

Wprowadzenie

Niniejszy skrypt zawiera materiał opracowany przez autora na potrzeby wykładu zatytułowanego „Metody gromadzenia i analizy danych”, prowadzonego w ramach pierwszego roku uzupełniających studiów magisterskich na wydziale Informatycznych Technik Zarządzania w Wyższej Szkole Informatyki Stosowanej i Zarządzania w Warszawie. W tytule skryptu podkreślono aspekt analizy danych, ponieważ temu to aspektowi głównie poświęcony jest zarówno wykład, jak i niniejszy skrypt. Przy opracowywaniu treści wykładu autor odwoływał się do swoich wieloletnich doświadczeń, zarówno teoretycznych, jak i praktycznych, z dziedziny analizy danych, zebranych w trakcie pracy w Instytucie Badań Systemowych Polskiej Akademii Nauk.

Zarówno wykład, jak i skrypt, przeznaczone są dla osób specjalizujących się przede wszystkim w zarządzaniu gospodarczym i zastosowaniu instrumentarium informatycznego do zarządzania. Dlatego też szczególnie nacisk położono na interpretację wyników otrzymywanych przy pomocy różnych technik analizy danych, znacznie bardziej niż na zawartość matematyczną lub charakterystykę numeryczną prezentowanych technik. Także i ze względu na ograniczenia objętościowe wykładu częste są w treści odwołania do łatwo dostępnych na rynku aplikacji i systemów, zawierających różne implementacje wielu technik analizy danych, wraz z odpowiednimi komentarzami.

Zasadniczy problem, na który chciałby mianowicie zwrócić uwagę autor skryptu, polega nie na braku możliwości stosowania tych różnych technik, ale wręcz odwrotnie – na nadmiernej łatwości odwoływania się do nich. I tak, minimalna zupełnie wiedza informatyczna wystarcza do tego, by posłużyć się, dla jakiegoś zbioru danych, którym dysponujemy, na przykład, podstawowymi metodami analizy danych, zrealizowanymi w ramach arkusza Excel Microsoftu (analiza regresji, analiza czynnikowa). Niebezpieczeństwo polega na możliwym braku krytycyzmu, a nawet zrozumienia, zarówno w odniesieniu do warunków stosowalności poszczególnych metod, jak i otrzymywanych przy ich pomocy wyników.

Stąd też potrzeba wykładu, który z jednej strony zapoznałby z szeroką gamą dostępnych technik i metod analizy danych, choćby na poziomie zrozumienia ich podstawowych założeń i mechanizmu działania, a z drugiej – pozwoliłby na umiejętne i świadome korzystanie z tych narzędzi do celów praktycznych, jakich nie brakuje zarówno w praktyce zarządzania gospodarczego i marketingu (na przykład identyfikacja i charakterystyka typów klientów, lub typów produktów), jak i w innych dziedzinach życia społecz-

WYKŁAD Z METOD ANALIZY DANYCH

nego i gospodarczego, z którymi przyjdzie zetknąć się absolwentom studiów magisterskich nakierowanych na informatyczne wsparcie zarządzania (badania opinii publicznej, edukacja i oświata, itp.). Jakkolwiek narzędzia informatyczne, z jakimi zapoznają się słuchacze niniejszego wykładu w trakcie studiów, pozwalają im w zasadzie na samodzielne opracowywanie aplikacji z dziedziny analizy danych, jednak należy spodziewać się, że w praktycznie prawie wszystkich przypadkach będą oni korzystali właśnie z gotowych programów i systemów, a zatem powinni być przygotowani do umiejętnego ich użytkowania.

Jan W. Owsieński

„Wszyscy zaprzeczają, ale każdy czeka, że się zdarzy”

Spis treści

Wprowadzenie

Wykład I

- I.1. Dana, informacja, wiedza
- I.2. Pozyskiwanie danych
- I.3. Podstawowe elementy teorii informacji i entropii Shannona

Wykład II

- II.1. Postacie danych – obiekty i zmienne
- II.2. Odległości i bliskości
- II.3. Relacje
- II.4. Zbiory rozmyte

Wykład III

- III.1. Kilka uwag o sensie, rozumieniu i postaci wiedzy
- III.2. Niektóre rodzaje zadań analizy danych
- III.3. Zadania analizy danych omawiane w ramach wykładu

Wykład IV

- IV.1. Porządkowanie obiektów – zarys zagadnienia
- IV.2. Podstawowe metody agregacji uporządkowań
- IV.3. Dalsze rozwinięcia metod porządkowania wielowymiarowego

Wykład V

- V.1. Zadanie analizy skupień i jego warianty
- V.2. Podstawowe grupy metod
- V.3. Algorytmy agregacji hierarchicznej
- V.4. Algorytmy p -średnich
- V.5. Metoda z globalną funkcją celu

WYKŁAD Z METOD ANALIZY DANYCH

V.6. Zastosowania w eksploracji danych

V.7. Przypadek jednowymiarowy

V.8. Przykład zastosowania

Wykład VI

VI.1. Analiza dyskryminacyjna

VI.2. Klasyfikacja

VI.3. Operacje na zmiennych

VI.4. Przekształcenia zbioru zmiennych

VI.5. Składowe główne i analiza czynnikowa

Wykład VII

VII.1. Analiza regresji

VII.2. Skalowanie wielowymiarowe

VII.3. Procedura analizy danych

Przykładowy regulamin zaliczenia wykładu i zestaw
zadań do zaliczenia (rok akademicki 2002/2003)

Przykładowe rozwiązania zadań

Literatura

WYKŁAD I:

Dana, informacja, wiedza – rozróżnienie, przykłady, znaczenie. Przykłady postaci danych i sposobów / sytuacji ich gromadzenia. Krótki zarys podstaw teorii informacji opartej na entropii Shannona.

I.1. Dana, informacja, wiedza

I.1.1. W niniejszym wykładzie będziemy starali się wyraźnie rozróżnić takie kategorie jak *dana*, *informacja* oraz *wiedza*. *Daną* może być mianowicie dowolny obiekt (liczba, słowo, obraz, element graficzny), w zależności od kontekstu (treści) i celu postępowania. W sensie niniejszego wykładu *informacją* jest pewna *cecha danych*, pozwalająca na dokonywanie konkretnych operacji na danych i wyciąganie z nich wniosków w ramach określonego kontekstu i celu postępowania. W istocie swojej informacja jest pewną interpretacją danej. *Wiedza* wreszcie jest końcowym efektem postępowania, określonego przez dane, informację, kontekst i cel.

I.1.2. Całość konkretnego kontekstu będziemy nazywali *uniwersum* (całością dostępnych i potencjalnych treści). Tak więc możemy mówić o, na przykład, uniwersum „zarejestrowanych na nośnikach trwałych utworów muzycznych”, uniwersum „produktów krajowych brutto”, uniwersum „wymiarów krawieckich (figur potencjalnych klientów)”, lub, nieco trywialniej, „uniwersum wyników LOTTO”. *Daną* będzie każdy element takiego uniwersum (przestrzeni), niekoniecznie dobrze określony. Zastrzeżenie dotyczące określoności odnosi się, w szczególności, do (1) dokładności (np. miara obwodu pasa wzięta na wdechu lub wydechu), (2) ustalenia wszystkich cech (*atrybutów*) odpowiedniego elementu uniwersum (np. brak daty urodzenia niektórych pań z próby), (3) różnego rozumienia (definicji) poszczególnych atrybutów.

I.1.3. Możemy mieć do czynienia z różnymi rodzajami celów postępowania, także w stosunku do posiadanych zbiorów danych. Jednakże w sekwencji „*dane*→*informacje*→*wiedza*” cel, wyrażony poprzez obrane kryterium postępowania, jest lepiej określony. Celem tym, ogólnie rzecz biorąc, jest „*synteza*”, czyli możliwość (i) wyrażania określonych treści, obejmujących zakres wielu (być może tysięcy, albo nawet milionów) danych, w sposób *skrótowy*, dostępny bez oglądu wszystkich tych danych, (ii) posługiwania się *relacjami* w obrębie uniwersum, pozwalającymi, w szczególności, na określanie konsekwencji lub zależności w obrębie danych oraz informacji. Zależy nam także na tym, by nasza wiedza była (iii) możliwie dokładna, pozwalając na ocenę i interpretację nowych danych, a nawet na ich przewidywanie.

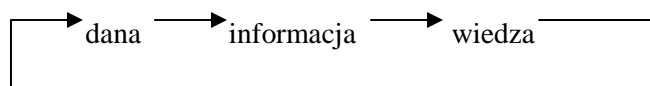
I.1.4. Ogólnie rzecz biorąc, rozróżnienie między *daną*, *informacją* i *wiedzą* nie jest ścisłe, choć można je formalnie i precyzyjnie uściślić w ramach pewnych podejść, zwłaszcza formalnych systemów matematycznych. Będziemy się niekiedy w trakcie wykładu odwoływać do takich systemów (na przykład w końcowej części Wykładu I – do teorii informacji i entropii Shannona-Weavera), ale raczej tylko ilustracyjnie lub przyczynkowo. Mimo jego nieprecyzyjności, pozostaniemy przy wprowadzonym intuicyjnie zrozumiałym podziale, przede wszystkim dlatego, że będziemy się nań powoływać dość często w trakcie wykładu, choćby w celu podkreślenia sensu i kierunku pewnych zasad i sposobów postępowania.

I.1.5. *Przykład I.1.* Niech danymi będą deklarowane (przez respondentów) miesięczne dochody gospodarstw domowych. Będą to zatem liczby typu 560, 1 300, 2 850, 3 570, 12 870, albo 54 620 (tutaj podane w PLN). Jeśli interesuje nas wyłącznie statystyka tych deklaracji, otrzymane dane (a więc takie, jak przykładowo przytoczone liczby) są wystarczającym materiałem do dalszych rozważań. Na ich podstawie można opracować odpowiednie statystyki (średnia, mediana, zakres, wariancja, itp.). Nie będzie jednak możliwe ustalenie innych zależności (relacji) poza, ewentualnie, relacjami między dochodami a licznosciami (lub udziałami) odpowiednich sub-populacji. Informacją w tym przypadku będą, na przykład, udziały poszczególnych sub-populacji o określonych przedziałach przychodów. Będą to zatem elementy wiedzy o rozwarstwieniu (dochodowym) całości społeczeństwa, która może być reprezentowana w postaci syntetycznej w formie krzywej Lorenza lub współczynnika Giniego. Aby móc dysponować bardziej znaczącą wiedzą, np. dotyczącą zależności między wiekiem członków gospodarstwa domowego a dochodami, liczbą członków gospodarstwa domowego a dochodami, albo wykształceniem członków gospodarstwa domowego a dochodami – musimy opierać się na znacznie bardziej rozwiniętej (wielowymiarowej) bazie danych. Będziemy wówczas mogli starać się uzyskać wiedzę typu relacji (reguły) „jeśli gospodarstwo domowe zamieszkuje miasto o liczbie mieszkańców powyżej 500,000, składa się z mniej niż 6 osób, a wszyscy dorośli członkowie gospodarstwa mają co najmniej wykształcenie średnie, to dochód w gospodarstwie na głowę jest co najmniej dwa razy wyższy od przeciętnego w kraju z prawdopodobieństwem równym 0.85”.

I.1.6. W *Przykładzie I.1* daną jest, na przykład, liczba 560. Staje się ona informacją, umieszczona w odpowiednim kontekście, składającym się, w tym przypadku, ze stwierdzenia, że (i) chodzi o sumę w PLN, (ii) która jest wysokością miesięcznych dochodów gospodarstwa domowego w określonym miesiącu, (iii) i ewentualnym podaniem innych okoliczności (miejsce zamieszkania, liczba osób, wiek, wykształcenie). Informacja zatem składa się, w pewnym sensie, z większej liczby danych, co wynika z konieczności umieszczenia danej w kontekście (określenia właściwego uniwersum), aby

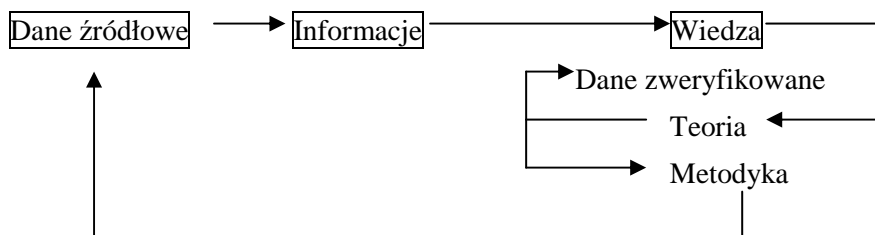
uczynić ją informacją. Tak określone informacje składają się na wiedzę, obejmującą znaczącą część kontekstu, a w szczególności – być może nawet całe uniwersum danych i jego cechy.

I.1.7. Należy jednak zauważyć, że układ $\text{dana} \rightarrow \text{informacja} \rightarrow \text{wiedza}$ działa w istocie w pętli sprzężenia zwrotnego, tj.

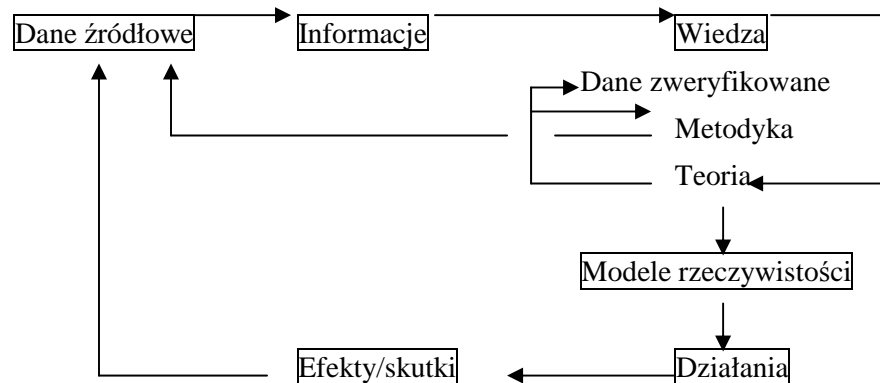


Wyraża się to w dwojaki sposób: (A) poprzez możliwość oceny poszczególnych danych w ramach istniejącej wiedzy (np. ewentualność odrzucenia danej jako „nieprawidłowej”, to znaczy – nie pasującej do istniejącej wiedzy, co niesie także za sobą, oczywiście, określone ryzyko – odrzucenia istotnej danej), a także włączania ich do istniejącego stanu wiedzy; (B) poprzez możliwość przewidywania dalszych danych, a zatem i możliwość projektowania sposobów zdobywania nowych danych o określonych własnościach. W istocie, sama możliwość interpretowania danych jako informacji jest uwarunkowana uprzednim istnieniem (i zastosowaniem) określonej wiedzy (w tym, na najprostszym poziomie, umiejętności czytania i liczenia).

I.1.8. Jak wynika z poprzednich rozważań, wiedza składa się z „zatwierdzonych” zbiorów danych (co najczęściej powinno oznaczać, że są to dane „prawdziwe” – użycie cudzysłowu oznacza tutaj wyłącznie, że prawdziwość danych jest sprawdzana – i sprawdzona – przy pomocy procedur również zawierających się w obrębie istniejącej wiedzy), relacji między nimi i procedur z nich wynikających. W ramach często używanych potocznych terminów można mówić, że „zatwierdzone” („sprawdzone”) zbiory danych, w odróżnieniu od „danych źródłowych”, stanowią właściwą wiedzę (np. tak zwana „wiedza encyklopedyczna”), relacje między nimi – „teorie”, zaś procedury – „metodykę”.



I.1.9. Po co nam jednak, w końcu, wiedza? Otóż – przynajmniej w ramach niniejszego wykładu – przyjmujemy, że wiedza zwiększa sprawność naszego działania. To dzięki *teoriom* (sprawdzonym) wiemy, jaki skutek odniosą nasze działania, a dzięki *metodyce* wiemy, jak pozyskać nowe, potrzebne nam, dane. *Teorie* są w istocie *modelami rzeczywistości*, jakimi się posługujemy w naszych działaniach. Jeśli modele te są prawdziwe (zgodne z rzeczywistością), to nasze działania będą skuteczne, jeśli nie – będą one bezowocne. Podobnie z *metodyką* – jeśli jest ona poprawna, to otrzymamy nowe dane, uzupełniające naszą wiedzę, a jeśli nie, to nasze pomiary i eksperymenty dostarczą tylko „szumu”.



W ten sposób opisana poprzednio pętla sprzężenia zwrotnego rozszerza się na całość naszego działania i poznania świata.

I.2. Pozyskiwanie danych

I.2.1. Dane są z zasady wynikiem zaprogramowanego postępowania, wynikającego z zastosowania istniejącej wiedzy. W odniesieniu do danych źródłowych, a więc takich, które są wprost (ewentualnie tylko po bardzo powierzchownej wstępnej obróbce) rezultatem obserwacji lub pomiaru, możemy mówić o dwóch zasadniczych rodzajach takiego postępowania, a mianowicie: (1) *rutynowe*, okresowe lub stałe gromadzenie danych (często automatyczne), stanowiące część działalności jakiejś organizacji czy instytucji (np. billingi telefoniczne, logi internetowe, transakcje przechodzące przez kasy), często zupełnie uboczną, choć zazwyczaj konieczną (np. prowadzenie księgowości), bądź też jej zasadniczą funkcję (np. pomiary zmiennych klimatycznych prowadzone w stacjach meteorologicznych, zdjęcia powierzchni Ziemi wykonywane z satelitów), i (2) *doraźne* zbieranie danych (nawet, jeśli w jakiejś mierze powtarzalne) dla określonego celu (tutaj najlepszym przy-

kładem są przeróżne sondaże opinii publicznej, większość badań marketingowych, a także wybory). (Specyficzna sytuacja, w której uzupełniamy tylko posiadane zbiory danych przy pomocy pojedynczych obserwacji, może być zazwyczaj sklasyfikowana w ramach jednego z powyższych dwóch sposobów postępowania.)

I.2.2. Powyższy, zgrubny na oko, ale w istocie dość precyzyjny podział, jest związany z innymi podziałami istotnymi dla sposobów pozyskiwania danych. I tak, rozróżniamy (a) automatyczne i (b) „ręczne” sposoby zbierania danych. Ta klasyfikacja jest dość ściśle, z kolei, związana z poziomem „obiektywności” i „dokładności” zbieranych danych, a mianowicie: dane zbierane automatycznie są częściej bardziej (i) obiektywne i dokładne niż zbierane „ręcznie” (które często mają tendencję do zawierania (ii) „subiektywności”), a w każdym razie mamy większą kontrolę nad ich definicjami i ewentualnymi błędami lub nieporozumieniami, podczas, gdy dane zbierane, na przykład, przy pomocy kwestionariuszy, mogą zawierać nieścisłości i błędy o niekontrolowalnym charakterze, wymagając kłopotliwych i kosztownych zabiegów zmierzających do sprawdzania ich prawidłowości i akceptowalności. To rozróżnienie nie jest, oczywiście, absolutne: nawet raporty kasowe mogą być w istocie wynikiem oszustwa, zaś automatycznie działające i wyskalowane mierniki mogą ulec uszkodzeniu powodującemu podawanie błędnych pomiarów, podczas gdy zawartość kwestionariusza może zawierać dane (albo i informacje) nie do osiągnięcia metodami automatycznymi.

I.2.3. Jest rzeczą oczywistą, że dla przeciętnego konsumenta mediów dane źródłowe w powyższym sensie nie są w zasadzie dostępne. Danymi (i to „źródłowymi”) dla większości z nas są właśnie doniesienia mediów, będące w istocie (w najlepszym przypadku) wynikiem daleko idących transformacji pewnych (właściwych) danych źródłowych, pozyskiwanych w jeden z powyższych sposobów.

I.2.4. Aby dane mogły spełnić swoją rolę w uzupełnianiu i tworzeniu wiedzy, muszą spełnić podstawowe warunki dotyczące *porównywalności* i *reprezentatywności*. Wymagania te są w sposób formalny zazwyczaj przedmiotem analizy statystyki matematycznej, jednak łatwo je również wyrazić w sposób jakościowy, intuicyjnie oczywisty. I tak, *porównywalność* jest zapewniona przez odwoływanie się do danych o tym samym sensie, czyli o tych samych definicjach (np. wysokość w metrach nad poziomem morza, raczej niż wysokości względne, chyba, że porównujemy obiekty – np. wzniesienia – dla których poziom odniesienia wysokości względnych jest ten sam, lub choćby w przybliżeniu ten sam), bądź choćby o definicjach na tyle jasnych i jednoznacznych, że można bez (większego) trudu dokonać transformacji danych pomiędzy różnymi definicjami (np. wielkość PKB dla

różnych krajów wyrażona w walutach tych krajów, przy znajomości odpowiednich kursów walutowych). Oczywiście, zapewnienie porównywalności musi uwzględniać takie aspekty jak czas (na ogół oznacza to porównywanie w jednakowym czasie dla różnych obiektów lub w różnych momentach czasu dla tego samego obiektu) i przestrzeń (np. porównywanie analogicznych jednostek przestrzennych). *Reprezentatywność* odnosi się do możliwości charakteryzowania większych populacji, niż objęte aktualnie posiadanym zbiorem danych. Dane te powinny mianowicie odpowiednio odzwierciedlać strukturę całej populacji, na przykład w przypadku badania opinii publicznej, podział całej (dorosłej) ludności Polski według takich kryteriów jak płeć, wiek, wykształcenie, charakter miejsca zamieszkania (wieś, małe miasto, średnie miasto, duże miasto), itp. Spodziewamy się, intuicyjnie, że tak rozumiana reprezentatywność jest blisko związana z proporcjonalnością: niewątpliwie uznalibyśmy za nieprawdziwe wyniki badania opinii publicznej, wiedząc, że stosunek liczby mężczyzn do kobiet wśród osób badanych wynosił 73:27, a pytanie dotyczyło długości okresu płatnego urlopu macierzyńskiego. Zagadnienie reprezentatywności jest przedmiotem badań ważnej dziedziny analizy statystycznej (której wynikiem jest, między innymi, powtarzające się w sposób niemal magiczny, stwierdzenie typu „wyniki pochodzą z badania przeprowadzonego na reprezentatywnej próbie 1032 osób” dotyczące badań opinii publicznej w Polsce, wskazujące, że liczba 1000 wystarcza do wyczerpania – w sensie reprezentacji – wszystkich warstw populacji ludności Polski). Nie zawsze jednak jesteśmy w tak luksusowej sytuacji, że możemy zapewnić sobie reprezentatywność – i to z wielu powodów, a przede wszystkim: (i) brak znajomości własności całej populacji i jej ewentualnych sub-populacji, (ii) brak środków do przeprowadzenia odpowiednich badań. Nie będziemy jednak zajmowali się bliżej zagadnieniami porównywalności i reprezentatywności w tym wykładzie – pierwsze jest w znacznej mierze kwestią zdrowego rozsądku, a w bardziej skomplikowanych przypadkach (np. porównywanie wartości PKB według siły nabywczej walut w odpowiednich krajach) jest przedmiotem specjalizowanych analiz, drugie zaś jest w większej mierze przedmiotem statystyki matematycznej i jej metod.

I.2.5. Najważniejszym aspektem możliwości przejścia od danych (np. liczb lub obrazów) do informacji i wiedzy jest definicja danych. Obejmuje ona takie elementy jak: przedmiot(y) danych (o czym to jest? lub: co to jest?), użyte jednostki (np. zakresy widma w obrazie i jego rozdzielczość), użyta metoda pozyskania danych i jej charakterystyki, czas, miejsce, itp. Braki w znajomości definicji danych prowadzą do ich fałszywej interpretacji, a zatem i do nieprawidłowych wniosków dotyczących informacji, a następnie i wiedzy. Podkreślmy, że bardzo często dane podawane w mediach do wiadomości publicznej mają na tyle słabo zarysowane definicje, że nieraz trudno jest

w ogóle podjąć jakąkolwiek sensowną interpretację. Dobrym przykładem w tym zakresie, nawiązującym do kwestii porównywalności, są wyniki badań popularności poszczególnych partii politycznych:

I.2.6. *Przykład I.2.* Instytucje zajmujące się badaniem opinii publicznej, powiedzmy, Demometr i ĆBOŻ, ogłosiły wyniki badania popularności partii politycznych („na jaką partię zagłosowałby/wałaby Pan/i, gdyby wybory parlamentarne odbyły się dzisiaj?”). Opublikowane w prasie wyniki podajemy w tabelce poniżej:

Partie:	KDL	SiP	OO	Samiswoi	PRL
<i>Demometr</i>	22%	14%	9%	13%	11%
<i>ĆBOŻ</i>	28%	13%	8%	15%	14%

Łatwo zauważyć różnice między tymi dwoma badaniami, które nie są jednak uderzające (proporcje są mniej więcej zachowane). Ważniejsze jednak są wątpliwości dotyczące (a) w ogóle znaczenia podanych liczb, (b) ich porównywalności. Sprowadzają się one w tym wypadku do odpowiedzi na następujące pytania: (i) jakie liczby partii mieli do wyboru respondenci w obu badaniach, czy była to ta sama liczba? (ii) i czy były to te same partie? (iii) czy 100% obejmuje tylko odpowiedzi zdecydowane – wskazania partii – czy również „brak opinii” (zauważmy, że podana suma „głosów” dla Demometru wynosi 69%, a dla ĆBOŻ – 78%)? (iv) czy kolejność partii w kwestionariuszach była taka sama? Odpowiedzi na te, i jeszcze bardziej szczegółowe pytania (np. czy w ramach badania zadawano jeszcze inne pytania i czy poprzedzały one to pytanie zasadnicze o preferencje dla partii?) stanowią właśnie o definicji danych, prowadzącej do poprawnej informacji i wiedzy, a jednocześnie o spełnianiu (lub nie) warunku porównywalności. Na zasadniczą część przynajmniej tych pytań dostalibyśmy od razu odpowiedź, gdyby prezentowana poprzednio tabelka odsetków „głosujących” na poszczególne partie została opublikowana w następującej postaci:

Partie:	KDL	SiP	OO	Samiswoi	PRL	X1	WU	Nie wiem	Suma
<i>Demometr</i>	22%	14%	9%	13%	11%	4%	3%	24%	100%
Nr partii	1	2	3	4	5	7	6	-	
<i>ĆBOŻ</i>	28%	13%	8%	15%	14%	-	-	22%	100%
Nr partii	1	3	4	2	5	-	-	-	

w której by podano wyniki dla wszystkich uwzględnianych w badaniu partii, odsetek odpowiedzi „nie wiem” („brak opinii”), oraz kolejność partii na kwestionariuszach obydwu agencji badania opinii publicznej.

I.3. Podstawowe elementy teorii informacji i entropii Shannona

I.3.0. Przedstawimy obecnie w zarysie wstęp do bodaj najpopularniejszej (jeśli nie faktycznie jedynej) formalnej teorii informacji, opartej na pojęciach i zależnościach z zakresu teorii prawdopodobieństwa.

I.3.1. Załóżmy, że interesuje nas pewna wielkość (zmienna), oznaczona X , przyjmująca skończoną (lub co najwyżej przeliczalną) liczbę wartości. Konkretnie wartości, jakie ta wielkość będzie w rzeczywistości przyjmowała (i o których będziemy wiedzieli), będą naszymi danymi (źródłowymi). Wielkość tę będziemy interpretować jako pewną zmienną losową, jakkolwiek dla większej części dalszych rozważań nie będzie to w zasadzie potrzebne. Każdej wartości zmiennej, oznaczonej x_i , $i \in I$, gdzie I jest zbiorem wartości zmiennej X , można przypisać pewne prawdopodobieństwo wystąpienia, oznaczone p_i . Zakładamy, że prawdopodobieństwa te są znane, co może w praktyce wynikać z obserwacji częstości odpowiednich zdarzeń w przeszłości, albo znajomości odpowiednich cech przedmiotu obserwacji (np. prawdopodobieństwo, że moneta upadnie orłem albo reszką w górę jest, o czym wiemy dobrze, równe mniej więcej 0.5 w obu przypadkach; jeśli założymy, że moneta jest jednorodna i nieskończenie cienka, to oba te prawdopodobieństwa na pewno będą równe 0.5).

I.3.2. Mamy, oczywiście, $\sum p_i = 1$, oraz istnienie odpowiednich momentów tego rozkładu prawdopodobieństwa, począwszy od $E(X)$, czyli wartości oczekiwanej. Jednak warunki związane z momentami wyższych rzędów nie będą dla nas istotne.

I.3.3. Naszym problemem jest „niepewność”, wynikająca z faktu, że nie wiemy, jakie wartości przyjmie X . Możemy powiedzieć, że posiadamy pewną „wiedzę” (znajomość rozkładu prawdopodobieństwa $P(X) = \{p_i\}_i$), ale w naszej konkretnej sytuacji brakuje nam danych i związanej z nimi informacji. Prezentowana teoria, odnosząca się do niepewności, nie zajmuje się jednak pojedynczymi zdarzeniami, tj. przyjmowaniem przez wielkość X poszczególnych wartości x_i . Jej przedmiotem zainteresowania jest całkowita miara „niepewności”, związana z całym rozkładem $P(X)$.

I.3.4. Sens tak sformułowanego zagadnienia można zilustrować następującymi przykładami: (A) jeśli na pięciu z sześciu ścianek kostki do gry wypiszemy cyfrę oczek równą 1, a na szóstej (dowolnej) – inną cyfrę oczek, równą, powiedzmy, 5, to wiadomość (dana), że wyrzucono jedno oczko nie przyniesie nam żadnej istotnej informacji, wiemy bowiem, że prawdopodobieństwo takiego zdarzenia jest równe $5/6$, i raczej spodziewamy się takiego obrotu (!) sprawy; (B) sytuacja jest zupełnie inna w przypadku zwykłej kostki do gry: wiadomość o wyrzuceniu konkretnej (każdej) liczby oczek ma

znacznie większą wartość w sensie informacji – tutaj wszystkie prawdopodobieństwa $p_i=1/6$, $i=1,\dots,6$, i nie spodziewamy się raczej wyrzucenia żadnej konkretnej liczby oczek; (C) naturalnie, jeśli mamy do czynienia z rozkładem, w którym dla jednego i , powiedzmy, bez żadnej straty ogólności, że dla $i=1$ (możemy zawsze nasze zdarzenia elementarne o indeksach $i = 1, 2, 3, \dots$, odpowiednio przenieść), mamy $p_i=1$, a dla pozostałych $i=2, 3, 4, \dots$ mamy $p_i = 0$, to niepewność związana z takim rozkładem prawdopodobieństwa jest oczywiście zerowa, a zatem i informacja niesiona przez dane o zajściu zdarzeń elementarnych – też zerowa. Widzimy zatem, że w istocie można przypisać określoną wartość niepewności całym rozkładom prawdopodobieństwa, i że jest to, z kolei, ściśle związane z „wartością informacyjną” zaistnienia poszczególnych zdarzeń realizujących zmienną losową o określonym rozkładzie, która to wartość informacyjna jest uzależniona od prawdopodobieństwa zajścia tych zdarzeń (por., np. Vetschera, 2000).

I.3.5. Budowę miary niepewności związanej z rozkładem $P(X)$ zaczniemy od formułowania warunków na poszukiwaną funkcję, przy czym warunki te powinny odpowiadać intuicjom związanym z miarą niepewności, odniesioną do całego $P(X)$, a następnie spróbujemy zobaczyć, jaka lub jakie funkcje spełnia(ją) te warunki.

I.3.6. Oznaczmy poszukiwaną funkcję, opisującą niepewność związaną z rozkładem $P(X)$, przez $H(P(X))$, z ewentualnym pominięciem oznaczenia X , tj. $H(P)$. Możemy już obecnie, na podstawie czysto jakościowych rozważań i przeświadczeń, sformułować cały szereg dość oczywistych postulatów względem funkcji $H(P)$. I tak,

- (i) chcemy, żeby funkcja $H(P)$ osiągała maksimum dla rozkładu jednostajnego (równomiernego), tj. wtedy, gdy $p_i=1/n \ \forall \ i \in I$, przy czym $n=\text{card } I$, tj. n jest liczbą możliwych wartości zmiennej X (nie możemy zatem mówić już o nieskończonej, lecz przeliczalnej liczbie tych wartości); sens tego założenia zilustrowaliśmy poprzednimi przykładami;
- (ii) chcemy, żeby funkcja $H(P)$ była funkcją ciągłą prawdopodobieństw tworzących $P(X)$, a więc, by nieskończenie małemu przyrostowi (dodatniemu i ujemnemu) wartości prawdopodobieństw towarzyszyła nieskończenie mała zmiana wartości $H(P)$; i to założenie wydaje się być zgodne ze zdrowym rozsądkiem;
- (iii) chcemy, żeby $H(P)$ była symetryczna względem argumentów, czyli poszczególnych prawdopodobieństw tworzących $P(X)$ – co jest w istocie równoważne niezależności od kolejności, w jakiej poszczególne zdarzenia (a właściwie ich prawdopodobieństwa)

pojawiają się w $H(P)$, a więc również całkiem rozsądne wymaganie;

- (iv) dość podobnym wymaganiem jest, by $H(P)$ była niezależna od sposobu „grupowania” prawdopodobieństw według podzbiorów zdarzeń elementarnych x_i składających się na X ; i tak, w przypadku, gdy mamy rozkład $P(X) = \{p(x_1) = 0.5, p(x_2) = 0.3, p(x_3) = 0.2\}$, to możemy przedstawić go w równoważny sposób jako dwa zależne od siebie rozkłady: $P(X) = \{p(x_1)=0.5, p(x=y)=0.5\}$ i $P(Y) = \{p(y_1)=0.6, p(y_2)=0.4\}$; chodzi więc o to, by wartość $H(P)$ nie zależała od tego rodzaju różnych reprezentacji, co jest znów zupełnie naturalnym wymaganiem;
- (v) chcemy, żeby $H(P)$ była addytywna względem poszczególnych zdarzeń elementarnych rozpatrywanego rozkładu prawdopodobieństwa, czyli $H(P) = \sum_i h(p_i)$, co powinno prowadzić do znacznego uproszczenia postaci funkcji $H(P)$;
- (vi) poza wymienionymi pięcioma formułuje się jeszcze różne inne wymagania, o nieco, lub całkiem wyraźnie, mniejszej intuicyjnej oczywistości; przytoczymy jeszcze tylko jeden z nich, żeby zilustrować ich charakter: jeśli mianowicie mamy dwa rozkłady prawdopodobieństwa, $P_n(X) = \{p_1, \dots, p_n\}$ oraz $P_m(Y) = \{q_1, \dots, q_m\}$, z których otrzymujemy rozkład $P_{nm}(X, Y) = \{p_1q_1, p_1q_2, p_1q_3, \dots, p_1q_m, p_2q_1, \dots, p_nq_1, \dots, p_nq_m\}$, to chcemy, żeby $H(P_{nm}(X, Y)) = H(P_n(X)) + H(P_m(Y))$; i to wymaganie jest w gruncie rzeczy usprawiedliwione niezależnością rozkładów $P_n(X)$ i $P_m(Y)$ względem rozkładu łącznego $P_{nm}(X, Y)$.

I.3.7. Dla tak określonych warunków na funkcję niepewności rozkładu prawdopodobieństwa, $H(P)$, na podstawie znanych (choć nie wszystkim) własności funkcji można wykazać (czego tutaj nie będziemy jednak robili), że postać tej funkcji powinna być następująca:

$$H(P) = k \sum_i p_i \log_a p_i \quad (1.1)$$

gdzie k jest pewną stałą, zaś a pewną podstawą logarytmu.

Należy jednak podkreślić, że do otrzymania postaci (1.1) wystarczają warunki (i), (ii), (iv) z punktu I.3.6, i że postać ta – taka sama – może zostać otrzymana przy pomocy innych zestawów warunków, na przykład (ii), (v), (vi), albo zestawów, w których występuje warunek (iii). Literatura podaje szereg takich zestawów „rozsądnych” warunków, prowadzących do otrzymania postaci (1.1) (por. Taneja, 2001; Feldman, 2002).

I.3.8. Ponieważ postać (1.1) pozostawia możliwość doboru stałych k i a , w ramach teorii informacji Shannona ustalono te dwie wielkości na podstawie prostego rozumowania jako $k=-1$ oraz $a=2$, czyli

$$H(P) = -\sum_i p_i \log_2 p_i. \quad (1.2)$$

Podstawą tego rozumowania jest następujący konkretny przypadek: załóżmy, że mamy rozkład dwupunktowy o możliwych dwóch zdarzeniach, x_1 i x_2 , przy czym $p_1 = p_2 = 1/2$. Jest to przypadek analogiczny do rzutu (idealną) monetą. Podstawiając do wzoru (1.2) dane dla tego przypadku otrzymujemy:

$$\begin{aligned} H(P) &= \\ &= -(1/2 \log_2 1/2 + 1/2 \log_2 1/2) = -(1/2 \cdot (-1) + 1/2 \cdot (-1)) = -(-1) = 1 \end{aligned} \quad (1.3)$$

co oznacza, że opisany przypadek odpowiada, dla przyjętych wartości k i a , „jednostce” informacji – taki symetryczny rozkład dwupunktowy niesie jednostkową informację. Odpowiada to w pełni intuicji: w sytuacji „symetrycznego albo-albo”, niezależnie od tego, które ze zdarzeń zajdzie, w wyniku tego jednego zdarzenia usunięta zostaje cała niepewność. Wiele mówiącym jest określenie otrzymanej w powyższy sposób jednostki mianem „bitu” (co odpowiada zaistnieniu 0 lub 1 na odpowiedniej pozycji zapisu binarnego).

I.3.9. Jest naturalne, że $H(P)=0$ wówczas, gdy mamy do czynienia z całkowitą pewnością, a więc, gdy $p_i=1$ dla jakiegoś $i \in I$, oraz $p_j=0$, dla $j \in I-i$. Jeśli $\text{card } I > 1$, to uzyskanie wartości zerowej $H(P)$ dla takiego przypadku ze wzorów (1.1) lub (1.2) wymaga dodatkowo wykazania, że $\lim_{x \rightarrow 0} x \log_2 x = 0$. Jeśli $\text{card } I = 1$, to sytuacja jest trywialna (tylko jedno możliwe, a zatem i pewne, zdarzenie). Jest ona jednak w pewnym sensie przybliżeniem wspomnianych poprzednio sytuacji, w których prawdopodobieństwo innych zdarzeń jest równe zeru, lub też „wystarczająco” bliskie zeru, aby je móc pominąć.

I.3.10. Można łatwo dowieść, że

$$H(P) \geq 0 \quad (1.4)$$

co jest istotną cechą dla miary niepewności (ewentualny kłopot z interpretacją wartości ujemnych miary niepewności).

I.3.11. Funkcję rozkładu prawdopodobieństwa $H(P)$, zwłaszcza w postaci (1.2), nazywa się powszechnie „entropią” lub „entropią Shannona”, który ją pierwszy wprowadził. Nazwa ta pochodzi od nazwy termodynamicznej (makroskopowej) funkcji stanu (mikroskopowego) – entropii – zdefiniowanej jako

$$S(E) = \log N(E) \quad (1.5)$$

gdzie $N(E)$ oznacza liczbę osiągalnych stanów mikroskopowych (położeń i prędkości) w funkcji energii E . Ponieważ zakładamy (co jest wystarczająco dobrze spełnione w warunkach mikroskopowych), że mikrostan odpowiadający tym samym poziomom energetycznym (np. w fizyce cząstek elementarnych) są równie prawdopodobne, więc możemy przyjąć, że prawdopodobieństwo zajścia stanu i -tego jest równe

$$p(i) = 1/N(E), \text{ dla wszystkich możliwych } i. \quad (1.6)$$

Łatwo zauważyć, że po wstawieniu zależności (1.6) do (1.2) otrzymujemy wyrażenie na entropię rozciągniętą na wszystkie możliwe stany mikroskopowe. To właśnie ta analogia z entropią termodynamiczną (analogia – ponieważ obserwujemy tylko identyczność zależności matematycznych, podczas, gdy ich interpretacja może być całkowicie różna) spowodowała, że C. E. Shannon (por. Shannon i Weaver, 1949) w ten sposób nazwał opracowaną przez siebie miarę niepewności dla rozkładów prawdopodobieństwa. Według często powtarzanej anegdoty namówił na to Shannona słynny matematyk John von Neumann (drogi Czytelniku: zapewne słyszałeś o teorii gier, której podstawy zawdzięczamy von Neumannowi, ale wspomnijmy także o *social choice theory*, do której odwołujemy się, jakkolwiek bardzo pobieżnie, również w niniejszym wykładzie), który stwierdził, że skoro nikt tak naprawdę nie wie, czym jest entropia termodynamiczna (poza tym, że można ją wyrazić matematycznie, jak to w grubym zarysie przedstawiliśmy), wprowadzenie entropii informacyjnej da Shannonowi znaczną przewagę w trwających (zresztą do dziś) dyskusjach dotyczących entropii termodynamicznej.

I.3.12. Często przywoływaną interpretacją entropii jest potraktowanie pojedynczego wyrażenia $-\log_2 p_i = h(p_i)$ jako reprezentacji elementarnej niepewności o charakterze „stopnia zaskoczenia” zdarzeniem i -tym (naturalnie, im mniejsze p_i , tym większe nasze zaskoczenie). W ramach tej interpretacji wyrażenie na $H(P)$ może być traktowane jako „średnia wartość zaskoczenia” (albo „średnia wartość zmniejszenia niepewności przez obserwację zdarzenia i -tego”) dla rozkładu prawdopodobieństwa $P(X)$. Ten rodzaj interpretacji łatwo utożsamić z potocznym rozumieniem pojęcia „informacji”: jeśli nie jesteśmy jakąś wiadomością zaskoczeni, to zapewne nie niesie ona ze sobą istotnej dla nas informacji („tak właśnie myślałem”, albo wręcz „byłem tego pewien”), nasza niepewność została zmniejszona w niewielkim stopniu, ponieważ i tak była ona niezbyt wysoka.

I.3.13. Inną jeszcze interpretacją entropii, znacznie bardziej formalną, do której zresztą w pewnej mierze odwołaliśmy się w omówieniu wyrażenia (1.3), jest określenie trudności zgadnięcia wartości przyjmowanej przez zmienną losową X . Zagadnienie to jest zresztą jednym z ważnych problemów teorii informacji i poświęca mu się często wiele uwagi.

Rozpatrzmy mianowicie następujący przykład rozkładu prawdopodobieństwa zmiennej X :

$$p_1 = 1/2, p_2 = 1/4, p_3 = 1/8, p_4 = 1/8 \quad (p_1 + p_2 + p_3 + p_4 = 1).$$

Zastanówmy się, ile pytań o odpowiedziach „tak-nie” będzie nam (średnio) potrzeba, żeby dowiedzieć się, którą z czterech wartości przyjęła zmienna losowa X ? (które ze zdarzeń elementarnych zaszło?) Analogicznie jak w wyprowadzeniu entropii założymy, że powyższe prawdopodobieństwa są nam znane. Oczywiście, posługując się wyłącznie zdrowym rozsądkiem, zaczniemy od pytania o x_1 . Średnio w co drugim przypadku powinniśmy mieć rację. Tak więc, w połowie przypadków wystarczy nam tylko jedno pytanie. Jeśli jednak nie zgadliśmy, zapytamy o x_2 . I znów będziemy mieli rację w połowie (takich) przypadków. Itd., itp. W rezultacie, otrzymujemy następujące wyrażenie na średnią liczbę pytań:

$$1/2(1) + 1/4(2) + 1/4(3) = 1,75,$$

przy czym przed nawiasami podaliśmy (średnie) częstości odgadnięcia, a w nawiasach podaliśmy liczby potrzebnych pytań. I otóż, okazuje się, że entropia przykładowego rozkładu prawdopodobieństwa, policzona według (1.2), jest także równa 1,75. Nie jest to bynajmniej zbieg okoliczności. Można mianowicie wykazać, że

$$H(P) \leq \text{średnia liczba pytań „tak-nie” potrzebna do zgadnięcia wartości } X \leq H(P)+1 \quad (1.7)$$

zakładając, że zgadujący realizuje optymalną strategię (czyli, na przykład, nie zaczyna zgadywać od zdarzeń o najniższych prawdopodobieństwach). Podobnie jak przy interpretacji entropii jako miary zaskoczenia, im bardziej wyrównane prawdopodobieństwa p_i , tym więcej trzeba (średnio) pytań, żeby dojść do wartości X .

I.3.14. Własność entropii omawiana w poprzednim punkcie jest ściśle związana z teorią kodowania, również rozwiniętą przez Shannona. W szczególności, zależność (1.7) stosuje się wprost do średniej długości kodów binarnych, jakie można zastosować do zakodowania stanów zmiennej X . W granicy, $H(P)$ jest średnią liczbą bitów, potrzebną do zakodowania (przechowywania) wartości zmiennej losowej X .

I.3.15. Ważną cechą entropii jest możliwość jej rozszerzania na wiele zmiennych. I tak, łączna entropia dwóch zmiennych losowych, X i Y , wyrażona jest poprzez

$$H(P(X,Y)) = - \sum_i \sum_j p(x_i, y_j) \log_2 p(x_i, y_j). \quad (1.8)$$

Jest to miara niepewności związana z łącznym rozkładem prawdopodobieństwa zmiennych X i Y , $P(X,Y)$. Można również zdefiniować entropię warunkową, w postaci

$$H(X/Y) = - \sum_i \sum_j p(x_i, y_j) \log_2 p(x_i/y_j). \quad (1.9)$$

Zauważmy, że w tym przypadku średnia („wartość oczekiwana”) jest otrzymana poprzez ważenie przez wartości prawdopodobieństwa łącznego, podczas, gdy „stopień zaskoczenia” odnosi się do prawdopodobieństwa warunkowego (na ile jesteśmy zaskoczeni, że x_i zaszło przy zaistnieniu warunku zdarzenia y_j). Korzystając z zależności na prawdopodobieństwo warunkowe i wstawiając otrzymane z niego $p(x_i, y_j) = p(x_i)p(y_j/x_i)$ do wyrażenia na wartość średnią (oczekiwaną) analogicznie do $H(\cdot)$ otrzymujemy bardzo wygodną zależność na entropię łączną:

$$H(P(X,Y)) = H(P(X)) + H(P(Y/X)) \quad (1.10)$$

co prowadzi do ważnych konsekwencji, a w tym, do

$$H(P(Y/X)) = H(P(X,Y)) - H(P(X)) \quad (1.11).$$

I.3.16. W ramach teorii opartej na pojęciu entropii informacyjnej definiuje się też pojęcie *informacji wzajemnej*, a mianowicie

$$I(P(Y;X)) = \sum_i \sum_j p(x_i, y_j) \log_2(p(x_i, y_j)/p(x_i)p(y_j)) \quad (1.12)$$

co, po odpowiednich przekształceniach, opartych na wprowadzonych już zależnościach, prowadzi do

$$\begin{aligned} I(P(Y;X)) &= \\ &= H(P(X)) - H(P(X/Y)) = H(P(Y)) - H(P(Y/X)) = \\ &= H(P(X)) + H(P(Y)) - H(P(X,Y)). \end{aligned} \quad (1.13)$$

Zależności te pokazują, dlaczego $I(\cdot)$ nazywana jest informacją wzajemną: jest to miara redukcji niepewności co do jednej zmiennej na podstawie znajomości drugiej. Pokazują one ponadto, że $I(\cdot)$ jest funkcją symetryczną względem zmiennych X i Y .

I.3.17. Pojęcie entropii można stosować do ciągłych rozkładów prawdopodobieństwa, ale także, z drugiej strony, można je stosować w sposób empiryczny wtedy, gdy nasza wiedza o odpowiednich zmiennych ma ograniczony charakter (dotyczy to częstej sytuacji, w której dysponujemy nie prawdopodobieństwami p_i , a tylko częstościami o ograniczonej, na przykład liczbą obserwacji, „ważności”).

Pojęcie to ma bardzo szerokie zastosowanie (jak widzieliśmy już – np. w teorii kodowania), ze względu na swoje teoretyczne własności związane z jego interpretacją dotyczącą zmniejszania niepewności czy dostarczania informacji na podstawie obserwacji poszczególnych zdarzeń („danych”).

WYKŁAD II

Podsumowanie wykładu I. Wprowadzenie do postaci danych – pojęcia podstawowe używane w wykładzie. Stosunek do probabilistyki. Wielowymiarowość i wielość charakteru danych. Obiekty i zmienne. Normalizacja i standaryzacja. Rodzaje zadań analizy danych.

II.0.1. W wykładzie I zapoznaliśmy się z rozróżnieniem danych, informacji i wiedzy, jako różnymi stopniami transformacji naszych obserwacji otaczającego świata. Przypomnieliśmy sobie, że niezależnie od przejścia dane \rightarrow informacje \rightarrow wiedza funkcjonuje sprzężenie zwrotne między wiedzą, a danymi, pozwalające na ocenę danych i przewidywanie lub efektywne poszukiwanie nowych danych. Wiedza zaś składa się ze zbioru „zatwierdzonych” (sprawdzonych) danych (wiedzy encyklopedycznej), relacji między danymi (obiektami, ich cechami lub podzbiorami), równoważnych „teoriom” lub choćby „hipotezom”, oraz metodyki, pozwalającej na ocenę pozyskanych danych i pozyskiwanie lub przewidywanie innych.

II.0.2. Niezależnie od tego zapoznaliśmy się z podstawami teorii informacji – entropii – Shannona, jako praktycznie jedynej do tej pory ścisłej metody postępowania pozwalającej na uzyskanie formalnie prawidłowej oceny „stopnia niepewności”, a przez to i „wartości informacji” związanej z różnymi rozkładami prawdopodobieństwa. Oznacza to w istocie możliwość przejścia od danych do informacji, a w każdym razie jej potencjalnej wartości, w obrębie pewnego specyficznego modelu danych i informacji. Ten prosty model, niezależnie od swojej formalnej poprawności, dostarcza efektywnych narzędzi, na przykład w zakresie podstaw teorii (optymalnego) kodowania, czyli zapisu informacji. Poza znaczeniem historycznym oraz zastosowaniami w specjalizowanych dziedzinach analizy danych, pojęcie entropii zyskało bardzo szerokie, choć nie zawsze formalnie uzasadnione, zastosowanie w wielu rodzajach zagadnień, i będziemy się na nie jeszcze w ramach niniejszego wykładu powoływali.

II.0.3. Tym niemniej, większa część niniejszego wykładu poświęcona będzie zagadnieniom wykraczającym poza model Shannona. Związane to jest przede wszystkim z zasadniczym założeniem teorii informacji i entropii Shannona, a mianowicie założeniem o dysponowaniu rozkładem prawdopodobieństwa (lub choćby dobrze uwarunkowanych częstości) zdarzeń i wnioskowaniu dopiero na jego podstawie. Nasze rozważania dotyczyć będą mianowicie w pewnym sensie sytuacji o krok wcześniejszej, a mianowicie sytu-

acji rozpoznawania struktury zbioru zebranych (najczęściej po raz pierwszy) danych. W szczególny sposób tej kwestii poświęcony jest punkt II.3 wykładu, w którym rozważamy niektóre rodzaje zadań analizy danych.

II.1. Postacie danych – obiekty i zmienne

II.1.1. Przejdziemy obecnie do zarysowania założeń, jakie przyświecać będą dalszym częściom wykładu. Dotyczyć one będą najpierw postaci danych, jakimi się będziemy zajmowali. Założymy przede wszystkim, że dane mogą być zapisywane w postaci elektronicznej, a w szczególności – w postaci cyfrowej, po ewentualnie odpowiednim przekształceniu (zakodowaniu). Nie jest to bynajmniej założenie silnie ograniczające, ponieważ faktycznie praktycznie wszystkie dane, z jakimi się możemy zetknąć, mogą być w ten sposób zapisane. Prawda, że w niektórych przypadkach (utwory muzyczne) zapis ten z trudem poddaje się analizie, ale nie jest ona niemożliwa, choć na ogół jest skomplikowana. Idąc dalej, jeśli dane można zapisać cyfrowo, to są one w istocie po prostu liczbami (lub zestawieniami liczb). Tak więc mamy do czynienia z liczbami, jakkolwiek ich interpretacje i definicje mogą być bardzo różnorodne. W wielu przypadkach liczby te będą zatem zapisem pewnej „pierwotnej postaci” danych (np. ocena stylu w skokach narciarskich pochodząca od jednego z sędziów), dokonany ze względów czysto technicznych, podczas gdy analiza może być prowadzona zarówno na poziomie tego zapisu technicznego, jak i dalszych zapisów, będących jego wynikiem (punktacja skoku uwzględniająca długość oraz agregację ocen za styl od różnych sędziów).

II.1.2. *Przykład II.1.* Dobrym przykładem zestawu danych, ilustrującym wiele zagadnień związanych z postaciami danych, może być „opis kliniczny pacjenta”. Opis taki może zawierać: (i) wiek, (ii) płeć, (iii) wzrost, (iv) wagę, (v) temperaturę ciała, (vi) ciśnienie tętnicze krwi, (vii) tętno, (viii) wynik badania radiologicznego (na przykład w postaci zdjęcia lub tylko stwierdzenia określonego stanu na podstawie zdjęcia, powiedzmy – „złamanie otwarte kości podudzia”), (ix) klasyfikację jednostki chorobowej, (x) ogólną ocenę stanu pacjenta dokonaną przez lekarza dyżurnego (np. w punktach od 0 – „martwy”, do 10 – „całkowicie zdrowy”, poprzez, powiedzmy, 3 – „stan podkrytyczny”, lub 8 – „do leczenia ambulatoryjnego”). Daną może być tutaj zarówno każda z pozycji opisu (jeśli jest zaobserwowana) lub też cały opis. Łatwo zauważyć znaczną różnorodność poszczególnych pozycji tego opisu. Scharakteryzujemy je nieco dokładniej w poniższej tabelce.

Tab. II.1.

Pozycja opisu	Zakres wartości	Charakter	Możliwe operacje	Inne uwagi
i. Wiek	0-120 w latach	Mierzalny, obiektywny, stały	Porównywanie, operacje arytmetyczne	Można zawęzić do kilku przedziałów
ii. Płeć	Dwie wartości	Binarny, obiektywny	„ta sama – inna”	
iii. Wzrost	0-230 w cm	Mierzalny, obiektywny, stały	Porównywanie, operacje arytmetyczne	Można zawęzić do kilku przedziałów
iv. Waga ciała	0-200 w kg	Mierzalny, obiektywny, stały	Porównywanie, operacje arytmetyczne	Można zawęzić do kilku przedziałów
v. Temperatura	35° – 42° C	Mierzalny, obiektywny, zmienny	Porównywanie, operacje arytmetyczne	Można zawęzić do kilku przedziałów
vi. Ciśnienie	Dwie liczby – między 0 a 350	Mierzalny, obiektywny, zmienny	Porównywanie, operacje arytmetyczne	Można zawęzić do kilku przedziałów
vii. Tętno	0-250	Mierzalny, obiektywny, zmienny	Porównywanie, operacje arytmetyczne	Można zawęzić do kilku przedziałów
viii. Wynik badania	Opis werbalny	Subiektywny	Bardzo ograniczone	
ix. Klasyfikacja kliniczna	Kody jednostek chorobowych	Dyskretny, częściowo subiektywny	„ta sama – inna”	Ograniczona porównywalność w obrębie zbliżonych jednostek
x. Stan pacjenta	0-10	Dyskretny, subiektywny	Porównywanie	

Zaznaczmy, że podane tutaj charakterystyki poszczególnych pozycji opisu (danych lub elementów danych), takie, w szczególności, jak „obiektywny”, czy „subiektywny”, powinny być traktowane z odpowiednią ostrożnością – nie są to żadne charakterystyki formalne, a tylko elementy oceny ewentualnych danych. Tym niemniej, jest oczywiste, że „tętno” jest daną znacznie bardziej obiektywną niż, powiedzmy, „stan pacjenta”, nawet jeśli pomiar tętna jest obciążony istotnym błędem. Podobnie, dopuszczenie możliwości „operacji arytmetycznych” w stosunku do, na przykład, wieku, nie oznacza, że dopuszczalne i sensowne są dowolne operacje (np. sumowanie wieku pacjentów, czy dzielenie wieku jednego pacjenta przez wiek drugiego), ale, że w ogólności można na tego rodzaju danych takie operacje wykonywać

WYKŁAD Z METOD ANALIZY DANYCH

(np. średnia wieku wszystkich pacjentów lub w ich określonych grupach, czy różnice wieku pacjentów).

II.1.3. Bodaj najważniejszym aspektem konstruowania tabeli cech, takiej, jak tu pokazano na przykładzie, jest określenie *definicji* poszczególnych cech (np. definicji diagnostycznych poszczególnych jednostek klinicznych). O znaczeniu definicji dla możliwości otrzymywania informacji, i dalej wiedzy – z danych, już wspominaliśmy. Zauważmy, że nawet w obrębie zarysowanej tutaj w przykładzie dziedziny i to w odniesieniu do, wydawałoby się, oczywistych wielkości (jak ciśnienie tętnicze) mamy do czynienia z bardzo szerokim wyborem, który jest zależny od obranego kontekstu (uniwersum). W tym przypadku możemy bowiem mieć, na przykład, do czynienia z jednym, chwilowym pomiarem ciśnienia, ale też i z pomiarem, który stanowi element zaplanowanego dłuższego postępowania, podczas którego ciśnienie będzie mierzone wielokrotnie (stąd pytanie: czy wielkość (jakby) stała – bo jednorazowa, czy też zmienna – bo będzie mierzona wielokrotnie). Inne aspekty wyboru w tym przypadku dotyczyć mogą uproszczonego kodowania (np. „niskie”, „normalne”, ...), bądź innych sposobów uproszczonego zapisu.

II.1.4. Jeśli pozycje opisu, tak jak to zilustrowano w powyższej tabeli, są odpowiednio dobrze określone i ustalona jest procedura (czasami jest to procedura „niejawna”, na przykład wówczas, gdy dziecko obserwuje znajomych sobie ludzi i w pewien sposób ich charakteryzuje – być może zresztą w sposób całkiem zbliżony do podanego w tabeli) ich zapisu lub zapamiętywania, to możemy już zestawiać tabelę, w której umieszczone zostaną obiekty naszej obserwacji, w tym przypadku – pacjenci:

Tab. II.2.

Id pacjenta	i	ii	iii	iv	v	vi	vii	viii	ix	x
Jan Kowalski s. Jana	38	0	176	77	37,2	140/70	72	ooox	123/1	8
Piotr Strak s. Józefa	82	0	169	63	36,7	108/60	61	oooo	246/2	9
Anna Bocian c. Jacka	67	1	168	82	38,2	220/95	76	ooxx	378/0	4
...										

II.1.5. Przytoczona powyżej przykładowa tabela opisów konkretnych pacjentów jest typową ilustracją postaci danych, z jakimi się najczęściej spotykamy. Możemy wówczas mówić o *macierzy danych*, w której wiersze odpo-

wiadają *obiektom* (obserwacjom), zaś kolumny – *zmiennym* (atrybutom, cechom).

Macierz taką oznaczać będziemy przez X , zaś jej wiersze, opisy obiektów, przez x_i , gdzie $i \in I$ są indeksami obiektów (obserwacji) zawartych w zbiorze $I = \{1, \dots, n\}$ (zauważmy analogię do oznaczeń przyjętych w opisie teorii informacji Shannona). Poszczególne elementy macierzy oznaczone będą x_{ik} , $k \in K = \{1, \dots, m\}$, gdzie k jest indeksem cech obiektu, a K – zbiorem tych cech. Jeśli zajdzie potrzeba posługiwania się, niezależnie od wektorów $x_i = \{x_{ik}\}_k$, również kolumnami macierzy X , to oznaczać je będziemy przez $x_{\cdot k} = \{x_{ik}\}_i$.

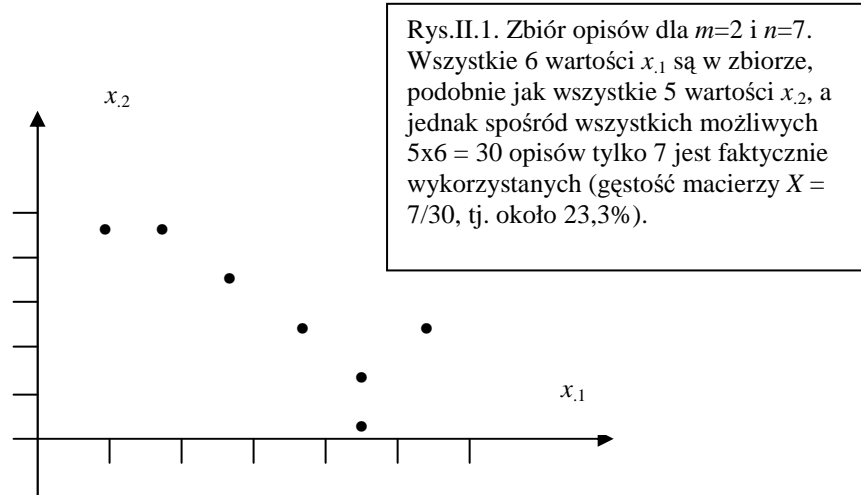
Tab. II.3.

Wiersze – obiekty (obserwacje):	Kolumny – cechy (zmienne):			
	$k=1$	$k=2$...	$k=m$
$i=1$	x_{11}	x_{12}	...	x_{1m}
$i=2$	x_{21}	x_{22}	...	x_{2m}
...
$i=n$	x_{n1}	x_{n2}	...	x_{nm}

Macierz X , której odpowiada zbiór indeksów obiektów I (dla uproszczenia będziemy zazwyczaj odwoływali się do tego ostatniego zbioru jako zbioru obiektów, chyba, że będą tego wymagały odpowiednie operacje rachunkowe), zawiera *zaobserwowane opisy obiektów*. Należy je jednak wyraźnie odróżnić od *możliwych opisów obiektów* (a zwłaszcza *wszystkich* możliwych opisów obiektów). I tak, na przykład, o ile w zbiorze opisów pacjentów przyjmowanych w szpitalnej izbie przyjęć można spodziewać się wystąpienia zarówno mężczyzn, jak i kobiet, czyli wszystkich możliwych wartości zmiennej „płeć”, to zapewne nie wystąpią w takim (ograniczonym) zbiorze niektóre wartości zmiennej „wzrost”, a także innych podobnych (mierzalnych, wielowartościowych) zmiennych. Jeśli jednak jest do pomyślenia, a nawet zdarza się niekiedy w praktyce, że zbiór opisów obiektów X zawiera wszystkie możliwe wartości wszystkich przyjętych zmiennych (zwłaszcza, jeśli są to zmienne dyskretne, nominalne lub porządkowe o niewielu wartościach), to wykluczamy w zasadzie taką możliwość w odniesieniu do całych zestawień wartości zmiennych, czyli opisów obiektów, które – nawet jeśli w całym zbiorze X wyczerpane byłyby wszystkie możliwe wartości poszczególnych zmiennych – mogą obejmować (i zazwyczaj obejmują) tylko niewielką część wszystkich możliwych opisów. Odpowiednią ilustrację przedstawia Rys. II.1.

W związku z powyższym wprowadzamy pojęcie przestrzeni opisów obiektów, \mathbf{X} , czyli $x_i \in \mathbf{X}$, a ponadto, jeśli pojmujemy X nie tylko jako macierz opisów, ale i jako zbiór dokonanych obserwacji, to mamy $x_i \in X \subseteq \mathbf{X}$. Prze-

strzeń \mathbf{X} jest iloczynem kartezjańskim przestrzeni (zbiorów wartości) poszczególnych zmiennych k , czyli $\mathbf{X} = \mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3 \times \dots \times \mathbf{X}_k \times \dots \times \mathbf{X}_m$, gdzie \mathbf{X}_k oznaczają zbiory możliwych wartości (przestrzenie) poszczególnych cech (zmiennych).



II.1.6. Macierzowa postać danych sprawia wrażenie bardzo sztywnej i ograniczonej. Jest ona faktycznie dopasowana do możliwości stosowania różnych technik obliczeniowych, mających na celu analizę danych i pozyskiwanie na ich podstawie informacji i wiedzy. Jednocześnie przecież można w niej zawrzeć, przynajmniej na poziomie „ideowym”, praktycznie prawie wszystkie używane rodzaje danych. Należy podkreślić, że jeśli jakiś rodzaj danych z trudem daje się przedstawić w postaci macierzowej, bądź reprezentacja ta jest tylko symboliczna, to możemy być pewni, że ten rodzaj danych może być przetwarzany i analizowany automatycznie tylko w bardzo ograniczonym stopniu.

Dobrym przykładem może być tutaj analiza obrazów. Jeśli mamy do czynienia (a tak jest zazwyczaj) z obrazem dwuwymiarowym, to może on być (i często jest) opisany przez: dwie zmienne położenia (pikseli) na płaszczyźnie), x_1 i x_2 , których zakres wartości jest zależny od wielkości obrazu i jego rozdzielczości, zmienną waloru (szarości), x_3 , o zakresie wynikającym z charakterystyki obrazu (np. 10 poziomów szarości – od 0 oznaczającego biel do 10 oznaczającego czern), oraz x_4 – dwu- lub trzy-cyfrowy kod koloru, naturalnego, lub sztucznego. Liczba obiektów odpowiada liczbie pikseli, czyli $\text{card}\mathbf{X}_1 \cdot \text{card}\mathbf{X}_2$. Jakkolwiek taki opis obrazu wydaje się sztuczny (naturalną „macierzą” jest tutaj raczej sam dwuwymiarowy obraz), jednak wykonywane operacje odpowiadają dokładnie interpretacji macierzowej.

Inne wątpliwości mogą się wiązać z opisami o – na pierwszy rzut oka – zmiennej długości, jak na przykład raporty kasowe odpowiadające poszczególnym transakcjom: każda przecież może dotyczyć innej liczby produktów, kupowanej przez poszczególnych klientów. Jeśli x_i mają odpowiadać poszczególnym takim transakcjom, wraz z opisem zakupów, to możliwych jest, w ramach „modelu macierzowego”, kilka rozwiązań. Pierwsze, najprostsze, polega na założeniu takiej liczby kolumn (m), która odpowiada największej liczbie różnych produktów w analizowanym zbiorze takich transakcji. Transakcje, które miałyby mniejszą liczbę pozycji, scharakteryzowane byłyby przez wektory x_i zawierające zera na dalszych pozycjach – poza pozycjami odpowiadającymi faktycznie dokonanym zakupom. Innym, stosowanym rozwiązaniem jest użycie macierzy X , w której liczba kolumn (m) odpowiada wszystkim sprzedawanym (rozdzielnym) produktom i wobec tego transakcje będą scharakteryzowane niezerowymi wartościami tylko w kolumnach odpowiadających poszczególnym produktom. Tak otrzymana macierz X będzie niewątpliwie bardzo „rzadka” (stosunek liczby niezerowych elementów macierzy do jej rozmiarów, tj. do wartości nm , będzie bardzo mały, zapewne wyrażnie poniżej 1%). Zauważmy, że ten rodzaj rozwiązań i – ogólniej – rozumowania stosuje się także, na przykład, do relacyjnych baz danych.

II.1.7. W Tab. II.1 przytoczyliśmy pobieżne charakterystyki poszczególnych zmiennych, czy cech, jakie zostały użyte w przykładzie. Określenie zmiennych jest niezwykle ważnym etapem analizy danych – etapem projektowania, jakkolwiek niekiedy może być również powtórzone już podczas właściwej analizy. Precyzyjne definicje zmiennych pozwalają na ich właściwą interpretację, ale także na ich prawidłowe przetwarzanie. Jest oczywiste, co już zaznaczyliśmy, że zmienna „wzrost” nie może być traktowana z rachunkowego punktu widzenia tak samo jak zmienna „stan pacjenta”, co zaznaczyliśmy już zgrubnie w Tab. II.1. Formalnie rzecz biorąc, rozróżnia się *skale* zmiennych, związane z dopuszczalnymi operacjami na tych zmiennych, podane w Tabeli II.4.

W nawiązaniu do określeń i charakterystyk podanych w Tabeli II.4 dodajmy, że skala *nominalna* odpowiada sytuacji, w której przypisujemy obiektom pewne, często wręcz symboliczne, nieporównywalne, wartości cech (np. „blondyn”, „szatyn”, „rudy”, „łysy”,...), o których możemy powiedzieć tylko tyle, że albo są takie same („to dwie blondynki”), albo, że są różne („to łysy i brunet”).

Skala *porządkowa* odpowiada sytuacji, w której jesteśmy co najwyżej w stanie uszeregować odpowiednie wartości cech („geniusz”, „talent”, „zdolny”, „przeciętniak”, „kujon”, „tępak”). Rozróżnienie między skalami *przebiegową* i *ilorazową* nie jest dla nas bardzo istotne, zresztą w istocie mamy

często do czynienia z sytuacjami pośrednimi i uznaniowymi, zwłaszcza, kiedy dane źródłowe są od razu przekształcane do celów dalszej obróbki.

Tab. II.4.

Skala	Dozwolone przekształcenia matematyczne	Dopuszczalne relacje	Dopuszczalne operacje arytmetyczne
<i>Nominalna</i>	$y=f(x)$, gdzie $f(x)$ – dowolne wzajemnie jednoznaczne przekształcenie	Równości ($x_{ik} = x_{i'k}$) Różności ($x_{ik} \neq x_{i'k}$)	Zliczanie zdarzeń (liczba relacji równości i/lub różności)
<i>Porządkowa</i>	$y=f(x)$, gdzie $f(x)$ – dowolna ściśle rosnąca funkcja	Powyższe, oraz Większości ($x_{ik} > x_{i'k}$) Mniejszości ($x_{ik} < x_{i'k}$)	Zliczanie zdarzeń (liczba relacji równości, różności, większości, mniejszości)
<i>Przedziałowa</i>	$y=a+bx$, przy czym $b>0$, zaś $x \in \mathbf{R}$; wartość zerowa przyjmowana konwencjonalnie	Powyższe oraz równości różnic i przedziałów ($x_{ik}-x_{i'k} = x_{jk}-x_{j'k}$)	Powyższe oraz dodawanie i odejmowanie
<i>Ilorazowa</i>	Praktycznie dowolne (często przyjmuje się 0 jako ograniczenie lewostronne)	Powyższe oraz równości ilorazów ($x_{ik}/x_{i'k} = x_{jk}/x_{j'k}$)	Powyższe oraz mnożenie i dzielenie

Za: Walesiak (2002).

II.1.8. Zgodnie z ostatnią uwagą z poprzedniego punktu formalne definicje skal powinny być traktowane z odpowiednim dystansem (co nie oznacza, że możemy sobie pozwolić, powiedzmy, na stwierdzenie, że „łysy”-„rudy” = „blondyn”-„szatyn”). Jednak, na przykład, niewątpliwie możemy stwierdzić, że różnica między „geniuszem” a „talentem” jest mniejsza niż między „geniuszem” a „kujonem”, a ważność tego stwierdzenia nie ogranicza się tylko do przytoczonego przykładu zmiennych porządkowych. W jeszcze większej mierze postulat „zdrowego rozsądku” stosuje się do skal przedziałowej i ilorazowej. W ogólności – interpretacja i dopuszczalne operacje rachunkowe zależne są od konkretnego znaczenia (definicji) poszczególnych zmiennych. Należy jedynie absolutnie przestrzegać raz przyjętych zasad, wynikających ze znaczenia poszczególnych zmiennych i ich konkretnych wartości.

II.1.9. Często spotyka się rozróżnienie między zmiennymi „jakościowymi” i „ilościowymi”. Jest to, znów, dość nieprecyzyjne rozróżnienie, wskazujące głównie na dopuszczalne operacje na zmiennych i ich wartościach. Można z góry powiedzieć, że zmienne ilościowe to zmienne mierzone przy pomocy skal – przedziałowej i ilorazowej, ale w niektórych przypadkach – także i porządkowej. Natomiast zmienne jakościowe są na ogół nominalne, choć mogą być także porządkowe. Tak więc możemy, na przykład, takie wyraźnie jakościowe określenia jak „celujący”, „bardzo dobry”, „dobry”, ... zakodować (być może tylko na naszą odpowiedzialność) w postaci liczb: 6, 5, 4, ... Jakkolwiek należy pamiętać, że w wielu przypadkach takie ilościowe kodowanie zmiennych jakościowych jest zabiegiem niesłychanie zgrubnym i upraszczającym, a zarazem wysoce arbitralnym, to jednak w praktyce tego rodzaju operacje są nierzadko wykonywane. W wielu takich sytuacjach występuje wyraźny „zgrzyt” niekonsekwencji, gdy kody – jakościowy (językowy) i ilościowy (liczbowy) nie przystają w sposób przekonywujący do siebie, a nie istnieje żadna formalna teoria pozwalająca na odpowiednio dokładne przeprowadzenie takiego kodowania bez poważnego stopnia arbitralności. Do zagadnienia tego będziemy zresztą jeszcze kilka razy wracali.

II.1.10. Podobnie, mówimy często o zmiennych „ciągłych” i „dyskretnych”, mając na myśli ciągłość i dyskretność (skokowość, nieciągłość) zbiorów wartości odpowiednich zmiennych. Zauważmy jednak, że nawet dla „teoretycznie” ciągłych zmiennych (wzrost, waga) faktyczne zbiory ich wartości są skończone, co wynika z zapisu wyników pomiaru i praktycznej przydatności skończonej długości zapisu (dokładności) tych wielkości, często niezbyt wielkiej (wzrost w centymetrach, waga w kilogramach, wiek w latach). Jednocześnie – zmienne „z natury” dyskretne mogą przyjmować stosunkowo wiele wartości (np. liczba lat nauki szkolnej). Tak więc rozróżnienie pomiędzy zmiennymi ciągłymi i dyskretnymi staje się – w praktyce – całkowicie względne, co nie oznacza, byśmy mieli zapominać o jego istnieniu (np. możliwość dokładniejszego pomiaru zmiennych ciągłych, inne zjawiska dotyczące pomiarów błędnych, itp.).

W nawiązaniu do poprzedniego punktu zaznaczmy jeszcze, że zmienne dyskretne są nierzadko kodami zmiennych jakościowych, co występuje szczególnie często w przypadku zmiennych *binarnych* (albo *zero-jedynkowych*), czyli przyjmujących tylko dwie wartości, na ogół 0 lub 1. Jeśli mianowicie nie są to wprost wartości (ilościowe) zmiennej (np. obecność lub brak pewnej cechy), to często są używane do kodowania jakościowych alternatyw (kobieta-mężczyzna, miasto-wieś, akceptacja-odrzuć, ...).

II.1.11. Innym aspektem opisu obiektów jest możliwość wystąpienia *braków* lub *luk* w danych. Może to oznaczać, że danej wielkości (cechy) nie udało się zmierzyć. Fakt ten jest, oczywiście, również pewną daną (niezależnie od

jej specyficzności), która można anonsować w macierzy danych w określony sposób (np., niezbyt oryginalnie: „-,„). Istnieje cały szereg prostych, ale także i dość skomplikowanych, technik pozwalających w lepszy lub gorszy sposób uwzględnić tę kwestię. Wrócimy jeszcze do niej w dalszych częściach wykładu.

Inna sytuacja występuje, gdy pewna zmienna (cecha) „nie stosuje się” do danego obiektu (np. kolor włosów do łysiego, przy czym nie mamy na myśli jego włosów z okresu młodości)¹. Wymaga to zazwyczaj odpowiedniego kodowania zmiennej – lub zmiennych – odnoszących się do tego zjawiska i przeważnie nie stanowi istotnego problemu, chyba, że poważnie zwiększa rozmiary zagadnienia przez dodawanie wartości zmiennych lub nowych zmiennych.

II.1.12. Dość naturalnym aspektem danych jest obecność różnorodnych *błędów*. Wynikają one z wielu różnych przyczyn (subiektywność, definicje, błędy przyrządów, zmiana warunków, ...). Poza stwierdzeniem faktu ich istnienia nie będziemy się nimi w tej chwili zajmowali – większość z nich powinna zostać w jakiś sposób uwzględniona w prezentowanych tutaj oraz znanych skądinąd metodach analizy danych, w tym zwłaszcza metodach statystyki matematycznej.

II.1.13. Poświęcimy także trochę miejsca kwestii *wewnętrznej spójności* danych. Otóż, jeśli w opisie pacjenta mamy zestawienie: wiek – 13 lat, wzrost – 154 cm, waga – 37 kg, to możemy uznać, że mamy do czynienia z bardzo szczupłą nastolatką lub nastolatkiem. Jednak, jeśli te same trzy zmienne mają wartości: wiek – 38 lat, wzrost – 186 cm, waga – 26 kg, to możemy spodziewać się jakiegoś poważnego odstępstwa, wynikającego bądź z błędu, bądź z zaistnienia niezwykle szczególnego przypadku. Jeszcze bardziej radykalnym przejawem niespójności byłoby, dla przykładu z Tab. II.1-2 wystąpienie następujących wartości zmiennych: **v.** 38.2°C, **vi.** 130/80, **vii.** 0, **x.** 0.

Z powyższych dwóch przykładów (dane „dziwne” i dane „absurdalne”) wynika, że niespójność danych jest w zasadzie funkcją już ustalonych *teorii* wiążących dane ze sobą (np. wiek-waga-wzrost). Jest to jeden z przejawów wpływu istniejącej wiedzy na ocenę danych.

Odpowiednie metody analizy danych, zresztą nie odbiegające w swoich zasadach od konsekwentnie stosowanego zdrowego rozsądku, zapewniają wykrycie większości sytuacji niespójności. Ich działanie (np. proste programy testowe) często są takie same, jak przy wykrywaniu ogólniejszej kategorii nieprawidłowości w danych, tj. błędów.

¹ „Jak się ma łysa śpiewaczka?” „Dobrze, ma dziś nową fryzurę” (Eugène Ionesco, „Łysa śpiewaczka”).

Trzeba jednak zarazem stwierdzić, że sam fakt wykrycia „dziwnej” danej nie tylko nie musi oznaczać, że jest ona błędna, ale wręcz może prowadzić do stwierdzenia bardzo ważnej informacji (entropia!!).

II.1.14. Jednym z istotnych aspektów odnoszących się do kwestii zmiennych i ich wartości, uwzględnianym w dużym zakresie w specjalizowanych pakietach programowych, jest *normalizacja* i *standaryzacja* zmiennych. Działania te związane są z wspomnianym poprzednio postulatem porównywalności. Nieco obszerniej potraktujemy ten temat przy omawianiu definicji odległości i bliskości. Tutaj tylko wspomnimy, że obie wspomniane operacje mają na celu, na przykład, sprowadzenie wartości różnych zmiennych do jednego, wspólnego przedziału. Ma to na celu zapewnienie porównywalności między zmiennymi. Na przykład, jeśli ktoś waży 136 kg, przy wzroście 203 cm, to po sprowadzeniu obu tych konkretnych wartości do przedziałów $[0,1]$, na podstawie minimalnych i maksymalnych wartości zmiennych wzrostu i wagi w badanej populacji, możemy nieco sensowniej powiedzieć jaka będzie pozycja obu wspomnianych wartości: czy waga, czy też wzrost będą (relatywnie, zatem) większe?

Innym ważnym skutkiem normalizacji jest *otrzymanie wielkości pozbawionych jednostek* (por. wzory (II.1)-(II.4), co jest w ogóle warunkiem porównywalności wielkości mierzonych w różnych jednostkach.

Istnieje kilka powszechnie przyjętych sposobów sprowadzania zmiennych do porównywalności. Przytoczymy tutaj kilka podstawowych, wraz z odpowiednim komentarzem. Przyjmiemy przy tym dodatkowe oznaczenie, a mianowicie wyniki bezpośrednich pomiarów, lub inne dane źródłowe, oznaczać będziemy χ_{ik} , zachowując oznaczenie x_{ik} dla wyników normalizacji czy standaryzacji, jako wielkości dalej przetwarzanych. I tak, możemy stosować następujące transformacje zmiennej k -tej opisu obiektów:

$$x_{ik} = \chi_{ik} / \max_j \chi_{jk} \quad (\text{II.1})$$

$$x_{ik} = \chi_{ik} / E_k \chi_{.k} \quad (\text{II.2})$$

$$x_{ik} = (\chi_{ik} - \min_j \chi_{jk}) / (\max_j \chi_{jk} - \min_j \chi_{jk}) \quad (\text{II.3})$$

$$x_{ik} = (\chi_{ik} - E_k \chi_{.k}) / V_k \chi_{.k} \quad (\text{II.4})$$

przy czym oznaczenie $E_k \chi_{.k}$ dotyczy centralnej statystyki rozkładu (empirycznego) zmiennej $\chi_{.k}$, najczęściej średniej, jakkolwiek w niektórych przypadkach może to być mediana, a nawet i moda, zaś $V_k \chi_{.k}$ oznacza wariancję albo inną sensowną miarę rozrzutu rozkładu. Transformację (II.1) najlepiej jest stosować wtedy, gdy wartości χ_{ik} zawarte są między 0 a pewną dodatnią wartością $\max_k \chi_{.k}$. Jeśli wartości zmiennej k -tej są dodatnie, ale nie zaczynają się od (lub w pobliżu) zera, lecz od pewnej większej wartości, $\min_k \chi_{.k}$, porównywalnej z $\max_k \chi_{.k}$, to lepiej jest użyć transformacji (II.3). Podobnie w

przypadku transformacji (II.2) – stosujemy ją do rozkładów o wartościach nieujemnych. Transformacja ta odnosi wartości przyjmowane w badanym zbiorze danych przez zmienną k -tą do średniej lub podobnej statystyki, a zatem nie przeprowadza, jak (II.1), zbioru wartości tej zmiennej do przedziału $[0,1]$. Jeśli x_{ik} po tej transformacji przybierze wartość 1, to znaczy, że jest równe, w zależności od przyjętego w (II.2) mianownika, średniej, medianie, lub modzie. Z kolei, transformacja (II.3) pozwala na ujęcie zbiorów wartości x_{ik} , które niekoniecznie są nieujemne (choć jej sens jest znacznie szerszy niż tylko rozszerzenie normalizacji na wartości ujemne, o czym już wspomnieliśmy). Wreszcie, transformacja (II.4), będąca właściwą standaryzacją, może być stosowana do dowolnych zakresów wartości x_{ik} , a jej sens polega na sprowadzeniu rozkładu wartości x_{ik} do scentrowanego rozkładu (o średniej lub medianie równej zero) o jednostkowym rozproszeniu (wariancja równa jeden).

Stosując różne sposoby normalizacji musimy pamiętać o znaczeniu interpretacyjnym transformacji, jakie posłużyły do otrzymania z pierwotnych x_{ik} dalej używanych x_{ik} .

II.1.15. Niezależnie od normalizacji i standaryzacji stosowane są inne, na ogół nieskomplikowane, zabiegi zmierzające do otrzymania x_{ik} w takiej postaci, w jakiej chcielibyśmy je dalej przetwarzać i analizować. Jeden z takich zabiegów był już wspomniany w Tab.II.1. Polega on na podziale zakresu wartości zmiennej na kilka przedziałów i wyrażanie wartości tylko w odniesieniu do tych przedziałów (np., wyrażenie wzrostu w odniesieniu do mężczyzn poprzez następujące, tutaj arbitralnie przyjęte, kategorie: „bardzo wysoki” – powyżej 190 cm, „wysoki” – między 178 cm a 190 cm, „średniego wzrostu” – między 168 cm a 177 cm, „niski” – między 159 cm a 167 cm, i „bardzo niski” – poniżej 159 cm).

Zabieg ten jest usprawiedliwiony na dwa sposoby: *po pierwsze* – bardzo często konkretna wartość zmiennej jest dla nas mniej interesująca (np. czy 184 cm czy 186 cm wzrostu) niż bardziej zgrubnie ocenione położenie w zakresie wartości tej zmiennej, a *po drugie* – również bardzo często określonym przedziałom wartości zmiennej można przypisać dość powszechnie właściwie rozumiane znaczenia (pojęcia), tak jak w przytoczonym tutaj przykładzie. Nierzadko ten rodzaj zabiegu można stosować do więcej niż jednej zmiennej naraz. I tak, grupę pojęć odnoszących się do relacji między wzrostem a wagą („bardzo chuda”, „chuda”, „szczupła”, „w normie”, „przykości”, „gruba”, „otyła”, „...”) definiuje się w sensie nie przedziałów zakresów wartości poszczególnych zmiennych, ale zakresów relacji między nimi. Posługiwanie się takimi uproszczonymi wartościami zmiennych, odpowiadającymi wartościom o charakterze lingwistycznym (wyrażeniom języka naturalnego), jest bardzo wygodne i pozwala na efektywne komunikowanie się z

ewentualnym użytkownikiem narzędzi lub wyników analizy danych. Jak zwykle, jednak, i ten kij też ma dwa końce: rzadko się zdarza, żeby granice przyjętych przedziałów były uzasadnione w jednoznaczny, obiektywny sposób (jak to ma miejsce, na przykład, w przypadku pojęć „ciało stałe”, „ciecz”, „gaz” – w odniesieniu do temperatury pewnej substancji). Za uproszczenie i możliwość łatwiejszego komunikowania się płacimy zazwyczaj arbitralnością określenia odpowiednich przedziałów i ewentualnymi, związanymi z tym, problemami. Należy w każdym razie być świadomym kwestii związanych z tego rodzaju przekształceniami zmiennych.

Inne przekształcenia obserwowanych wartości mają na ogół charakter techniczny (np. odniesienie do pewnych przyjętych norm), związany z zagadnieniami merytorycznymi dziedzin, z których pochodzą analizowane dane, i nie będziemy się nimi tutaj zajmowali.

II.1.16. W tym miejscu wspomnimy o możliwości stosowania metod teorii prawdopodobieństwa, a zatem i statystyki matematycznej, do zbiorów danych o charakterze podobnym do zarysowanego tutaj. Otóż kluczowe ograniczenie stosowalności tych metod jest związane z wymaganiami tych metod co do rodzaju danych. Muszą to być dane „porządne”, a więc dane o jednokowym charakterze dla różnych zmiennych. Nawet mianowicie jeśli metody statystyki matematycznej mogą być stosowane dla różnych rodzajów zmiennych (np. ciągłych, dyskretnych, binarnych), to z zasady wszystkie rozpatrywane zmienne powinny mieć ten sam charakter. Co więcej, zastosowanie metod statystyki matematycznej jest dalej jeszcze ograniczone przez fakt, że istniejące wyniki z tej dziedziny z reguły wymagają założenia podobieństwa lub analogiczności, albo wręcz identyczności rozkładów prawdopodobieństwa dla różnych zmiennych. Istnieją co prawda wyniki dla mniej wymagających założeń, jak również możliwości przekształcania zmiennych, prowadzące w wielu przypadkach do otrzymania analogicznych rozkładów prawdopodobieństwa, ale otrzymany w ten sposób zakres „swobody” w analizowaniu różnych cech zbiorów danych jest bardzo ograniczony.

Stąd, a także ze względu na szczupłość wykładu, odwoływać się będziemy w jego trakcie wyłącznie do metod, mających charakter „manipulacji” danych, zmierzających do otrzymania ich ewentualnych intuicyjnie interpretowalnych struktur („informacji” i „wiedzy”), jakkolwiek metody te często znajdują również zastosowanie w statystyce matematycznej, w ramach bardziej formalnie uzasadnionych podejść.

Z drugiej strony, należy pamiętać, że metody, które nie odwołują się - w sposób uzasadniony, a nie tylko poprzez użycie sztafażu pojęć i definicji – do teorii prawdopodobieństwa i twierdzeń statystyki matematycznej, mają ograniczone zastosowanie, jeśli idzie o wnioski wyciągane z ich wyników (stosowalność do danego zbioru danych i ewentualnie bardzo zbliżonych

zbiorów danych). Jeśli nasza praca ma prowadzić do ogólniejszych wniosków, mających zastosowanie szersze niż wyprowadzone dla pewnego konkretnego zbioru danych to powinniśmy zatem starać się znaleźć uzasadnienie, które musi odwoływać się do takich pojęć jak charakter rozkładów (gęstości, prawdopodobieństwa) i ich cechy, bądź przeprowadzić odpowiednie testy dla innych zbiorów danych o podobnym charakterze, mogących uzasadnić szerszy charakter naszych wniosków.

II.1.17. W tym miejscu także odpowiednia wydaje się uwaga dotycząca zasad stosowania wielu istniejących pakietów analizy danych, w tym zarówno najbardziej standardowych (jak np. Excel), jak i specjalizowanych (jak np. SPSS). Trzeba mieć mianowicie na względzie fakt, że pakiety te, jakkolwiek na ogół bardzo rozwinięte i wyposażone w bardzo wiele funkcji i opcji, nie dają możliwości odpowiednio elastycznego traktowania poszczególnych zmiennych w zależności od ich interpretacyjnego sensu. Jest to całkowicie domena projektanta-użytkownika.

Niezależnie od tego ograniczenia elastyczności musimy pamiętać przy każdorazowym zastosowaniu takich aplikacji pakietowych, że odpowiednie definicje zmiennych mogą w nich w ogóle nie pasować do naszych potrzeb, a w każdym razie wymagać szczególnej ostrożności interpretacyjnej.

II.2. Odległości i bliskości

II.2.1. Kiedy nasze dane są już (albo mogą być) przedstawione w postaci macierzowej, można zacząć wykonywać na nich pewne podstawowe operacje. W istocie, znaczna część tych operacji wykonywana jest przez nas wszystkich na poły nawet świadomie w każdej niemal chwili. Mówimy tu, w szczególności, o postrzeganiu *podobieństw* i *różnic* między obiektami.

I tak, jeśli mamy troje pacjentów opisanych, powiedzmy, przez następujące wartości, odpowiednio, wzrostu, wagi i temperatury ciała:

(1) 168 67 36,9 (2) 170 69 37,1 (3) 186 92 38,9

to naturalnie zauważymy, że opisy dwóch pierwszych są w istocie bardzo podobne, podczas, gdy trzeci różni się wyraźnie od dwóch pierwszych. Ta nasza dość oczywista intuicja może zostać w prosty sposób potwierdzona rachunkowo: policzmy mianowicie sumę (wartości absolutnych) różnic pomiędzy poszczególnymi wartościami zmiennych dla par tych hipotetycznych pacjentów. Otrzymamy wówczas:

$$[(\text{pacjent 1})-(\text{pacjent 2})] = |168-170| + |67-69| + |36,9-37,1| = 2+2+0,2 = \mathbf{4,2};$$

i analogicznie:

$$[(\text{pacjent 1})-(\text{pacjent 3})] = 18+25+2 = \mathbf{45};$$

$$[(\text{pacjent 2})-(\text{pacjent 3})] = 16+23+1,8=\mathbf{40,8}$$

(zauważmy, że pominęliśmy tutaj milczeniem fakt, że dodawaliśmy do siebie centymetry, kilogramy oraz stopnie Celsjusza, czego można było łatwo uniknąć, stosując normalizację). Jak widać, otrzymaliśmy liczby, odzwierciedlające zróżnicowanie między, z jednej strony, pacjentem (1) i (2) oraz, z drugiej, pacjentami (1) i (3), a także (2) i (3), różniące się o rząd wielkości, co doskonale usprawiedliwia naszą intuicję, że pacjenci (1) i (2) są „podobni”, zaś pacjent (3) – wyraźnie od nich różny.

II.2.2. Analogiczne operacje, pozwalające na ocenę (w tym także automatyczną) podobieństw i różnic między obiektami w ich większych zbiorach, prowadzą do zdefiniowania *odległości* i *podobieństw* między obiektami. Wielkości te są wyznaczane w sposób praktycznie taki sam, jak to zilustrowaliśmy na przykładzie, z, naturalnie, wieloma wariantami o charakterze technicznym, które jednak mają swoje uzasadnienie interpretacyjne.

II.2.3. Odległość między obiektami $i, j \in I$ (albo, ogólniej, $x \in \mathbf{X}$), oznaczać będziemy d_{ij} , a jeśli będziemy się odwoływać wprost do opisów obiektów, to $d(x_i, x_j)$, bądź, powiedzmy, $d(x, y)$. Podobieństwo (bliskość) tych obiektów będzie oznaczona s_{ij} , albo $s(x_i, x_j)$, albo wreszcie $s(x, y)$. O obu tych funkcjach założymy, że przeprowadzają pary punktów z przestrzeni \mathbf{X} w *nieujemne liczby rzeczywiste*. Zakładamy bowiem, że nie ma sensu mówić o ujemnych odległościach bądź podobieństwach. Podobnie, w znacznej większości przypadków będziemy zakładali, że zarówno odległości, jak i bliskości, są *symetryczne* względem obiektów, tj. odległość „od i do j ” jest równa odległości „od j do i ” ($d_{ij}=d_{ji}$), i analogicznie dla bliskości ($s_{ij}=s_{ji}$), co wydaje się na pierwszy rzut oka oczywiste.

Jakkolwiek założenie symetryczności wydaje się faktycznie oczywiste, nie jest ono bynajmniej takim, a w każdym razie nie we wszystkich praktycznych przypadkach jest spełnione. Dotyczy to np. odległości, mierzonej w sensie czasu przejazdu samochodem, między dwoma punktami miasta. W godzinach rannego szczytu dojazdów do pracy tak mierzona odległość d_{ij} między położonym na peryferiach i a ulokowanym w centrum j będzie znacznie większa niż d_{ji} . Innym, znanym z życia przypadkiem jest odległość w jednowymiarowej i ukierunkowanej („strzałka czasu”) przestrzeni czasu między Bożym Narodzeniem a Wielkanocą: od Bożego Narodzenia do Wielkanocy jest na ogół 3-4 miesiące, podczas, gdy od Wielkanocy do Bożego Narodzenia – 8-9 miesięcy.

II.2.4. W literaturze przyjmuje się często różne dodatkowe (poza nieujemnością i symetrią) warunki na odległości i podobieństwa, które wywodzą się ze specyficznych cech przestrzeni \mathbf{X} , dla których są one formułowane (np. tzw.

nierówność trójkąta, tj., że dla wszystkich $i, j, i' \in I$ [a faktycznie – dla wszystkich, dowolnych $x_i, x_j, x_{i'} \in \mathbf{X}$] zachodzi nierówność $d_{ij} + d_{ji'} \geq d_{i'i}$, spełniana przez geometrię Euklidesową i przez większość rozważanych przestrzeni). My natomiast, głównie ze względu na zupełnie dowolny charakter rozważanej przez nas przestrzeni \mathbf{X} , nie będziemy zakładali żadnych innych warunków, w tym także wspomnianej nierówności trójkąta, poza nieujemnością i symetrią.

II.2.5. Jeśli jednak dla danego zbioru obiektów definiujemy jednocześnie odległość i podobieństwo (bliskość), to wymagania interpretacyjne narzucają, dla dowolnych trzech obiektów i, i', i'' warunek

$$d_{ii'} \leq d_{ii''} \Leftrightarrow s_{ii'} \geq s_{ii''} \quad (\text{II.5})$$

czyli – jeśli odległość między dwoma danymi obiektami jest większa niż pomiędzy dwoma innymi, to ich bliskość musi być mniejsza, i odwrotnie, co jest zrozumiałe.

II.2.6. Przy wyznaczaniu odległości między skróconymi opisami pacjentów w przykładzie z punktu II.2.1 posłużyliśmy się następującym, dość naturalnym schematem:

- (a) wyznacz odległości (w tym przypadku różnice) dla poszczególnych zmiennych k (wzrost, waga, temperatura), i następnie
- (b) zagraj te m odległości do jednej wartości d_{ij} .

Schemat ten obowiązuje także i w znacznej większości istniejących definicji odległości i podobieństw. Niekiedy, w sytuacjach, w których analizowane zmienne mają dokładnie ten sam charakter (lub zostały do takiego samego charakteru sprowadzone przez normalizację i/lub standaryzację), stosowane są definicje, w których nie da się rozdzielić łatwo etapów (a) i (b). Na razie jednak zajmiemy się krótkim przeglądem definicji, w którym założymy, że taki rozdział można przeprowadzić. Jest on zwłaszcza istotny w sytuacjach o charakterze ogólniejszym niż te, w których mamy do czynienia ze zmiennymi o jednolitych własnościach, to znaczy, gdy jednocześnie staramy się uwzględnić w opisach obiektów, a także i w odległościach czy bliskościach między nimi, zmienne „jakościowe” i „ilościowe”, nominalne i ciągłe, itp.

II.2.7. Odległości dla poszczególnych zmiennych (*odległości cząstkowe*). Dla tych wielkości stosować będziemy oznaczenia d_{ij}^k oraz s_{ij}^k . Poniżej podajemy kilka określeń, jakie bodaj najczęściej występują w praktycznych zastosowaniach.

$$d_{ij}^k = 0, \text{ jeśli } x_{ik} = x_{jk}, \text{ oraz } = 1, \text{ jeśli } x_{ik} \neq x_{jk} \quad (\text{II.6})$$

$$d_{ij}^k = |x_{ik} - x_{jk}|. \quad (\text{II.7})$$

Definicja (II.6) stosowana jest przede wszystkim do zmiennych nominalnych, dla których możemy tylko stwierdzić identyczność lub różność cech dwóch obiektów. Może one być także stosowana do zmiennych porządkowych, jeśli tak wynika z naszej interpretacji znaczenia tych zmiennych. Druga z definicji może być stosowana do wszystkich zmiennych, które w ogólności nazywamy, niezbyt precyzyjnie, jak to już zaznaczyliśmy poprzednio, „ilościowymi”.

Powyższe dwie definicje wydają się wyczerpywać większość określeń faktycznie stosowanych w praktyce. Inne, często podawane w literaturze, są w istocie tylko ich wariantami. I tak, pewnym wariantem definicji (II.6) jest niekiedy spotykane określenie

$$d_{ij}^k = 0, \text{ jeśli } x_{ik}=x_{jk}=1, \text{ oraz } = 0 \text{ w pozostałych przypadkach, (II.8)}$$

stosowane zazwyczaj do zmiennych binarnych, tj. $X_k = \{0,1\}$. Przypisanie wartości zero odległości tylko wtedy, gdy wartości cechy dla obu obiektów równe są 1, odpowiada uznaniu, że wartość 0 tej cechy ma charakter np. „braku pomiaru” lub w ogóle „braku cechy” i wobec tego jej taka sama wartość nie wskazuje na „podobieństwo” dwóch obiektów pod tym względem (co, jak widać, jest całkowicie sprawą interpretacji).

Z kolei, definicja (II.7) występuje często w postaci

$$d_{ij}^k = |x_{ik}-x_{jk}| / |\max_i x_{i'k}-\min_i x_{i'k}| \quad (\text{II.9})$$

wprowadzającej element normalizacji do definicji odległości, analogicznie do poprzednio rozpatrywanej normalizacji dla zmiennych. Podobnie jak i dla zmiennych, możemy mieć do czynienia z różnymi wariantami sposobów normalizacji odległości, w szczególności o mianownikach innych niż w (II.9).

II.2.8. Bliskości (podobieństwa) dla poszczególnych zmiennych (*bliskości cząstkowe*). W wielu zastosowaniach i w wielu podejściach (np. niektórych opartych na teorii prawdopodobieństwa) pojęcie bliskości par obiektów jest co najmniej równie naturalne, jak pojęcie odległości. W szczególności, łatwo zauważyć, że definicja (II.6) może być niezwykle łatwo przekształcona tak, by wyrażała bliskość, a mianowicie:

$$s_{ij}^k = 1, \text{ jeśli } x_{ik}=x_{jk}, \text{ oraz } = 0, \text{ jeśli } x_{ik} \neq x_{jk} \quad (\text{II.10})$$

gdzie po prostu zamieniliśmy miejscami 1 i 0 z definicji (II.6). Nieco gorzej sytuacja przedstawia się ze zmiennymi „ilościowymi”, dla których, na poziomie poszczególnych zmiennych, odległość wydaje się być bardziej naturalnym pojęciem, a to ze względu na zagadnienie *skali*, a więc znów – konieczności pewnego rodzaju normalizacji lub odniesienia. I tak, jeśli ze-

chcemy wyznaczyć bliskość przy pomocy definicji podobnej do (II.7), to musimy się posłużyć dodatkową wielkością, określającą skalę (a także pozwalającą uniknąć ujemnych wartości bliskości), np.

$$s_{ij}^k = \max_{i', j'} |x_{i'k} - x_{i''k}| - |x_{ik} - x_{jk}|. \quad (\text{II.11})$$

Analogicznie możemy zdefiniować s_{ij}^k w nawiązaniu do definicji (II.9):

$$s_{ij}^k = 1 - (|x_{ik} - x_{jk}| / |\max_i x_{ik} - \min_i x_{ik}|). \quad (\text{II.12})$$

W ogólności, jeżeli z kontekstu zadania wynika, że dla niektórych zmien-nych bardziej naturalne jest określenie odległości, a dla innych – bliskości, i w takich różnych postaciach dokonywane są odpowiednie operacje, to możemy stosować pewne standardowe przekształcenia, na przykład, jeśli zachowane są warunki normalizacji, to

$$s_{ij}^k = 1 - d_{ij}^k \text{ oraz } d_{ij}^k = 1 - s_{ij}^k \quad (\text{II.13})$$

choć możliwe są również inne transformacje odległość-bliiskość.

II.2.9. Agregacja. Poznaliśmy już, w przykładzie z punktu II.2.1, najprostszy sposób agregacji odległości (lub bliskości) cząstkowych, czyli ich sumowanie. I tak, najczęściej spotykanymi sposobami agregacji, tutaj przedstawionymi wyłącznie dla odległości, są:

$$d_{ij} = \sum_k d_{ij}^k \quad (\text{II.14})$$

$$d_{ij} = (\sum_k d_{ij}^k) / m \quad (\text{II.15})$$

$$d_{ij} = (\sum_k (d_{ij}^k)^a)^{1/a} \quad (\text{II.16})$$

$$d_{ij} = \min_k d_{ij}^k \quad (\text{II.17})$$

$$d_{ij} = \max_k d_{ij}^k \quad (\text{II.18})$$

przy czym definicja (II.16), zwana uogólnioną odległością Minkowskiego, dla której zazwyczaj zakłada się $a \geq 1$, obejmuje takie bardzo znane i popularne definicje (zwłaszcza, jeśli odległości cząstkowe są oparte na zmien-nych ilościowych, najchętniej ciągłych) jak odległość Euklidesowa, oparta na twierdzeniu Pitagorasa ($a=2$), czy odległość „taksówkowa” („city block” albo „Manhattan”), dla $a=1$. Dwie ostatnie przytoczone definicje, a więc (II.17) i (II.18), są granicznymi, względem a , przypadkami definicji (II.16).

II.2.10. Należy także pamiętać o tym, że zdarzają się sytuacje, w których pewne odległości bądź bliskości są dane „gotowe”, nie zaś za pośrednic-

twem „położen” obiektów x_i w przestrzeni \mathbf{X} . Tak może być wówczas, gdy, na przykład, jednym z elementów opisu obiektów jest macierz odległości szosowych lub kolejowych między danymi miejscowościami (lokalizacjami obiektów), rozpatrywanymi w zagadnieniu, które to odległości, jak wiadomo, nie wynikają bynajmniej wprost z położenia geograficznego (a tym bardziej, gdyby te odległości były wyrażone, powiedzmy, nie w jednostkach długości, lecz czasu). W takich przypadkach traktujemy te „dane” odległości czy bliskości tak samo jak inne odległości czy bliskości cząstkowe d_{ij}^k , pamiętając o odpowiednich regułach dotyczących normalizacji, sensu interpretacyjnego, itp.

II.2.11. Podane tutaj definicje odległości i bliskości dalece nie wyczerpują zakresu wykorzystywanych definicji, zarówno tych znanych szerzej i używanych wielokrotnie, w różnych pracach i różnych zastosowaniach, jak – zwłaszcza – definicji specjalizowanych, często wprowadzanych dla danego konkretnego zastosowania. Jakkolwiek określenie odległości i bliskości między obiektami (obserwacjami) wydaje się – i słusznie – być etapem wstępnym dalszej, „właściwej”, analizy danych, to jednak w wielu przypadkach staje się zagadnieniem samym w sobie (niezależnie od znaczenia merytorycznego tego etapu, które pozostaje zawsze bardzo wysokie). I tak, w niektórych zastosowaniach analizy danych wymagających wprowadzenia odległości, na przykład w badaniach genetycznych, wyznaczanie (obliczanie) wartości odległości często jest wykonywane przy pomocy specjalizowanych algorytmów optymalizacyjnych (np. wtedy, gdy odległość jest zdefiniowana jako minimum przy określonych ograniczeniach z pewnego obszernego zbioru możliwych „dróg” między obiektami). W takich przypadkach odległość nie tylko ma zasadnicze znaczenie merytoryczne dla dalszego ciągu analizy, ale i nakład obliczeniowy, związany z jej wyznaczeniem, może być decydujący dla pracochłonności całej procedury.

II.2.12. *Przykład II.2.* Dla zilustrowania znaczenia sposobu wyznaczania odległości i nietrywialności tego zagadnienia przytoczymy przykład z praktyki analitycznej autora (ewentualnie zainteresowanych wynikami, zresztą ciekawymi i wiele mówiącymi, odsyłam do oryginalnych publikacji: Owsieński i Zadrożny, 2000, 2001).

Analizowanym zbiorem danych były głosowania Sejmu RP kadencji 1993-1997, dokładnie: 96 pierwszych głosowań tego Sejmu. Macierz danych składała się z niecałych 460 obiektów (wierszy), odpowiadających poszczególnym posłom (bez nazwisk i przynależności partyjnej). Zmiennymi były kolejne głosowania ($k=1, \dots, 96$, $m=96$ zmiennych), a wartościami tych zmiennych dla poszczególnych obiektów-posłów były zachowania w trakcie głosowań. Rozróżniano, zgodnie z protokołem kancelarii Sejmu, pięć „wartości” tych zachowań, czyli pięć wartości zmiennych (dla wszystkich 96

zmiennych zbiór tych wartości był, naturalnie, taki sam). Te pięć wartości to: 1. „za”, 2. „przeciw”, 3. „wstrzymanie się od głosu”, 4. „nie branie udziału w głosowaniu [mimo obecności na sali obrad]”, 5. „nieobecność”. Zadaniem analityków było podzielenie posłów na grupy według podobieństwa ich zachowań w całym zestawie 96 głosowań (bez uwzględniania, jak wspomniano, ich przynależności partyjnej).

Jednym z pierwszych zagadnień było zdefiniowanie odległości między posłami w 96-wymiarowej przestrzeni zachowań podczas głosowań (czyli różnic zachowań względem wszystkich 96 głosowań naraz). Ponieważ wszystkie zmienne mają identyczny charakter, więc zasadniczą sprawą było określenie odległości cząstkowej dla każdej z nich.

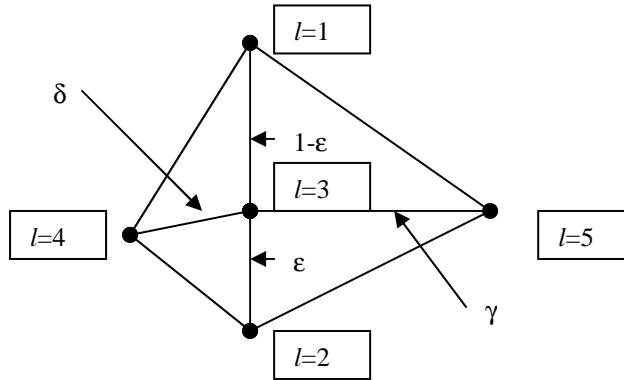
Wiemy, że wszystkie przestrzenie opisów cząstkowych, X_k , $k=1, \dots, 96$, składają się z pięciu wartości. Wartości te oznaczmy przez l , $l=1, \dots, 5$. Ponieważ wszystkie zmienne są identyczne co do charakteru, możemy w ogólności rozpatrywać – na zasadzie definicji odległości cząstkowych – po prostu odległości między poszczególnymi wartościami l , oznaczone, dla uproszczenia, $d(l, l')$. Aby móc takie praktyczne definicje wprowadzić, musimy oprzeć się na pewnych intuicyjnie oczywistych założeniach, odpowiadających przedmiotowi analizy.

Pierwszym takim (oczywistym chyba) założeniem jest, że $d(l, l)=0$ dla wszystkich $l=1, \dots, 5$, czyli, że jeśli w jakimś głosowaniu k -tym dwóch posłów zachowało się tak samo ($l=l$), to dla tego głosowania odległość cząstkowa między nimi wynosi zero, niezależnie od tego, jaki to był rodzaj zachowania.

Następnym założeniem było, że największą możliwą wartością odległości cząstkowej jest 1, przy czym występuje ona dla $d(„za”, „przeciw”) = d(1, 2) = d(„przeciw”, „za”) = d(2, 1)$. Założenie to łączy element merytoryczny (największa możliwa rozpiętość zachowań to głosowanie przeciwstawne) z technicznym (normalizacja do przedziału $[0, 1]$).

W tym miejscu należy zaznaczyć, że wstrzymanie się od głosu jest w większości przypadków w istocie opowiedzeniem się po stronie „przeciw” (w większości głosowań wymagana jest większość „za”), jakkolwiek dokładne określenie znaczenia tego zachowania wymagałoby stwierdzenia (i) rodzaju głosowania (obowiązująca Konstytucja RP przewiduje aż pięć (tak!) rodzajów większości w Parlamencie), oraz (ii) konkretnej sytuacji (np., czy wynik głosowania nie był z góry przesądzony na skutek określonej „arytmetyki partyjnej”). Ponieważ nie było możliwe przeprowadzenie takiej analizy dla wszystkich 96 głosowań, założono, że odległość $d(1, 3) = d(3, 1) = 1 - \epsilon$, zaś $d(2, 3) = d(3, 2) = \epsilon$, gdzie $\epsilon \in [0, 1]$ jest pewnym parametrem, który może być zmieniany (jakkolwiek pozostaje stały dla wszystkich zmiennych), zaś jego wartość wyjściowa została określona na $1/2$.

Sposób potraktowania pozostałych odległości, wraz z poprzednio już skomentowanymi, pokazano na poniższym schematycznym rysunku, zgodnie z którym odległości $d(3,4)$ i $d(3,5)$ są zależne od innych dwóch parametrów, oznaczonych δ i γ . Ponadto, co wynika z pierwszego z uczynionych założeń, sumy par sąsiednich odległości między $l=3$ a pozostałymi wartościami zachowań nie mogą przekroczyć 1 (tj. $d(3,1)+d(3,4)\leq 1$, $d(3,1)+d(3,5)\leq 1$, $d(3,5)+d(3,2)\leq 1$, $d(3,2)+d(3,4)\leq 1$).



Analiza prowadzona była dla kilku wybranych wartości parametrów ϵ , δ i γ , poczynając od $\epsilon=1/2$ oraz $\delta=\gamma=0$, w celu sprawdzenia zależności otrzymywanych wyników od tych parametrów, a zatem i definicji oraz interpretacji różnicowań między zachowaniami posłów (zwłaszcza znaczenia zachowań takich jak obecność bez głosowania i nieobecność).

Jak widać zatem, wyznaczanie odległości może być niezmiernie istotnym elementem procedury analizy danych, w dodatku nietrywialnym zarówno z punktu widzenia merytorycznego, jak i technicznego.

II.2.13. Zauważmy jeszcze, że macierz danych $X = \{x_{ik}\}_{ik}$ może być „widziana” również w perspektywie transpozycji, tj. obrócona o 90° w lewo. W takiej perspektywie wiersze stają się kolumnami (obiekty – zmiennymi) i odwrotnie – kolumny stają się wierszami (zmienne – obiektami). Odległości zdefiniowane dla obiektów będą w takiej sytuacji policzone dla zmiennych. Jest to szczególnie istotne wobec faktu, że istnieją definicje, wywodzące się ze statystyki, które zostały sporządzone dla zmiennych (por. Wykład I: entropia i informacja, teoria Claude Shannona). Najbardziej znaną wśród nich jest prawdopodobnie korelacja (współczynnik korelacji), mający interpretację podobieństwa (oryginalnie: podobieństwa zmiennych), czyli

$$s_{ij} = \frac{\sum_k (x_{ik}x_{jk} - x_i^{\text{sr}}x_j^{\text{sr}})}{\sum_k (x_{ik}^2 - (x_i^{\text{sr}})^2) \sum_k (x_{jk}^2 - (x_j^{\text{sr}})^2)}, \quad (\text{II.19})$$

wyrażona tutaj właśnie dla obiektów, a nie dla zmiennych.

II.3. Relacje

II.3.1. Będziemy potrzebowali w wykładzie kilku dodatkowych pojęć, które obecnie wprowadzimy. Zaczniemy od *relacji* (por. Tab. II.4). Relacje są związkami między wielkościami, stanowiącymi w pewnym sensie rozszerzenie lub uzupełnienie pojęcia funkcji (tzw. funkcjami zdaniowymi). Dwie wielkości, oznaczone x i y , pozostają w relacji R , dokładnie: wielkość x jest w relacji R z wielkością y , co zapisujemy xRy , jeśli jest spełniony pewien warunek, nazywany warunkiem relacji, lub, w skrócie, relacją. Przykładami konkretnych relacji (dla określonych wielkości) są: „ x jest dwa razy mniejsze od y ” oraz „ x jest mniejsze od y ”.

II.3.2. Relacje mogą spełniać (lub nie) dla wszystkich elementów $x, y, z \in \mathbf{X}$ określone charakterystyczne własności, a mianowicie:

Relacja R jest *zwrotna* jeśli $\forall x \in \mathbf{X}$ mamy xRx (II.20a)

Relacja R jest *symetryczna* jeśli $\forall x, y \in \mathbf{X}$ mamy $xRy \Rightarrow yRx$ (II.20b)

Relacja R jest *przechodnia* jeśli $\forall x, y, z \in \mathbf{X}$ mamy $xRy \wedge yRz \Rightarrow xRz$. (II.20c)

O relacji R , która jest jednocześnie zwrotna, symetryczna i przechodnia mówimy, że jest relacją *równoważności*.

II.3.3. I tak, oczywiście, relacja równości („ $x=y$ ”) jest relacją równoważności, ponieważ spełnia (II.20a,b,c). Natomiast relacja większości („ $x>y$ ”) nie jest zwrotna ani symetryczna, ale jest za to przechodnia.

II.3.4. Załóżmy, że mamy określony podział zbioru X i odpowiadającego mu zbioru I , na podzbiory $A_q \subseteq I$, $q=1, \dots, p$, przy czym $\cup_q A_q = I$ (podzbiory A_q wyczerpują cały zbiór I) oraz $A_q \cap A_{q'} = \emptyset$, $q \neq q'$ (przecięcie dwóch różnych podzbiorów A_q jest puste). Podział taki oznaczmy P i powiemy, że jest on podziałem właściwym. Wówczas dla $i, j \in I$ możemy określić relację, równoznaczną z podziałem P , a mianowicie relację R przynależności do tego samego skupienia A_q . Taka relacja jest, oczywiście, relacją równoważności.

II.3.5. Zależności definiujące relacje mogą być przedstawione w postaci wyrażeń algebraicznych. I tak, załóżmy, że chcemy przedstawić relację większości między obiektami o indeksach i, j . Wprowadzamy zmienną binarną $y_{ij} \in \{0, 1\}$ taką, że $y_{ij} = 1$ wtedy, kiedy obiekt i jest większy niż j , oraz $y_{ij} = 0$ w przeciwnym przypadku. Skoro tak, to

$$y_{ii} = 0 \quad (\text{II.21a})$$

$$y_{ij} + y_{ji} \leq 1 \quad (\text{II.21b})$$

$$y_{ij} + y_{jl} - y_{il} \leq 1. \quad (\text{II.21c})$$

W zupełnie analogiczny sposób, relację stanowiącą o podziale P ($y_{ij}=1$ jeśli i oraz j należą do tego samego skupienia, $y_{ij} = 0$ w przeciwnym przypadku) możemy przedstawić jako:

$$y_{ii} = 1 \quad (\text{II.22a})$$

$$y_{ij} - y_{ji} = 0 \quad (\text{II.22b})$$

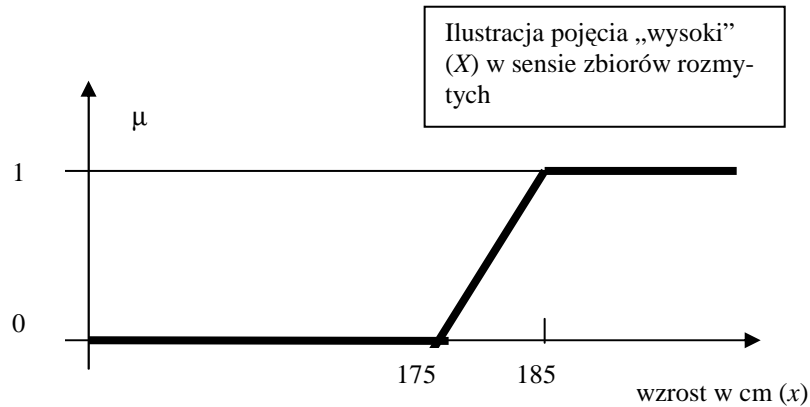
$$y_{ij} + y_{jl} - y_{il} \leq 1. \quad (\text{II.22c})$$

II.4. Zbiory rozmyte

II.4.1. Zbiory rozmyte są rozszerzeniem klasycznego pojęcia zbioru. Do wprowadzenia zbiorów rozmytych posłużymy się pojęciem funkcji $\mu_X(x)$ definiującej zbiór (funkcji charakterystycznej zbioru), a mianowicie $\mu_X(x)=1$ jeśli $x \in X$, oraz $\mu_X(x)=0$ jeśli $x \notin X$, dla określonego „uniwersum” \mathbf{X} , z którego pochodzą elementy x i w stosunku do którego definiujemy zbiór X , przy czym mamy tutaj $X \subseteq \mathbf{X}$. Oznacza to, że element x może albo należeć do zbioru X ($\mu_X(x)=1$), albo do niego nie należeć ($\mu_X(x)=0$) i nie przewiduje się żadnych innych sytuacji. Funkcja charakterystyczna zbioru przyjmuje tylko dwie wartości.

II.4.2. Nieco ogólniej, funkcja charakterystyczna zbioru μ może jednak być również określona w taki sposób, by przyjmować nie tylko wartości ze zbioru $\{0,1\}$, ale z całego odcinka $[0,1]$, np. $\mu_X(x)=0,6$. Mówimy wówczas, że $\mu_X(x)$ jest funkcją (stopnia) przynależności elementów x , należących do uniwersum \mathbf{X} , do pewnego zbioru X , który to zbiór będziemy wówczas właśnie nazywali *zbiorem rozmytym*. W odniesieniu do definicji przynależności elementów do zbiorów przytoczony przykład (tj. $\mu_X(x)=0,6$) będzie oczywiście, oznaczał, że element x należy do zbioru X w „stopniu” 0,6. Jakie to jednak może mieć znaczenie?

II.4.3. Zbiory rozmyte są mianowicie w największej mierze odpowiednią reprezentacją pojęć używanych w mowie potocznej, takich jak „szczupła”, „wysoki”, „wysoka wygrana”, „tani”, itp. Ilustracją niech będzie poniższy rysunek, pokazujący wartości $\mu_X(x)$ dla wartości x wzrostu w centymetrach (uniwersum \mathbf{X} może być, powiedzmy, zbiór potencjalnych wartości wzrostu w centymetrach od 0 do 250), określające zbiór X odpowiadający pojęciu „wysoki”:



Można, oczywiście, dyskutować, czy osoba o wzroście 175 cm jest „wysoka”, ale jeśli ktoś ma więcej niż 185 cm, to chyba może bez żadnych wątpliwości być określony jako „wysoki”. Dlatego też, np. $\mu_X(180)=0,5$, co oznacza, że dla tego wzrostu równie dopuszczalne jest stwierdzenie, że odpowiada on wysokiemu wzrostowi, jak i że jeszcze nie.

Zauważmy, że nie dokonaliśmy tutaj bynajmniej żadnego „obiektywnego” określenia co to znaczy „wysoki”, ale po prostu zapisaliśmy dość precyzyjnie matematycznie pewien sensowny (jakkolwiek niewątpliwie arbitralny) sposób rozumienia tego pojęcia.

II.4.4. Łatwo się zorientować, że zbiory rozmyte mogą być bardzo przydatne przy wygodnym dla przetwarzania cyfrowego definiowaniu kodowania zmiennych „jakościowych” i ich poszczególnych wartości („bardzo wysoka”, „wysoka”, „średniego wzrostu”, ..., bądź, jak w popularnym „teście” siły przy pomocy gruszki pięściarskiej: „Cienki Bolek”, „Niezdara”, „Laluś”, „Macho”, „Paker”, „Siłacz”, „Champion”), także w tych przypadkach, gdy dana cecha łączy ze sobą więcej niż jedną zmienną (cytowany już przykład „otyłego”, „przy kości”, „proporcjonalnego”, „szczupłego”..., określeń łączących wzrost, wagę i budowę ciała). Podkreślmy przy tym, że takie „rozmyte” kodowanie ma największy sens wówczas, gdy odpowiednie zmienne „jakościowe” są porządkowe i dają się bez większych wątpliwości przełożyć na odpowiednie skale „ilościowe”.

II.4.5 Zbiory rozmyte mogą mieć także zastosowanie w przypadku określania bądź szukania podziału P zbioru I o określonych własnościach (podzbiory zdefiniowane A_q jako zbiory rozmyte), i faktycznie jest to jedno z ich najczęstszych zastosowań w analizie danych. Dotyczy ono, naturalnie,

obiektów x_i , co do których nie ma pewności, do którego podzbioru powinny należeć.

II.4.6. W sytuacji dopuszczenia zbiorów rozmytych jako podzbiorów A_q powstaje pytanie, czy jesteśmy w stanie dla nich sformułować warunki analogiczne do warunków (II.22), definiujące tym razem podział P na rozmyte podzbiory A_q ? Okazuje się, że jest to zadanie trudne i w zasadzie odstępkuje się od takich warunków, pozostawiając na ogół jeden tylko, intuicyjnie dość oczywisty warunek, a mianowicie

$$\sum_q \mu_q(i) = 1 \quad (\text{II.23})$$

gdzie $\mu_q(i)$ jest funkcją przynależności obiektów o indeksach i (przeszliśmy na oznaczenia związane z indeksami ze zbioru I) do rozmytych podzbiorów A_q tworzących podział P . Warunek (II.23) oznacza po prostu, że dany obiekt musi być w całości „rozdzielony” pomiędzy podzbiory rozmyte A_q (w szczególności, dla zbiorów nierozmytych, mamy warunek (II.23) spełniony przez to, że dla wszystkich $i \in I$, $\mu_q(i)=1$ dla pewnego q i $\mu_q(i) = 0$ dla pozostałych q).

WYKŁAD III

Kilka słów o sensie, rozumieniu i postaci wiedzy. Rodzaje zadań analizy danych. Zadania analizy danych prezentowane w wykładzie.

III.1. Kilka uwag o sensie, rozumieniu i postaci wiedzy

III.1.1. Celem analizy danych jest uzyskanie na podstawie danych wiedzy. Jak wspominaliśmy, trzema zasadniczymi elementami wiedzy są: (i) zaakceptowany zbiór („prawidłowych”) danych („encyklopedia”), (ii) metodyka postępowania, przede wszystkim w zakresie pozyskiwania nowych danych, ale także ich przetwarzania (a więc właśnie analizy danych), (iii) teorie bądź hipotezy robocze, jakie zdołano w obrębie danej dziedziny ustalić i jakimi się ona posługuje.

III.1.2. Istotą teorii (hipotez) jest *synteza*. Oznacza to, że jeśli tylko teoria jest prawdziwa, możemy przy jej pomocy wyrazić w stosunkowo prosty sposób znaczny zbiór danych (powiązania między nimi), albo pewne cechy takiego zbioru (być może obejmującego wszystkie dane).

Innym niezmiernie ważnym aspektem efektywnej teorii jest *możliwość przewidywania*. Faktycznie, jeśli teoria reprezentuje powiązania między danymi, to znając jedno, możemy wyznaczyć (w szczególności: przewidzieć) inne. I tak, teoria powszechnego ciążenia Izaaka Newtona, wyrażona w postaci wzoru na siłę przyciągania grawitacyjnego (przypomnij sobie, łaskawy Czytelniku, z kursu fizyki), wiąże ze sobą masy, odległość i siłę wzajemnego oddziaływania, pozwalając na pominięcie niektórych pomiarów (np. siły), a z drugiej strony – pozwalając na wyznaczenie wartości tych pomiarów.

III.1.3. Teorie (hipotezy) mają na ogół postać pewnych modeli bądź reguł, mówiących, najogólniej, że „jeśli A, to B” (w przytoczonym przykładzie: jeśli masy dwóch ciał wynoszą m_1 i m_2 , a odległość między ich środkami ciężkości wynosi d , to siła ich wzajemnego przyciągania wynosi F i jest wyznaczona przy pomocy odpowiedniego wzoru). To właśnie takie zależności pozwalają na efektywne działanie („jeśli zrobię to [A], to uzyskam taki [na pewno, albo z odpowiednio wysokim prawdopodobieństwem] wynik [B]”), i dlatego wiedza jest podstawą efektywnego działania.

III.1.4. Zaznaczmy jednak, że uzyskanie w praktyce modeli (reguł) o powyższej postaci, w więc o postaci będącej w istocie zapisem zależności *przyczynowo-skutkowej* jest niezmiernie trudne i możliwe tylko w przypadku, gdy jesteśmy w stanie przeprowadzić *eksperyment aktywny*, to znaczy stworzyć

warunki odpowiadające przyczynom A i obserwować, czy zajdzie skutek B (eksperymenty fizyczne, znacznie rzadziej np. medyczne, w których znacznie trudniej jest zdefiniować dokładnie A, a także często i B). W znakomitej większości sytuacji możemy jedynie obserwować *współwystępowanie* zdarzeń związanych z A i B, a następnie wnioskować o regule „jeśli A, to B” nie w sensie przyczynowo-skutkowym, ale tylko w odniesieniu do współwystępowania (np. korelacja).

III.1.5. Modele lub reguły są wyznaczane na podstawie pewnych zbiorów danych, a następnie sprawdzane przy pomocy innych zbiorów. Sprawdzanie (weryfikacja) teorii (modeli) jest niezwykle ważnym elementem metodyki każdej dziedziny wiedzy. To właśnie dlatego niektóre dziedziny, np. ekonomii, są często krytykowane z punktu widzenia teoriopoznawczego, ponieważ wypowiediane przez ich reprezentantów stwierdzenia (nawet, jeśli jakoś uzasadnione przez określone zbiory danych) są albo z trudem, albo w ogóle nie weryfikowalne. W tej sytuacji nie możemy w ogóle mówić o istnieniu teorii, albo co najwyżej – o istnieniu konkurencyjnych hipotez (np. „dodatkowy pieniądz na rynku będzie generował popyt, co pozwoli na zwiększenie produkcji i zatrudnienia i w efekcie pozwoli gospodarce na wyjście z recesji” zestawione z „dodrukowane pieniądze spowodują wyłącznie zwiększenie inflacji, być może tylko przejściowe, a żeby efekt produkcyjny był trwały, trzeba będzie stale drukować puste pieniądze”). Sama konstrukcja hipotez i teorii powinna pozwalać na ich weryfikację, a w szczególności – na ich odrzucenie na podstawie danych (postulat „falsyfikowalności” Karla R. Poppera). Teorie, których nie można odrzucić (skonstruować doświadczenia, sprawdzającego ich prawdziwość), nie mogą być uznane za naukowe (czyli: uzasadnione). To samo, oczywiście, dotyczy modeli i reguł.

III.1.6. Zależności otrzymywane w wyniku analizy danych, jeśli prowadzą do otrzymania modeli albo reguł o wspomnianej poprzednio postaci, to najczęściej mają one charakter tzw. „modeli empirycznych”, czyli opartych wprost (i być może wyłącznie) na danych, bez głębszej analizy merytorycznej danego zjawiska, w odróżnieniu od „modeli teoretycznych”, lub wprost „teorii”. To często powoływane rozróżnienie ma na celu podkreślenie, że „teorie” lub „modele teoretyczne” wyjaśniają dane zjawiska i ewentualnie mogą dostarczyć podstaw do przewidywania (zależności przyczynowo-skutkowe), co nie jest bynajmniej takie proste w przypadku modeli empirycznych, o czym już wspominaliśmy.

Nie negując bynajmniej zasadności przytoczonego rozróżnienia (niewątpliwie współwystępowanie jest zupełnie czymś innym niż zależność przyczynowo-skutkowa), należałoby jednak zaznaczyć, że w zasadzie zawsze modele empiryczne są produktami wcześniejszych etapów funkcjonowania pętli sprzężenia zwrotnego dane-informacja-wiedza z punktów I.1.7-9, podczas

gdy modele teoretyczne – dalszych etapów jej funkcjonowania. Jest oczywiste, że zanim zostanie opracowany („odkryty”) model teoretyczny, musi zostać zgromadzona wiedza o charakterze empirycznym, opierająca się wprost na obserwacjach, na danych. Bardzo często zależności o charakterze empirycznym prowadzą jednak bezpośrednio do hipotez teoretycznych, a dalej do „porządných” teorii.

Z drugiej strony cały szereg modeli zakorzenionych w nauce (takich jak, na przykład, prawo powszechnej grawitacji Newtona) ma nadal w dużej mierze charakter empiryczny, oparty na obserwacji, a nie teoretyczny (ciągle jeszcze nie wiemy, dlaczego wzór na siłę grawitacyjną jest prawdziwy).

Wreszcie, należy zastrzec się, że modele empiryczne, nawet jeśli są formułowane w postaci pewnych ustalonych struktur modelowych (jak to się często dzieje w ekonometrii), oznacza to na ogół tylko, że możliwe jest porównywanie (jakości) odpowiednich modeli czy reguł, jakkolwiek w niektórych przypadkach struktury takie mają choćby częściowe usprawiedliwienie teoretyczne (np. modele przepływów w handlu zagranicznym oparte na zasadzie grawitacji).

III.1.7. Spotyka się również rozróżnienie między modelami „deskryptywnymi” i „normatywnymi”. Modele deskryptywne mają za zadanie *opisywać rzeczywistość* (tj. jej odpowiedni wycinek) przy pomocy wykrytych za pomocą analizy danych reguł i zależności, podczas gdy modele normatywne mają pokazywać jak należy tę *rzeczywistość zmienić* – jaki powinien być stan lub kierunek zmian oraz jak to osiągnąć (dobór instrumentów). Oczywiście, modele normatywne (często optymalizacyjne), muszą posiadać odpowiednie elementy deskryptywne, żeby nie być oderwanymi od realiów i faktycznie wiarygodnie projektować (lepszą, naturalnie) przyszłość. To rozróżnienie o tyle nie jest interesujące dla niniejszego wykładu, że zajmujemy się w nim wyłącznie analizą danych pochodzących z rzeczywistych, istniejących zjawisk i procesów, a więc wyłącznie, w myśl tego rozróżnienia - modelami deskryptywnymi.

III.1.8. Nie zawsze w wyniku analizy danych jesteśmy w stanie otrzymać dobrze uwarunkowane struktury, modele lub reguły (porządne teorie). Może to być skutkiem wielu różnych przyczyn (zbyt ubogi zbiór danych, nadmierne skomplikowany proces, który je generuje, istotne błędy pomiaru itp.). Staramy się jednak uzyskać struktury (sumaryczne charakterystyki danych), które choćby w pewnej mierze przybliżają się do dobrych, syntetycznych i prognostycznych modeli. Także i z tego powodu procedura analiza danych obejmuje szereg etapów i zróżnicowanych podejść, o których będziemy mówili w dalszym ciągu wykładu.

III.2. Niektóre rodzaje zadań analizy danych

III.2.1. Sposób traktowania zaobserwowanych danych zależy w dużej mierze od tego, na jakim etapie pętli sprzężenia zwrotnego, omawianej w Wykładzie I (I.1.7-I.1.9), aktualnie jesteśmy. Zilustrujemy to obecnie na przykładzie pewnej grupy zadań analizy danych.

III.2.2. I tak, w szczególności, często stajemy przed zagadnieniem podziału zbioru obserwacji (obiektów) I na podzbiory, grupujące obiekty o różnych charakterystykach, które to podzbiory mogą być interpretowane jako odpowiadające różnym „typom”, „modelom”, czy „procesom”. Na przykład w badaniach marketingowych możemy być zainteresowani podziałem populacji klientów na podzbiory odpowiadające różnym „typom” klientów, przy czym typy te są określone przez różne preferencje, siłę nabywczą, czy zachowania konsumpcyjne. Zatrzymamy się na tym zagadnieniu po pierwsze dlatego, że jest ono bodaj najczęstszym na wstępnym etapie analizy danych, a po drugie – dlatego, że pozwala ono na dobrą prezentację możliwych zadań analizy danych.

III.2.3. Tak jak w II.3 i II.4 podzbiory zbioru I oznaczmy A_q , $A_q \subseteq I$, $q=1, \dots, p$. Jeśli spełnione są warunki $A_q \cap A_{q'} = \emptyset$, $q \neq q'$, oraz $\cup_q A_q = I$, czyli, że podzbiory są rozłączne, a ich suma mnogościowa wyczerpuje zbiór I , to mówimy, że zdefiniowany został podział właściwy zbioru I , oznaczony $P=\{A_q\}_q$ (w dalszym ciągu wykładu, jeśli P będzie oznaczało inne pojęcie niż podział właściwy zbioru obiektów I , w szczególności – inne jego podziały – to będzie to specjalnie zaznaczone).

III.2.4. Przy przyjętych oznaczeniach możemy rozróżnić trzy zasadnicze rodzaje zadań odnoszących się do tak pojętej struktury (zbioru) danych:

Zadanie analizy skupień: mamy dany zbiór I , np. w postaci macierzy, jak w Tab. II.2. Staramy się rozpoznać podział P utworzony w taki sposób, że obiekty znajdujące się w tych samych podzbiorach (skupieniach, „klastrach”) A_q są możliwie podobne, zaś obiekty znajdujące się w różnych podzbiorach – możliwie różne. Nie posługujemy się przy tym żadną inną („aprioryczną”) informacją poza zawartością macierzy danych.

Zadanie dyskryminacji: mamy dany zbiór obiektów I oraz jego podział P , o którym na ogół zakładamy, że jest „prawidłowy” lub nawet „optymalny” w sensie poprzednio omówionego zadania analizy skupień. Staramy się utworzyć możliwie proste funkcje („reguły”), rozdzielające zbiory obiektów A_q , a właściwie zbiory B_q , które odpowiadają zbiorom A_q , ale stanowią wyczerpujący podział całego uniwersum, tj. zbioru \mathbf{X} . I tak, w najprostszym przypadku, dla $p=2$ (czyli $P=\{A_1, A_2\}$) w przestrzeni \mathbf{X} chcemy w jak najprostszy sposób utworzyć taką funkcję F , że jeśli dla danego $x \in \mathbf{X}$

mamy $F(x) < 0$, to $x \in B_1$, $A_1 \subseteq B_1$, jeśli zaś $F(x) \geq 0$, to $x \in B_2$, $A_2 \subseteq B_2$. Przytaczamy ten przykład nie tylko ze względu na jego prostotę zapisu, ale i dlatego, że faktycznie rozwiązywane efektywnie zadania dyskryminacji są niezwykle proste (niewielka liczba skupień, proste funkcje rozdzielające – najczęściej linie proste). Jedynie dla bardzo regularnych rozkładów gęstości obserwacji (prawdopodobieństwa) udało się otrzymać zależności dyskryminacyjne, w tych przypadkach przybierające zresztą niezwykle regularne postacie.

Zadanie klasyfikacji: mamy dany zbiór I oraz podział P i dla nowych obserwacji (obiektów) $x \in \mathbf{X}$, których poprzednio nie było w zbiorze I , chcemy te nowe obserwacje przypisać skupieniom tworzącym podział P , bądź też stwierdzić, że nie da się takiego przypisania dokonać (odrzuć obserwacji bądź utworzenie nowego skupienia). Jeśli dysponujemy funkcją dyskryminacji, to zadanie klasyfikacji staje się trywialne, ale, jak wspomnieliśmy, funkcje takie są wyznaczalne tylko dla bardzo ograniczonej klasy zadań, tak więc klasyfikacja musi się odbywać przy pomocy specjalizowanych metod, na ogół bardzo zbliżonych do metod analizy skupień (możliwość tworzenia nowych skupień, ewentualność likwidacji poprzednio istniejących).

III.2.5. Należy dodać, że bardzo często, wraz zadaniem podziału zbioru obserwacji na podzbiory (skupienia, klasy) występuje zadanie identyfikacji reprezentantów skupień (klas). Najczęściej chodzi o obiekty, należące do przestrzeni (uniwersum) \mathbf{X} , które najlepiej oddają charakter wszystkich obiektów w danym skupieniu. Mówimy wówczas na ogół o „typach” lub „modelach” właściwych dla klas. Najprostszym rozwiązaniem tego zagadnienia wydaje się być wyznaczenie skupień A_q , a następnie, dla każdego skupienia, reprezentanta w postaci średniej,

$$x^q = \frac{1}{\text{card}A_q} \sum_{i \in A_q} x_i.$$

Takie rozwiązanie jest w istocie bardzo powszechnie stosowane. Jednak łatwo zauważyć cały szereg trudności, jakie się od razu pojawiają, a zwłaszcza dwie zasadnicze: (i) tak wyznaczone x^q zapewne nie będzie należało do zbioru X (co jest często narzucanym warunkiem), a nawet nie musi należeć do zbioru \mathbf{X} , co jest praktycznie zawsze niezbędne do zaakceptowania reprezentanta; (ii) optymalne wyznaczanie skupień i ich reprezentantów to nie są w ogólności niezależne procedury, zwłaszcza jeśli chcemy skupienia i reprezentantów wyznaczyć jednocześnie.

III.2.6. Reprezentantami skupień mogą być nie tylko pojedyncze obiekty, ale podzbiory skupień, ich określone funkcje, a w szczególności – modele, na przykład modele regresji. Możemy wówczas mówić o modelach zjawisk

odpowiadających poszczególnym skupieniom (podzbiorom zbioru X), albo regionom w przestrzeni X . Takimi modelami są, na przykład, modele zachowań konsumenckich różnych grup konsumentów („jeśli konsument przeznaczając miesięcznie na żywność od w do z złotych, to udział produktów cukierniczych w tych wydatkach wyniesie $t\%$, gdzie w, z, t są charakterystykami pewnego skupienia w przestrzeni zachowań konsumenckich”), wyznaczonych na podstawie badań danych o konsumentach.

III.2.7. Wymienione trzy obszerne grupy zadań (analiza skupień, dyskryminacja, klasyfikacja) mogą być interpretowane w ramach omawianej pętli sprzężenia zwrotnego danych, informacji i wiedzy na różnych etapach jej działania (analiza skupień przy bardzo niewielkiej wiedzy początkowej, dyskryminacja przy obszerniejszej, i klasyfikacja przy jeszcze większym zasobie wiedzy). Niezależnie jednak od tego podziału zadań analizy danych istnieje cały szereg podziałów, zwłaszcza związanych z postacią struktury zbioru danych, a przeto i rodzaju wiedzy, jaki posiadamy na ich temat, lub chcemy osiąść. Tak więc, poza podziałem na grupy, możemy poszukiwać: (i) uporządkowania lub uporządkowań obiektów; (ii) modeli uzależniających jedne aspekty (zmienne) od innych, lub jedne obiekty od innych; (iii) czynników wyjaśniających w uproszczony sposób złożoność zbiorów danych scharakteryzowanych wieloma zmiennymi; (iv) przekształceń przestrzeni X prowadzących także do prostszych opisów obiektów; itp.

Poszczególne rodzaje zadań mogą być ze sobą powiązane. I tak, możemy być zainteresowani podziałem na wewnętrznie spójne podzbiory obiektów (skupienia) i jednocześnie wyznaczeniem modeli w obrębie tych skupień, bądź ustaleniem uporządkowań w przestrzeni niewielkiej liczby niezależnych od siebie wzajemnie czynników.

Zadania te i metody ich rozwiązywania będą zasadniczym przedmiotem niniejszego wykładu.

III.3. Zadania analizy danych omawiane w ramach wykładu

III.3.1. W ramach wykładu omówimy następujące rodzaje zadań analizy danych:

Porządkowanie (rangowanie) względem wielu zmiennych

Analizę skupień

Analizę dyskryminacyjną

Klasyfikację

Analizę regresji

Analizę czynnikową

Skalowanie wielowymiarowe

III.3.2. Niektóre z tych dziedzin omówimy szerzej, podczas gdy inne niejako tylko zasygnalizujemy. I tak, szerzej omówimy porządkowanie względem wielu zmiennych, analizę skupień, analizę regresji oraz analizę czynnikową, podczas gdy o pozostałych grupach zadań przedstawimy tylko pobieżną informację.

WYKŁAD IV:

Porządkowanie obiektów. Zastosowania. Porządkowanie względem wielu zmiennych – trudności, istniejące podejścia i metody i ich cechy.

IV.1. Porządkowanie obiektów – zarys zagadnienia

IV.1.1. Często mamy do czynienia z porządkowaniem obiektów w postaci różnych „rankingów”, prezentowanych w mediach. Są to, na przykład, rankingi popularności polityków, rankingi szkół, wyniki zawodów sportowych itp. W każdym z tych przypadków poszczególne obiekty (w tym i ludzie) są ustawione w znaczącej kolejności, „od góry do dołu”, „od największych do najmniejszych”, lub odwrotnie. W każdym z nich również miejscu zajmowanemu w tym ustawieniu – uporządkowaniu – przypisana jest pewna wartościująca interpretacja („szkoła i jest lepsza niż szkoła j ”, „drużyna j jest lepsza niż drużyna j' ”). Niekiedy dopuszcza się możliwość usytuowania więcej niż jednego obiektu na tej samej pozycji rankingu czy uporządkowania (dwa srebrne medale, miejsce *ex aequo*, ta sama liczba punktów, itp.). Dodajmy, że rankingi takie cieszą się dużą popularnością, zwłaszcza, kiedy dotyczą interesujących obiektów – a na ogół są organizowane i publikowane po to, by zainteresowanie przyciągnąć. Tym bardziej powinniśmy się przyrzeć zasadom, na jakich są one oparte, albo na jakich powinny być oparte.

IV.1.2. Przede wszystkim łatwo zauważyć, że kwestia sporządzenia takich rankingów w praktycznie wszystkich przypadkach nie jest bynajmniej trywialna. Wynika to, poza ewentualnymi innymi czynnikami (takimi jak błędy, zrozumienie znaczenia zmiennych itp.), z jednej podstawowej przyczyny, a mianowicie istnienia więcej niż jednej zmiennej, opisującej obiekty, a mającej (przynajmniej w teorii) wpływ na wynik rankingu.

I tak, na przykład, w zupełnie prostej, wydawałoby się, sytuacji rankingu polityków względem popularności, każda z osób występujących w rankingu opisana jest faktycznie przez trzy zmienne ($m=3$): $k=1$: procent respondentów danego badania opinii publicznej, deklarujących *sympatię* dla danego polityka, $k=2$: procent respondentów deklarujących *antypatię* do tej osoby, oraz $k=3$: procent osób stwierdzających, że tej osoby nie znają. Oczywiście, dla każdej uwzględnionej osoby ($i \in I$) te trzy wartości powinny się sumować do 100%, tj. $\forall i \in I: x_{i1} + x_{i2} + x_{i3} = 1$, co ewentualnie umożliwia nam wyeliminowanie jednej z tych wartości (np. x_{i3}), jako zależnej od dwóch pozostałych. Tym niemniej, pozostają nam dwie zmienne i, jak to wiemy z czytelnictwa odpowiednich rankingów, nie muszą się one wcale „grzecznie” za-

chowywać, to znaczy, wcale nie jest tak, że jeśli $x_{i1} > x_{j1}$, to $x_{i2} < x_{j2}$, zwłaszcza w przypadku tzw. polityków „kontrowersyjnych” (proszę zauważyć, że jednym z zasadniczych przymiotów uniwersytetu, od czasów platońskiej Akademii, jest apolityczność).

IV.1.3. W tej sytuacji jedynymi pewnymi, jednowymiarowymi rankingami są wyniki (niektórych – mierzalnych) zawodów sportowych (a więc, np. kolejność na mecie według osiągniętych czasów lub liczba punktów w strzelaniu sportowym), choć i tutaj malkontenci dopatrziliby się wątpliwości (liczy się ramię czy też biust zawodnika-zawodniczki??). Wojskowy czy harcerski ranking: „według wzrostu” jest też dość pewny. Ale już wyniki ligi piłkarskiej mogą dostarczyć przedmiotu do ożywionej dyskusji: 1. sposób punktowania wyników meczów: trzy czy dwa punkty za zwycięstwo? a także 2. relacja między punktami a stosunkiem bramek i ocena wyniku liczonego w bramkach: np. wartość goli strzelonych na wyjeździe w Lidze Mistrzów, bądź wątpliwości dotyczącego tego, czy brać pod uwagę różnicę, czy też raczej stosunek bramek. Nie wspominamy tutaj, naturalnie, takich „zawodów”, jak konkursy łyżwiarstwa figurowego czy konkursy muzyczne (np. odbywający się co pięć lat w Warszawie wyścig fortepianowy im. Fryderyka Szopena).

IV.1.4. Jakkolwiek sporządzanie uporządkowań, bądź rankingów, wydaje się być marginesem dziedziny analizy danych, nie jest nim jednak, a to dla trzech przyczyn: (i) jest bardzo rozpowszechnione, i wykorzystywane wcale nie tylko w ramach mediów dla sprawdzenia lub osiągnięcia popularności, stanowiąc także podstawę wielu innych, znacznie bardziej zaawansowanych podejść i metod; (ii) jest w istocie pewną metodą identyfikacji modeli o określonym charakterze. I tak, np., jeśli wiadomo, że od wielu lat mistrzem Polski w boksie na żelazne rękawice z zamkniętymi oczami jest zawodnik i , który zazwyczaj pokonuje w finale zawodnika j na punkty, to mamy do czynienia z pewnym modelem (regułą), i fakt, że w danym, kolejnym, roku w finale turnieju boksu na żelazne... itp. zawodnik i pokonał przez zejście śmiertelne zawodnika j , a dotychczasowi finaliści odpadli (w podobny sposób) już w ćwierćfinale, stanowi bardzo istotną informację – obalającą model, i być może stanowiącą podstawę do stworzenia nowego modelu; (iii) jak to już zauważyliśmy, sporządzanie dobrze uzasadnionych uporządkowań stanowi zagadnienie nietrywialne i jest punktem wyjścia wielu wartościowych konstrukcji teoretycznych i algorytmów.

IV.1.5. Istnieje bardzo wiele sposobów obejścia lub uwzględnienia problemu wielu zmiennych w sporządzaniu rankingów lub uporządkowań. Niektóre z nich skomentujemy wstępnie, przed przystąpieniem do dokładniejszej prezentacji metod, poniżej:

- (1) *uwzględnienie tylko jednej zmiennej*; ten wybieg jest najczęściej stosowany w rankingach publikowanych w mediach, przy czym trudno jest ustalić, jaka jest zasadnicza przyczyna tego faktu: założenie, że czytelnicy (widzowie) nie pojmą bardziej skomplikowanej konstrukcji rankingu, gdyby był on oparty na wielu zmiennych, czy też brak kompetencji dziennikarzy; w tym przypadku ranking oparty jest tylko na jednej zmiennej, $k=1$, a pozostałe zmienne, $k>1$, traktowane są wyłącznie jako zmienne (cechy) „dodatkowe”, o charakterze ilustracyjnym, rozważane jako podstawa do porządkowania tylko wtedy, gdy dla właściwej zmiennej porządkującej mamy dla dwóch obiektów te same wartości; w ten sposób właśnie sporządzane są rankingi popularności polityków; w ten również sposób prezentowane są np. listy „500 przedsiębiorstw” („Rzeczpospolita”, „Polityka”, ...), ułożone według jednej zmiennej (np. wielkości rocznych obrotów), z pozostałymi zmiennymi będącymi tylko dodatkową ilustracją cech przedsiębiorstw;
- (2) *sporządzenie jednej zmiennej stanowiącej prosty agregat uwzględnianych w ten sposób zmiennych*; niekiedy stosuje się takie proste agregaty, zazwyczaj obejmujące jednak nie więcej niż dwie, rzadko trzy zmienne wyjściowe, zaś agregaty te otrzymywane są jako suma, iloczyn, itp. odpowiednich zmiennych; najoczywistszym przykładem są tu skoki narciarskie (suma not za długość i styl), bądź jazda figurowa na łyżwach (suma czy średnia not za zawartość techniczną programu i jego wykonanie), choć i w innych dziedzinach używa się niekiedy tego podejścia;
- (3) *sporządzenie agregatu większej liczby zmiennych poprzez bardziej skomplikowane zabiegi* (zliczanie miejsc w rankingach według poszczególnych zmiennych, zastosowanie wag dla poszczególnych zmiennych i ich sumowanie, inne przekształcenia).

IV.1.6. Porządkowanie obiektów $i \in I$ względem wielu zmiennych $k \in K$ jest ściśle związane z kilkoma ważnymi dziedzinami matematyki stosowanej, w tym przede wszystkim *optymalizacji wielokryterialnej* oraz *teorii wyborów kolektywnych (społecznych)*, tzw. „*social choice theory*”. Obie te dziedziny mają rozwiniętą teorię (aksjomatyka, własności, twierdzenia) i obszerne instrumentarium (metody, algorytmy, warunki). Nie będziemy się jednak zajmowali szerzej żadną z tych dziedzin ani ich wnioskami w odniesieniu do agregacji uporządkowań, zarówno ze względu na raczej pragmatyczne ukie-

nie od tego wspomniane dziedziny powinny być przedmiotami osobnych wykładów.

Omówimy obecnie pokrótce poszczególne ze zilustrowanych poprzednio dość wrywkowo podejść.

IV.2. Podstawowe metody agregacji uporządkowań

IV.2.1. Porządkowanie *leksykograficzne* – to pierwsze i zarazem najprostsze z poprzednio omawianych podejść. Postępowanie w tej metodzie porządkowania względem wielu zmiennych jest następujące: Najpierw układamy kolejność zmiennych według ich ważności dla naszego porządkowania, poczynając od zmiennej najważniejszej. Tak więc zmienna pierwsza jest ważniejsza od drugiej i ogólnie zmienna k -ta jest ważniejsza od $k+1$ -ej. (Dla ustalenia uwagi założymy, że porządkujemy według malejących wartości poszczególnych zmiennych. Zauważmy, że nie ogranicza to w niczym ogólności naszych rozważań, ponieważ zawsze możemy przejść od porządkowania „w dół” do porządkowania „w górę” – i odwrotnie – przez zmianę znaku wszystkich wartości zmiennej dla $i \in I$.) Następnie dokonujemy porządkowania według zmiennej $k=1$, to znaczy według wartości x_{i1} . Kiedy natykamy się na sytuację, w której $x_{i1}=x_{j1}$, to sprawdzamy wartości x_{i2} oraz x_{j2} i, analogicznie jak dla $k=1$, jeśli $x_{i2} > x_{j2}$, to i poprzedza j w uporządkowaniu. Gdybyśmy i dla $k=2$ natknęli się na równe wartości danych ($x_{i2}=x_{j2}$), to przechodzimy do zmiennej $k=3$ i sprawdzamy, czy $x_{i3} > x_{j3}$. Itd., itp. Nazwa porządkowania leksykograficznego pochodzi od kolejności (ewentualnego) rozważania zmiennych, zgodnie z kolejnością czytania wierszami, od lewej do prawej. To podejście ma szereg zalet: przede wszystkim jest intuicyjnie niezwykle proste, nie wymaga żadnych dodatkowych operacji na wartościach x_{ik} , które są porównywane tak, jak je otrzymano początkowo, a poza tym jest bardzo proste algorytmicznie (jakkolwiek sprawami algorytmów na poziomie porządkowania i sortowania oraz ich optymalizacji nie będziemy się tutaj zajmowali).

IV.2.2. Istotną wadą tego podejścia, jeśli jest ono stosowane do zmiennych przybierających wiele wartości w obrębie zbioru X (a więc przede wszystkim zmiennych ilościowych formalnie rzecz biorąc ciągłych) jest faktyczna marginalizacja zmiennych poza pierwszą. Tak się dzieje, gdy próbujemy zastosować porządkowanie leksykograficzne do tworzenia rankingów podmiotów ekonomicznych, na przykład na podstawie wielkości obrotów, zysku netto, zwrocie na majątku, liczbie zatrudnionych, itp. Pierwsza z uwzględnianych zmiennych całkowicie wyznacza uporządkowanie, które w teorii miało uwzględniać więcej zmiennych, chyba, że jakimś zupełnym przypad-

kiem dwa obiekty będą miały tę samą wartość tej pierwszej, decydującej zmiennej, i wówczas, dla tych dwóch obiektów, odwołamy się do zmiennej drugiej.

Tak więc porządkowanie leksykograficzne, jeśli ma być faktycznie porządkowaniem względem wielu zmiennych, powinno być stosowane do danych o obiektach opisywanych przez zmienne, które przyjmują niewielką liczbę wartości. Wówczas konieczność porównywania wartości dla kolejnych zmiennych zdarza się bardzo często. Dobrym przykładem takiej sytuacji może być próba ustalenia kolejności uczniów w klasie lub w szkole według stopni na świadectwie, pod warunkiem, że potrafimy uszeregować przedmioty według ich ważności. Ponieważ mamy do dyspozycji (na świadectwie) tylko pięć ocen, a ponadto rozkłady ocen są najczęściej silnie skoncentrowane na wartościach bliskich „środkowi” (dostateczny, dobry) więc „remisy” będą bardzo częste. Dla bardzo wielu uczniów trzeba będzie sięgać po oceny nie tylko z drugiego, ale i trzeciego, czwartego itp. przedmiotu, aby móc ustalić ich kolejność. Nie wykluczy to przy tym możliwości zaistnienia „remisów” także i dla uporządkowania według wszystkich zmiennych (przedmiotów).

Istnieją wszakże również stosunkowo proste metody, pozwalające na uwzględnienie wszystkich zmiennych bez ich przekształcania, znane już od dawna i stosowane, jakkolwiek niekoniecznie w ogólnie pojętej dziedzinie rangowania wielowymiarowego, także w odniesieniu do zmiennych quasi-ciągłych, a w każdym razie przybierających wiele wartości.

IV.2.3. I tak, pierwsza z tutaj prezentowanych metod jest pochodną metody wyznaczania wyników głosowania, zaproponowanej jeszcze w połowie XVIII wieku przez Bordę. Stosowana jest ona, jako metoda uzupełniająca, w sędziowaniu łyżwiarstwa figurowego (metoda *sumy miejsc*).

Dla każdej zmiennej $k=1, \dots, m$, opisującej obiekty, wyznaczamy ranking (uporządkowanie), otrzymując w ten sposób m uporządkowań. Na podstawie tych uporządkowań możemy każdemu obiektowi w każdym uporządkowaniu k przypisać „miejsce”, liczone od miejsca 1, zajmowanego, zgodnie z poprzednim założeniem, przez obiekt $i(k)$ taki, że $x_{i(k)} = \max_j x_{jk}$, czyli obiekt charakteryzujący się największą wartością danej zmiennej. Założymy dodatkowo, co jest dość naturalne, że miejsca mogą być „dzielone” (ten sam numer miejsca) przez obiekty o tych samych wartościach dla danej zmiennej. Miejsce ostatnie, n , przypisane będzie – jeśli dla danej zmiennej wszystkie wartości x_{ik} są różne – obiektowi, dla którego wartość danej zmiennej jest najmniejsza. Jeśli zatem oznaczmy miejsca obiektów i w rankingach częściowych otrzymanych dla poszczególnych zmiennych k przez o_{ik} , będące liczbami naturalnymi, to muszą one spełniać warunki

$$\forall i, j \in I: o_{ik} < o_{jk} \Leftrightarrow x_{ik} > x_{jk} \wedge o_{ik} = o_{jk} \Leftrightarrow x_{ik} = x_{jk}.$$

Łatwo się już w tym momencie domysleć, że ranking zagregowany zostanie otrzymany na podstawie wartości wyrażeń $\sum_k o_{ik} = \omega_i$, a więc „sumy miejsc” poszczególnych obiektów $i \in I$ uzyskanych w rankingach według poszczególnych zmiennych $k \in K$, w analogiczny sposób, w jaki na podstawie pojedynczych x_{ik} otrzymaliśmy o_{ik} . Zasadnicza różnica polega jednak na tym, że miejsce pierwsze w tym zagregowanym rankingu zajmie obiekt (lub obiekty) o *najmniejszej* wartości ω_i .

IV.2.4. Można się – przez chwilę – zastanawiać, dlaczego taki algorytm porządkowania jest uważany za nie tylko właściwy, ale i korzystny. Łatwo jednak zauważyć, że

- (i) przede wszystkim omijamy przy jego pomocy bez żadnego kłopotu trudne zagadnienie nieporównywalności skal poszczególnych zmiennych (możemy rozważać, powiedzmy, zmienne o skali $[0,1]$ i $[-283, +100]$, bez konieczności sprowadzania ich do porównywalności);
- (ii) agregacja jest bardzo prosta (suma miejsc) i intuicyjnie dość oczywista;
- (iii) ten rodzaj agregacji, jak to dalej zobaczymy, pozwala na znaczną elastyczność jej traktowania;
- (iv) wszystkie operacje, jakie są w ramach tego algorytmu wykonywane, są bardzo proste (porównywanie, sumowanie, itp.).

IV.2.5. Zatrzymajmy się obecnie na chwilę nad zagadnieniami, jakie wyłaniają się przy agregacji uporządkowań dla wielu zmiennych, a na jakie już w pewnej mierze natknęliśmy się, omawiając przedstawione uprzednio dwa podejścia (leksykograficzne i sumy miejsc). I tak,

- (1) najpierw podkreślmy, że porządkowanie nie może być stosowane do zmiennych nominalnych; jedynym dopuszczalnym rozwiązaniem jest w tym przypadku sporządzanie osobnych uporządkowań dla poszczególnych wartości zmiennych nominalnych, ponieważ nie możemy dokonywać porównań pomiędzy tymi wartościami;
- (2) projektowane i używane metody muszą uwzględniać zróżnicowanie skal wielkości poszczególnych zmiennych, tak, aby faktycznie agregacja uporządkowań dokonywana była stosunku do struktur porównywalnych (w metodzie leksykograficznej omijamy ten problem przez odnośnienie się do kolejnych zmiennych wyłącznie w sytuacji „remisu” dla poprzedniej, rozpatrywanej zmiennej, zaś w metodzie sumy miejsc

rozważamy miejsca w rankingach według zmiennych, a nie wartości dla tych zmiennych);

- (3) jednocześnie, przy sporządzaniu rankingów względem wielu zmiennych może być jednak ważnym, jakie wartości przybierają zmienne dla poszczególnych obiektów (na przykład, w odniesieniu do ostatnio prezentowanej metody Bordy, czyli sumy miejsc, w sytuacji gdy co prawda wśród trzech rozważanych zmiennych obiekt i był dwa razy [nieco] „gorszy” od obiektu j , ale w jednym przypadku był od niego „lepszy” i wówczas różnica między nimi była bardzo duża – to czy, i jeśli tak, jak uwzględniać ten fakt?);
- (4) w wielu przypadkach chcielibyśmy sterować świadomie wagami (ważnościami) przykładanymi do poszczególnych zmiennych i dobrze byłoby, żeby metody pozwalały na takie sterowanie, nawet jeśli nie jest ono konieczną częścią składową metody (w metodzie leksykograficznej sprawa ta została załatwiona przez sztywne przypisanie kolejności zmiennych, oznaczające bardzo duże zróżnicowanie wag kolejnych zmiennych, natomiast w metodzie sumy miejsc możliwe jest otrzymywanie wartości ω_i jako sumy ważonej (kombinacji liniowej) miejsc otrzymanych dla poszczególnych uporządkowań cząstkowych według kolejnych rozpatrywanych zmiennych, tj.

$$\omega_i = \sum_k w_k o_{ik},$$

gdzie w_k są wagami (ważnościami) przypisywanymi poszczególnym zmiennym k , spełniającymi często narzucany dość oczywisty warunek

$$\sum_k w_k = 1;$$

- (5) dodatkowo chcielibyśmy (a są to wymagania o charakterze technicznym), żeby wynik otrzymany był jako efekt stosunkowo łatwo pojmowalnej procedury, która może bez problemów zostać przedstawiona, np. nieprofesjonalnym decydom, a zarazem, co jest z tym blisko związane, żeby jego otrzymanie nie wymagało jakichś żmudnych obliczeń;

IV.2.6. Naturalną konsekwencją chęci uporządkowania obiektów względem wielu zmiennych jest sporządzenie jednej zmiennej, będącej sumą ważoną (kombinacją liniową) wszystkich uwzględnianych zmiennych. Na możliwość realizacji takiej podstawy do agregacji uporządkowań zwracaliśmy już uwagę w kontekście metody sumy miejsc (postulat (4) w punkcie IV.2.5). Obecnie zajmiemy się „czystą” postacią tej metody. Odwołuje się ona mianowicie nie do wartości miejsc o_{ik} , ale wprost do wartości danych x_{ik} , na podstawie których otrzymuje się zagregowaną wartość charakteryzującą obiekty i , oznaczoną w_i , a mianowicie:

$$w_i = \sum_k w_k x_{ik},$$

i dla tak otrzymanych wartości w_i dokonuje się końcowego porządkowania, $w_i \rightarrow o_i$.

Łatwo zauważyć, że tak pojęte i zastosowane wagi w_k muszą uwzględniać, poza samymi wartościami związanymi z założoną lub „obiektywną” ważnością poszczególnych zmiennych, także kwestię normalizacji zmiennych (wstępnego wyrównania wag wyrażonych niejawnie poprzez skale wartości zmiennych). Możemy zatem mówić o tych wagach w_k jako iloczynach $w_k = w_k^* \cdot w_k^{\text{norm}}$, gdzie w_k^* jest wagą właściwą (współczynnikiem ważności) zmiennej k , zaś w_k^{norm} jest jej parametrem normalizacji (na przykład: $w_k^{\text{norm}} = 1/\max_j x_{jk}$), sprowadzającym do porównywalności.

Dalszą konsekwencją konieczności normalizacji jest utrata – w pewnym stopniu przynajmniej – „naturalności” wag przyłożonych wprost do wartości zmiennych. Dotyczy to mianowicie postulatu (3) z punktu IV.2.5: po dokonaniu normalizacji trudno jest nam się zorientować, jaką rolę odgrywa w ostatecznym uporządkowaniu dana zmienna i różnice wartości według tej zmiennej między poszczególnymi obiektami. A przecież to właśnie chęć zachowania proporcji między wartościami dla poszczególnych zmiennych jest zasadniczym usprawiedliwieniem zastosowania metody ważonych zmiennych, w odróżnieniu od metody ważonych rankingów, tj. metody sumy miejsc. Widzimy zatem, że w doborze odpowiedniej metody agregacji uporządkowań musimy iść na pewnego rodzaju kompromisy.

Zauważmy jeszcze, że metoda ważonej sumy miejsc różni się od metody ważonych zmiennych „tylko” kolejnością postępowania, bowiem w pierwszej z nich mamy:

$$x_{ik} \rightarrow o_{ik} \rightarrow [w_i] \rightarrow \omega_i \rightarrow o_i$$

zaś dla drugiej:

$$x_{ik} \rightarrow [w_i] \rightarrow w_i \rightarrow o_i .$$

IV.2.7. Każda z już przedstawionych i dalej jeszcze rozpatrywanych metod ma w świetle przedstawionych warunków i wymagań pewne zalety i wady. Będziemy je komentowali przy pomocy zaprezentowanego poniżej akademickiego przykładu.

IV.2.8. *Przykład IV.1.* Załóżmy, że dysponujemy następującymi wartościami danych, opisujących pewien zbiór obiektów (w tym wypadku – kandydatki do przyjęcia do pracy na stanowisko asystentki dyrekcji w pewnej niedużej, ale dynamicznej firmie):

<i>Kandydatki</i>	<i>Wiek</i>	<i>Wykształcenie¹</i>	<i>Doświadczenie²</i>	<i>Języki³</i>	<i>Komputer⁴</i>	<i>Wrażenie ogólne⁵</i>
Inwencja A.	27	2	2	1	2	3
Porcja B.	25	1	2	2	2	3
Kwintessencja C.	24	3	1	2	1	4
Deucja D.	19	1	0	1	2	5
Kollacja E.	31	4	3	1	4	3
Redukcja F.	28	2	2	3	2	3
Wallencja G.	57	3	5	1	1	2

Uwagi do skal i znaczenia poszczególnych wartości liczby punktów dla przywołanych zmiennych:

¹*Wykształcenie*: 1 – średnie, 2 – średnie z „kursami”, 3 – licencjackie lub inżynierskie, 4 – pełne wyższe;

²*Doświadczenie*: 0 – pierwsza praca; 1 – do dwóch lat pracy u jednego pracodawcy; 2-5 – dłuższy staż pracy u kilku (ale niezbyt wielu) pracodawców;

³*Języki* (obce, choć dobra znajomość polskiego nie jest od rzeczy): 0 – brak znajomości; 1 – podstawowa znajomość jednego (angielskiego) języka; 2-4 – lepsza znajomość jednego lub większej liczby języków;

⁴*Komputer*: 0 – brak znajomości; 1 – podstawowe opanowanie z Windows^{MS}; 2 – praktyczne podstawy Word^{MS} i Excel^{MS}; 3-5 – większy zakres umiejętności i praktyki;

⁵*Wrażenie ogólne*: 1-6 punktów na podstawie rozmowy z kandydatką (np. pewność siebie).

Zmienna wieku została przywołana, ponieważ wystąpiła w ogłoszeniu („...poszukujemy kandydatki w wieku 25-29 lat...”), ale (1) nie była w istocie traktowana w sposób absolutny, a zarazem (2) trzeba było tę zmienną jakoś uwzględnić. Można było, na przykład, zakodować wiek kandydatek w ten sposób, żeby przyznawać punkty według następującego schematu: 2 punkty, jeśli $x_{i1} \in [25, 29]$, 1 punkt, jeśli $x_{i1} \in [21, 24] \cup [30, 36]$, oraz 0 punktów w pozostałych przypadkach. (Zauważmy, że nie interesuje nas logika kryjąca się za tymi wymaganiami dotyczącymi wieku, podobnie, jak nie wnikamy w powody przyznawania tych, a nie innych liczb punktów na podstawie „Wrażenia ogólnego”.)

WYKŁAD Z METOD ANALIZY DANYCH

Należy zauważyć, że dla takiej punktowej skali wszystkich zmiennych (abstrahując od dość trywialnego faktu, że uszeregowanie jest tutaj odwrotne od założonego w opisie metod) możemy być usprawiedliwieni, jeśli po prostu zsumujemy punkty dla poszczególnych zmiennych (metoda często stosowana przy różnego rodzaju egzaminach, z czym tutaj, faktycznie, mamy do czynienia, a równoważna sumie po zmiennych z równymi wagami). W wyniku takiego działania otrzymujemy liczby punktów pokazane w pierwszej kolumnie poniższej tabeli. Kolejne kolumny pokazują wyniki dla innych metod (w nawiasach kwadratowych pokazano miejsca w wynikowych uporządkowaniach, z uwzględnieniem miejsc *ex aequo*):

<i>Kandydatki</i>	<i>Suma punktów</i>	<i>Suma miejsc¹</i>	<i>Leksyko-graficznie²</i>	<i>Suma ważona (1)³</i>	<i>Suma ważona (2)⁴</i>
Inwencja A.	12 [3]	15 [3]	[5]	1,9 [4]	2,0 [5]
Porcja B.	12 [3]	15 [3]	[6]	1,85 [5]	2,0 [5]
Kwintessencja C.	12 [3]	15 [3]	[3]	2,0 [3]	2,2 [4]
Deucja D.	9 [4]	18 [5]	[7]	1,25 [6]	1,7 [6]
Kollacja E.	16 [1]	12 [1]	[1]	2,8 [1]	2,85 [1]
Redukcja F	14 [2]	13 [4]	[4]	2,3 [2]	2,4 [2]
Wallencja G.	12 [3]	16 [2]	[2]	2,3 [2]	2,35 [3]

¹ uwzględniając miejsca *ex aequo*;

² miejsca w uporządkowaniu leksykograficznym według kolejności z poprzedniej tabeli, z przesunięciem zmiennej wieku na ostatnie miejsce

³ wagi: wykształcenie: 0,25; doświadczenie: 0,2; języki: 0,2; komputer: 0,15; wrażenie ogólne: 0,1; wiek: 0,1

⁴ wagi: wykształcenie: 0,2; doświadczenie: 0,2; języki: 0,2; komputer: 0,15; wrażenie ogólne: 0,2; wiek: 0,05

IV.2.9. Jakie wnioski można wysnuć z tego przydługiego przykładu? (proszę zauważyć, że ten przykład, niezależnie od swojej akademickości, nie jest „złośliwy”, tzn. nie zawiera żadnych rozmyślnie ułożonych „dziwności”). A więc, po kolei, nie pomijając wniosków pozornie trywialnych:

- (1) wyniki otrzymane przy pomocy różnych metod różnią się między sobą;
- (2) istnieje zatem możliwość manipulacji (np. osoba prowadząca wywiad, najwyraźniej – dla nieznanych przyczyn – wysoko oceniająca

- p. Deucję, mogłaby próbować przeforsować metodę leksykograficzną, przy założeniu, że pierwszą rozpatrywaną zmienną będzie właśnie „Wrażenie ogólne”);
- (3) niewątpliwie to właśnie w ramach porządkowania leksykograficznego najłatwiej jest dokonywać manipulacji (wystarczy, by pewien obiekt był „najlepszy” według choćby jednej zmiennej, to nawet jeśli jest „najgorszy” według pozostałych, można go ustawić na najwyższej pozycji rankingu); wówczas jednak taka manipulacja będzie najbardziej widoczna;
 - (4) pewne możliwości manipulacji daje także metoda sumy ważonej, ale w tym przypadku osiągnięcie wyniku podobnego do porządkowania leksykograficznego wymagałoby w istocie zaprzeczenia sensu metody sumy ważonej; możliwości manipulacji w ramach metody sumy ważonej wzrastają istotnie w przypadkach, w których chodzi o stosunkowo niewielkie „przesunięcia” w końcowym uporządkowaniu z punktu widzenia koniecznych zmian w wartościach odpowiednich wag;
 - (5) jednakże w ramach „rozsądnych” metod i procedur, jeśli tylko zbiór danych przejawia odpowiednio wyraźny charakter (w tym przypadku: niewątpliwie wiodąca pozycja Kollacji), możliwe jest uzyskanie wyniku charakterze „obiektywnym”;
 - (6) metoda sumy miejsc nie wymaga określania wag, co jest jej niewątpliwą zaletą (chyba, że wagi w sposób „naturalny” wynikają z jakichś innych przesłanek), ale za to metoda sumy ważonej pozwala na analizę „wrażliwości” rozwiązań na rozkład wag (np. odpowiedź na ważne często pytanie: „czy istnieje taki rozkład wag, przy którym obiekt i znajdzie się na pierwszym miejscu rankingu, a jeśli tak, to jaki?”); jeśli rozwiązanie (uporządkowanie zagregowane) jest bardzo wrażliwe na rozkład wag $\{w_k\}_k$, czyli niewielka zmiana wartości wag powoduje istotne zmiany w uporządkowaniu, to nasze uporządkowanie zagregowane jest słabo uwarunkowane i może być łatwo poddane w wątpliwość.

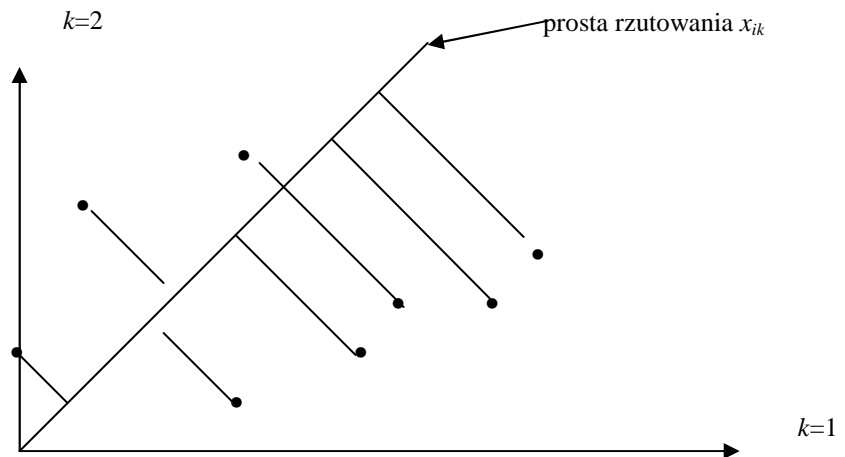
IV.3. Dalsze rozwinięcia metod porządkowania wielowymiarowego

IV.3.1. Omówiliśmy poprzednio właściwie wszystkie zasadnicze, stosowane praktycznie, metody agregacji uporządkowań, jakkolwiek, po pierwsze, nie we wszystkich przypadkach rankingi częściowe występowały w postaci jawnej (np. w algorytmie sumy ważonej nie mieliśmy do czynienia z upo-

rządkowaniami według poszczególnych zmiennych), a po drugie – nie omówiliśmy niemalże wcale zagadnień szczegółowych, odnoszących się do kwestii stanowiących przedmiot *optymalizacji wielokryterialnej*, czy tzw. *social choice theory*, a więc związanych z formalnymi własnościami metod.

IV.3.2. Przejdziemy obecnie do omówienia kilku postaci zadania porządkowania wielowymiarowego i związanych z nimi metod, w których właśnie uporządkowania cząstkowe występują wyłącznie w sposób niejawny.

IV.3.3. Przykład IV.2. Pierwszą z omawianych postaci zadania zilustrujemy następującym przykładem: niech obiekty $i=1,...,8$ będą opisane dwiema zmiennymi (dwie zmienne pozwalają na ilustrację na płaszczyźnie), $k=1,2$, przy czym wartości x_{ik} są jak pokazane czarnymi punktami na poniższym rysunku:



Wartości zmiennej zagregowanej otrzymujemy poprzez rzutowanie położeń punktów x_{ik} na pewną prostą, zaś położenia rzutów na tej prostej stanowią podstawę do uporządkowania wielowymiarowego (w tym przypadku: dwuwymiarowego).

IV.3.3. Łatwo się zorientować, że – czego jednak nie będziemy dowodzili – zilustrowane tutaj postępowanie jest w istocie równoważne sumie ważonej po wartościach zmiennych. Wagi przypisane poszczególnym zmiennym odzwierciedlane są przez nachylenie prostej rzutowania (w przypadku większej liczby zmiennych: nachylenie odpowiedniej hiperpłaszczyzny rzutowania). Jednakże w tym przypadku wagi zostały wyspecyfikowane nie w spo-

sób jawny, poprzez ich zadanie przez analityka, lub specjalistę z danej dziedziny, lecz jako wynik pewnego postępowania analitycznego, mającego przynajmniej pozór obiektywności. Dalej, jeśli pamiętamy, że rzutowanie jest czynnością, której wynikiem jest najkrótsza droga do danego tworu geometrycznego (tutaj: do prostej, lub hiperpłaszczyzny), to możemy całe postępowanie wyrazić w następującej, nieco uogólnionej postaci, mającej charakter zadania programowania matematycznego:

dla zadanych $x_i = \{x_{i1}, \dots, x_{im}\}$ znaleźć takie $o_i, i=1, \dots, n$, dla których osiągniemy

$$\min \sum_i d(x_i, o_i) \quad (\text{IV.1a})$$

przy ograniczeniu o ogólnej postaci

$$f(o_1, \dots, o_n) = 0 \quad (\text{IV.1b})$$

określającym „położenie” punktów o_i (np. przynależność do prostej) i ewentualnie jeszcze innych ograniczeniach, wyznaczających charakter poszukiwanych wartości o_i .

W zadaniu (IV.1) wagi ukryte są w ograniczeniu (IV.1b), zazwyczaj w postaci odpowiednich współczynników (zauważmy, że moglibyśmy również wprowadzić wagi w jawnej postaci wprost do sumy w minimalizowanym wyrażeniu z (IV.1a)). Najważniejsze jednak jest to, że postać zadania (IV.1) daje nam możliwość „optymalizacji” wag: przecież współczynnik nachylenia prostej z Przykładu IV.2 mógł nie być zadany z góry, ale stanowić nieznaną parametr zadania! I taki jest w istocie sens sformułowania (IV.1): szukamy uporządkowania wielowymiarowego, które możliwie najlepiej odzwierciedla położenia punktów x_i w przestrzeni \mathbf{X} , zaś ewentualne wagi, jeśli w ogóle warto w tym kontekście odwoływać się do tego pojęcia, są wynikiem odpowiedniego postępowania. Jeśli natomiast zależy nam na wprowadzeniu jawnych wag, to, jak wspomnieliśmy, można je uwzględnić w minimalizowanej sumie z (IV.1a).

IV.3.4. Przyjmując możliwość manipulowania wagami zmiennych przy pomocy ograniczenia (IV.1b) musimy pamiętać jednak, że nie może ono mieć dowolnego charakteru. Jeśli określa ono prostą w przestrzeni dwóch zmiennych, x_1 i x_2 , to nachylenie tej prostej, wyznaczone przez współczynnik a z równości $x_2 = ax_1 + b$, powinno być dodatnie ($a > 0$). I analogicznie dla równania odpowiedniej hiperpłaszczyzny przy $m > 2$. Wynika to z jednorodności warunków uporządkowania (porządkujemy według wszystkich zmiennych „w tym samym kierunku”).

IV.3.5. Jedna uwaga na zakończenie omawiania tego podejścia, opartego na rzutowaniu. Otóż, zarówno na rysunku, stanowiącym ilustrację, jak i we wzorach (IV.1) zakładamy, że wszystkie występujące w nich zmienne zosta-

ły znormalizowane, ponieważ w przeciwnym przypadku otrzymane w sposób niejawnny wagi zmiennych byłyby silnie zależne od skal, w jakich zostały one wyrażone.

IV.3.5. W wielu przypadkach sporządzenie macierzy wartości x_{ik} , prowadzącej do uporządkowań cząstkowych, jawnych, bądź niejawnnych, jest trudne, albo wręcz niemożliwe. W szczególności, w badaniach marketingowych zakłada się dość często, że otrzymanie takich wartości na podstawie odpowiedzi na pytania ankietowe jest mocno obciążone ograniczonymi możliwościami respondentów (brak porównywalności, zmienność skal itp.). W tych sytuacjach najczęściej odwołujemy się do *porównań parami*. Oznacza to, że zamiast wartości x_{ik} otrzymujemy od respondentów wartości („precedencji”) y_{ij} , których interpretacja zazwyczaj jest taka, że $y_{ij}=1$, jeśli i jest „lepsze” od j , oraz $y_{ij}=0$ w przeciwnym przypadku. Łatwo zauważyć, że tak określone wartości wynikające z porównań parami są analogiczne do zmiennych występujących w warunkach (II.21), definiujących relację większości. W sytuacji, na przykład, zbierania opinii konsumentów na temat pewnej grupy produktów, ze względu na ich cechy (zmienne) k nie możemy jednak wymagać od respondentów, aby przestrzegali warunków określonych przez (II.21), zwłaszcza, jeśli mają do czynienia z większą liczbą obiektów, n . Skoro jednak tak, to nie możemy wprost na podstawie otrzymanych tą drogą wartości y_{ij} , powiedzmy – średnich po respondentach – ustalać kolejności obiektów w uporządkowaniu. Faktycznie, jesteśmy zmuszeni szukać na podstawie y_{ij} otrzymanych z badania ankietowego wartości y_{ij}^* , spełniających warunki relacji większości, w ten sposób, że y_{ij}^* różnią się możliwie niewiele od wyjściowych y_{ij} .

Należy przy tym zauważyć, że wartości wyjściowe, empiryczne y_{ij} mogą być otrzymane jako średnie po poszczególnych respondentach, a także jako średnie po poszczególnych zmiennych. Niewątpliwie otrzymujemy wówczas wartości $y_{ij} \in [0,1]$ raczej niż $\in \{0,1\}$. Tym niemniej, możliwe jest przeprowadzenie operacji minimalizacji funkcji odzwierciedlającej odbieganie zbioru wartości y_{ij}^* od y_{ij} i uzyskanie w ten sposób uporządkowania zagregowanego. To postępowanie używane jest w niektórych metodach badania rynku.

Odpowiednie zadanie ma postać formalną zadania programowania matematycznego, a mianowicie, na przykład, następującego zadania programowania liniowego binarnego:

$$\sum_{i,j} (y_{ij}^* y_{ij} + (1 - y_{ij}^*)(1 - y_{ij})) \rightarrow \max \quad (\text{IV.2a})$$

przy znanych już ograniczeniach

$$y_{ij} \in \{0,1\} \quad (\text{IV.2b})$$

$$y_{ii} = 0 \quad (\text{IV.2c})$$

$$y_{ij} + y_{ji} \leq 1 \quad (\text{IV.2d})$$

$$y_{ij} + y_{jl} - y_{il} \leq 1, \quad (\text{IV.2e})$$

które to ograniczenia wymuszają uporządkowanie bez „remisów” ($y_{ii}=0$). Sens funkcji celu (IV.2a) polega na tym, by wyższym niż $\frac{1}{2}$ wartościom empirycznym $y_{ij}^* \in [0,1]$ przypisać w możliwie jak największej liczbie przypadków $y_{ij}=1$ (pierwszy ze składników pod sumą) i odwrotnie, wartościom y_{ij}^* mniejszym od $\frac{1}{2}$ przypisać $y_{ij}=0$ (drugi ze składników pod sumą). Za-uważmy, że gdyby pominąć drugi ze składników pod sumą, to optimum trywialne uzyskalibyśmy dla $y_{ij}=1 \ \forall i < j \in I$. Stąd konieczność uwzględnienia drugiego ze składników.

IV.3.6. Warto jeszcze na chwilę zatrzymać się nad zadaniem (IV.2), prowadzącym do otrzymania uporządkowań na podstawie porównań parami.

Po pierwsze, wartości wyjściowe y_{ij}^* mogą wynikać nie koniecznie z uśredniania pojedynczych „precedencji” podanych przez respondentów jakichś badań marketingowych, ale raczej wynikać z takich „precedencji” wyznaczonych lub otrzymanych dla poszczególnych zmiennych (y_{ij}^k). Wówczas dopuszczalne staje się także otrzymywanie y_{ij}^* jako sumy ważonej po pojedynczych y_{ij}^k . W ten sposób zaproponowana metoda staje się w sposób jawny metodą agregacji wielowymiarowej.

Po drugie, w nawiązaniu do ostatniej uwagi z punktu IV.3.5, możemy funkcję celu (IV.2a) przedstawić w postaci parametrycznej:

$$\sum_{i,j} (ry_{ij}^* + (1-r)(1-y_{ij}^*)(1-y_{ij})) \rightarrow \max \quad (\text{IV.2a}')$$

w której $r \in [0,1]$. Zgodnie z uwagą z punktu IV.3.5, dla $r=1$, a więc pominięcia drugiego ze składników pod sumą, otrzymalibyśmy trywialne rozwiązanie o postaci $y_{ij}=1 \ \forall i < j \in I$. Takie rozwiązanie jest po prostu zachowaniem uporządkowania, jakie przyjęliśmy w celu zapisania naszego zadania (kolejność obiektów zgodnie z wyjściową numeracją). Z kolei, dla $r=0$, trywialnym, łatwo przewidywalnym rozwiązaniem jest $y_{ij}=0 \ \forall i < j \in I$, czyli uporządkowanie będące dokładnie odwróceniem kolejności przyjętej przy zapisie danych. Gdybyśmy zadanie sparametryzowane, (IV.2a'), (IV.2b-e) rozwiązywali dla coraz niższych wartości r począwszy od 1, to otrzymywane kolejno rozwiązania polegałyby na wprowadzaniu do początkowej kolejności obiektów zmian (zamian miejsc i przesunięć), począwszy od takich, które są w największym stopniu uzasadnione danymi precedencjami („najkonieczniejszych”). Dla $r=1/2$ otrzymalibyśmy poszukiwane rozwiązanie optymalne, zaś dla $r < 1/2$ kolejne otrzymywane rozwiązania polegałyby na dalszych modyfikacjach, oddalających uporządkowanie od optymalnego w kierunku kolejności odwrotnej od wyjściowej ($y_{ij}=0 \ \forall i < j \in I$). Procedura taka pozwoliłaby na śledzenie „siły” precedencji, a także na analizę wrażliwości upo-

rządkowania optymalnego (zakres r , w jakim ono obowiązuje, różnice w stosunku do poprzedzających i następnych uporządkowań).

Pewne dalsze, obszerniejsze informacje na temat zastosowania modeli i technik optymalizacyjnych, podobnych do przytoczonych tutaj, w zagadnieniu porządkowania wielowymiarowego można znaleźć w pracach Owsieński, Zadrozny (1986a,b, 1988).

IV.3.7. Wiele powszechnie używanych aplikacji, w tym również arkuszy kalkulacyjnych, zawiera pewne możliwości realizacji niektórych metod porządkowania. Są to jednak w przeważającej większości możliwości niezmienne ograniczone, praktycznie dotyczące porządkowania według jednej zmiennej (ewentualnie porządkowania leksykograficznego). Tym niemniej, szersze możliwości tych aplikacji, w tym dotyczące tworzenia agregatów zmiennych (na przykład ich kombinacji liniowych), pozwalają na realizację innych również metod agregacji uporządkowań. Niezależnie od tego, zawarte w nich algorytmy optymalizacji liniowej pozwalają na dokonywanie wielu operacji i formułowanie różnych konkretnych zadań, na przykład takich, jak opisane w punkcie IV.3.6, bądź związanych z poszukiwaniem zestawów wag zmiennych, w_k , o określonych własnościach. Tak więc, jakkolwiek aplikacje takie nie zawierają na ogół żadnych jawnych metod porządkowania według wielu zmiennych, mogą być używane do tego celu dzięki różnym innym opcjom w nich zawartym.

WYKŁAD V

Analiza skupień: zadanie i jego warianty. Podstawowe grupy metod. Charakterystyka poszczególnych grup metod. Zastosowania.

V.1. Zadanie analizy skupień i jego warianty

V.1.1. Jak już wspomniano (Wykład III), zadanie analizy skupień polega na tym, żeby podzielić zbiór (indeksów) obiektów, I , na podzbiory („skupienia”), oznaczone A_q , $q=1, \dots, p$, w taki sposób, by obiekty umieszczone w tych samych skupieniach były możliwie do siebie podobne (bliskie), zaś obiekty umieszczone w różnych skupieniach – możliwie dalece różniły się między sobą.

V.1.2. Tak ogólnie sformułowane zadanie pozostawia wiele jego elementów do dalszego zdefiniowania. Dotyczy to przede wszystkim pojęć związanych z kryterium podziału, a więc „możliwie podobne” i „możliwie różne”, a także warunków nakładanych na podział, w tym, w szczególności – warunków nakładanych na skupienia oraz ewentualnie ustalenie liczby skupień p , bądź pozostawienie tej liczby jako jednej z niewiadomych, będących również elementem poszukiwanego rozwiązania zadania.

Dotykamy tutaj kwestii „obiektywności” ewentualnego podziału zbioru I , czyli „rzeczywistego istnienia” odpowiednich podzbiorów (skupień). Nie wchodząc w szczegóły tego w istocie dość skomplikowanego zagadnienia stwierdzmy tylko, że dla potrzeb niniejszego wykładu zakładać będziemy, że podział „istnieje”, jeśli spełnia warunki (kryteria) ogólnego zadania analizy skupień w sensie, jaki im nadała osoba prowadząca badanie. Co prawda można uznać to za wycofanie się na pozycję „definicji operacyjnej”, ale trzeba pamiętać, że analiza skupień jest dziedziną na wskroś pragmatyczną.

V.1.3. Wybór i definicja kryteriów, odnoszących się do określeń „możliwie podobne” i „możliwie różne”, stanowią o różnorodności metod analizy skupień. Podkreślmy w tym miejscu, że ta różnorodność głównie odnosi się do „poziomów percepcji”, czyli – czy podobieństwa i różnice są postrzegane (i następnie wykorzystywane algorytmicznie) na (i) poziomie poszczególnych obiektów, (ii) poziomie całych skupień, czy też (iii) całego podziału. Sposób postrzegania podobieństw i zróżnicowań, i wynikający z tego opis zadania stanowi w pewnym sensie „model” zadania analizy skupień, który prowadzi

do opracowywania odpowiadających „percepcji” czy „modelowi” sposobów rozwiązywania.

Nie istnieje obecnie żadna ogólna teoria, która pozwalałaby na stwierdzenie, że któraś z grup metod, wyodrębnionych w opisany powyżej sposób, jest „właściwa”, zaś inna – nie, ze względu na fakt, że spełniają one różne postulaty, stawiane metodom, wynikające z ogólnego sformułowania zadania analizy skupień, jakkolwiek istnieją poważne przesłanki dotyczące oceny poszczególnych cech metod.

V.1.4. Spotkaliśmy się już poprzednio z warunkami nakładanymi na podział właściwy, tj. $P = \{A_q\}_q$, dla którego obowiązują warunki $\cup_q A_q = I$ oraz $A_q \cap A_{q'} = \emptyset$, $q \neq q'$, czyli, że skupienia wyczerpują cały zbiór I i że są rozłączne. O ile pierwszy z tych warunków jest oczywisty i trudno go pominąć, o tyle drugi bywa traktowany na różne sposoby. I tak, w szczególności, niekiedy rozwiązywane jest zadanie, w którym drugi z warunków jest pominięty, a za to nałożone są inne warunki, bądź też sensowność rozwiązania jest zapewniana przez zastosowane kryterium jakości podziału. Specyficzna, także mniej rygorystyczna, postać przytoczonych warunków obowiązuje w przypadkach, w których poszukuje się skupień A_q w postaci zbiorów rozmytych, o czym również już wspominaliśmy.

Podobnie, była już mowa o możliwości rozwiązywania zadania analizy skupień, w którym poza podzbiorami A_q szukamy także „reprezentantów” tych podzbiorów, bądź to w postaci poszczególnych obiektów (należących do zbioru X , albo nie należących do tego zbioru, a tylko do przestrzeni \mathbf{X}), bądź w postaci ich zbiorów (będących z kolei najczęściej podzbiorami skupień), bądź w postaci pewnych modeli lub reguł (np. modeli regresji w odniesieniu do skupień lub reguł klasyfikacyjnych), a także innych sposobów reprezentowania skupień.

V.1.5. Zaznaczmy także, że nie będziemy się tutaj zajmowali niektórymi zadaniami pokrewnymi zadaniu analizy skupień, w tym, w szczególności, zadaniem „mieszaniny rozkładów”, które należy do dziedziny rachunku prawdopodobieństwa i statystyki matematycznej, niezależnie od wielu punktów wspólnych z rozważanym tutaj zadaniem analizy skupień – zwłaszcza w zakresie metod.

W zadaniu mieszaniny rozkładów zakładamy mianowicie, że nasza macierz (zbiór obserwacji) X jest wynikiem realizacji pewnej liczby (p) zmiennych losowych \mathbf{x}_q o różnych rozkładach prawdopodobieństwa, $f_q(x)$, $x \in \mathbf{X}$, a zatem otrzymany zbiór obserwacji, X , jest wobec tego wynikiem działania (zbiorem realizacji) „mieszaniny” tych zmiennych losowych, na przykład w postaci $f(x) = \sum_q w_q f_q(x)$, gdzie w_q są wagami, z jakimi poszczególne zmienne (ich gęstości rozkładu) wchodzi do mieszaniny.

Poszukujemy, na podstawie danych X , rozkładów $f_q(x)$ oraz wag w_q . Zauważmy, że zadanie to jest w zasadzie „mocniejsze” niż podstawowe zadanie analizy skupień, bowiem poszukujemy tutaj nie tyle (i nie tylko) podzbiorów A_q , ale przede wszystkim rozkładów prawdopodobieństwa, które je wygenerowały, a więc w pewnym sensie „modeli” odpowiadających skupieniom (same skupienia są zatem drugorzędnym przedmiotem tego zadania). Z drugiej jednak strony zadanie mieszaniny rozkładów nie obejmuje takich aspektów, często ujmowanych w analizie skupień, jak wspomniana uprzednio kwestia identyfikacji reprezentantów skupień. Poza tym, niezależnie od osiągnięć teoretycznych i metodycznych w zakresie rozwiązywania zadania mieszaniny rozkładów – na przykład możliwości wykrywania rozkładów o tych samych średnich (czyli „skupień” – zbiorów obserwacji – zajmujących region przestrzeni \mathbf{X} wokół tego samego punktu) – istniejące wyniki i techniki ograniczają się do bardzo wąskiej klasy zagadnień: rozkłady normalne lub jednorodnie, niewielka liczba rozkładów, najczęściej tylko dwa, wszystkie zmienne o tych samych – co do charakteru – rozkładach, itp.

Jednocześnie, podkreślmy, dość słabo uzasadnione teoretycznie techniki z zakresu analizy skupień stały się podstawą wielu metod z zakresu statystyki matematycznej, w tym właśnie technik rozwiązywania zadania mieszaniny rozkładów.

V.1.6. W tym miejscu zamieścimy kilka uwag historycznych. Uważa się na ogół, że dziedzina analizy skupień pojawiła się wraz ze studencką pracą polskiego antropologa Jana Czekanowskiego z początku XX wieku, w której analizował on związki (podobieństwa) kilkunastu czaszek neandertalczyków i próbował wykorzystać do tego celu prostą, opracowaną przez siebie, metodę analizy. Sprowadzała się ona do takiej permutacji kolumn i wierszy macierzy obserwacji X , aby na sąsiednich pozycjach umieścić obiekty i zmienne podobne do siebie. W opracowanej później, bardziej sformalizowanej wersji metoda przybrała nazwę „dendrytu Czekanowskiego”. Angielskojęzyczna nazwa dziedziny („*cluster analysis*”) pojawiła się w okresie międzywojennym za sprawą Tryona, specjalisty od zastosowań matematyki w psychologii (jej dyscypliny ilościowej – psychometrii). Prace Tryona do chwili obecnej są klasyką nie tylko w psychometrii, ale także, na przykład, w badaniach rynkowych. Po II Wojnie Światowej grupa matematyków wrocławskich, wywodzących się z przedwojennej szkoły lwowskiej (której reprezentantem był, między innymi, Stefan Banach), pod kierunkiem Hugona Steinhausa, opracowała metodę nazwaną później „dendrytem wrocławskim”. Metoda ta stała się, w rozlicznych wersjach, podstawą dużej grupy metod analizy skupień, a także i technik stosowanych w wielowymiarowej statystyce matematycznej.

V.2. Podstawowe grupy metod

V.2.1. Zgodnie z poprzednio sformułowaną uwagą, metody analizy skupień podzielimy według poziomu percepcji, na którym postrzegamy i analizujemy podobieństwa i różnice, wspomniane w podstawowej postaci zadania analizy skupień (a przeto i „modelu” zadania analizy skupień), i według których optymalizujemy podział P zbioru I . I tak, podzielimy metody na:

- metody hierarchicznej agregacji i hierarchicznego podziału,
- metody centrowania i reallokacji (metody centrów, także metody k -średnich, w naszym przypadku: metody p -średnich),
- metody analizy gęstości i podziału przestrzeni \mathbf{X} ,
- metody odwołujące się do globalnych postaci funkcji celu.

V.2.2. Metody hierarchicznej agregacji i hierarchicznego podziału, bodaj najliczniejsza grupa metod, oparte są na postrzeganiu podobieństwa i różnicowania między obiektami na poziomie poszczególnych obiektów, ewentualnie obiektów względem poszczególnych skupień. Polegają one na iteracyjnym łączeniu obiektów w skupienia według najmniejszej odległości między nimi, bądź na podziale zbiorów obiektów na mniejsze według największych odległości. W wyniku tak zarysowanego postępowania otrzymuje się pewną hierarchię (drzewo hierarchii), odpowiadającą agregacjom bądź podziałom. W dalszej części niniejszego wykładu omówimy nieco dokładniej główną, klasyczną grupę algorytmów agregacji hierarchicznej.

V.2.3. Metody centrowania i reallokacji (p -średnich) oparte są na idei reprezentantów skupień o określonych własnościach – na przykład średnich ze skupienia (stąd nazwa). Ich zasadą działania jest powtarzana iteracja: dla pewnego zadanego podziału P na skupienia A_q znajdź reprezentantów tych skupień, x_q^a , a następnie przyporządkuj tym reprezentantom obiekty ze zbioru X , tworząc w ten sposób nowe skupienia A'_q , a zatem i nowy podział, P' . Potem znów znajdujemy reprezentantów, tym razem dla A'_q , itd. itp. Iteracje te trwają aż do spełnienia pewnego kryterium stopu. W ramach tej grupy metod szczególnie łatwo realizowane są metody oparte na zbiorach (skupieniach) rozmytych. Zauważmy, że podstawą tej grupy metod jest postrzegania bliskości i odległości w ramach całych – ale osobnych – skupień. Podstawy działania tych metod zostaną również przedstawione w wykładzie.

V.2.4. O metodach analizy gęstości i podziału przestrzeni \mathbf{X} wspomnimy tylko w niniejszym wstępie. Nie tworzą one żadnej wyraźnej, dobrze określonej grupy metod, którą można by opisać w jednolity sposób. Dlatego też przytoczymy, dla przykładu, dwa rodzaje zasad, jakimi posługują się meto-

dy tutaj zaliczone. I tak, wiele metod analizy gęstości posługuje się tak zwaną funkcją gęstości $g(x,d)$, która może być określona jako

$$g(x,d) = \quad (V.1)$$

= liczba obiektów $\in I$, znajdujących się w odległości $\leq d$ od $x \in X$,

przy czym najczęściej x odpowiada jednemu z obiektów. Łatwo zauważyć, że

- (i) $g(x,d)$ jest funkcją rosnącą w d , od 1 (lub 0, w zależności od tego, czy x jest jednym z obiektów należących do I , czy też nie), aż do osiągnięcia wartości n , i to niezależnie od x ;
- (ii) ewentualne wykorzystanie $g(x,d)$ w analizie skupień opierać się będzie na charakterze tych funkcji i ich porównaniu dla poszczególnych x ; funkcjom szybko rosnącym dla niewielkich wartości d (zauważmy dalej, że możliwe jest w pewnej mierze normalizowanie $g(x,d)$, oparte na statystyce odległości w obrębie zbioru X) „powinna” odpowiadać przynależność do (większych skupień), zaś funkcjom rosnącym powoli – przynależność do małych skupień lub położenie (w „przerwach”) poza skupieniami.

Zasadniczym problemem metod opartych na funkcjach gęstości, które z całą pewnością niosą w sobie znaczącą informację o strukturze zbioru X , jest brak prostych zasad konstruowania na ich podstawie skupień oraz na ogół bardzo duży nakład obliczeniowy, zarówno w sensie potrzebnej objętości pamięci (funkcje $g(.,.)!$), jak i liczby wykonywanych operacji. W istocie, metody te najczęściej oparte są na zestawie operacji o charakterze heurystycznym.

Metody podziału przestrzeni X w pewien sposób również odwołują się do gęstości obiektów. Polegają one, w uproszczeniu, na założeniu określonego sposobu podziału przestrzeni X , a następnie wykonywanie operacji na tym podziale, w zależności od wyników analizy gęstości czy liczebności obiektów. Najprostszym – i najczęstszym – przykładem takiego postępowania jest podział przestrzeni X na wielowymiarowe „kostki” na podstawie podziału zakresów wartości poszczególnych zmiennych x_k na przedziały (proste zwłaszcza w przypadku zmiennych ciągłych) lub inne podzbiory zbioru wartości zmiennych. Można założyć, że taki początkowy podział X powinien dawać liczbę „kostek” porównywalną z liczbą obiektów n . Żeby „mieć pewność”, że nie popełnimy żadnego większego błędu, liczba kostek może być odpowiednio większa od n , na przykład kilkukrotnie. Działanie algorytmu polega, w dużym skrócie, na wykrywaniu sąsiadujących ze sobą kostek pustych (lub „prawie” pustych, czyli uznanych za „puste”) oraz sąsiadujących ze sobą kostek o dużym wypełnieniu obiektami („pełnych”). Te ostatnie, oczywiście, powinny odpowiadać skupieniom. Następnie kostki o odpo-

wiednich cechach (na przykład: najpierw kostki puste, a następnie, kolejno, kostki sąsiadujące o najwyższych liczbach obiektów zawartych w nich) są łączone, aż do spełnienia pewnego kryterium stopu (aż do ustania „sensowności” łączenia).

Aby móc podjąć decyzję co do tego, czy dana kostka jest raczej „pusta”, czy raczej „pełna”, posługujemy się na ogół prostymi wskaźnikami opartymi na podstawowych statystykach zbioru danych w podziale na kostki. Wskaźniki te odwołują się zazwyczaj do najczęściej zakładanych hipotetycznych rozkładów obiektów w przestrzeni X . I tak, zwykle przyjmujemy, że wszystkie kostki mają tę samą „objętość”, czyli ten sam iloczyn liczebności zbiorów wyznaczających kostki wzdłuż różnych zmiennych (wynikających z podziału zbioru wartości zmiennej na równe liczby wartości dla zmiennych dyskretnych i/lub jakościowych, oraz podziału na równe odcinki dla zmiennych ciągłych albo quasi-ciągłych). Oznacza to, w istocie, milczące założenie, że rozkład obiektów w przestrzeni X jest równomierny. To najczęstsze założenie nie jest naturalne, biorąc pod uwagę, że analiza skupień stosowana jest na wstępnych etapach analizy danych, kiedy o zbiorze X i jego własnościach wiemy bardzo niewiele, albo zgoła prawie nic. Jeśli liczba takich jednakowych kostek wynosi an , gdzie a jest pewną liczbą naturalną, to przy założeniu rozkładu równomiernego możemy spodziewać się średnio w każdej kostce obecności $1/a$ obiektów. Odchylenie standardowe od tej średniej (które łatwo możemy wyznaczyć przy powyższych założeniach na podstawie znanych zależności z rachunku prawdopodobieństwa!) pozwoli nam na ocenę, które kostki należy uznać za „pełne”, a więc mogące stanowić części skupień, a które za „puste”, a więc stanowiące elementy regionów przestrzeni poza skupieniami (pamiętajmy, że jeśli założymy zbyt wysoką wartość a , to nawet i kostki nie zawierające w ogóle obiektów nie będą mogły w pierwszych krokach procedury być uznane za „puste”).

Metody podziału przestrzeni mają wiele cech wspólnych z wieloma obecnie rozwijanymi metodami „drażenia” lub „eksploracji” danych (ang. *data mining*), w szczególności: opieranie się wyłącznie na współrzędnych obiektów, a nie ich odległościach czy bliskościach, oraz uproszczony sposób dokonywania podziałów. Podobnie jak poprzednio omawiane metody funkcji gęstości również i metody podziału przestrzeni oparte są w zasadniczy sposób na doborze różnych heurystyk. Jednakże mają one (oczywiście) znacznie mniejsze wymagania co do pamięci, a w niektórych realizacjach również mniejszą znacznie liczbę operacji (od pewnego momentu operacje wykonywane są już nie na obiektach, lecz na kostkach, i to na coraz większych kostkach). Głównym problemem takich algorytmów jest ich realizacja dla bardzo dużych zbiorów X w sensie liczby n , gdyż wówczas liczba kostek rośnie w sposób istotny, niekiedy negując sens innych uproszczeń, właściwych tej grupie metod.

V.2.5. Nieliczna grupa metod, posługujących się *globalnymi funkcjami celu*, reprezentuje perspektywę całości podziału. Mamy tu na myśli fakt, że przedmiotem analizy (i optymalizacji) nie są poszczególne odległości czy nawet skupienia, ale faktycznie całość podziału P naraz. Perspektywa ta przybiera w praktyce zawsze postać pewnej funkcji celu, opisanej na całym podziale (funkcji jakości podziału). Zadanie polega zatem na jawnej lub niejawnej optymalizacji względem tej funkcji, przy zachowaniu odpowiednich ograniczeń (na przykład takich, jakie przytoczyliśmy opisując pojęcie relacji). Optymalizacja ta polega na doborze obiektów do skupień i – w niektórych przypadkach – także liczby skupień. Te grupę metod zilustrujemy na przykładzie metody autora wykładu.

Należy jednak pamiętać, że sam fakt opisanego zagadnienia przy pomocy funkcji celu, obejmującej cały zbiór X oraz podział P nie wystarcza do tego, byśmy mogli w sposób uzasadniony mówić o „globalności” odpowiednich metod, czy odpowiadających im modeli zadania. I tak, charakterystycznym przykładem są tutaj metody p -średnich, dla których, jak zobaczymy, można – i robi się to – sformułować „całościową” funkcję celu w powyższym sensie, co jednak nie oznacza, że w ogólności metody te posługują się optyką globalną, chyba, że wprowadzimy do nich dodatkowe warunki. Kwestię tę poruszymy przy omawianiu metod p -średnich.

V.2.6. Trzeba podkreślić, że wiele, albo nawet większość obecnie realizowanych w postaci odpowiednich pakietów metod analizy skupień łączy w sobie więcej niż jedną z wyżej wymienionych zasad. I tak, często występuje połączenie najbardziej chyba rozpowszechnionych metod, czyli algorytmów agregacji hierarchicznej i algorytmów p -średnich. W tym połączeniu algorytm agregacji hierarchicznej służy zazwyczaj do ustalenia pewnego rozwiązania (albo pewnych jego cech), które stanowi następnie punkt startowy dla algorytmu p -średnich.

V.2.7. Podkreślimy również, że w obecnej chwili do rozwiązywania zadania analizy skupień, lub jego konkretnych postaci szczegółowych, stosuje się, poza algorytmami bezpośrednio realizującymi zasady różnych grup metod, a zatem i odpowiednich modeli zadania, także rozwijane ostatnio metaheurystyki (w zadaniach programowania matematycznego - poszukiwanie tabu oraz symulowane wyżarzanie, a poza tym najczęściej sztuczne sieci neuronowe i algorytmy genetyczne). W niniejszym wykładzie nie będziemy opisywali zastosowań tych metaheurystyk, ponieważ nie są one związane bezpośrednio z modelami zadań analizy skupień, a także najczęściej są „dopasowywane” do odpowiednich, rozwiązywanych zadań. Głównym mianowicie przedmiotem wykładu w zakresie analizy skupień jest zrozumienie jej sensu oraz jej podstawowych modeli i związanych z nimi sformułowań, prowadzących bezpośrednio do odpowiednich metod. Założeniem wykładu

jest możliwość odwoływania się do istniejących – na przykład w powszechnie dostępnych pakietach statystycznych lub arkuszach kalkulacyjnych – aplikacji realizujących metody analizy skupień (i inne metody), lecz z ich właściwym zrozumieniem.

V.3. Algorytmy agregacji hierarchicznej

V.3.1. Zasada działania algorytmów agregacji hierarchicznej jest następująca:

- ♦ dla zbioru X opisów obiektów x_i o liczności n , któremu odpowiada zbiór indeksów I , mamy dane (wyznaczone) odległości d_{ij} między obiektami;
 - ♦ ustalamy początkowy podział zbioru I , oznaczony P^0 , który jest identyczny ze zbiorem I , to znaczy, że każdy obiekt jest jednocześnie skupieniem, a zatem $i=q$, $A_q=q$, oraz $p=n$;
 - ♦ wprowadzamy macierz odległości między skupieniami, oznaczonych $D_{qq'}$, przy czym początkowo, zgodnie z założeniem dotyczącym początkowego podziału P^0 , przyjmujemy $D_{qq'} = d_{ij}$;
 - ♦ wprowadzamy oznaczenie kroku algorytmu, t , początkowo $t=0$;
 - ♦ (*) dla danego podziału P^t , a zatem i macierzy odległości między skupieniami $D_{qq'}^t$, podział następny, P^{t+1} , otrzymywany jest w ten sposób, że znajdujemy $\min_{qq'} D_{qq'}^t$, a więc najmniejszą odległość między skupieniami w podziale P^t , i łączymy dwa odpowiadające tej najmniejszej odległości skupienia w jedno;
 - ♦ w ten sposób otrzymujemy podział P^{t+1} , w którym $p(P^{t+1}) = p(P^t) - 1$;
 - ♦ po dokonaniu połączenia dokonujemy modyfikacji macierzy odległości $D_{qq'}^t$, w odniesieniu do odległości pozostałych skupień względem obu skupień, które zostały połączone;
- $t:=t+1$, a następnie wracamy do (*).

Postępowanie takie kończy się, w sposób naturalny, po $n-1$ krokach, to znaczy dla $P^{n-1}=\{I\}$, czyli podziału stanowionego przez zaledwie jedno skupienie, obejmujące wszystkie obiekty.

Nazwa tej grupy algorytmów jest teraz oczywista: zarysowane postępowanie prowadzi do utworzenia struktury hierarchii (drzewa hierarchii), która powstaje poprzez łączenie w kolejnych krokach (odpowiadających kolejnym poziomom hierarchii) poszczególnych skupień, aż do otrzymania jednego skupienia. Zilustrowano to schematycznie na Rys. V.1 dla zbioru danych o $n=8$, w którym, najwyraźniej, istnieją dwie grupy (względnie) po-

dobnych do siebie obiektów ($\{1,2,3\}$ oraz $\{4,5,6,7\}$), a także obiekt od nich wszystkich „odstający” (8).

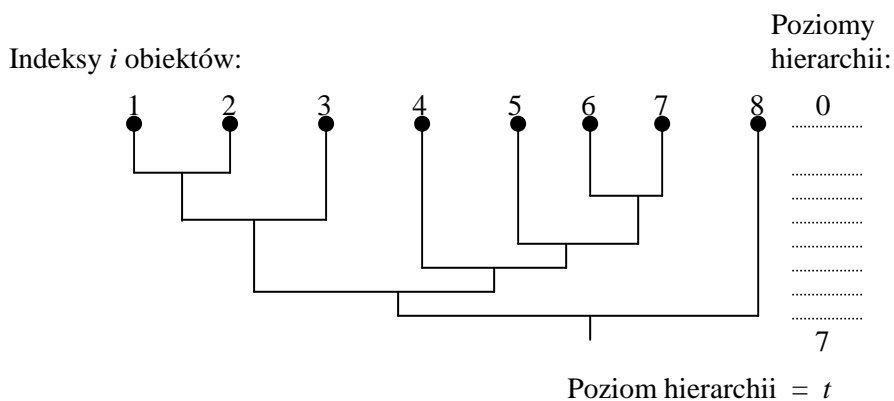
Poza diagramem hierarchii, zilustrowanym na Rys. V.1, istnieją inne reprezentacje wyników algorytmów agregacji hierarchicznej, zwłaszcza płaskie odwzorowania geometrycznej struktury zbioru danych (obiektów) w postaci grafów pokazujących kolejne połączenia, odpowiadające najkrótszym odległościom w zbiorze. Zawierają one bardziej syntetyczną informację niż diagramy hierarchii (pokazują bardziej „naocznie” strukturę przestrzenną ewentualnych skupień), ale są nieco trudniejsze w interpretacji (nie zawsze możliwe jest podanie odpowiednich charakterystyk liczbowych na diagramie i interpretacja jest w dużej mierze pozostawiona odbiorcy).

V.3.2. Łatwo zauważyć, że przedstawiona procedura jest w istocie procedurą łączenia najbliższych sobie skupień (w początkowych krokach: po prostu obiektów). Jest ona niezwykle prosta i intuicyjnie zrozumiała, a nawet wręcz oczywista.

Niezależnie od tego nie jest ona jednak bynajmniej jednoznacznie określona przez podany tutaj opis. Na bazie tego opisu możemy mianowicie stworzyć wiele różnych algorytmów, nie mówiąc o możliwościach, jakie się wyłaniają wtedy, gdy jeszcze dodatkowo (nieco) zmodyfikujemy tę procedurę.

Zasadniczym punktem, pozwalającym na (istotnie!) różne realizacje algorytmu agregacji hierarchicznej, jest mianowicie *sposób modyfikacji odległości między skupieniami po dokonaniu połączenia najbliższych skupień*.

Rys. V.1. Schemat działania algorytmów agregacji hierarchicznej.



V.3.3. Przeanalizujemy obecnie ten kluczowy punkt algorytmów agregacji hierarchicznej. Oznaczmy indeksy łączonych w pewnym kroku algorytmu skupień przez q^* , q^{**} , przy czym $q^* < q^{**}$. Załóżmy, że nowo utworzone skupienie będzie miało zatem indeks q^* . Rozmiar macierzy $\{D_{qq^*}\}$ ulega zmniejszeniu (zauważmy, że pomijamy tutaj konsekwentnie indeks kroku procedury) – znika z niej, oczywiście, wyraz $D_{q^*q^{**}}$, a ponadto wszystkie inne wyrazy, będące odległościami między (powyższe założenie o indeksie nowego skupienia) obiektem q^{**} a innymi obiektami ze zbioru X (zbioru indeksów I). Natomiast wyrazy D_{q^*q} , dla $q \neq q^*$, muszą zostać zmodyfikowane, jako odległości od nowego skupienia.

Istnieje szereg reguł wyznaczania nowych odległości D_{q^*q} , przede wszystkim, choć niekoniecznie tylko, na podstawie poprzednich odległości D_{q^*q} , $D_{q^{**}q}$ oraz $D_{q^*q^{**}}$. Większość z nich jest ujęta w postaci tak zwanego wzoru Lance'a-Williamsa-Jambu (przytoczonego tutaj w oryginalnej, prostszej postaci Lance'a-Williamsa), w którym po lewej stronie mamy nową odległość, a po stronie prawej – odległości z poprzedniego kroku:

$$D_{q^*q} = \alpha D_{q^*q} + \beta D_{q^{**}q} + \gamma D_{q^*q^{**}} + \delta |D_{q^*q} - D_{q^{**}q}|, \quad (V.2)$$

zaś poszczególne warianty są wyznaczone przez wartości współczynników α , β , γ , oraz δ , występujących w tym wzorze. Podstawowe warianty przedstawione są w poniższej tabelce:

Nazwa algorytmu	α	β	γ	δ
Najbliższego sąsiedztwa	1/2	1/2	0	-1/2
Najdalszego sąsiedztwa	1/2	1/2	0	1/2
Mediany	1/2	1/2	-1/4	0
Średniej grupowej	$\frac{n_{q^*}}{n_{q^*} + n_{q^{**}}}$	$\frac{n_{q^{**}}}{n_{q^*} + n_{q^{**}}}$	0	0
Środka ciężkości	$\frac{n_{q^*}}{n_{q^*} + n_{q^{**}}}$	$\frac{n_{q^{**}}}{n_{q^*} + n_{q^{**}}}$	$-\frac{n_{q^*}n_{q^{**}}}{n_{q^*} + n_{q^{**}}}$	0
J. H. Warda	$\frac{n_q + n_{q^*}}{n_q + n_{q^*} + n_{q^{**}}}$	$\frac{n_q + n_{q^{**}}}{n_q + n_{q^*} + n_{q^{**}}}$	$-\frac{n_q}{n_q + n_{q^*} + n_{q^{**}}}$	0

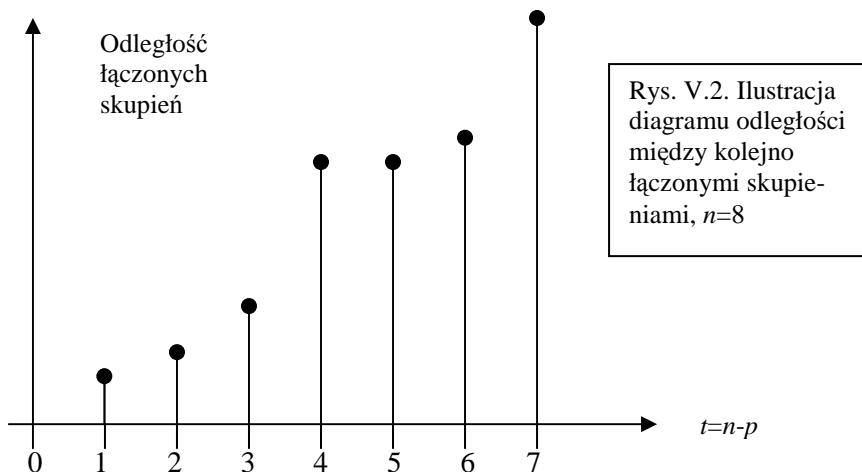
gdzie n_q oznacza liczbę elementów skupienia o indeksie q .

W istocie najpopularniejszymi – i mającymi najpoważniejsze zaplecze teoretyczne są dwa pierwsze algorytmy, w których – co łatwo sprawdzić przez zastosowanie odpowiednich wzorów z tabeli – nowa odległość między skupieniami (D_{q^*q}) powstaje przez wzięcie, odpowiednio, minimum albo maksimum z uprzednio rozważanych dwóch odległości, czyli D_{q^*q} i $D_{q^{**q}}$.

V.3.4. Nietrudno spostrzec, że algorytmy agregacji hierarchicznej w rzeczywistości nie prowadzą (wprost) do rozwiązania zadania analizy skupień, ale raczej dostarczają materiału do jego rozwiązywania: zamiast podziału P , będącego poszukiwanym przez nas rozwiązaniem, dostarczają ciągu podziałów, P' , otrzymanych na podstawie przesłanek o charakterze lokalnym (odległości między poszczególnymi obiektami, w dalszych krokach algorytmu traktowane jak odległości między skupieniami).

Jeśli „ufamy” tym algorytmom i poszukujemy, na przykład, podziału o określonej krotności p , to możemy wprost z ciągu $\{P'\}$ wybrać interesujący nas podział. Jak to łatwo sprawdzić, także na Rys. V.1, podział na p skupień otrzymamy w wyniku $n-p$ -tego kroku algorytmu.

W przeciwnym przypadku, po zastosowaniu wybranego algorytmu agregacji hierarchicznej i otrzymaniu całego drzewa hierarchii, sięgamy do innego jakiegoś wskaźnika, który ma nam podpowiedzieć: (i) który (lub które) z podziałów P' można zaakceptować, a także (ii) czy w ogóle istnieje wśród nich podział akceptowalny.



Wskaźniki takie odwołują się najczęściej do miar zbliżonych do wariancji („zwartości” otrzymanych skupień). Jednak bodaj najprostszym jest diagram kolejnych odległości łączonych skupień (Rys. V.2). Jest rzeczą oczy-

wistą, że diagram ten ma charakter rosnący (albo przynajmniej niemalejący). Może on przy tym wykazywać pewne cechy szczególne, na przykład: dla jakiegoś t (czyli p , poprzez $t=n-p$) możemy czasem zaobserwować wyraźny „skok” wartości odległości łączonych skupień. Jeśli tak, to należy zaakceptować podział o p sprzed „skoku” jako rozwiązanie (poważniejszy problem powstaje w sytuacji, gdy takich skoków – przynajmniej w ocenie analityka – jest więcej niż jeden). Jeśli natomiast krzywa jest gładka i nie wykazuje istotnych różnicowań przyrostów, to powstaje podejrzenie, że w ogóle struktura badanego zbioru nie ma charakteru odpowiadającego podziałowi na skupienia (podzbiory).

I tak, na przykład, na podstawie schematycznego diagramu z Rys. V.2 można by zaakceptować podział dla $t=3$, czyli o liczności skupień $p=n-3=8-3=5$, albo też podział dla $t=6$, czyli o $p=n-6=8-6=2$.

V.3.5. Algorytmy agregacji hierarchicznej są bodaj najczęściej implementowane w różnych pakietach obróbki danych. Wynika to z ich prostoty i łatwości interpretacji. Takie aplikacje zawierają na ogół możliwość otrzymania automatycznie diagramu drzewa hierarchii, pod warunkiem, oczywiście, że liczba obiektów (a przeto i poziomów hierarchii) nie jest zbyt duża (na ogół nie ma możliwości prezentacji innych graficznych odzwierciedleń wyników działania tych algorytmów). Jest to dodatkowa zaleta takich implementacji, poza samą realizacją algorytmu.

Wiele gotowych algorytmów agregacji hierarchicznej, zaimplementowanych w popularnych pakietach typu arkuszy kalkulacyjnych bądź statystycznych jest wyposażonych w testy statystyczne badające istotność podziału na skupienia (na przykład przy pomocy testu χ^2) i na tej podstawie najczęściej podają one wynik w postaci podziału P . Pamiętać jednak należy, że aplikacje te nie są w stanie sprawdzić, czy wszystkie założenia, jakie muszą być spełnione, aby móc bez zastrzeżeń zastosować takie testy, są faktycznie dla danego zbioru obiektów spełnione (na przykład wspomniana wielokrotnie kwestia jednorodności zmiennych). Dlatego też warto zawsze samemu przyjrzeć się proponowanym przez pakiety rozwiązaniom i ocenić je krytycznie.

V.4. Algorytmy p -średnich

V.4.1. Jak już wspomnieliśmy, algorytmy tej grupy, konkurującej co do popularności z poprzednio omówionymi, oparte są na iteracjach, składających się z dwóch kroków: wyznaczania średnich lub centrów skupień oraz przyporządkowywania im obiektów, czyli faktycznie ustalania kolejnych podziałów. Postępowanie takie kończy się po stwierdzeniu osiągnięcia odpo-

wiedniego ustabilizowania otrzymywanych struktur, bądź ich cyklicznego powtarzania.

V.4.2. Przyjmijmy, bez ograniczania ogólności naszych rozważań, że startujemy od zbioru „reprezentantów” skupień, oznaczonych x_q^a , $q=1, \dots, p$, gdzie p jest z góry założone. Dla tych położeń reprezentantów w przestrzeni \mathbf{X} (zauważmy, że nie muszą oni być obiektami, będącymi elementami X) przypisujemy im wszystkie (pozostałe) obiekty ze zbioru X na podstawie pewnego kryterium. W praktycznie wszystkich przypadkach kryterium tym jest odległość $d(x_q^a, x_i)$. Tak więc, kolejne obiekty o indeksach $i = 1, \dots, n$ przypisywane są temu reprezentantowi x_q^a , dla którego odległość $d(x_q^a, x_i)$ jest najmniejsza. W ten sposób powstaje podział P , składający się z otrzymanych w powyższy sposób skupień A_q , będący prostą konsekwencją przyjętego na początku zbioru $\{x_q^a\}_q$.

Następnie, dla poszczególnych skupień wyznaczamy ich reprezentantów. Dokonujemy tego na podstawie zastosowania pewnego wybranego kryterium, opisanego na podzbiorach obiektów, stanowiących skupienia. Bardzo często kryterium to ma charakter sumy odległości od środka lub wariancji i wobec tego wyznaczony nowy reprezentant, jako odpowiadający minimum takiego kryterium, jest średnią po elementach skupienia.

Łatwo zauważyć, że przy powyższych założeniach, w szczególnym przypadku może się okazać, że otrzymani nowi reprezentanci są tacy sami jak poprzednio. Jeśli tak, to procedura, oczywiście, kończy działanie.

V.4.3. Kluczowymi parametrami tej grupy metod są kryteria przydziału do skupienia i wyznaczania reprezentantów. To one decydują o różnorodności algorytmów w tej grupie. Niezależnie od tego, szerokie zastosowanie znajdują tutaj zbiory rozmyte (skupienia rozmyte), co jest związane – poza uzasadnieniami o charakterze interpretacyjnym – z łatwością numeryczną optymalizacji przy wprowadzeniu skupień w postaci funkcji przynależności $\mu_q(x)$.

V.4.4. Wszystkie jednak metody p -średnich można zinterpretować jako posługujące się (pozornie) globalną funkcją celu o postaci

$$Q(P) = \sum_q W(x_q^a, A_q), \quad (\text{V.3})$$

gdzie $W(x_q^a, A_q)$ jest wspomnianym kryterium, najczęściej o charakterze wariancji, a poszukujemy podziału $P = \{x_q^a, A_q\}_q$, który będzie minimalizował powyższe wyrażenie, spełniając jednocześnie warunki podziału. Standardową postacią $W(x_q^a, A_q)$ jest suma różnic lub odległości w stosunku do reprezentanta, na przykład:

$$W(x_q^a, A_q) = \sum_{i \in A_q} d^2(x_q^a, x_i), \quad (\text{V.4})$$

przy czym odległość w powyższym wzorze jest na ogół tożsama z jedną z podstawowych definicji, wprowadzonych w ramach wykładu, bardzo często jest mianowicie odległością Euklidesową. Niezależnie jednak od przyjętej definicji szczegółowej, zarówno samej odległości $d(x_q^a, x_i)$, jak i wyrażenia $W(x_q^a, A_q)$, łatwo jest zauważyć, że jeśli tylko wartość p , czyli liczba skupień, jest także poddana optymalizacji, to z pewnością optimum (minimum wartości $Q(P)$) jest osiągane dla $p=n$, $q=i$, $x_q^a=x_i$, $A_q=i$. W tym podziale ($P=I$) każdy obiekt jest osobnym skupieniem i zarazem reprezentantem tego skupienia, a dla większości używanych sformułowań $Q(P)$ wartość tej funkcji osiąga zero.

Dlatego też w metodach p -średnich *zakłada się z góry liczbę p* i dla niej dokonuje optymalizacji według zarysowanego algorytmu.

W nawiązaniu do wzoru (V.3) i przykładowej definicji (V.4) zauważmy jeszcze, że zasadnicza procedura metody p -średnich, składająca się z dwóch iteracyjnie powtarzanych kroków, jest w istocie równoważna minimalizacji funkcji (V.3) względem, kolejno, skupień A_q dla zadanych reprezentantów x_q^a , a następnie – względem reprezentantów x_q^a dla zadanych skupień A_q .

V.4.5. Innym problemem wszystkich algorytmów z grupy p -średnich jest ich lokalność, także w obrębie ustalonego p , to znaczy – zależność otrzymywanych rozwiązań od punktów startowych. Wynika ona z istnienia (w ogólności, dla dowolnych zbiorów danych) wielu minimów lokalnych funkcji $Q(P)$, nawet dla zadanego p . Trudność tę omija się w obecnej chwili, dzięki względnie szybkiej zbieżności procedury centrowania i reallokacji (czyli szybkiemu osiągnięciu rozwiązań $\{x_q^a, A_q\}_q$) oraz szybkości działania współczesnych procesorów, poprzez zadawanie wielokrotnie różnych punktów startowych i tworzenie „statystyki” rozwiązań. Te z rozwiązań, które odpowiadają najmniejszym wartościom $Q(P)$ i/lub które powtarzają się wielokrotnie, są akceptowane. Należy podkreślić, że często różnice wartości $Q(P)$ dla różnych rozwiązań są na granicy precyzji prowadzonych obliczeń.

V.4.6. Kwestia zależności od p jest traktowana w tej grupie algorytmów podobnie jak w przypadku algorytmów agregacji hierarchicznej – przez analizę wartości pewnego dodatkowego kryterium, wyliczanych dla optymalnych podziałów, otrzymanych dla kolejnych wartości p . Uwagi sformułowane w punkcie V.3.4 znajdują także i tutaj zastosowanie.

V.4.7. Bardzo ważną cechą algorytmów p -średnich jest możliwość ich wykorzystania, w sposób zupełnie naturalny, jako algorytmów klasyfikujących. Jeśli mianowicie w pewnym momencie mamy do czynienia z dodatkowymi obiektami, spoza początkowego zbioru X , to możemy je włączyć do procedury w trakcie jej działania na etapie przypisywania obiektów x_i do reprezentantów x_q^a i tworzenia w ten sposób następnego podziału $P=\{A_q\}$, jako obiekty o indeksach $i>n$. Możemy to również uczynić po zakończeniu

działania dla danego zbioru obiektów. Musimy przy tym, oczywiście, założyć, że zasadnicze warunki rozwiązywanego zadania się nie zmieniają (np. liczba skupień p).

V.4.8. Jak wspominaliśmy, reprezentantami skupień nie muszą wcale być średnie wartości w skupieniu, bądź obiekty ze skupienia, najbliższe średnim, jakkolwiek te dwie sytuacje spotyka się najczęściej:

$$x_q^a = \frac{1}{cardA_q} \sum_{i \in A_q} x_i, \quad (V.5)$$

czyli

$$x_{qk}^a = \frac{1}{cardA_q} \sum_{i \in A_q} x_{ik}, \quad (V.5')$$

bądź też, ponieważ definicja (V.5) w ogólności może wyprowadzić x_q^a poza zbiór X , a nawet poza przyjętą przestrzeń \mathbf{X} ,

$$x_q^a = \arg \min_{x \in X} d(x, \frac{1}{cardA_q} \sum_{i \in A_q} x_i). \quad (V.6)$$

Można, oczywiście, posługiwać się innymi jeszcze definicjami, analogicznymi do powyższych. Jednakże metoda p -średnich stwarza również możliwość odwoływania się do zupełnie innego rodzaju reprezentacji, na przykład, modeli regresji względem określonych zmiennych, otrzymanych dla poszczególnych skupień (por. Wykład VI).

W tym miejscu wspomniemy jednak o innej ważnej możliwości, która nie wynika wprost z definicji (V.5) i (V.6). Łatwo mianowicie zauważyć, że jeśli mamy do czynienia ze zmiennymi nominalnymi, to (V.5') nie może być stosowane. Nie możemy bowiem z zasady wyciągać średnich z „wartości” zmiennych nominalnych. Na przykład, jeśli 26% obiektów w skupieniu to kobiety, a 74% to mężczyźni, to zastosowanie (V.5') jest w ogóle niemożliwe, podczas gdy zastosowanie (V.6) względem $x \in \mathbf{X}$ da w wyniku stwierdzenie, że reprezentantem jest mężczyzna (o cechach określonych przez pozostałe zmienne), co nie zawsze będzie satysfakcjonującym wynikiem. W dodatku, w sytuacjach, gdy mamy do czynienia z wieloma zmiennymi nominalnymi, a określenia przynajmniej niektórych z nich mają charakter subiektywny („brunet”, „przystojny”,...), brak satysfakcji z reprezentowania skupień przez (V.6) może się pogłębiać. Ewentualnym wyjściem jest „histogram (częstościowy) skupienia” (He i in., 2003, Huang, 1997, 1998). Oznaczmy mianowicie przez $n_q(kl)$ liczbę obiektów należących do skupienia q , które przyjmują dla zmiennej k wartość l . Oczywiście,

$$\sum_l n_q(kl) = \text{card } A_q \quad \forall k.$$

Wspomniany histogram skupienia jest zbiorem wektorów wartości $\{n_q(kl)/\text{card } A_q\}_l$ dla wszystkich zmiennych $k=1, \dots, m$, czyli częstości występowania poszczególnych wartości l w skupieniu. Tak więc x_q^a nie jest już wektorem tylko m wartości dla poszczególnych zmiennych, ale zbiorem wektorów o zróżnicowanej długości (liczby wartości dla poszczególnych zmiennych są w ogólności różne). Tym niemniej, cały opis metody p -średnich pozostaje w mocy, pod warunkiem odpowiedniego zdefiniowania odległości występującej w (V.4), co jednak nie przedstawia żadnych trudności.

Zauważmy, że dla przytaczanego poprzednio przykładu zmiennej binarnej („kobieta”-„mężczyzna”) metoda p -średnich z histogramem częstościowym skupienia nie daje żadnych korzyści (mamy tylko dwie możliwości do wyboru i przypisanie do skupień jest także binarne). Korzyści z zastosowania tej metody są jednak zauważalne dla zmiennych nominalnych o większej liczbie wartości.

V.4.9. Poświęćmy obecnie nieco uwagi zastosowaniu zbiorów rozmytych. Zbiory rozmyte, o których wspomnieliśmy, że znajdują dość naturalne zastosowanie w analizie skupień, szczególnie dobrze „pasują” do metody p -średnich.

Przypomnijmy, że oznaczenie $\mu_A(x)$ odnosi się do stopnia przynależności elementu (obiektu) x do pewnego zbioru (rozmytego) A , przy czym stopień przynależności $\in [0,1]$. Tak więc, jeśli nasz podział jest rozmyty (P_μ), to znaczy, że składa się ze skupień rozmytych, możemy powyższe oznaczenie uprościć do $\mu_q(i)$, czyli przynależności obiektów o indeksach $i \in I$ do skupień o indeksach q , $q=1, \dots, p$.

Wspominaliśmy już o podstawowym warunku sensowności podziału rozmytego, czyli

$$\sum_{q=1}^p \mu_q(i) = 1 \quad \forall i \in I, \quad (\text{V.7})$$

oznaczającym, że każdy obiekt i musi być dokładnie „cały” „rozdzielony” między skupienia rozmyte o indeksach q . Ponieważ poszukiwane wartości $\mu_q(i)$ będą wyznaczone przy pomocy metod automatycznych, musimy w sposób jawny określić i inne ograniczenia na nie, a mianowicie:

$$\mu_q(i) \in [0,1] \quad \forall q, i \quad (\text{V.8})$$

oraz

$$0 < \sum_{i=1}^n \mu_q(i) < n. \quad (\text{V.9})$$

Warunek (V.7) jest definicyjnym warunkiem zbiorów rozmytych, natomiast warunek (V.9) ustala, że nie zajmujemy się zbiorami pustymi (lewa nierówność), ani też zbiorami obejmującymi wszystkie obiekty w całości (prawa nierówność), tj. $A_q = I$.

Procedura optymalizacji jest dokładnie taka sama, jak w przypadku klasycznych skupień A_q tworzących nierozmyty podział P . Kolejno zatem znajdujemy reprezentantów x_q^a i odpowiadające im skupienia (tutaj rozmyte) A_q , aż do spełnienia pewnego warunku stopu. Z jedną przecież bardzo ważną różnicą: dzięki nieco zmodyfikowanej dla tej wersji metody postaci funkcji celu, a mianowicie

$$Q(P_\mu) = \sum_{q=1}^p \sum_{i=1}^n (\mu_q(i))^\alpha d(x_i, x_q^a), \quad (\text{V.10})$$

w której α jest pewnym (dobieranym przez użytkownika) parametrem, na ogół $\alpha \in [1, \infty)$, oraz właściwościom tej funkcji, zwłaszcza zaś jej różniczkowalności, możliwe jest uzyskanie jawnych, analitycznych postaci rozwiązań dla zadań minimalizacji względem x_q^a oraz A_q (por. ostatni akapit punktu V.4.4). I tak, dla $\alpha > 1$ i ustalonego x_q^a optymalnym rozwiązaniem względem A_q , czyli, w tym przypadku, wartości $\mu_q(i)$, jest

$$\mu_q(i) = \left[\sum_{q'=1}^p (d(x_i, x_{q'}^a) / d(x_i, x_q^a))^{1/(\alpha-1)} \right]^{-1}. \quad (\text{V.11})$$

Łatwo zauważyć, dlaczego założenie o $\alpha > 1$ było konieczne. Niezależnie jednak od czysto arytmetycznych powodów zaznaczmy, że dla $\alpha = 1$ otrzymuje się również bezpośrednio rozwiązanie, które jest po prostu identyczne z procedurą minimalizacji odległości od reprezentanta dla standardowych skupień (i podziałów) nierozmytych.

Jednocześnie, dla zadanych wartości $\mu_q(i)$, otrzymujemy w sposób jawny optymalnego (w istocie „średniego”, w sensie średniej ważonej, jako optimum analogicznego do średniej, minimalizującego wyrażenie na wariancję) reprezentanta:

$$x_q^a = \left(\sum_{i=1}^n (\mu_q(i))^\alpha x_i \right) / \sum_{i=1}^n \mu_q(i). \quad (\text{V.12})$$

Zależności (V.11) i (V.12) definiują klasyczną procedurę rozmytej wersji metody p -średnich, którą tutaj podajemy za pracą Bezdek i in. (1986). Jakkolwiek istnieją inne jeszcze wersje tej metody z zastosowaniem zbiorów rozmytych, podana tutaj jest podstawową, najprostszą i najczęściej używaną.

Nietrudno się przy tej okazji zorientować, że zastosowanie zbiorów rozmytych, którego pierwotną motywacją była „niepewność” co do przynależności obiektów do poszczególnych skupień, okazało się dodatkowo (a być może i zasadniczo) uzasadnione prostotą algorytmiczną (jawne wzory zamiast przeglądu zbioru danych).

V.4.10. Metoda p -średnich jest prawie równie popularnym elementem różnych pakietów analizy danych, co algorytmy agregacji hierarchicznej. Istniejące aplikacje najczęściej same, automatycznie, dobierają najlepsze p na podstawie odpowiednich statystyk. Z drugiej strony, punkty startowe są ustalane przy pomocy prostej wstępnej analizy zbioru danych, często z zastosowaniem uproszczonych algorytmów agregacji hierarchicznej, bądź (częściej) podziału przestrzeni. Przy stosowaniu takich aplikacji należy zwrócić szczególną uwagę właśnie na otrzymywane liczby skupień oraz na wrażliwość rozwiązań względem tego i innych ewentualnie dostępnych założeń (na przykład definicji odległości).

V.4.11. Nakład obliczeniowy algorytmów agregacji hierarchicznej oraz p -średnich najłatwiej ocenić na podstawie potrzebnej do obliczeń wielkości pamięci, ponieważ wyznacza ona także i liczbę wykonywanych operacji arytmetycznych (w zawartych dalej kilku uwagach nie uwzględniamy wielu możliwości uproszczeń algorytmicznych bądź algorytmów specjalizowanych). Ponieważ algorytmy agregacji hierarchicznej polegają na porównywaniu odległości między obiektami, a następnie ich modyfikacji i porównywaniu jako odległości między skupieniami, więc wymagają one pamięci o objętości $O(n^2)$, jako, że liczba odległości między n obiektami wynosi $n(n-1)/2$. Liczba operacji może zatem sięgać $O(n^3)$ (otrzymanie całości drzewa hierarchii wymaga wykonania n kroków procedury). Jeśli natomiast idzie o algorytmy p -średnich (w wersji nierozmytej), to poprzestają one na macierzy X , a w każdym kroku wyznaczanych jest n odległości (od reprezentantów). Algorytmy te są na ogół szybko zbieżne (często wystarczy kilka lub kilkanaście iteracji), co powoduje, że dla dużych n mają istotną przewagę numeryczną nad algorytmami agregacji hierarchicznej. Problemem, jak już zauważyliśmy, jest konieczność rozpoczynania procedury z różnych punktów startowych, co może wspomnianą przewagę znacznie zmniejszyć. Pozostaje jednak korzyść związana z wielkością pamięci.

V.5. Metoda z globalną funkcją celu

V.5.1. Obie zaprezentowane tutaj grupy algorytmów oparte były na niezwykle prostych i intuicyjnie zrozumiałych zasadach. Poza tym, wykazują one cały szereg innych cech pozytywnych, na przykład – zwłaszcza w odniesieniu do szybkości działania oraz pewnych własności teoretycznych (na przykład odtwarzanie przez algorytmy agregacji hierarchicznej optymalnych struktur grafowych opartych na zadanym zbiorze obiektów, bądź własności podziału przestrzeni X i reguły klasyfikacji w przypadku algorytmów p -średnich).

V.5.2. Jednakże obie te grupy metod mają jedną zasadniczą wadę: nie dostarczają wprost rozwiązania (nawet przybliżonego) zasadniczego zadania analizy skupień. Do jego otrzymania potrzebne jest zastosowanie dodatkowych kryteriów, pozwalających na wybór pomiędzy różnymi rozwiązaniami, generowanymi w sposób jawny lub nie przez te metody, głównie zależnymi od p .

V.5.3. Dlatego też od dość wczesnego etapu rozwoju analizy skupień – od końca lat sześćdziesiątych – poszukiwano metody (a zatem i modelu) analizy skupień, pozwalającej na rozwiązanie jej zadania w całości. Poszukiwania te przybierały w zasadzie wyłącznie postać formułowania odpowiednich postaci funkcji celu. Zaznaczmy przy tym, że znakomita większość formułowanych funkcji celu i odpowiadających im zadań, na ogół o charakterze zadań programowania matematycznego, nadal była uzależniona od wartości p .

V.5.4. Przedstawimy tutaj w skrócie jedno z niewielu konsekwentnych podejść, dających w wyniku pełne rozwiązanie zadania analizy skupień. Podejście to zostało opracowane przez autora niniejszego wykładu (Owsinski, 1991).

V.5.5. Wprowadzamy ogólną funkcję celu zadania w postaci:

$$Q^*(P) = Q^D(P) + Q_s(P), \quad (V.13)$$

w której $Q^D(P)$ jest funkcją, opisującą (wszystkie) odległości pomiędzy skupieniami (obiektami w różnych skupieniach), zaś $Q_s(P)$ jest funkcją, opisującą (wszystkie) bliskości wewnątrz skupień (bliskości pomiędzy obiektami w tych samych skupieniach). Funkcję $Q^*(P)$, oczywiście, maksymalizujemy. Analogiczną („dualną”) funkcją celu jest

$$Q_*(P) = Q_D(P) + Q^S(P), \quad (V.14)$$

w której, z kolei, $Q_D(P)$ wyraża odległości obiektów należących do tych samych skupień, podczas gdy $Q^S(P)$ – bliskości obiektów należących do różnych skupień. Zauważmy, że $Q_D(P)$ może być po prostu wyrażeniem,

wykorzystywanym jako funkcja celu w metodach p -średnich. Naturalnie, $Q^*(P)$ jest minimalizowana.

V.5.6. Tak sformułowana globalna funkcja celu – globalna, ponieważ przy odpowiednich, sensownych postaciach jej dwu składników odwzorowujemy w sposób właściwy podstawowe sformułowanie zadania analizy skupień i uzyskujemy niezależność od liczby skupień, p – może w szczegółach być bardzo różnie formułowana. Pozwala to na odzwierciedlenie dużego wachlarza sytuacji merytorycznych, odpowiadających różnym postaciom składników funkcji globalnej.

Przykładową postacią, dającą duże możliwości algorytmiczne (por. następne punkty) jest, bodaj najprostsza:

$$Q^D(P) = \sum_{q=1}^{p-1} \sum_{q'=q+1}^p \sum_{i \in A_q} \sum_{j \in A_{q'}} d_{ij}, \quad (\text{V.15})$$

będąca sumą wszystkich odległości między obiektami położonymi w różnych skupieniach, wraz z

$$Q_S(P) = \sum_{q=1}^p \sum_{i \in A_q} \sum_{j \in A_q, j>i} s_{ij}, \quad (\text{V.16})$$

czyli sumą wszystkich bliskości między obiektami położonymi w tych samych skupieniach.

Jak widać, w sformułowaniu tym posługujemy się jednocześnie odległościami i bliskościami, zgodnie zresztą z dość wyraźną „filozofią” metody, widoczną w ogólnych wzorach (V.13) i (V.14). Pozostaje zatem odpowiednie zdefiniowanie przejścia między odległością a bliskością i vice versa, o czym wspominaliśmy przy omawianiu tych pojęć.

V.5.7. Zagadnieniem zasadniczym pozostaje jednak sposób optymalizacji (minimalizacji, maksymalizacji) tych różnych konkretnych postaci funkcji $Q^*(P)$ lub $Q^*(P)$. Faktycznie, zagadnienie to było jedną z zasadniczych przyczyn porażek większości propozycji istotnie globalnych funkcji celu: po prostu nie istniały odpowiadające im efektywne algorytmy optymalizacji. W większości przypadków algorytmy okazywały się przynajmniej potęgowe względem n , o wysokim wykładniku potęgi, jeśli nie wykładnicze. A jednocześnie – algorytmy agregacji hierarchicznej oraz p -średnich charakteryzują się prostotą i skończoną lub szybką zbieżnością.

V.5.8. W szczególności, najmocniejszym przykładem jest bodaj następujące sformułowanie zadania analizy skupień:

$$Q(P) = \sum_{i,j} (y_{ij} d_{ij} + (1-y_{ij})(M-d_{ij})) \rightarrow \min \quad (\text{V.17})$$

gdzie y_{ij} są zmiennymi definiującymi podział, spełniającymi warunki podane w punkcie II.3.5 wykładu (wzory II.22), zaś M jest pewną liczbą, decydującą o skali funkcji celu (dla normalizowanych wartości odległości będzie to 1). Łatwo zauważyć, że mamy tu do czynienia z zadaniem programowania liniowego o ograniczeniach określonych przez (II.22), dla którego istnieją efektywne algorytmy i pakiety programowe o wysokiej sprawności obliczeniowej. Zarazem jednak można obliczyć, że musimy w tym zadaniu uwzględnić $n(n-1)(n-2)/6$ ograniczeń określonych wzorami (II.22), co oznacza, na przykład, dla 500 obiektów (zadanie „w dolnej strefie stanów średnich”), liczbę ponad 20 milionów ograniczeń. Bez algorytmów specjalizowanych i procesorów o najwyższych wydajnościach rozwiązywanie tego zadania jest niemożliwe.

V.5.9. Jeśli składniki $Q_*(P)$ i $Q^*(P)$ spełniają pewne warunki, a w szczególności – są monotoniczne względem operacji agregacji skupień, i to przeciwnie monotoniczne (na przykład: jeden składnik rośnie wraz z agregacją skupień, podczas gdy drugi maleje, co jest spełniane przez sformułowanie oparte na wzorach (V.15, 16)), to możliwe jest, mimo wszystko, zaprojektowanie prostego algorytmu (sub)optimalizacji zaproponowanej tutaj funkcji globalnej. Algorytm ten odwołuje się, mianowicie, do różnic wartości $Q_*(P)$ lub $Q^*(P)$ wynikających z operacji agregacji.

Prześledźmy ten proces na przykładzie $Q^*(P)$. Wprowadźmy mnożnik $r \in [0,1]$ i zmieńmy nieco definicję $Q^*(P)$:

$$Q^*(P,r) = r Q^D(P) + (1-r)Q_S(P) \quad (V.18)$$

i załóżmy, że rozpoczynamy działanie algorytmu od $r=1$. Mamy wówczas

$$Q^*(P,1) = Q^D(P).$$

Jeśli tak, to oczywiście, ze względu na własności $Q^D(P)$ (przede wszystkim monotoniczność względem p), optimum dla $Q^*(P,1)$ jest równoznaczne z podziałem $P=I$. Przy agregacji skupień (najpierw, oczywiście, obiektów) mamy do czynienia z monotonicznym zachowaniem obu składników optymalizowanej funkcji celu. Pozwala nam to na wybór tej konkretnej agregacji (pary skupień), dla której przyrost jest najszybszy (a więc zmiana – zmniejszenie – odpowiadającej mu wartości r najmniejsza), zgodnie z warunkiem agregacji dla podziału P :

$$r Q^D(P) + (1-r)Q_S(P) = r (Q^D(P)-\Delta^D) + (1-r)(Q_S(P)+\Delta_S), \quad (V.19)$$

(gdzie Δ^D i Δ_S oznaczają odpowiednie przyrosty składników funkcji celu w stosunku do ich wartości dla P) skąd, po prostej redukcji, dostajemy zależność

$$-r\Delta^D + (1-r)\Delta_S = 0 \Rightarrow r = \Delta_S / (\Delta_S + \Delta^D), \quad (V.20)$$

z której możemy wyznaczyć kolejne, następne wartości parametru r , odpowiadające kolejnym lokalnie optymalnym wartościom $Q^*(P)$ i postaciom P , wynikającym z kolejnych agregacji. Jak się łatwo zorientować, rozwiązanie (sub)optimalne otrzymywane jest, kiedy wartość r przejdzie przez $\frac{1}{2}$. Dalsze agregacje, dla r poniżej $\frac{1}{2}$, prowadzą już do nieoptymalnych rozwiązań (agregacji skupień zbyt odległych od siebie).

V.5.10. Przedstawiona w zarysie procedura łączy zalety globalności, względem zarówno zawartości A_q , jak i ich liczności p , z prostotą algorytmu, który polega na łączeniu skupień na zasadach podobnych do algorytmów agregacji hierarchicznej. Otrzymane rozwiązanie jest co prawda suboptymalne, ale dysponowanie globalną funkcją celu pozwala na ewentualne poszukiwanie innych rozwiązań, poza otrzymanymi z procedury, a dającymi potencjalnie lepsze wartości funkcji celu. Mogą temu służyć, w szczególności, proste algorytmy oparte na idei wymiany obiektów między skupieniami.

V.6. Zastosowania w eksploracji danych

V.6.1. Dziedzina eksploracji danych („drążenia” danych, ang. *data mining*) jest wynikiem potrzeby (ale zarazem i możliwości) analizowania bardzo wielkich zbiorów danych, będących wynikiem rozwoju informatyki, a w szczególności – internetu. Automatycznie zbierane dane, na przykład przez operatorów telefonicznych, bądź internetowych (logi), są gromadzone w zbiorach o licznosciach sięgających miliardów obserwacji (obiektów). Inną cechą tych danych jest, zazwyczaj, niemożność przedstawienia ich w postaci „porządnej” macierzy X (inherentnie zmienna długość odpowiednich rekordów, braki w danych), choćby ze względu na rozmiary tych zbiorów.

V.6.2. Celem eksploracji danych jest uzyskanie wiedzy na podstawie tak wielkich zbiorów danych, przy czym z góry zakładamy, że wiedza ta będzie miała charakter przybliżony.

V.6.3. Jest rzeczą oczywistą, że w takich sytuacjach stosować można tylko metody niezwykle uproszczone. I tak, eksploracja danych posługuje się w istocie wyłącznie uproszczonymi metodami analizy skupień. W stosunku do metod, jakie mogłyby znaleźć (i faktycznie znajdują) zastosowanie w eksploracji danych, można z góry sformułować szereg założeń, jakie powinny one spełniać:

- tylko jeden, najwyżej dwa (a w każdym razie bardzo ograniczona liczba), przeglądów całego zbioru obiektów,
- posługiwanie się położeniami obiektów, x_i , raczej niż ich (jawnymi) odległościami lub bliskościami, a jeśli odległości i bliskości są faktycz-

nie wyznaczane, to tylko na bieżącym etapie algorytmu, bez przechowywania (w żadnym przypadku) macierzy odległości lub bliskości,

- bezpośrednia analiza dokonywana sekwencyjnie (obiekt po obiekcie), albo co najwyżej w odniesieniu do stosunkowo niewielkich podzbiorów rozpatrywanego zbioru,
- możliwość – ograniczona – zastosowania modelu klasyfikacyjnego, aż do momentu wykrycia znacznych odstępstw od tego modelu.

V6.4. Można się łatwo zorientować, że ten zestaw założeń wyklucza wykorzystanie globalnych funkcji celu i orientuje w kierunku zastosowania algorytmów lokalnych, to znaczy przede wszystkim algorytmów agregacji hierarchicznej, podziału przestrzeni oraz algorytmów typu p -średnich. W istocie, metody (heurystyki) eksploracji danych posługują się szeregiem prostych zabiegów znanych z dziedziny analizy skupień, w tym zwłaszcza z zakresu algorytmów agregacji hierarchicznej (zasada najbliższego sąsiada) oraz podziału przestrzeni (nie rozróżnianie obiektów, które spełniają pewne warunki bliskości, i poszukiwanie rozwiązania w postaci takich właśnie podzbiorów nierozróżnialnych obiektów). Popularnością cieszą się również algorytmy pochodzące z grupy metod p -średnich, zwłaszcza wobec faktu, że posługują się one praktycznie wyłącznie położeniami obiektów, a nie odległościami między nimi (na każdym kroku wyliczane są tylko odległości wszystkich obiektów od odpowiadających im reprezentantów skupień), oraz możliwości stosowania ich w postaci algorytmów klasyfikacyjnych.

V.7. Przypadek jednowymiarowy

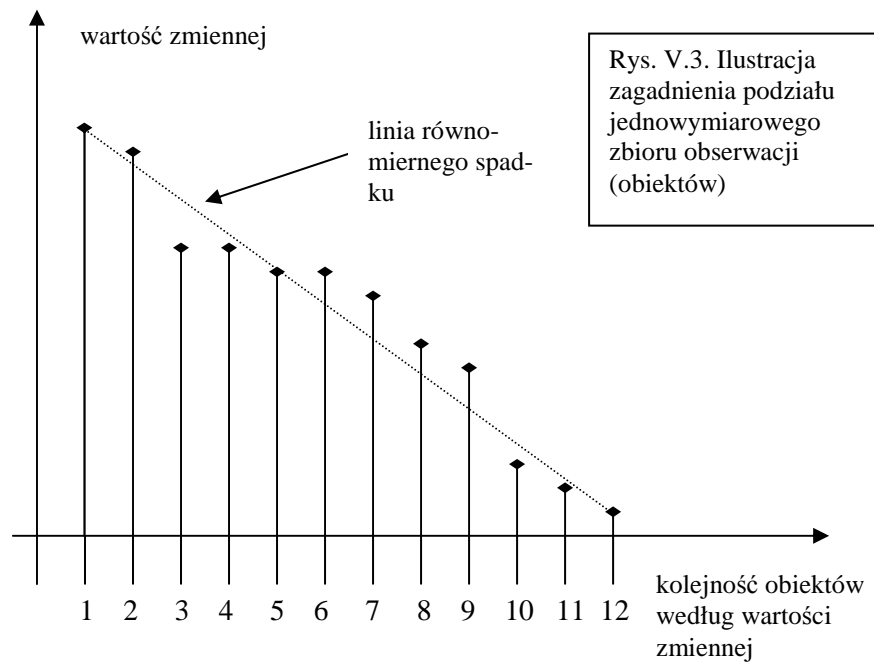
V.7.1. Skomentujemy obecnie pokrótce przypadek szczególny, w którym dane są obiekty scharakteryzowane tylko jedną zmienną. Zajmiemy się nim, dość zresztą pobieżnie, z dwóch przyczyn: po pierwsze, wiele aplikacji typu arkuszy kalkulacyjnych zawiera (niekiedy wyłącznie, w sensie algorytmów, które można by zaliczyć do analizy skupień) taką właśnie możliwość, a więc podziału na „segmenty” pewnego ciągu uporządkowanych wartości, i po drugie – jest to w istocie przypadek szczególny, także z punktu widzenia stosowanych do niego metod.

V.7.2. Takim standardowym zastosowaniem zadania podziału ciągu wartości na „segmenty” jest wspomniany powyżej przypadek podziału pewnego uporządkowania bądź rankingu. I tak, mając, na przykład, przedsiębiorstwa uporządkowane „od największego do najmniejszego” chcielibyśmy w jakiś łatwy (automatyczny) i zrozumiały sposób podzielić je na kategorie, odpowiadające, powiedzmy, pojęciom „największe”, „bardzo duże”, „duże”, ...,

itp., na której to podstawie moglibyśmy, powiedzmy, do dalszej analizy zdefiniować odpowiednie zbiory rozmyte odpowiadające tym pojęciom.

V.7.3. W rozważanym przypadku znacznie większe zastosowanie niż na ogół w dziedzinie analizy skupień znajdują klasyczne metody statystyki matematycznej, co jest prostą konsekwencją faktu, że mamy do czynienia tylko z jedną zmienną, a zatem na pewno analizowane zmienne są jednorodne. W istocie, najefektywniejsze metody podziału jednowymiarowego oparte są na modelach statystyki, a w szczególności – na założeniach i wynikach dotyczących rozkładów prawdopodobieństwa wartości danej zmiennej w populacji odpowiadającej badanemu zbiorowi obiektów (który albo jest całością tej populacji, albo próbą z niej, niekoniecznie o charakterze losowym).

V.7.4. Najprostszym postępowaniem w rozważanym przypadku wydaje się być na pierwszy rzut oka podział według odstępów między wartościami. Zilustrujemy to przykładem, pokazanym na Rys.V.3.



Mamy $n=12$, a ponieważ jest tylko jedna zmienna, wartości jej dla poszczególnych obiektów oznaczmy x_i , $i=1,...,12$. Założymy przy tym, że numeracja obiektów odpowiada ich kolejności w uporządkowaniu (czyli $x_i \geq x_{i+1}$),

zgodnie z rysunkiem. Nie zmniejszamy to ogólności naszych rozważań, ponieważ zawsze możemy zmienić kolejność indeksów obiektów w zbiorze I .

Wprowadźmy obecnie ciąg wartości różnic $\Delta x_i = x_i - x_{i+1}$, $i=1, \dots, n-1$. Możemy procedurę podziału zbioru $I = \{1, \dots, 12\}$ oprzeć na tych wartościach. W pierwszym kroku znajdujemy $\max_i \Delta x_i$ i dokonujemy pierwszego podziału zbioru I pomiędzy $\arg \max_i \Delta x_i$ oraz $\arg \max_i \Delta x_i + 1$, co odpowiada podziałowi w miejscu, gdzie różnica między kolejnymi wartościami zmiennej dla kolejnych obiektów jest największa (na Rys. V.3 – powiedzmy, między $i=2$ a $i=3$, co daje po pierwszym kroku podział na dwa skupienia: $\{1, 2\}$ i $\{3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$). W drugim kroku szukamy największej różnicy w obrębie istniejących skupień (na Rys. V.3, powiedzmy, Δx_9), itd.

W ten sposób realizujemy procedurę podziału hierarchicznego (wspomnieliśmy już o istnieniu takich procedur), opartą na różnicach między kolejnymi obiektami w uporządkowaniu. Możemy ją, analogicznie (choć w odwrotnym kierunku) jak procedury agregacji hierarchicznej, doprowadzić aż „do końca”, to znaczy do otrzymania skupień będących pojedynczymi obiektami. Podobnie jak i w tamtym przypadku powstanie wówczas problem ustalenia, w którym miejscu tej procedury otrzymujemy „właściwe” rozwiązanie.

Dość prostą i intuicyjnie narzucającą się odpowiedź może być porównanie z pokazaną na Rys. V.3 „linią równomiernego spadku”, czyli prostą poprowadzoną przez punkty x_1 oraz x_n (tutaj: x_{12}). Łatwo zauważyć, że jeśli różnica Δx_i jest większa niż wynikająca z „równomiernego spadku” (odpowiednie linie się przecinają), to istnieje podejrzenie, że w tym miejscu należałoby dokonać podziału.

V.7.5. Zarysowana procedura jest w istocie dość ogólna i może zostać zaakceptowana, z jednym wszakże bardzo ważnym warunkiem: że jako „linię równomiernego spadku” będziemy używali linii wynikającej z naszej wiedzy dotyczącej faktycznego rozkładu wielkości obserwowanej zmiennej w badanej populacji (w szczególności może to być także owa prosta równomiernego spadku, przytoczona w przykładzie). Zauważmy bowiem, że bardzo często mamy do czynienia z rozkładami, w których wielkości skrajne (bądź wielkości na jednym ze skrajów, na przykład największe) występują rzadziej niż pozostałe, i to w sposób bardzo regularny (na przykład rozkład normalny, rozkład Poissona, itp.). Jeśli mamy zatem wiedzę, pozwalającą, przynajmniej w przybliżeniu, założyć sensownie rodzaj rozkładu, a zatem i odpowiadającą mu linię spadku wartości, to linia taka może służyć za podstawę do dokonywania podziałów w ramach procedury podobnej do poprzednio zarysowanej.

Uważaj Czytelnik zorientował się już jednak zapewne, że tego rodzaju postępowanie niesie ze sobą ryzyko wewnętrznej sprzeczności: jeśli jest praw-

dą, że w danej populacji rozkład wartości zmiennej ma założony charakter, to jaki sens mają wprowadzone na jego podstawie podziały, a zatem i skupienia? Jest to zagadnienie dokładnie analogiczne do przytaczanego już swojego czasu przy omawianiu stosunku metod analizy danych do metod statystyki matematycznej i rachunku prawdopodobieństwa.

Z powyższych powodów stosuje się na ogół w zadaniach podziału jednowymiarowego metody najprostsze, podobne do zilustrowanej przy pomocy Rys. V.3, a więc oparte na porównywaniu różnic. Podkreślimy jednak, że często są one odpowiednio modyfikowane ze względu na specyficzne cechy konkretnych zadań.

Posługując się algorytmami podziału uporządkowań w ramach różnych pakietów oprogramowania zwróćmy baczniejszą uwagę na założenia, jakie są przy tym czynione i czy nasz przypadek w rzeczywistości spełnia te założenia.

V.7.6. Na zakończenie jeszcze jedna uwaga: odwołał się tutaj do algorytmu podziału hierarchicznego, analogicznego w pewnym sensie do agregacji hierarchicznej, lecz niejako „odwrotnego”. W istocie, zastosowanie algorytmu podziału hierarchicznego do tego zadania jest niejako naturalne, znacznie bardziej odpowiednie niż w ogólnym przypadku wielowymiarowego zadania analizy skupień.

V.8. Przykład zastosowania

V.8.1. Zarysujemy obecnie przykład konkretnego zastosowania metod analizy skupień, taki, w którym zastosowanie to dało ważne i dość zaskakujące wyniki. Ze względu na ograniczoną objętość wykładu ograniczymy się do dość pobieżnego zarysu tego projektu badawczego i jego wyników.

V.8.2. W latach 1980-1984 prowadzono w Instytucie Badań Systemowych PAN prace dotyczące Bełchatowskiego Okręgu Przemysłowego (BOP). Prace te koncentrowały się na zagadnieniach rolnictwa w obszarze oddziaływania BOP. Chodziło głównie o skutki rolnicze wielkiej odkrywki węgla brunatnego („największa dziura w ziemi w Europie” i związany z nią tzw. lej depresyjny, czyli obniżenie poziomu wód gruntowych, oraz zanik niektórych wód powierzchniowych, również cieków), a także elektrowni (zanieczyszczenia powietrza). Prowadzono, między innymi (projekt obejmował także model optymalizacyjny rolnictwa regionalnego), analizę dostępnych danych dotyczących rolnictwa na poziomach gospodarstw, wsi i gmin.

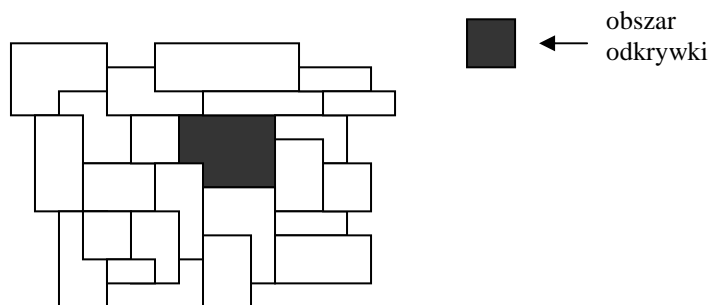
Ostatecznym celem miało być określenie wielkości i charakteru inwestycji niezbędnych dla utrzymania produktywności i opłacalności działalności rolniczej na tych terenach (jednym ze zlecniodawców był Wojewódzki

Zarząd Inwestycji Rolniczych w Piotrkowie Trybunalskim). Analiza danych miała zatem z jednej strony dostarczyć podstaw do budowy odpowiedniego modelu, a z drugiej – pozwolić na sprawdzenie pewnych zasadniczych hipotez, dotyczących charakteru zachodzących procesów (por., np. Owsński i Hołubowicz, 1986, Owsński, 1987).

V.8.3. W szczególności, analizowano dane o poszczególnych wsiach (obiek-
tach) w okolicy odkrywki. Dla każdej wsi (w zależności od wariantu badania – od $n=18$ do 31 wsi-obiektów) dysponowano $m=8-15$ zmiennymi, takimi jak, na przykład: średnia bonitacja (jakość) gleb ornych, średnia jakość użytków zielonych (łąk i pastwisk), udział powierzchni użytków zielonych, plony z hektara wybranych roślin uprawnych, obsada bydła, liczba traktorów na 1 hektar, itp.

Jedną z hipotez roboczych, wynikających z badań agrotechnicznych i ekonomicznych, było istnienie określonych strat na obszarach użytków zielonych, wynikające z wpływu leja depresyjnego, przy jednoczesnym braku strat na gruntach ornych, na których uprawiano rośliny znacznie mniej wrażliwe na dostępność wody gruntowej niż łąki i pastwiska.

Niezależnie od wspomnianych poprzednio zmiennych rolniczych, każda wieś była scharakteryzowana położeniem względem odkrywki i wzajem względem siebie. Znacznie uproszczony schemat przestrzenny układu wsi pokazano na Rys.V.4.

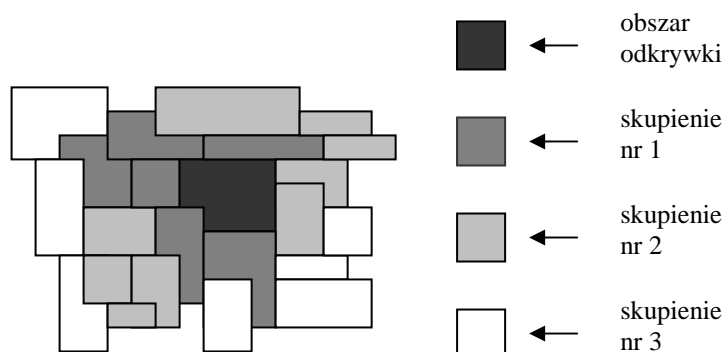


Rys.V.4. Uproszczony schemat przestrzenny układu wsi na obszarze oddziaływania BOP.

V.8.4. Przeprowadzono szereg przeliczeń analizy skupień przy pomocy metody autora wykładu, opisaną w punkcie V.5. W początkowych przeliczeniach nie uwzględniano w ogóle jako zmiennej (ani odległości) położenia wsi względem odkrywki. Tym niemniej otrzymywano konsekwentnie obraz o znaczącej strukturze przestrzennej, podobny do pokazanego przykładowo

na Rys. V.5 (otrzymywano za każdym razem 3-4 skupienia, najczęściej, jak to pokazano na Rys. V.5 – trzy).

Wysnuto przypuszczenie, że ten wyraźny układ przestrzenny z odkrywką jako jego ośrodkiem wynika z wpływu strat ponoszonych na użytkach zielonych. Przeprowadzono wobec tego dalsze przeliczenia, w których wpływ stanu użytków zielonych był zminimalizowany. Jednak i te przeliczenia dały w wyniku podobną strukturę przestrzenną wsi.



Rys.V.5. Typowy układ przestrzenny wsi w obszarze oddziaływania BOP otrzymany w wyniku analizy skupień (bez uwzględniania położenia jako zmiennej).

V.8.5. Trzeba było wobec tego uznać, że zjawisko to ma inny charakter, niż wynikający ze wspomnianej hipotezy. Przeprowadzono dodatkowe badania, które doprowadziły do ustalenia przyczyny przestrzennej regularności, pojawiającej się w wynikach analizy skupień. Była nią chęć (niektórych) rolników uzyskania odszkodowań za straty spowodowane przez odkrywkę, prowadząca do fałszowania niektórych danych, zwłaszcza dotyczących plonów roślin polowych. Ponieważ było jasne, że szansa uzyskania odszkodowań rośnie w miarę zbliżania się do kopalni i zwiększania się faktycznie odnotowywanych strat w wyniku oddziaływania leja depresyjnego, zafałszowania te również wzrastały w pobliżu odkrywki.

V.8.6. W ten sposób, przy pomocy analizy skupień, wspartej dodatkowymi badaniami, uzyskano wiedzę dotyczącą zachowań (części) rolników w pobliżu kopalni odkrywkowej. Wiedza ta została uzyskana niejako „po drodze”, w wyniku postępowania nakierowanego na inny cel. Jakkolwiek mu-

Jan W. Owsinski

siała ona zostać empirycznie potwierdzona, można powiedzieć, że jej uzyskanie odbyło się praktycznie bez kosztów.

Pamiętajmy jednak, że bardzo często analiza skupień nie tylko nie prowadzi do tak spektakularnych wyników, ale w ogóle nie daje podstaw do sensownego, dobrze uwarunkowanego podziału zbioru obserwacji na podzbiory.