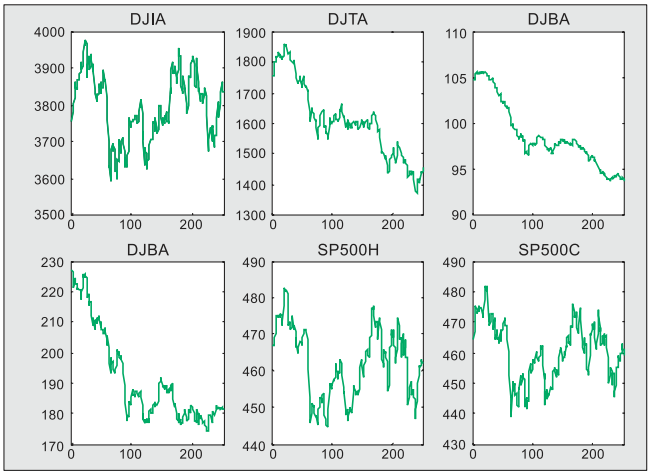
Przykład zaprezentowany w poprzednim rozdziale dotyczył stosunkowo niewielkiego wolumenu danych i grupowanie krajów według analizowanych kryteriów ostatecznie można by przeprowadzić „ręcznie". Poniżej pokazano przykład analizy dużo większego wolumenu. W tabeli 4 zamieszczono historyczne notowania 8 indeksów

Tabela 4. Dane z roku 1998 dotyczące wartości wybranych indeksów giełdy nowojorskiej oraz wielkości dziennych wolumenu obrotów

DATE	DJIA	DJTA	DJBA	DJUA	SP500H	SP500L	SP500C	NYVD000
03.01	3756	1751	104	227	466	464	464	269161
...	...	...	...	...	...	...	...	...
30.12	3834	1455	93	181	462	459	459	256210
War.	7691	14675	12	225	87	94	89	2 10^9

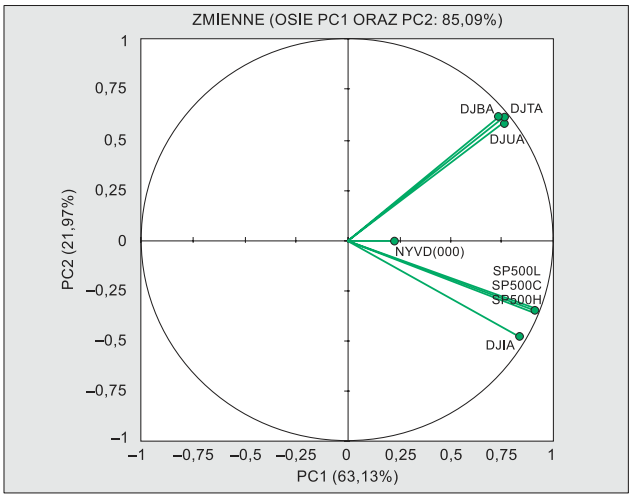
giełdy nowojorskiej w każdym dniu notowań w 1998 roku. Uwzględniono 4 wybrane indeksy przemysłowe typu Dow Jones, 3 wybrane indeksy typu S&P 500, skupiające największe notowane na giełdzie firmy oraz jedną wartość NYVD z dziennym wolumenem obrotów (w tysiącach $). Pokazano jedynie pierwszy i ostatni dzień notowań z całkowitej ich liczby 252. W ostatnim wierszu pokazano, jak w poprzednim przykładzie, wartości wariancji poszczególnych atrybutów. Interesuje nas, czy w ciągu wybranego roku, analizując jedynie posiadane dane o notowaniach, można odkryć podobne do siebie okresy aktywności giełdy.

Z samej analizy zmienności poszczególnych indeksów (rys. 4) trudno odpowiedzieć na postawione pytanie. Posłużymy się więc analizą PCA.

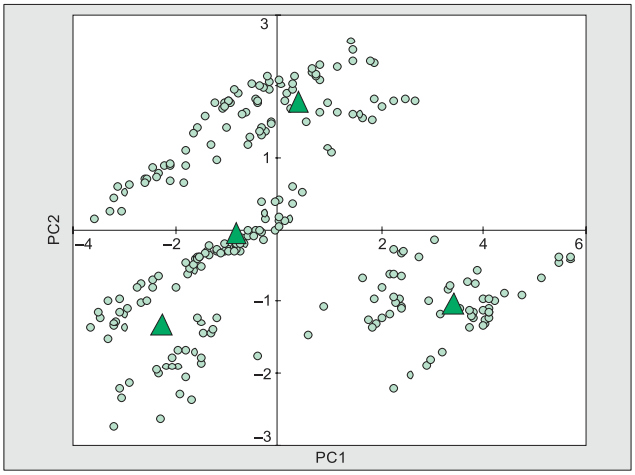


Rys. 4. Wykresy roczne zmienności wybranych indeksów giełdy nowojorskiej

Wynika z niej (z braku miejsca nie zamieszczono wyników analogicznych do tabeli 3), że już dwie składowe główne opisują ponad 85% wariancji pierwotnych danych. Na rys. 5 potwierdza się, że indeksy tego samego rodzaju są w większości mocno ze sobą skorelowane oraz, że wartość wolumenu obrotów nie jest w tej analizie istotna.



Rys. 5. Wykres zmiennych. Położenie wektorów ładunków względem dwóch pierwszych składowych głównych



Rys. 6. Wykres obserwacji. Notowania w poszczególnych dniach zrzutowane na dwie pierwsze składowe główne

Na rys. 6 pokazano dane z obserwacji (notowań w danym dniu) na płaszczyźnie dwóch pierwszych składowych głównych. Wyraźnie widać tam 4 grupy danych. Dla większej czytelności rysunku nie zaznaczono na nim przy każdym z punktów etykiety dnia, którego on dotyczy. Trójkątami zaznaczono środki wyliczonych metodą *k*-średnich czterech grup.

* * *

Bardzo często wielowymiarowe obserwacje zawierają (z informacyjnego punktu widzenia) nadmiarową liczbę atrybutów. Wartości obserwacji nie są równomiernie rozłożone wzdłuż wszystkich kierunków wielowymiarowego układu współrzędnych, ale wykazują koncentracje w pewnych podprzestrzeniach. Znalezienie tych podprzestrzeni (z reguły o dużo niższym wymiarze, niż oryginalna przestrzeń) pozwala dostrzec prawidłowości niewidoczne lub niemożliwe do zauważenia w oryginalnym układzie.

LITERATURA

[1] Koronacki J., Ćwik J.: *Statystyczne systemy uczące się*, WNT, 2005
[2] Koronacki J., Mielniczuk J.: *Statystyka dla studentów kierunków technicznych i przyrodniczych*, WNT, 2004
[3] Jolliffe I. T.: *Principal Component Analysis*, Springer Verlag, 2002
[4] Jackson J. E.: *A User's Guide to Principal Components*, Wiley & Sons, 1991
[5] Smith L. I.: *A tutorial on Principal Components Analysis*, 2002, http: //kybele. psych. cornell. edu/~edelman/Psych-465-Spring-2003/PCA-tutorial. pdf
[6] GUS, http: //www. stat. gov. pl/gus/prod_bud_inw_PLK_HTML. htm

Artykuł recenzowany