Wstępne przetwarzanie i analiza danych

Marcin Szeląg

Zakład ISWD, Instytut Informatyki, Politechnika Poznańska

16.10.2019

Plan prezentacji

Pomiary i dane

Dane

Gromadzenie danych – odwzorowywanie obiektów z dziedziny zainteresowania na reprezentację symboliczną za pomocą pewnej **procedury pomiarowej**, która kojarzy wartości zmiennych z właściwościami obiektów.

Dane – c.d.

"Surowe" dane będące podstawą procesu odkrywania wiedzy często charakteryzują się następującymi cechami:

- błędy pomiarowe,
- brakujące wartości w danych,
- zniekształcenia podczas próbkowania,
- błędy przy wprowadzaniu danych,
- ...

Skale pomiarowe

- Skala porządkowa – np. stopień dolegliwości
- Skala ilorazowa – np. waga, cena
- Skala przedziałowa – np. temperatura w stopniach Celsjusza lub Fahrenheita, czas kalendarzowy
- Skala symboliczna (nominalna) – kolor włosów, wyznanie, miejsce zamieszkania

Istotność formułowania jedynie takich twierdzeń statystycznych odnośnie danych, dla których wartość "prawda" jest niezmiennikiem dozwolonych przekształceń skali pomiarowej (np. mediana vs. średnia dla skali porządkowej).

Pomiary i odwzorowania

Pomiary reprezentujące – służą reprezentowaniu związków w systemie empirycznym (zależności między wartościami zmierzonymi przekładają się na zależności między wartościami rzeczywistymi). Cel – zrozumienie (opisanie) tego, co się dzieje w systemie.

Pomiary niereprezentujące (operacyjne) – zarówno definiowanie właściwości, o którą chodzi, jak i przypisywanie do niej wartości, np. jakość życia w medycynie lub pracochłonność w inżynierii oprogramowania jako funkcje (agregaty) różnych składowych. Cel – przewidzenie tego, co się będzie działo w systemie.

Przekształcanie danych

Czasami dogodne jest zmodyfikowanie "surowych" danych przed ich analizą. Zauważmy, że między postacią modelu a naturą danych jest pewna dualność. Np. w przypadku regresji w celu sprawdzenia czy zmienna Y jest kwadratem zmiennej X można najpierw dokonać przekształcenia $U = X^2$, a następnie poszukiwać współczynników zależności liniowej pomiędzy Y i U.

Stosowane przekształcenia danych (pula funkcji):

- pierwiastek kwadratowy,
- odwrotność,
- logarytm,
- podnoszenie do potęgi całkowitej dodatniej,
- ...

Postać (schemat) danych

- Dane standardowe, w postaci macierzy (tabeli) o rozmiarze $m \times n$, gdzie m – liczba wierszy (rekordów, obserwacji, instancji, itp.), n – liczba kolumn (zmiennych, cech, atrybutów, pól, itp.).
- Dane wielorelacyjne – zbiór danych składający się z wielu powiązanych ze sobą tabel (relacji). Np. baza danych z listą płac może przechowywać dwie tabele – jedną z informacją o pracownikach (nazwisko, nazwa działu, wiek, pensja) i drugą z informacją o działach (nazwa działu, budżet, kierownik). Denormalizacja danych wielorelacyjnych (np. połączenie po nazwie działu) jest niepożądana – powoduje utratę elastyczności (wybranie a priori jednego z możliwych połączeń danych) oraz spadek efektywności – powtarzanie licznych wartości.

Postać (schemat) danych – c.d.

- Szeregi czasowe – kolejne wartości odpowiadają pomiarom w kolejnych punktach czasu – ważny jest porządek obserwacji (kolejność nie jest dowolna!).
- Ciąg znaków – sekwencja symboli pewnego skończonego alfabetu, np. tekst, sekwencja DNA/RNA. Typ uporządkowany, dla którego niekoniecznie odpowiednia jest standardowa postać macierzowa.
- Sekwencja zdarzeń – ciąg par postaci {zdarzenie, czas wystąpienia}, np. telekomunikacyjny dziennik alarmów.
- Struktura hierarchiczna – złożony schemat danych. Np. zbiór danych dzieci pogrupowany w klasy, które są pogrupowane w roczniki, które są pogrupowane w szkoły, itd.

Przekleństwo wymiarowości

Przekleństwo (klątwa) wymiarowości (*curse of dimensionality*) – w miarę wzrostu liczby wymiarów (zmiennych) liczba obiektów (obserwacji) potrzebnych do wiarygodnego oszacowania parametrów lub funkcji rośnie wykładniczo.

Metody radzenia sobie z przekleństwem wymiarowości:

- wybór podzbioru istotnych zmiennych o liczności $n' \ll n$,
- przekształcenie oryginalnych n zmiennych na nowy zbiór n', tak by $n' \ll n$.

Problem rozmiaru zbioru danych

Duża liczba przykładów trenujących, niezbędna do uzyskania w drodze indukcyjnego uczenia się dostatecznie dokładnej hipotezy (modelu), może stać się jednocześnie jedną z najpoważniejszych przeszkód.

Koszt obliczeniowy zależy co najmniej liniowo od rozmiaru przetwarzanego zbioru danych. Stąd też niekiedy zachodzi konieczność ograniczenia liczby przykładów uwzględnianych podczas uczenia się bądź też stosowania innych technik, np. okienkowania.

Ocena jakości danych

- GIGO (ang. garbage in – garbage out) – wprowadzanie błędnych danych powoduje uzyskiwanie błędnych wyników.
- Eksploracja danych często zajmuje się wtórną analizą danych.
- Na jakość danych składają się:
  - jakość pojedynczych rekordów i pól,
  - całościowa jakość zbioru danych.
- Źródła błędów pomiarowych:
  - niedbalstwo,
  - usterki przyrządów,
  - nieadekwatna definicja mierzonej wielkości,
  - niemożność wykonania dokładnego pomiaru.

Ocena jakości danych – c.d.

Rodzaje błędów pomiarowych:

- *nieprecyzyjność* – duża zmienność (wariancja) wartości mierzonej w powtarzanych pomiarach,
- *nieścisłość* = *obciążenie* – różnica pomiędzy wartością średnią z powtarzanych pomiarów a "wartością prawdziwą".

Niekiedy precyzję określamy terminem *wiarygodność*. Z kolei *ścisła* (poprawna) procedura pomiarowa charakteryzuje się nie tylko małą zmiennością, ale daje również wyniki bliskie temu, co uważamy za "wartość prawdziwą" zmiennej (ma małe obciążenie).

Poprawność procedury pomiarowej jest wymagana w celu wyciągania prawdziwych wniosków co do rzeczywistości opisanej przez dane.

Ocena jakości danych – c.d.

Jakość zbioru danych:

- Wnioskowanie o populacji na podstawie *reprezentatywnej* próby losowej
- Statystyka – szacowanie wartości parametrów rozkładu na podstawie statystyk (wartości obliczonych z próby). Takie oszacowania są przydatne gdy są dokładne.
- Wraz ze wzrostem rozmiaru próby rośnie precyzja i spada obciążenie szacowania wartości parametrów rozkładu.
- Obciążenie wynikające z zastosowania niewłaściwej (zniekształconej) próby, np. próba pracowników biurowych w Nowym Jorku w celu oszacowania przeciętnej wagi ludzi mieszkających w Nowym Jorku czy też wybór klientów z wykorzystaniem szeregu kroków selekcyjnych.
- W eksploracji danych, w przeciwieństwie do statystyki, często nie mamy kontroli nad sposobem gromadzenia danych $\Rightarrow$ zbiór danych jest próbą *możliwą* (*dogodną*), stąd konieczność ostrożnego formułowania wniosków z analizy.

Ocena jakości danych – c.d.

Jakość zbioru danych:

- Predykcja przyszłości a *dryft populacji* – np. rodzaj zakupów dokonywanych przez klientów w pewnym sklepie może się zmieniać z czasem, np. z powodu zmian w kulturze społecznej okolicznego sąsiedztwa czy w odpowiedzi na inicjatywy handlowe (populacja nie jest statyczna!).
- Czasami dla bardzo dużych zbiorów danych możemy być zmuszeni do samodzielnego właściwego spróbkowania zbioru.
- Zniekształcenie próbek może być traktowane jako szczególny przypadek danych niekompletnych – brakuje całych rekordów aby próba była reprezentatywna.
- Gdy w rekordach brakuje pojedynczych pól, przynajmniej widać, że brakuje danych.

Ocena jakości danych – c.d.

Jakość zbioru danych:

- Obserwacje (punkty) oddalone (ang. outliers) – przedmiot zainteresowania przy wykrywaniu anomalii i usterek (poszukujemy wzorców opisujących te anomalie). Przy budowaniu globalnego modelu wspomagającego zrozumienie danych, punkty oddalone mogą utrudniać zadanie (np. w przypadku metody najmniejszych kwadratów w regresji liniowej). Stąd też niekiedy zachodzi konieczność ich identyfikacji i usunięcia przed budową modelu.
- Ponieważ osoby wydobywające wiedzę z danych rzadko kontrolują gromadzenie tych danych, istotna jest świadomość zagrożeń wynikających ze słabej jakości danych.

"Dane kiepskiej jakości utrudniają klarowne myślenie i racjonalne podejmowanie decyzji. Dane obciążone, i wywodzone z nich zależności, mogą mieć poważne konsekwencje, jeśli chodzi o formułowanie praw i reguł".

Wstępne przetwarzanie danych

Wstępne przetwarzanie danych

Wstępne przetwarzanie danych obejmuje czyszczenie i przekształcanie danych w celu ich przygotowania do eksploracji.

Szacuje się, że wstępne przetwarzanie danych to 70-80% procesu odkrywania wiedzy.

Motywacje wstępnego przetwarzania (obróbki) danych są następujące:

- Eksploracja danych często zajmuje się danymi, które nie były używane od lat $\Rightarrow$ "przeterminowane" (zbędne lub przestarzałe) wartości
- Brakujące wartości (ang. missing values)
- Obecność punktów oddalonych (ang. outliers)
- Nieodpowiedni format danych
- Błędne dane – niezgodne z przyjętymi zasadami, wiedzą dziedzinową lub zdrowym rozsądkiem. Błędy mogą wynikać np. z niedoskonałości procedur pomiarowych, ze skończonej liczby miejsc dziesiętnych w reprezentacji liczb, itd.

Przykładowy zbiór danych do czyszczenia

ID klienta	Kod pocztowy	Płeć	Dochód	Wiek	Stan cywilny	Kwota transakcji
1001	10048	M	75 000	*C*	M	5000
1002	J2S7K7	K	−40 000	40	W	4000
1003	90210		10 000 000	45	S	7000
1004	6269	M	50 000	0	S	1000
1005	55101	K	99 999	30	R	3000

Przykładowy zbiór danych do czyszczenia – c.d.

ID klienta	Kod pocztowy	Płeć	Dochód	Wiek	Stan cywilny	Kwota transakcji
1001	10048	M	75 000	*C*	M	5000
1002	J2S7K7	K	−40 000	40	W	4000
1003	90210		10 000 000	45	S	7000
1004	6269	M	50 000	0	S	1000
1005	55101	K	99 999	30	R	3000

Zaobserwowane problemy w przykładowym zbiorze danych:

- kod pocztowy klienta 1002, który nie jest typowym kodem pięciocyfrowym w USA – kod St. Hyancinthe w Quebecu (Kanada)
- kod pocztowy klienta 1004 – prawdopodobnie brakuje początkowego zera (pole numeryczne) – kod pocztowy z regionu Nowej Anglii

Przykładowy zbiór danych do czyszczenia – c.d.

ID klienta	Kod pocztowy	Płeć	Dochód	Wiek	Stan cywilny	Kwota transakcji
1001	10048	M	75 000	*C*	M	5000
1002	J2S7K7	K	−40 000	40	W	4000
1003	90210		10 000 000	45	S	7000
1004	6269	M	50 000	0	S	1000
1005	55101	K	99 999	30	R	3000

Zaobserwowane problemy w przykładowym zbiorze danych:

- brakuje płci klienta 1003
- dochód roczny brutto ma potencjalnie 3 nieprawidłowe wartości:
  - klient 1002 – dochód ujemny, który musi być błędem,
  - klient 1003 – obserwacja odstająca, chociaż możliwa (kod pocztowy 90210 wskazuje na Beverly Hills),
  - klient 1005 – potencjalnie 99 999$ może być kodem dla niewłaściwie wprowadzonych danych, np. dla wartości brakującej (pozostałe dochody są zaokrąglone do 5000$),

Przykładowy zbiór danych do czyszczenia – c.d.

ID klienta	Kod pocztowy	Płeć	Dochód	Wiek	Stan cywilny	Kwota transakcji
1001	10048	M	75 000	*C*	M	5000
1002	J2S7K7	K	−40 000	40	W	4000
1003	90210		10 000 000	45	S	7000
1004	6269	M	50 000	0	S	1000
1005	55101	K	99 999	30	R	3000

Zaobserwowane problemy w przykładowym zbiorze danych:

- wiek klienta 1001 – wartość pola C najprawdopodobniej odnosi się do wcześniejszej kategoryzacji, gdzie przypisano danemu klientowi klasę C,
- wiek klienta 1004 – prawdopodobnie wiek nieznany, zakodowany przez 0 (np. klient, który odmówił podania wieku),
- znaczenie symboli w kolumnie stan cywilny – problem niejednoznaczności i wieloznaczności (np. czy S oznacza samotny czy w separacji?).

Brakujące wartości w danych

Problem stale pojawiający się w eksploracji danych. Metody radzenia sobie z tym problemem są następujące:

- Usunięcie rekordów zawierających brakujące dane – ryzyko usunięcia dużej ilości danych ("marnotrawienie danych"), niebezpieczeństwo usunięcia wzorców (brakujące dane mogą pojawiać się nie bez powodu!)

Brakujące wartości w danych – c.d.

- Wstawienie wartości zastępczych, dobranych na różne sposoby ("produkowanie" / "fabrykowanie" danych):
  - wartość stała dobrana przez analityka,
  - moda lub średnia wartość atrybutu,
  - moda lub średnia wartość atrybutu, znajdowana na podstawie rozkładu wartości wśród przykładów należących tylko do tej samej klasy decyzyjnej co analizowany przykład,
  - wartość losowa z obserwowanego rozkładu zmiennej w próbce $\Rightarrow$ po uzupełnieniu danych miary środka i rozrzutu będą zbliżone do tych miar dla danych oryginalnych,
  - zbiór wszystkich możliwych wartości danego atrybutu,
  - podzbiór wartości atrybutu wraz z informacją o stopniach możliwości ich realizacji,
  - estymacja wartości najbardziej prawdopodobnej, przy danych pozostałych wartościach zmiennych (np. estymacja bayesowska, regresja, drzewa i reguły decyzyjne).

Brakujące wartości w danych – c.d.

Bez nałożenia dodatkowych ograniczeń na proces generowania wartości zastępczych, istnieje ryzyko wprowadzenia błędów w danych (nietypowe współwystąpienia wartości zmiennych).

Niektóre algorytmy eksploracji danych mogą radzić sobie z wartościami brakującymi na bieżąco, w trakcie działania (np. algorytmy generujące reguły decyzyjne) – uwzględnianie brakujących wartości zintegrowane z algorytmem odkrywania wiedzy.

Wykrywanie błędnych klasyfikacji (kategorii)

W przypadku zmiennych jakościowych warto wykonać histogramy w celu identyfikacji kategorii (klas) o małej liczności, które mogą sugerować błędne / niekonsekwentne klasyfikacje.

Nazwa	Liczność
USA	1
Francja	1
Stany Zjednoczone	156
Europa	46
Japonia	51

Przekształcanie danych

Zmienne numeryczne mają na ogół zakresy, które różnią się od siebie. Dla pewnych algorytmów eksploracji danych różnice zakresów będą powodować, iż zmienne z większym zakresem będą miały większy wpływ na wyniki eksploracji. Stąd też konieczne może być znormalizowanie zmiennych:

- Normalizacja min-max:

$$X^* = \frac{X - min(X)}{max(X) - min(X)}$$

  Znormalizowane wartości należą do przedziału $[0, 1]$ chyba, że zostaną napotkane nowe wartości danych, które leżą poza początkowym przedziałem.

- Standaryzacja:

$$X^* = \frac{X - \mu(X)}{\sigma(X)}$$

  Wartości mniejsze (odpowiednio większe) od średniej po standaryzacji będą miały wartość ujemną (odpowiednio dodatnią). Z reguły wartość zmiennej po standaryzacji należy do przedziału od -4 do 4, ze średnią 0.

Graficzne metody identyfikacji punktów oddalonych

Punkty (obserwacje) oddalone (ang. ouliers) – wartości skrajne położone blisko granic zakresu danych lub sprzeczne z ogólnym trendem pozostałych danych.

Punkty oddalone mogą potencjalnie reprezentować błędy w danych. Niektóre metody i wskaźniki statystyczne są wrażliwe na obecność tych punktów $\Rightarrow$ niestabilność wyników eksploracji dla oryginalnych danych.

Metody identyfikacji obserwacji oddalonych (zmienne numeryczne):

- histogramy,
- wykresy rozrzutu (2D, 3D),
- wykresy pudełkowe (ramkowe) (ang. box plot).

Numeryczne metody identyfikacji punktów oddalonych

- Standaryzacja + oddalenie od wartości średniej o więcej niż $\pm 3\sigma$ – słaba metoda, gdyż punkty oddalone wpływają na średnią i odchylenie standardowe
- Statystyczne metody wykrywania punktów oddalonych mniej wrażliwe na obecność tych punktów, np. *rozstęp międzykwartylowy* $IQR = Q_3 - Q_1$ (środkowe 50% danych); punkty oddalone – wartość $\leq Q_1 - 1.5IQR$ lub $\geq Q_3 + 1.5IQR$ (poza "wąsami" na wykresie pudełkowym)

Przypomnienie wybranych wiadomości ze statystyki

Kowariancja:

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Kowariancja bada zgodność dwóch "sygnałów". Wartość $cov(x,y) \in (-\infty, \infty)$.

Współczynnik korelacji liniowej Pearsona:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{cov(x,y)}{\sigma(x)\sigma(y)}$$

Korelacja Pearsona wykrywa tylko zależności liniowe. Im $|r_{xy}|$ jest bliższe 1, tym większa zależność liniowa między x i y (prosta lub odwrotna). Z kolei duża zależność wcale nie musi oznaczać dużej korelacji (zależności nieliniowe!). Wartość $r_{xy} \in [-1, 1]$.

Dla metody najmniejszych kwadratów współczynniki a i b można wyznaczyć analitycznie. Niech dane będzie n obserwacji x_i, y_i. Wówczas rozwiązaniem problemu optymalizacji

$$\min_{a,b} \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

są wartości:

$$a = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2},$$

$$b = \bar{y} - a\bar{x}.$$

Eksploracyjna analiza danych
(ang. EDA – Exploratory Data Analysis)

Eksploracyjna analiza danych

Celem eksploracyjnej analizy danych (ang. EDA – Exploratory Data Analysis) jest zrozumienie i wizualizacja zbioru danych w celu postawienia hipotez dotyczących zależności obecnych w tych danych. Zatem w przeciwieństwie do statystycznego testowania hipotez, hipotezy nie są dostępne a priori i jedynie weryfikowane na zbiorze danych, lecz tworzone podczas analizy.

EDA pozwala na:

- pogłębienie wiedzy o zbiorze danych,
- sprawdzenie wzajemnych relacji pomiędzy atrybutami,
- identyfikację ciekawych podzbiorów danych do dalszej analizy,
- rozwinięcie wstępnej idei możliwych powiązań pomiędzy atrybutami i zmienną celu, jeżeli takie istnieją.

Eksploracyjna analiza danych – c.d.

- Pierwszy krok EDA to poznanie zbioru zmiennych (semantyka, dziedzina), na których opisane są dane oraz przyjrzenie się kilku / kilkunastu przykładowym rekordom danych.
- Dzięki znajomości znaczenia zmiennych można wyrobić sobie wstępny pogląd co do wzajemnych powiązań (korelacji) pomiędzy tymi zmiennymi.
- Obecność zmiennych skorelowanych może spowodować wyolbrzymienie jakiejś części danych lub nawet niestabilność modelu danych. Stąd też, powinno się unikać dostarczania do eksploracji danych zmiennych skorelowanych.

Korelacja pomiędzy zmiennymi

Wykrywanie korelacji pomiędzy zmiennymi:

- zmienne ilościowe (numeryczne) – wykresy rozrzutu dla par zmiennych, wyznaczenia wartości współczynnika korelacji liniowej r_{xy}; gdy widoczna korelacja, można pokusić się o wyznaczenie wartości współczynników prostej postaci $y = ax + b$ w celu określenia zaobserwowanej zależności,
- zmienne jakościowe (symboliczne) – wzorce w tablicach kontyngencji (test χ^2).

Należy usunąć z oryginalnego zbioru danych po jednej zmiennej z każdej pary skorelowanych zmiennych.

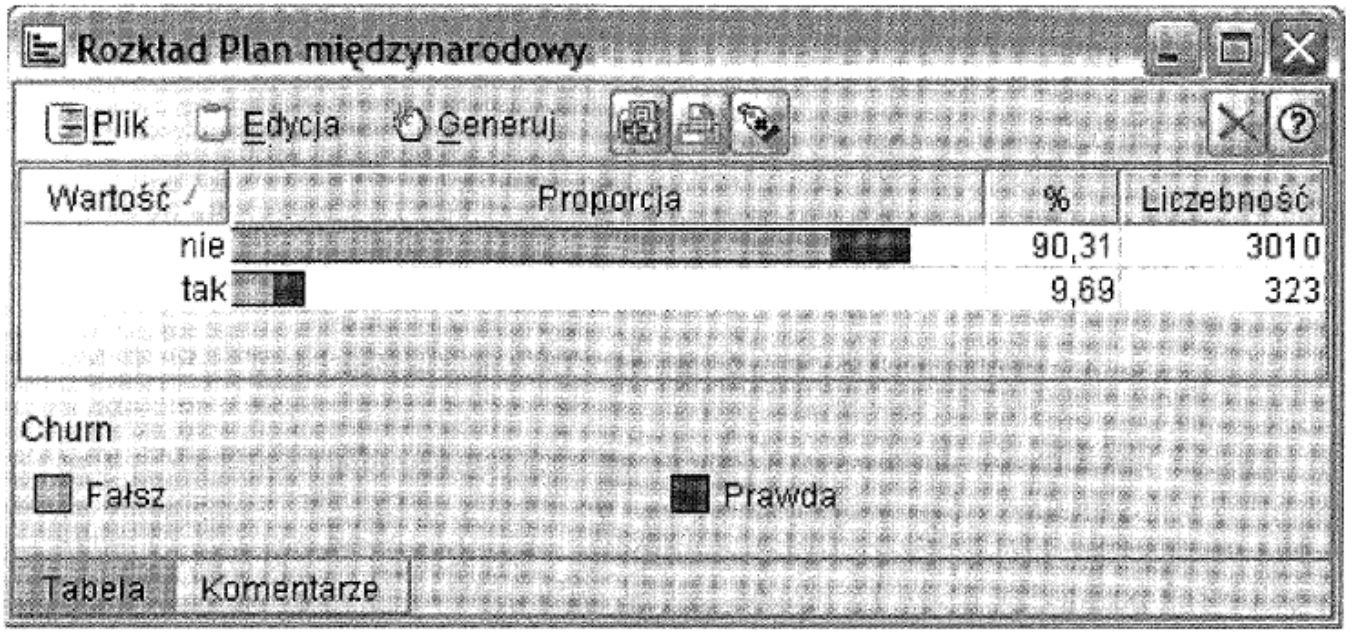
Badanie zmiennych jakościowych

Badanie zależności pomiędzy zmiennymi jakościowymi a kategoryczną zmienną celu:

- wykresy słupowe,
- wykresy słupkowe o równej długości słupków – lepsze przedstawienie proporcji,
- tablice kontyngencji – ilościowe określanie relacji pomiędzy daną zmienną jakościową a zmienną celu,
- badanie dwukierunkowych relacji między zmiennymi jakościowymi względem zmiennej celu (wykresy warunkowe).

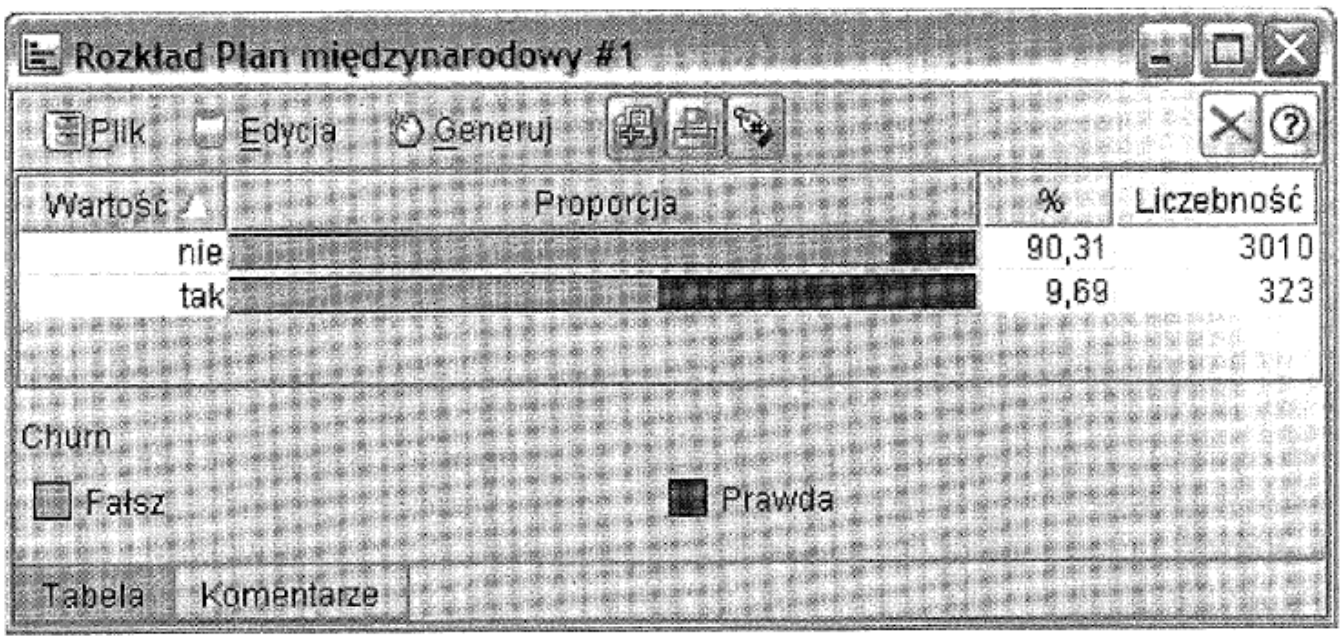
Badanie zmiennych jakościowych – przykład

Wykres słupkowy



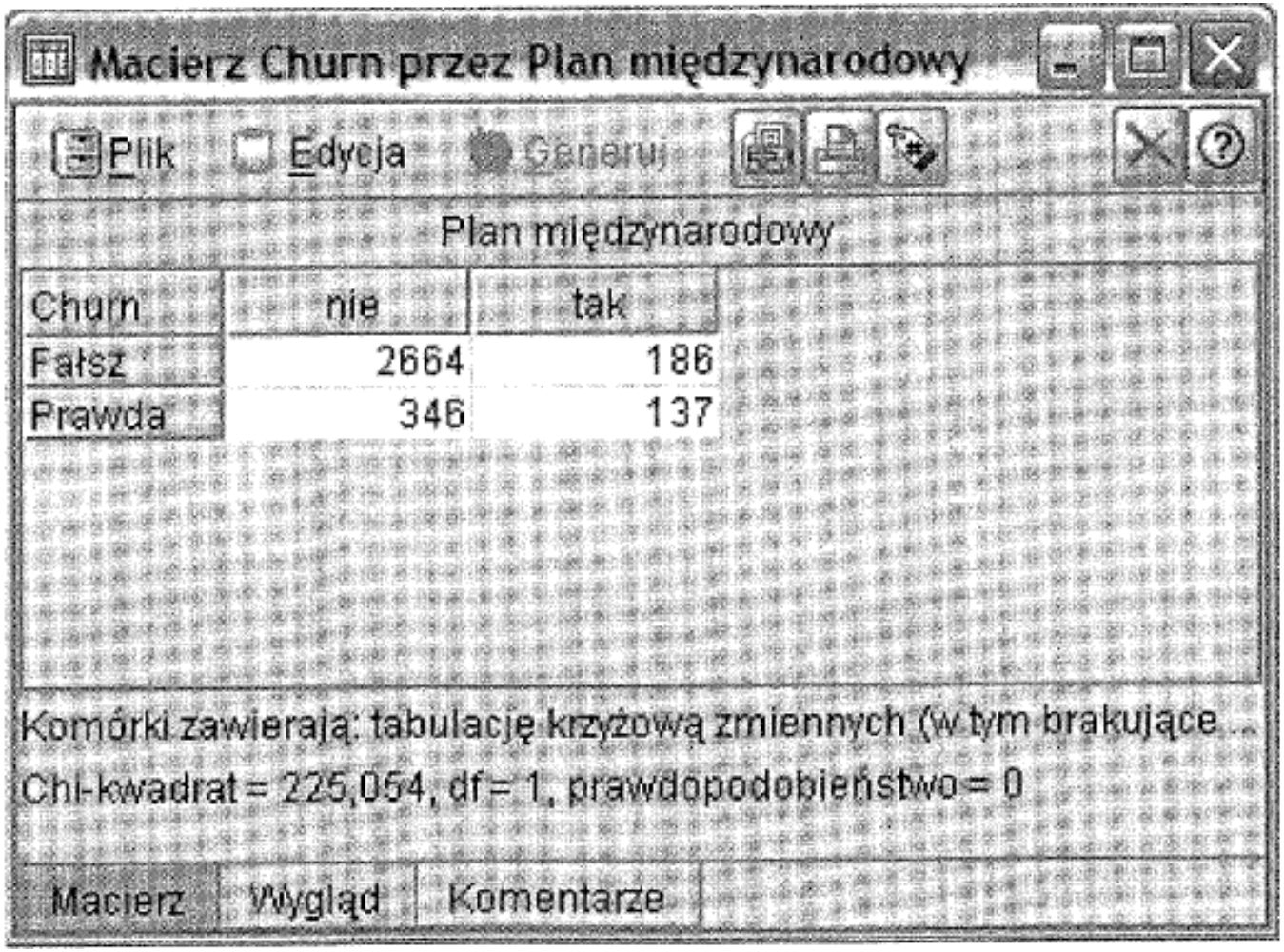
Badanie zmiennych jakościowych – przykład

Wykres słupkowy o równej długości słupków



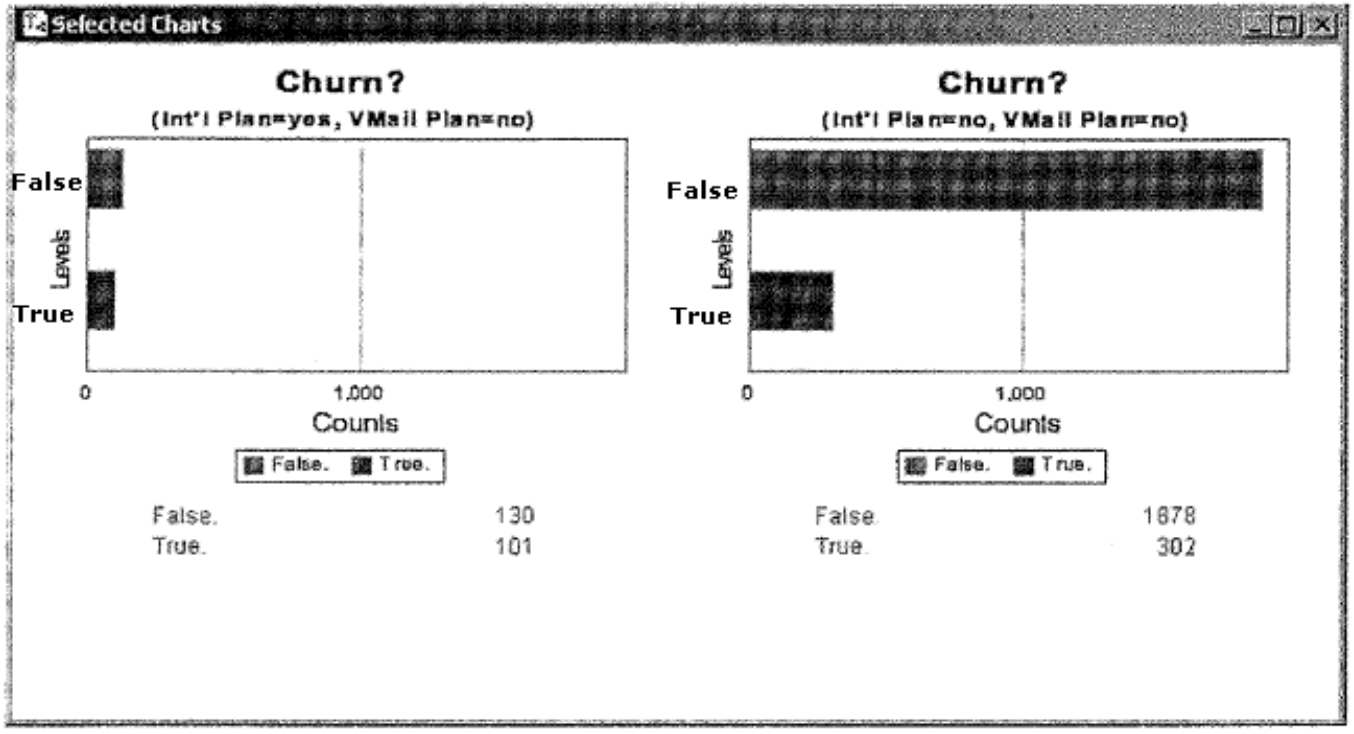
Badanie zmiennych jakościowych – przykład

Tablica kontyngencji



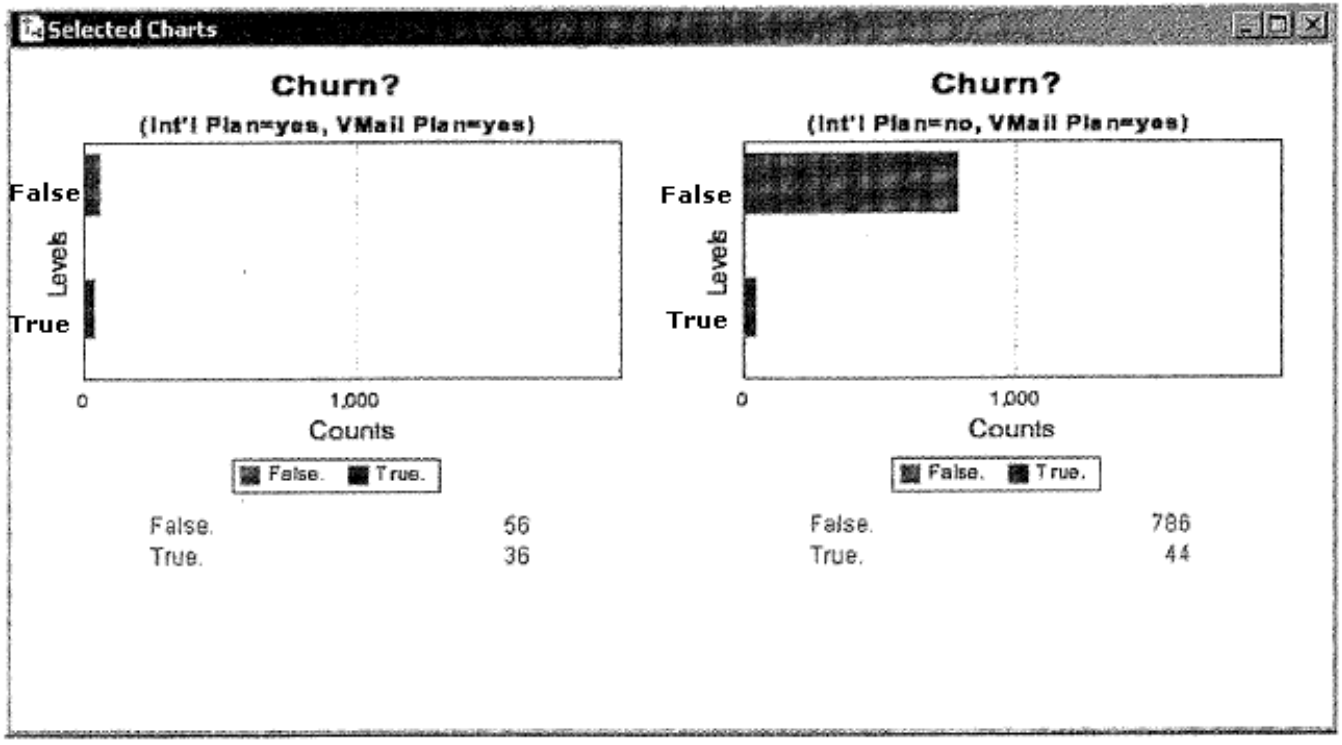
Badanie zmiennych jakościowych – przykład

Wykres warunkowy (klienci bez planu poczty głosowej)



Badanie zmiennych jakościowych – przykład

Wykres warunkowy (klienci z planem poczty głosowej)



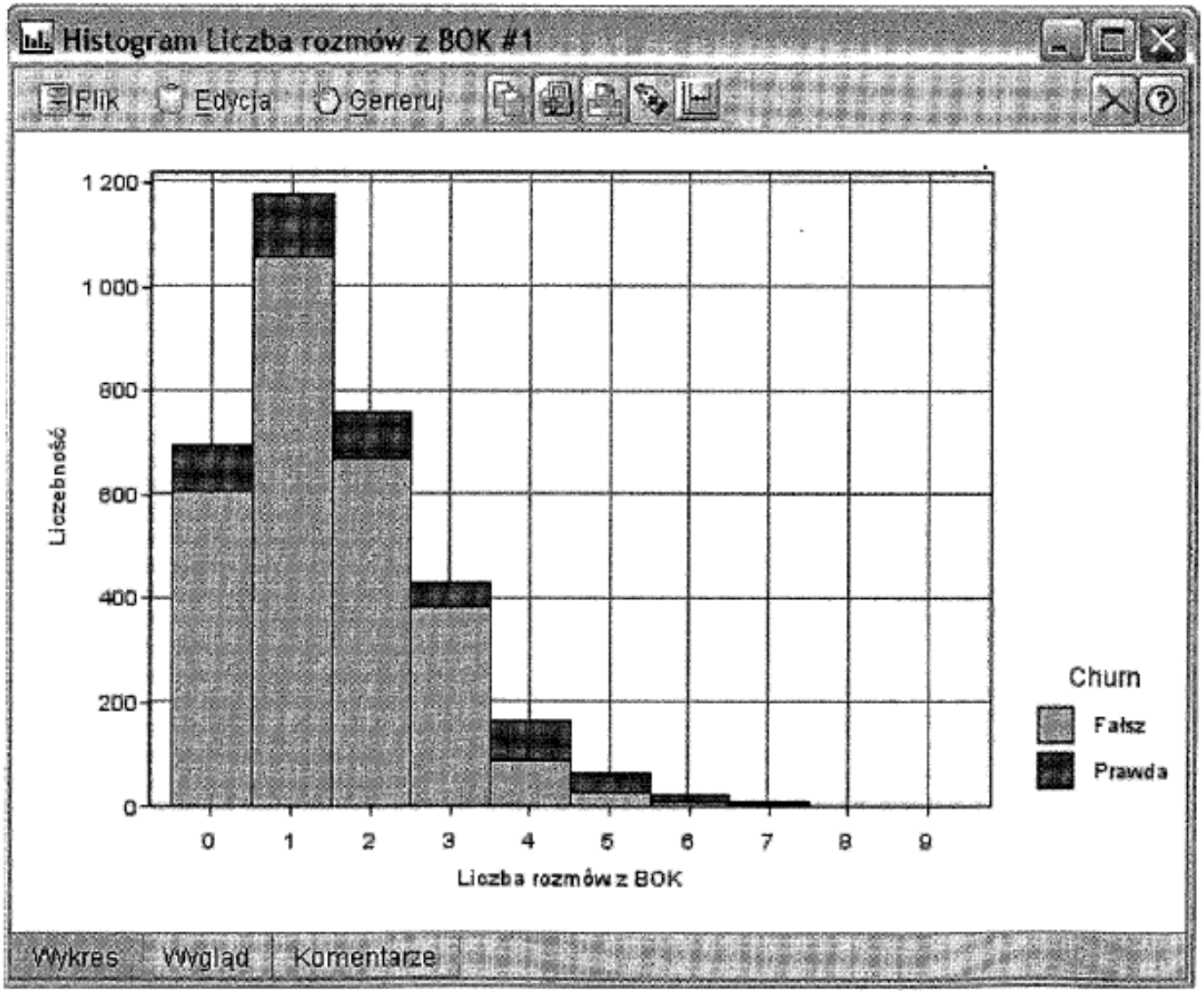
Badanie pojedynczych zmiennych ilościowych

Podsumowywanie każdej zmiennej ilościowej (numerycznej):

- min, max, miary środka (położenia): średnia, mediana i moda (dominanta), Q_1 – pierwszy (dolny) kwartyl, Q_3 – trzeci (górny) kwartyl, miary zmienności: wariancja, odchylenie standardowe, odchylenie przeciętne, rozstęp międzykwartylowy $IQR = Q_3 - Q_1$, rozstęp = max - min, skośność (asymetria) $= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^{\frac{3}{2}}}$,
- histogram – dobór przedziałów, możliwość skontrolowania danych, np. pod kątem obserwacji oddalonych; wygładzanie histogramu za pomocą estymacji jądrowej; możliwość nałożenia rozkładu kategorycznej zmiennej celu, a następnie normalizacji histogramu,
- wykres pudełkowy.

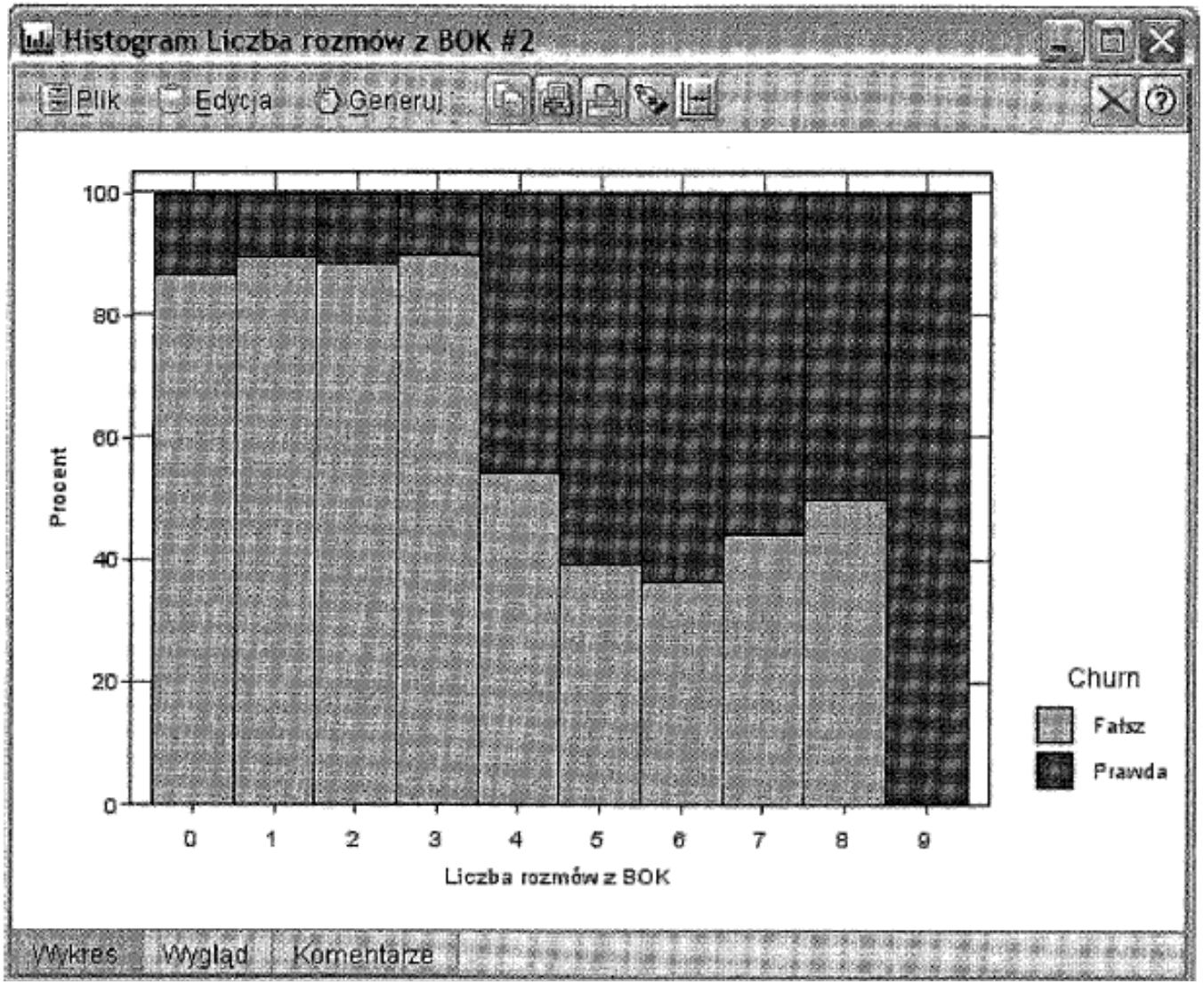
Badanie pojedynczych zmiennych ilościowych

Histogram z nałożonym rozkładem kategorycznej zmiennej celu



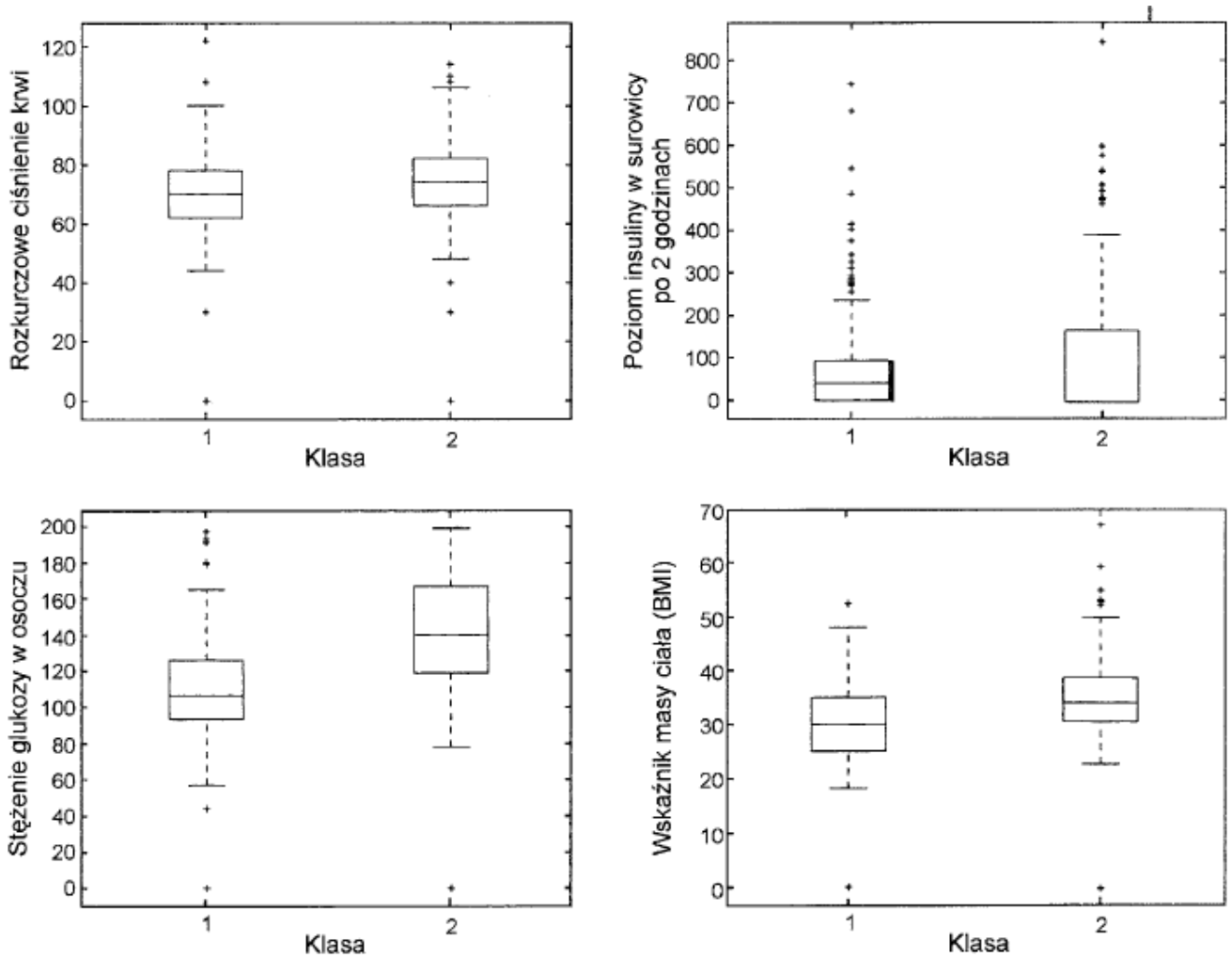
Badanie pojedynczych zmiennych ilościowych

Znormalizowany histogram z nałożonym rozkładem kategorycznej zmiennej celu



Badanie pojedynczych zmiennych ilościowych

Wykres pudełkowy (1 - cukrzycy, 2 - zdrowi)



Wygładzanie histogramu za pomocą estymacji jądrowej

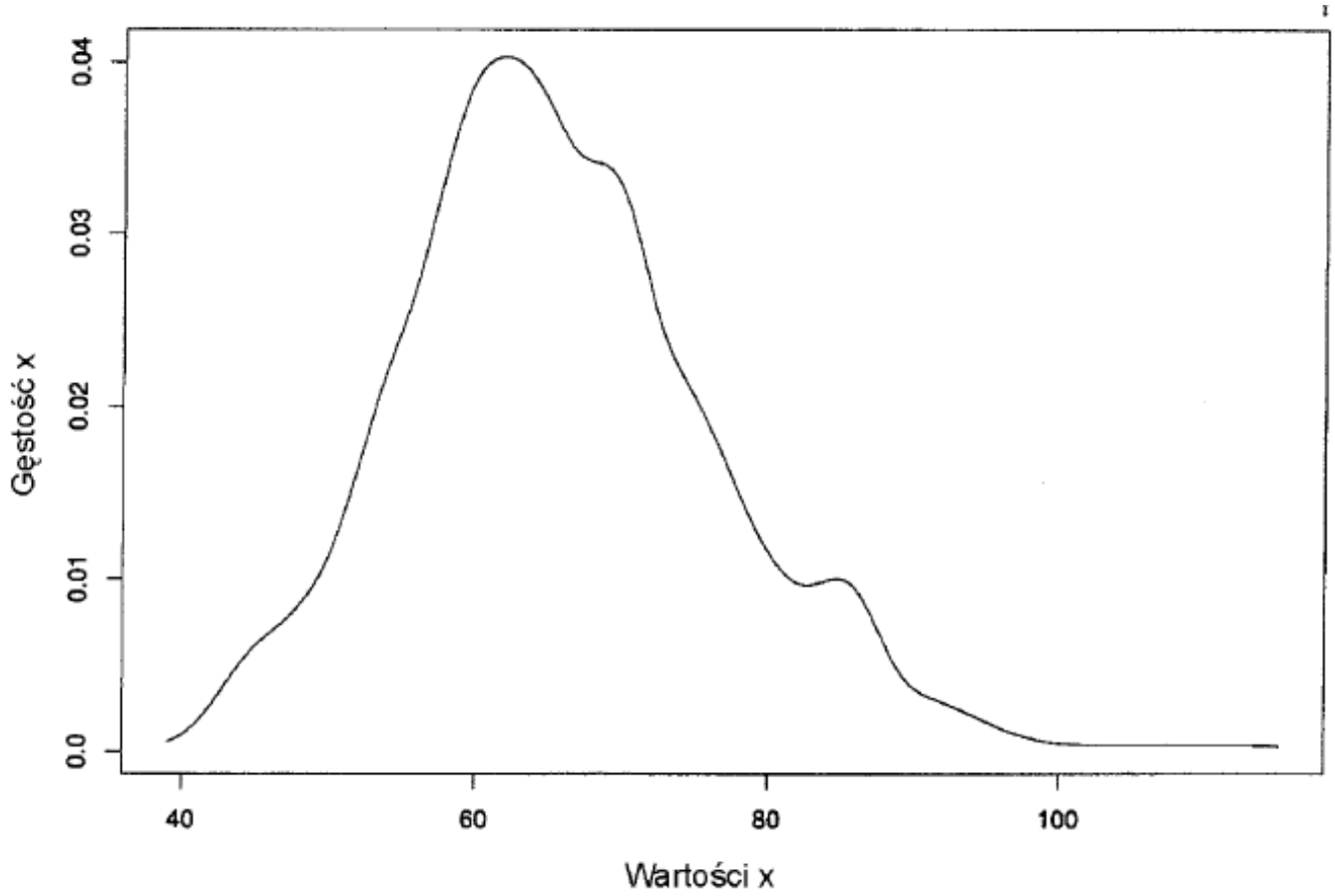
Idea – wyrównywanie wkładu każdego obserwowanego punktu danych z lokalnym sąsiedztwem tego punktu. Dla pojedynczej zmiennej X, o wartościach ze zbioru $\{x_1, x_2, \ldots, x_n\}$, wkład punktu danych x_i w szacowanie punktu x zależy od odległości między x_i i x:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K(\tfrac{x-x_i}{h}),$$

gdzie: K – funkcja jądrowa o szerokości pasma h taka, że $\int K(t)dt = 1$, w celu zapewnienia, że szacowanie $\hat{f}(x)$ całkuje się do 1. Na funkcję jądrową K wybiera się zazwyczaj gładką jednomodalną funkcję z wierzchołkiem w zerze, np. krzywą normalną $K(t,h) = Ce^{-\frac{1}{2}(\frac{t}{h})^2}$, gdzie C – stała normująca, $t = x - x_i$, h – odchylenie standardowe. Jakość estymacji jądrowej zależy mniej od kształtu K, a bardziej od wartości parametru h. Małe wartości h powodują "kolczastość" oszacowania gęstości $\hat{f}(x)$, natomiast duże wartości h powodują nadmiernie wygładzenie $\hat{f}(x)$.

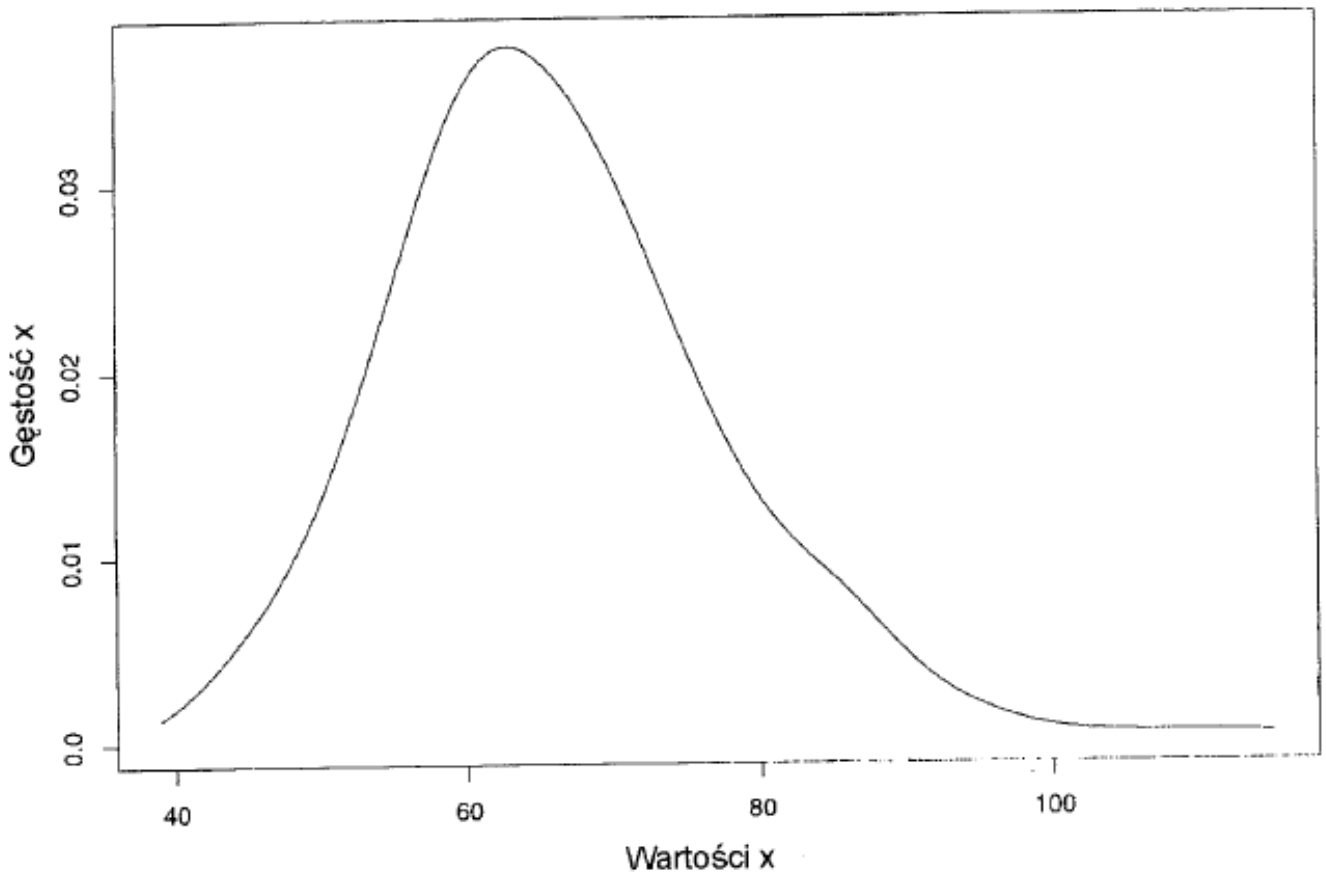
Wygładzanie histogramu za pomocą estymacji jądrowej

Estymacja jądrowa wagi (w kg) 856 kobiet



Wygładzanie histogramu za pomocą estymacji jądrowej

Bardziej wygładzona estymacja jądrowa wagi (w kg) 856 kobiet



Narzędzia do przedstawiania dwóch zmiennych ilościowych

Do przedstawiania zależności pomiędzy dwoma zmiennymi numerycznymi służą m.in.:

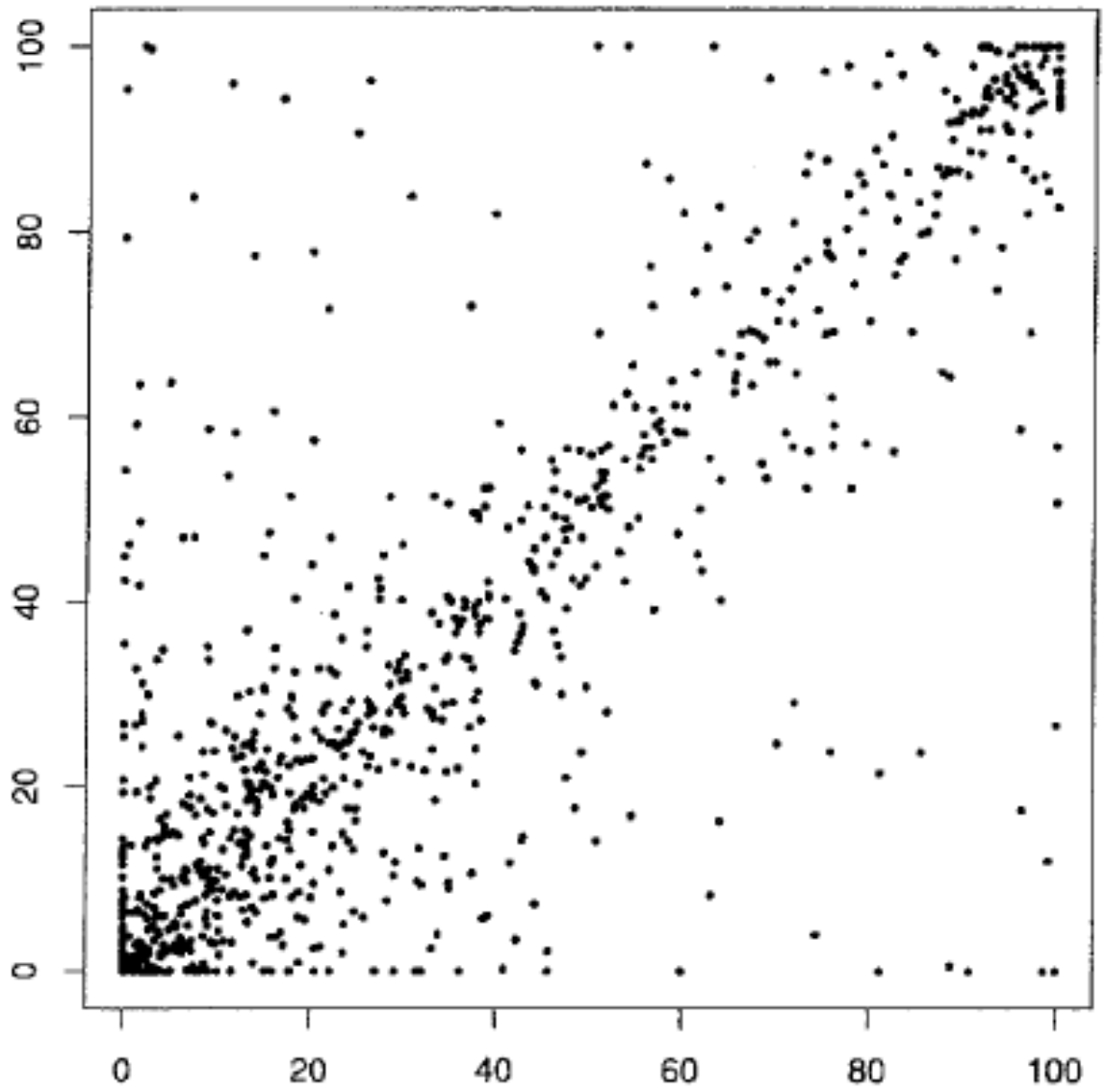
- (dwuwymiarowe) wykresy rozrzutu,
- wykresy warstwicowe.

Wykres warstwicowy może okazać się przydatny w sytuacji gdy wykres rozrzutu jest nieczytelny ze względu na dużą liczbę nakładających się obserwacji (tzw. naddrukowanie).

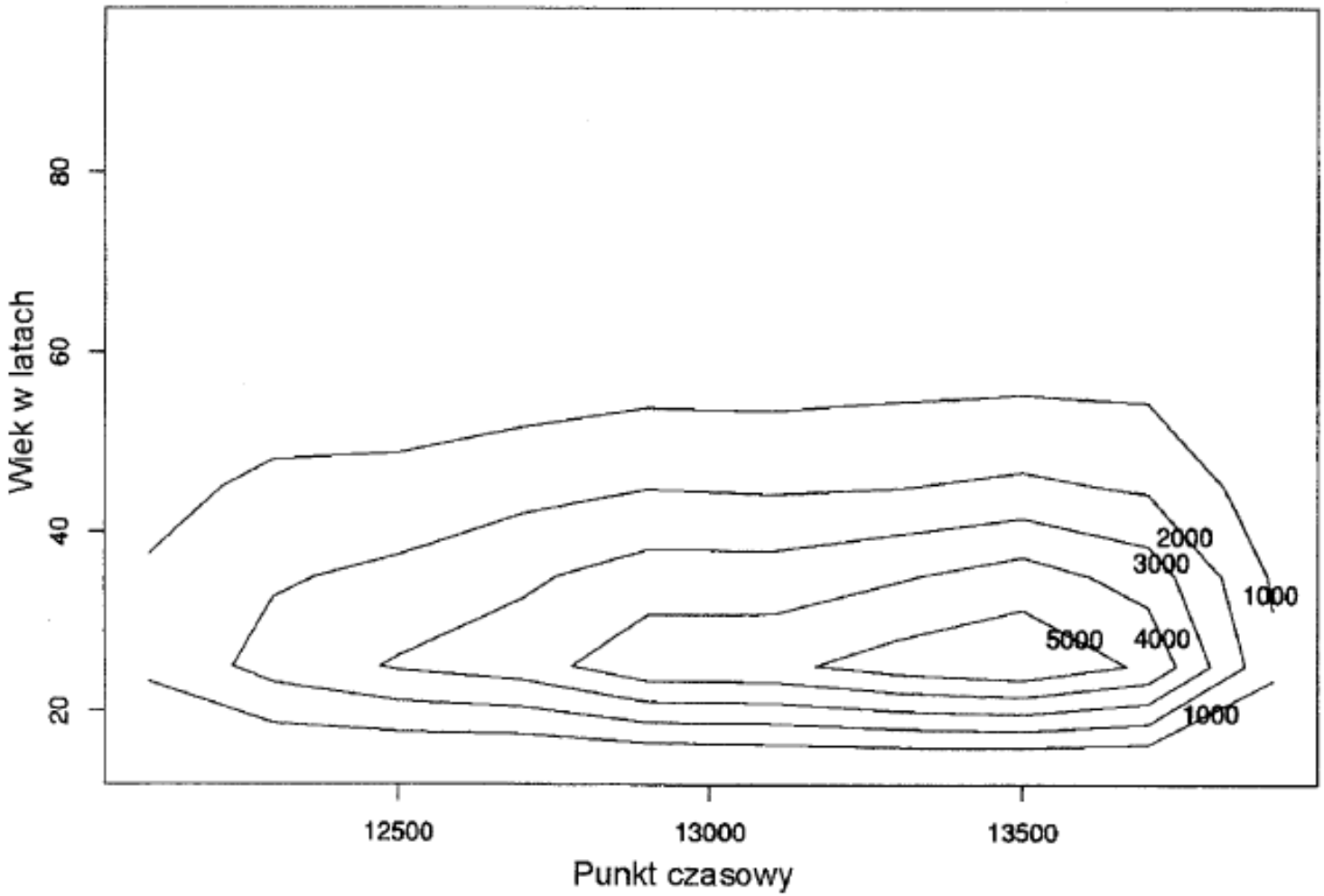
W celu stworzenia wykresu warstwicowego w dwóch wymiarach konieczne jest skonstruowanie estymacji dwuwymiarowej gęstości $\hat{f}(x, y)$ (dwuwymiarowe uogólnienie metody estymacji jądrowej).

Narzędzia do przedstawiania dwóch zmiennych ilościowych

Przykład wykresu rozrzutu dla 2 zmiennych bankowych



Przykładowy wykres warstwicowy

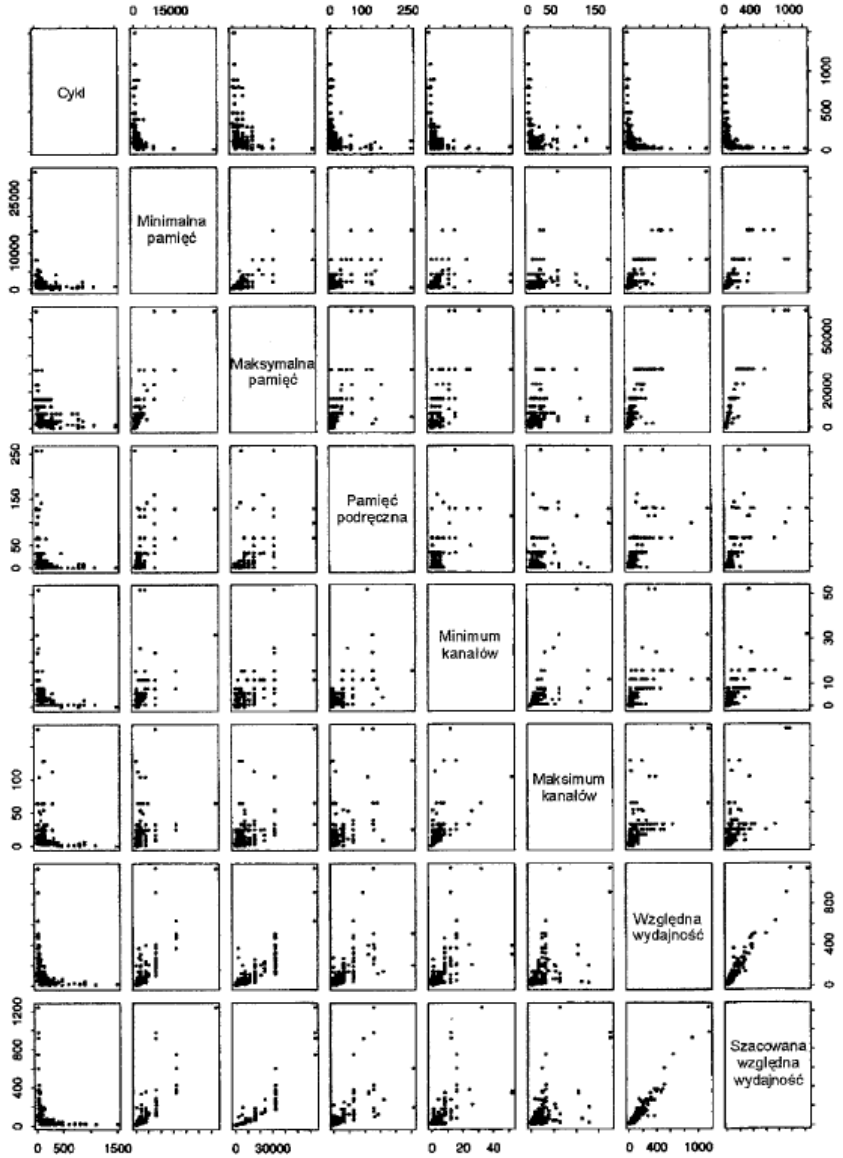


Narzędzia do przedstawiania > 2 zmiennych ilościowych

- Macierz wykresów rozrzutu. Rozwiązanie "wielokrotnie dwuwymiarowe" – rzutowanie danych wielowymiarowych na wiele wykresów dwuwymiarowych (z ignorowaniem części wymiarów).
- Trójwymiarowe wykresy rozrzutu jako obrazy interaktywne – możliwość automatycznego lub ręcznego obracania wykresu i zmiany kąta spoglądania na wykres.
- Wykresy warunkowe – ustalenie pary zmiennych i tworzenie serii diagramów składowych (np. wykresów rozrzutu, histogramów, itd.) w zależności od poziomów jednej lub więcej innych zmiennych.

Narzędzia do przedstawiania > 2 zmiennych ilościowych

Macierz wykresów rozrzutu dla danych dot. różnych CPU



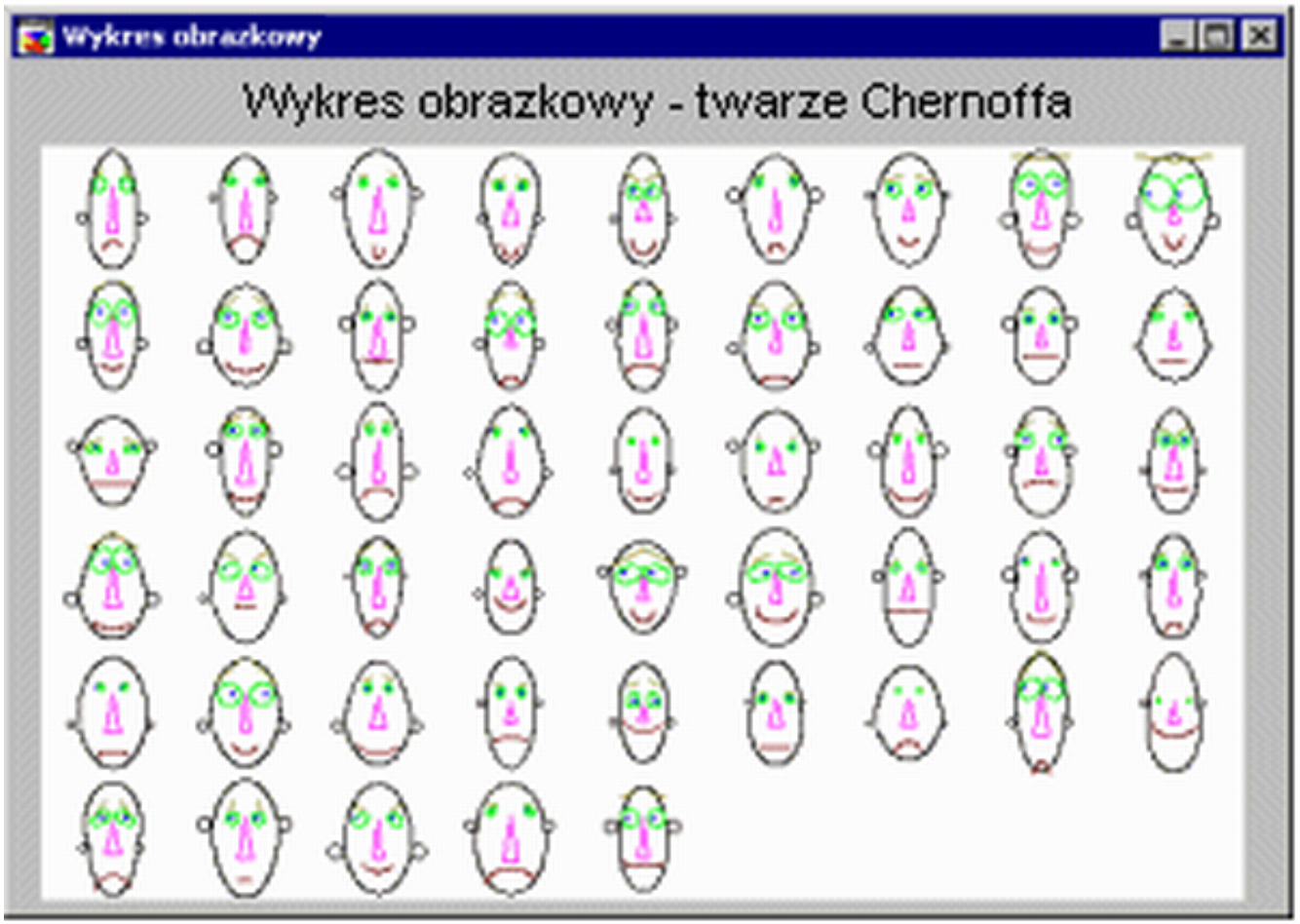
Narzędzia do przedstawiania > 2 zmiennych ilościowych

- Ikony – małe diagramy, w których rozmiary różnych cech są determinowane wartościami zmiennych; przykłady: ikony gwiazdowe, twarze Chernoffa; skuteczność dla stosunkowo małej liczby przykładów i zmiennych – konieczność obejrzenia każdego przykładu przez człowieka.
- Wykres współrzędnych równoległych – zmienne jako równoległe osie, a obiekty jako linie łamane, łączące zmierzone wartości cech obiektów; możliwość zastosowania kolorów i różnych stylów linii.

Narzędzia do przedstawiania > 2 zmiennych ilościowych

Przykładowy wykres gwiazdowy (12 własności chemicznych dla 53 próbek minerałów)

Twarze Chernoffa



Narzędzia do przedstawiania > 2 zmiennych ilościowych

Przykładowy wykres współrzędnych równoległych

Dyskretyzacja – wprowadzenie

Dyskretyzacja – proces zamiany atrybutów liczbowych na atrybuty symboliczne typu porządkowego. Polega na podziale oryginalnej dziedziny atrybutu liczbowego na pewną liczbę przedziałów i przypisaniu tym przedziałom kodów symbolicznych. W zasadzie dyskretyzacja jest tyle czynnością przygotowującą dane do dalszej analizy co eksplorującą te dane.

Dyskretyzacja jest wykonywana m. in. w celu:

- utworzenia histogramu,
- zastosowania algorytmów eksploracji wymagających zmiennych dyskretnych,
- uproszczenia danych kosztem pewnej utraty informacji,
- podziału próby na podpopulacje,
- wstępnego oglądu danych wielowymiarowych, np. przez zdyskretyzowanie jednej zmiennej i zbadanie jaką średnią przyjmują pozostałe zmienne dla obserwacji z poszczególnych przedziałów zmiennej zdyskretyzowanej.

Dyskretyzacja – wprowadzenie

Wiele różnych podejść do dyskretyzacji:

- nadzorowana vs. nienadzorowana,
- globalna vs. lokalna (podział z punktu widzenia atrybutów),
- dynamiczna vs. statyczna (podział z punktu widzenia doboru parametrów).

Możliwe strategie dyskretyzacji obejmują m. in.:

- podział dziedziny atrybutu na przedziały o równej szerokości,
- podział dziedziny atrybutu na przedziały o porównywalnej liczbie obserwacji,
- ChiMerge – zachowuje podobieństwo względnych częstości klas decyzyjnych w podprzedziałach,
- minimalizacja entropii warunkowej klas decyzyjnych (ang. Class Entropy Discretization) – wersja lokalna, wersja wykorzystująca zasadę MDL (algorytm Fayyada i Iraniego, 1993), wersja globalizowana,
- podział dziedziny atrybutu na grupy rekordów o podobnej wartości zmiennej celu – modyfikacje alg. analizy skupień.

Uwagi końcowe

- Wzorce odkryte w danych podczas eksploracyjnej analizy danych niekoniecznie muszą rzucać dużo światła na badane dane. Odkrycie anomalii, niedostatków w danych czy też w sposobie (procesie) ich gromadzenia może być równie cenne.
- W oparciu o przeprowadzoną wizualizację danych, np. za pomocą histogramów czy wykresów rozrzutu, możliwe jest wybranie interesującego podzbioru danych do dalszych badań.
- "Każdy ogromny zbiór danych zawiera dane podejrzane. Ogromny zbiór danych, który sprawia wrażenie nieskażonego niekompletnością, zniekształceniami, błędami pomiarowymi . . . , powinien wzbudzać nie ulgę, ale podejrzenie. Jedynie kiedy rozpoznamy i zrozumiemy niedociągnięcia w danych, możemy podjąć kroki łagodzące ich wpływ. Jedynie wtedy możemy być pewni, że odkryte struktury i wzorce odzwierciedlają to, co rzeczywiście dzieje się w świecie." – "Eksploracja danych", Hand, Mannila, Smyth, WNT 2005

Zakończenie

Dziękuję za uwagę.

Materiały dodatkowe:

- `http://www.kdnuggets.com/data_mining_course/dm12-data-preparation.ppt`.
- `https://sci2s.ugr.es/keel/pdf/algorithm/congreso/fayyad1993.pdf`.