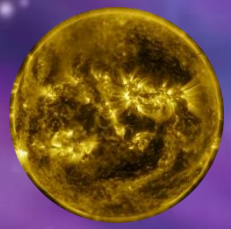
METODY INŻYNIERII WIEDZY

KNOWLEDGE ENGINEERING AND DATA MINING



TRANSFORMACJE I JAKOŚĆ DANYCH

Adrian Horzyk

Akademia Górniczo-Hutnicza

Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej

Katedra Automatyki i Inżynierii Biomedycznej, Laboratorium Biocybernetyki

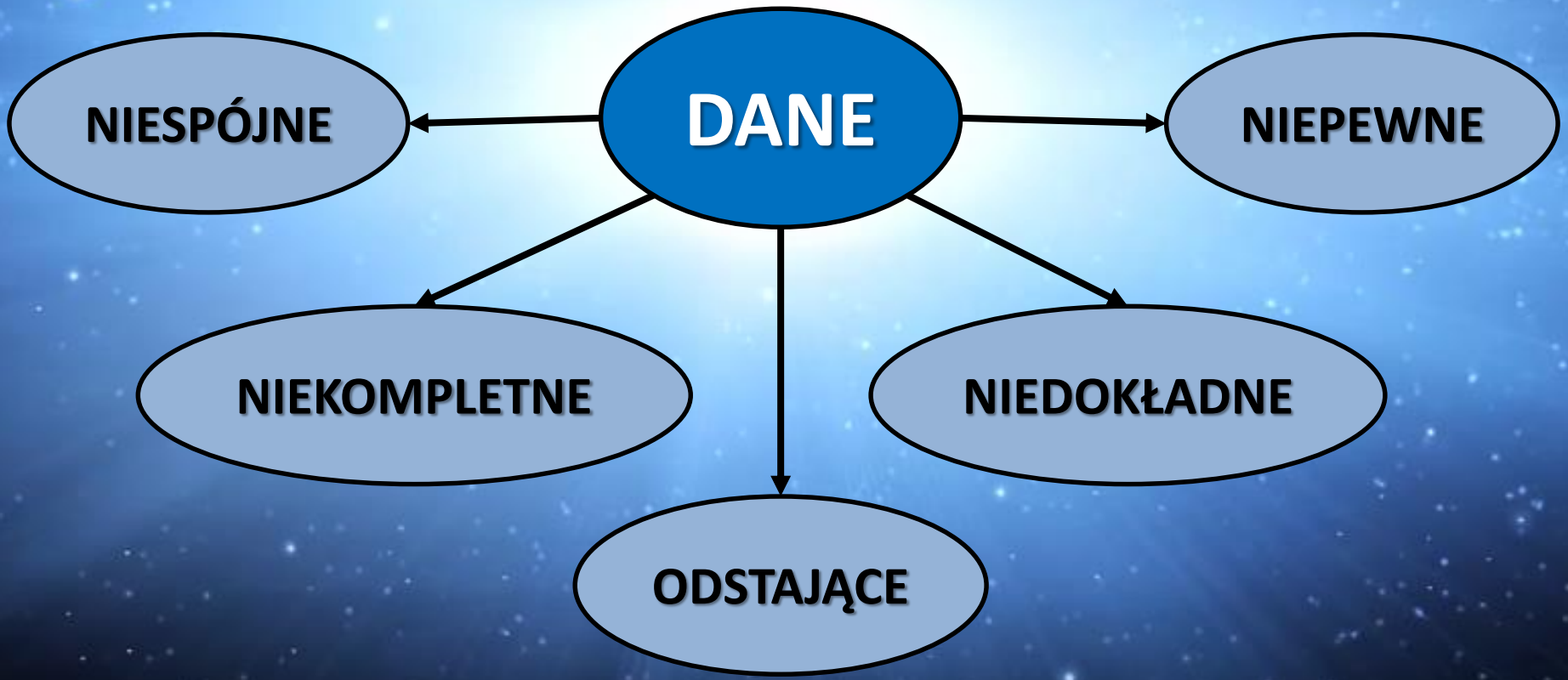
30-059 Kraków, al. Mickiewicza 30, paw. C3/205

horzyk@agh.edu.pl, Google: Adrian Horzyk

PROBLEM JAKOŚCI DANYCH

DATA QUALITY PROBLEMS

Dane mogą być niekompletne, niepewne, niedokładne, odstające lub niespójne. To powoduje różne trudności w ich przetwarzaniu zgodnie ze sloganem: „Śmieci na wejściu – śmieci na wyjściu.”



PROBLEM JAKOŚCI DANYCH

Dane niepewne – to dane, których poprawność jest niepewna i trudna do zweryfikowania.

Dane niekompletne – to dane, które dla co najmniej jednego atrybutu lub elementu sekwencji czy innej struktury nie mają określonej wartości.

Dane niedokładne – to dane o ograniczonej precyzji lub wyrażone w sposób symboliczny albo rozmyty.

Dane niespójne – to dane przypisujące jednemu obiektowi więcej niż jedną wartość dla przynajmniej jednego atrybutu, tzn. różne wartości powiązane są z tymi samymi obiektami.

Dane odstające – to dane znacznie różniące się od pozostałych, co może świadczyć o tym, że są błędne lub wyjątkowe.

PRZETWARZANIE DANYCH O OGRANICZONEJ JAKOŚCI

Przetwarzanie niekompletnych danych:

- z pominięciem niekompletnych rekordów (obiektów, krotek),
- po usunięciu atrybutów (kolumn) wprowadzających niekompletność do rekordów, jeśli niekompletność powodowana jest przez niewielką ilość atrybutów,
- po zastąpieniu brakujących danych danymi domyślną, średnią, medianą (wartością środkową), modą (wartością najczęstszą) dla określonego atrybutu,
- po zastąpieniu brakujących danych wartościami najbardziej prawdopodobnymi, wyznaczonymi na podstawie najbardziej podobnych obiektów, np. stosując metodę kNN,
- po zbudowaniu modelu dla kompletnych danych następuje próba przyporządkowania brakujących rekordów do którejś z grup/klas na podstawie zbudowanego modelu.

WSTĘPNA TRANSFORMACJA DANYCH

INITIAL DATA PREPROCESSING

to różnego rodzaju operacje zamiany wartości danych polegające na przeskalowaniu, normalizacji lub standaryzacji danych lub ich transformacji na postać uproszczoną pod kątem rozwiązywanego zadania, np. dyskryminacji.

STANDARYZACJA - STANDARDIZATION

Standaryzacja – to powszechnie stosowana w statystyce operacja polegająca na przeskalowaniu danych każdego elementu zbioru względem wartości średniej oraz odchylenia standardowego zgodnie z wzorem:

$$y_i = \frac{x_i - m}{\sigma}$$

$x = [x_1, x_2, \ldots, x_N]$ – to N-elementowy wektor danych źródłowych,

$y = [y_1, y_2, \ldots, y_N]$ – to N-elementowy wektor danych po standaryzacji,

m – to wartość średnia wyznaczona z tych danych,

σ – to odchylenie standardowe.

W wyniku standaryzacji otrzymujemy wektor cech, którego wartość średnia jest zerowa, natomiast odchylenie standardowe jest równe jedności.

Nie należy stosować dla danych o odchyleniu standardowym bliskim zeru!

NORMALIZACJA - NORMALIZATION

Normalizacja – to przeskalowanie danych względem wielkości skrajnych (min i max) danego wektora danych najczęściej do zakresu [0, 1] (czasami do [-1, 1]) zgodnie z następującą zależnością:

$$y_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

$x = [x_1, x_2, \ldots, x_N]$ – to N-elementowy wektor danych źródłowych,

$y = [y_1, y_2, \ldots, y_N]$ – to N-elementowy wektor danych po normalizacji.

Normalizacja jest wrażliwa na wartości odstające i o dużym rozrzucie, gdyż wtedy właściwe dane zostaną ściśnięte w wąskim przedziale, co może znacząco utrudnić ich dyskryminację!

Przeprowadzenie normalizacji jest czasami niezbędne do zastosowania metody, która wymaga, aby dane wejściowe lub wyjściowe mieściły się w pewnym zakresie, np. stosując funkcje sigmoidalną lub tangens hiperboliczny.

PROBLEM DANYCH ODSTAJĄCYCH

Dane odstające (outliers) – to takie dane, które nie pasują do modelu danych reprezentowanych przez pozostałe dane.

Dane odstające mieszczą się często poza przedziałem zmienności pozostałych danych dla jednego lub więcej atrybutów.

Czasami dane odstające to nietypowa kombinacja danych, która mieści się w granicach zmienności poszczególnych atrybutów, lecz jest na tyle dziwna, że nie jest zgodna z pozostałymi takim kombinacjami, np. dla problemów klasyfikacji.

Dane odstające mogą powstawać na skutek błędów, anomalii (np. pomiarowych) lub zjawisk szczególnych (czasami interesujących).

Nie istnieje ścisła matematyczna definicja danych odstających, gdyż zależy zwykle od charakteru danych oraz subiektywnej oceny.

Dane odstające zazwyczaj się usuwa lub zastępuje.

Mediana jest dosyć odporna na dane odstające, lecz zwykła średnia nie.

Stosuje się średnią winsorowską, w której wybrane skrajne obserwacje zastępuje się wartościami odpowiednio minimalnymi i maksymalnymi z pozostałych danych.

KORELACJE I KOWARIANCJE

Korelacja Pearsona – obliczana jest jako stosunek kowariancji wektorów x i y do iloczynu odchyleń standardowych:

$$p_{xy} = \frac{cov(x, y)}{std(x) \cdot std(y)}$$

Korelacja rangowa Spearmana wykorzystuje dodatkowo wektor rang oryginalnego zbioru obserwacji x lub y:

$$ps_{xy} = \frac{cov(r(x), r(y))}{std\big(r(x)\big) \cdot std(r(y))}$$

Przykład:

Jeśli wektor x składa się z następujących wartości:

$$x_1 = 2,2;\ x_2 = 1,3;\ x_3 = 1,7;\ x_4 = 2,2;\ x_5 = 4,2;\ x_6 = 3,8$$

To w wyniku sortowania uzyskamy:

$$x_2 = 1,3;\ x_3 = 1,7;\ x_1 = 2,2;\ x_4 = 2,2;\ x_6 = 3,8;\ x_5 = 4,2$$

Przypisując poszczególnym obserwacjom (danym) rangi wynikające z ich kolejności:

$$r_2 = 1;\ r_3 = 2;\ r_1 = 3,5;\ r_4 = 3,5;\ r_6 = 5;\ r_5 = 6$$

W przypadku takich samych wartości wartość rangi jest średnią z ich kolejności (r_1 i r_4).

Otrzymujemy więc następujący zbiór rang przypisanych do danych:

$$r_1 = 3,5;\ r_2 = 1;\ r_3 = 2;\ r_4 = 3,5;\ r_5 = 6;\ r_6 = 5$$

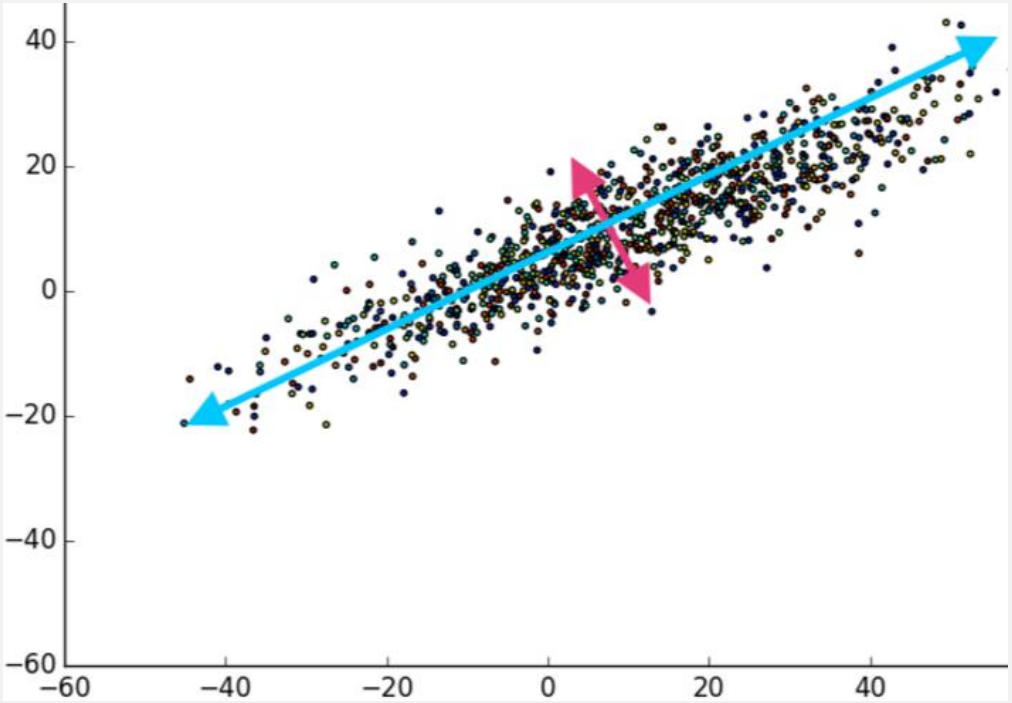
PCA – PRINCIPAL COMPONENT ANALYSIS

PCA – to metoda wstępnego przetworzenia danych polegająca na takim obróceniu ortogonalnego układu współrzędnych tak, żeby maksymalizować wariancję dla kolejnych współrzędnych: 1, 2, …

Na podstawie macierzy kowariancji konstruujemy nową przestrzeń obserwacji danych, w której największą zmiennością charakteryzują się początkowe czynniki (najpierw wyznaczone współrzędne).

Większa wariancja / zmienność umożliwiaja metodom klasyfikacji osiągnąć lepszą dyskryminację.

Ponadto PCA umożliwia uproszczenie danych o te czynniki / współrzędne, które charakteryzują się najmniejszą zmiennością.

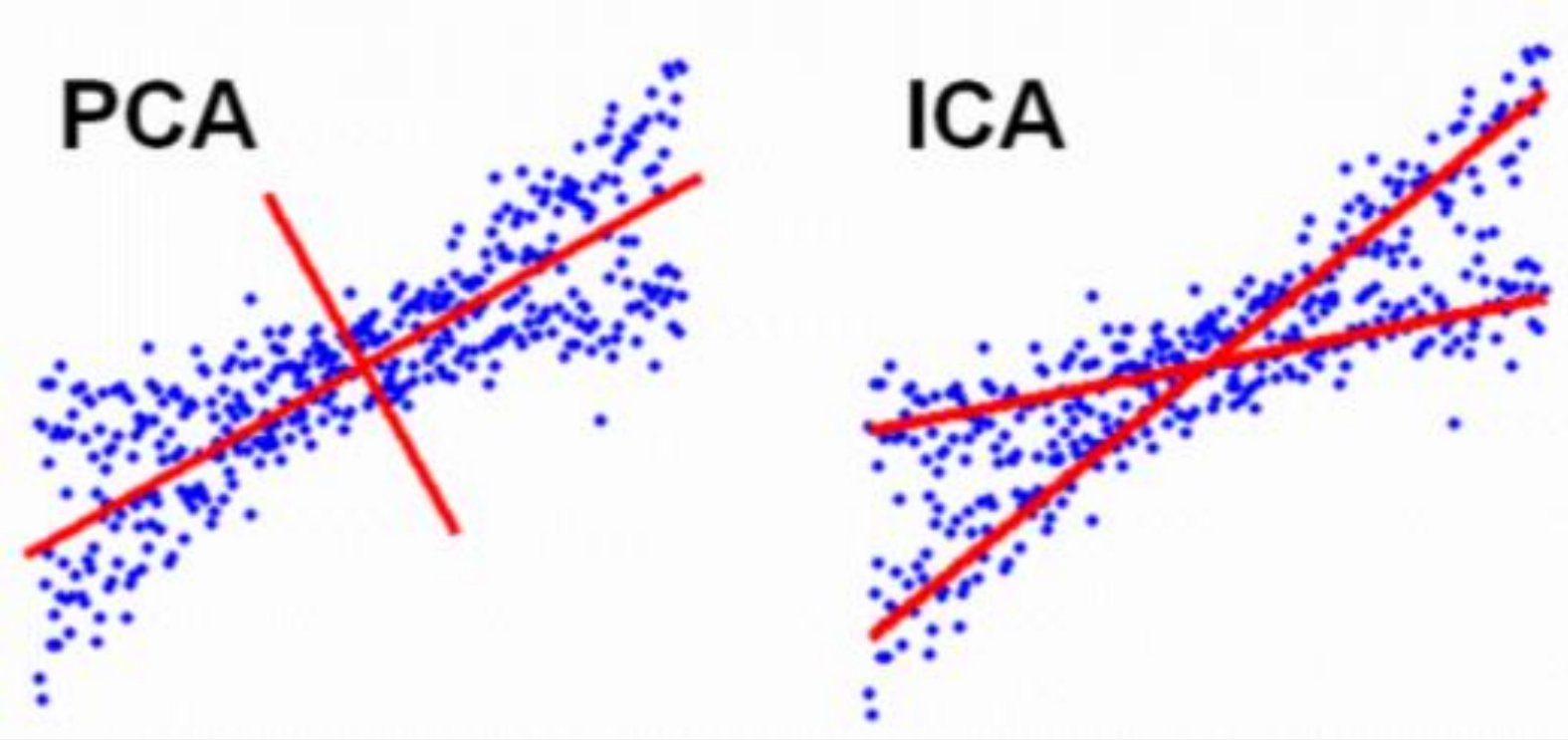


ICA – INDEPENDENT COMPONENT ANALYSIS

ICA – to statystyczna metoda podobna do PCA, które zadaniem jest odnalezienie współrzędnych niezależnych opisujących dane o największej zmienności (wariancji).

ICA również umożliwia redukcję wymiaru danych.

Daje zwykle lepsze wyniki niż PCA.



ICA – ALGORYTM

Szybki ICA algorytm wykorzystujący koncepcję negentropii:

1. Wypośrodkuj/Przesuń dane x, w taki sposób, żeby ich średnia była równa zeru:

$$x = x - x_m \qquad x_m = E\{x\}$$

2. Wyczyść x żeby maksymalizować nie Gaussowskie charakterystyki (PCA z filtracją):

$$z = V \Lambda^{-1/2} V^T x \qquad V \Lambda V^T = E\{x\, x^T\}$$

3. Weź losowy wektor początkowy w, ||w|| =1

4. Aktualizuj w (maksymalnie w kierunku nie Gaussowskim)

$$w = E\{z * g(w^T z)\} - E\{g'(w^T z)\}\, w$$

$$g(y) = \tanh(a_1 y) \text{ lub } g(y) = y * \exp(-y^2/2) \qquad \text{gdzie } 1 < a_1 < 2$$

$$w = w / \|w\|$$

5. Jeśli nie jest zbieżne wróć do punktu 4.

6. Uzyskaj niezależną współrzędną s:

7. $s = [\, w_1\ w_2 \ldots w_n \,]\, x$