# Metody zapewniania prywatności w kontekście eksploracji reguł ilościowych

Jacek Sosnowski

3 września 2014

Politechnika Warszawska
Wydział Elektroniki i Technik Informacyjnych
Instytut Informatyki

# Spis treści

# 1 Wstęp

Otaczająca wszystkich rzeczywistość coraz częściej przenika się ze światem wirtualnym. Innymi słowy, Internet oraz komputery i narzędzia z nimi związane, już dawno stały się niezbędnym elementem codzienności. Próby przewidzenia dalszego rozwoju obecnej sytuacji spędzają sen z powiek naukowcom i wizjonerom. Niemniej jednak już teraz obserwuje się ciekawy skutek ekspansji technik komputerowych. Jest nim powstawanie społeczeństw informacyjnych. Takim mianem określa się cywilizację, w której szczególnym dobrem ekonomicznym staje się informacja. To ona pośrednio zaczyna być podstawą dochodu narodowego, a przez to głównym źródłem utrzymania wielu jednostek. Obecnie powszechnie akceptowalne jest stwierdzenie, że informacja jest bardzo cenna. Z tego powodu następuje ciągły rozwój usług informatycznych. Według autorów [14] z socjoekonomicznego punktu widzenia, sektor informacji dzieli się na: produkcję, przetwarzanie oraz przemysł dystrybucji informacji. Ostatni z nich obejmuje teleinformatykę, a więc przesyłanie danych.

Wyniki najnowszych badań statystycznych mogą wskazywać na prawdziwość doniesień o rozwoju wspomnianego typu społeczeństwa nawet w pięknym kraju nad Wisłą. Według danych GUS [15] dla Polski *„w 2012 r. liczba firm z sektora ICT wzrosła w stosunku do 2009 r. o 25,6 % (...) natomiast liczba pracujących w tym sektorze – o 11,6 %”*. Dodatkowo w tym samym okresie przychody netto dla tego sektora zwiększyły się o 30,8 %. Oznacza to wzrost zainteresowania świadczonymi usługami.

Efektem ubocznym dojrzewania społeczeństwa informatycznego jest wytwarzanie ogromnej ilości danych. Im więcej aspektów codziennego życia jest wspomagane przez technologie, tym więcej danych można zebrać. Przykładowo, jeszcze kilka dekad temu zwykłe zakupy były prostą wymianą dóbr (z wykorzystaniem środka płatniczego). Dzisiaj każda taka transakcja jest rejestrowana w systemie komputerowym, a następnie zapisywana w przygotowanej bazie danych. Jest to osiągalne dzięki postępowi techniki. A przede wszystkim dzięki łatwo dostępnej i taniej pamięci dyskowej. Realne jest więc gromadzenie dużych ilości danych, co czyni większość dzisiejszych instytucji.

Z upływem czasu coraz tańsza oraz wydajniejsza staje się również moc obliczeniowa. Taki kierunek zmian wprost świetnie wpisuje się w potrzeby współczesnej cywilizacji. Gdyż daje możliwość efektywnej analizy potężnych baz. Zadaniem tym zajmuje się dziedzina eksploracji danych. Dzięki jej metodom możliwe jest odkrycie interesujących zależności oraz nieznanej struktury zawartej w zebranych danych. Można też pokusić się o stwierdzenie, że ta gałąź informatyki pozwala na wydobycie „wiedzy” ukrytej w posiadanych bazach.

Ponadto fascynujące jest, że nie trzeba posiadać na własność żadnej bazy. Żeby wydobyć wiedzę z konkretnego rodzaju danych, można taką kolekcję wypożyczyć wyłącznie w celu eksploracji. Należy jednak liczyć się z tym, że właściciel może nie mieć zaufania i nie zgodzi się na udostępnienie. Powodem mogą być wrażliwe informacje lub zwyczajna niechęć do ujawniania danych pojedynczych jednostek zapisanych w zbiorze. Przykładowo może to być baza zawierająca informacje o klientach sieci komórkowej. Nawet jeśli nie będzie wśród nich wiadomości o personaliach, to i tak ujawnienie indywidualnych rekordów byłoby niestosowne. W takiej sytuacji przychodzi z pomocą dyscyplina zachowywania prywatności w eksploracji (ang. *privacy preserving data mining*). Pozwala ona na przeprowadzenie procesu wydobywania wiedzy jednocześnie zapewniając prywatność danym indywidualnym.

Najbardziej popularną metodą eksploracji danych jest nieustannie wyszukiwanie reguł asocjacyjnych. Istnieje wiele algorytmów rozwiązujących to zadanie. Niestety większość z nich wymaga, by dane wejściowe były zorganizowane w postaci tabeli o wartościach wyłącznie binarnych. Natomiast współcześnie gromadzone dane zdecydowanie częściej mają bogatsze dziedziny wartości. Często spotykane są zbiory, które zawierają liczby rzeczywiste. W takim przypadku unikalnych próbek teoretycznie może być nieskończenie wiele. O takie właśnie zbiory opierają się ilościowe reguły asocjacyjne. Przykładem może być reguła:

$$waga : [80, 100] \wedge gorne\_cisnienie : [120, 140] \Rightarrow ryzyko = 1$$

Powyższa reguła jest oczywiście sztucznym przykładem, ale wiedza zawarta w takich implikacjach może być bezcenna. Dlatego celem niniejszej pracy jest zbadanie możliwości eksploracji ilościowych reguł asocjacyjnych oraz próba zapewnienia prywatności tego procesu. W tym celu połączone zostaną siły trzech dyscyplin eksploracji danych: analizy skupień, odkrywania reguł asocjacyjnych oraz zachowania prywatności. Cała praca nie aspiruje do miana przeglądu tych dziedzin. Przedstawia natomiast specyficzny tok zdobywania wiedzy i technologii dla rozwiązania problemu przedstawionego jako cel pracy. Dokumentuje tylko i wyłącznie te fragmenty teorii eksploracji danych, które aktywnie i bezpośrednio wpłynęły na rozwiązanie końcowe. Wszelkie nawiązania, które nie wyczerpują w pełni tematyki, ale dla kompletności rozważań znajdują się w tekście, zostały opatrzone komentarzem wskazującym dokładniejsze źródło.

# 2 Eksploracja reguł asocjacyjnych

Eksploracja reguł asocjacyjnych (ang. association rule mining) jest jednym z ważniejszych filarów eksploracji danych (ang. data mining). Głównym celem tej dziedziny informatyki jest wydobycie z dużych zbiorów danych informacji wyższego poziomu abstrakcji. Oznacza to odkrycie relacji czy struktur zawartych w analizowanym zestawie, najlepiej jeśli będą nieznane oraz interesujące. Manualna analiza nawet niewielkich baz danych może sprawiać problemy, a wraz ze wzrostem ich wymiarów, takie badanie staje się praktycznie niemożliwe dla człowieka. Dodatkowo najbardziej interesujące relacje ujawniają się dopiero w liczniejszych kolekcjach danych. Dobrym wprowadzeniem do wspomnianej tematyki jest słynna już praca [3]. Jednak dla kompletności rozważań zostanie przedstawiony tu bardzo krótki zarys teoretyczny problemu eksploracji danych.

Baza danych $D$ to zbiór transakcji (próbek) $D = \{T_1, T_2, T_3, \ldots, T_n\}$ . Każda próbka $T_j$ ($j \in [1, n]$) zawiera ustaloną liczbę składowych opisanych zbiorem atrybutów $I = \{i_1, i_2, i_3, \ldots, i_k\}$.

Odkrycie związków pomiędzy atrybutami umożliwia eksploracja reguł asocjacyjnych. Ogólnie przez pojęcie reguły rozumie się wyrażenie postaci:

$$X \Rightarrow Y$$
$$X \in I,\, Y \in I,\, X \cap Y = \emptyset$$

Gdzie X oraz Y to zdarzenia. Przy czym kiedy w bazie występuje X to z pewnym poziomem ufności wystąpi też Y. Ten poziom nazywany jest wiarygodnością, lub ufnością (ang. confidence). Biorąc pod uwagę definicję bazy, każde ze zdarzeń, zarówno X jak i Y to podzbiory atrybutów $I$. Oznacza to, że reguła asocjacyjna opisuje relację pomiędzy występowaniem atrybutów $X$, a pojawianiem się atrybutów $Y$.

## 2.1 Rodzaje atrybutów

Mówiąc potocznie, bazę danych można potraktować jako tabelę. Wtedy każdy wiersz to rekord, próbka, lub po prostu element bazy. Natomiast każda kolumna nazywana jest cechą, atrybutem albo własnością obiektów przechowywanych w bazie. Każda cecha może zawierać wartości różnego typu, które można podzielić na kilka klas:

### 2.1.1 Atrybuty kategoryczne

Atrybuty kategoryczne, dzieli się na *nominalne* oraz *porządkowe*[1]:

[1] Informacje zaczerpnięte między innymi z [7]

1. **Atrybuty nominalne** (dyskretne, skończone, wyliczeniowe) – wartości tworzą przestrzeń skończoną i zazwyczaj niewielką. Nie istnieje porządek pomiędzy wartościami atrybutu.

2. **Atrybuty binarne** – są szczególnym przypadkiem atrybutów nominalnych, ponieważ przyjmują tylko dwie wartości. Zazwyczaj jest to 0 i 1. Najczęściej (choć nie zawsze) oznaczają występowanie bądź brak danego atrybutu w transakcji. W [3] w rozdziale „Formal Model" można znaleźć formalną definicję atrybutów binarnych w kontekście bazy danych.

3. **Atrybuty porządkowe** (ang. ordinal attributes) – podobnie jak dla nominalnych, dziedzina jest skończona oraz niezbyt liczna, ale istnieje możliwość uporządkowania wartości tych atrybutów. Znajomość porządku nie implikuje jednak znajomości odległości pomiędzy wartościami. Przykładem może być następująca przestrzeń bardzo mało, niewiele, dużo, bardzo dużo. Bez dodatkowych informacji ilościowych opisujących poszczególne dostępne wartości nie można określić relacji odległości pomiędzy próbkami tego atrybutu.

### 2.1.2 Atrybuty numeryczne

**Atrybuty ciągłe** (*numeryczne*) – definicja przestrzeni wartości dla takich atrybutów oparta jest na dobrze określonym zbiorze liczbowym, np. liczby rzeczywiste, czy całkowite. Z tego powodu ilość tych wartości jest nieskończona (nie można przewidzieć ile różnych próbek znajduje się akurat w konkretnej bazie danych). Dodatkowo znany jest porządek pomiędzy wartościami oraz zazwyczaj dystans między nimi (według wybranej miary odległości). Takie atrybuty najczęściej są spotykane przy opisach właściwości fizycznych obiektów, czego przykładami mogą być: masa, temperatura, długość, wzrost, itp. Już ta krótka lista ukazuje, że pomimo teoretycznie nieskończonej liczby wartości jaką prezentują atrybuty numeryczne, czasami istnieje możliwość uściślenia dziedziny. Typowym wzorem jest wiek, który z punktu widzenia bazy danych może przyjmować dowolne dodatnie wartości, lecz w rzeczywistości nie spotykamy się z ludźmi żyjącymi 200 czy 300 lat. Tak zwany kontekst danych przechowywanych przez atrybut może pomóc przy manualnym grupowaniu przechowywanych w nim wartości.

## 2.2 Wsparcie i ufność

Z eksploracją reguł asocjacyjnych nierozłącznie związane są dwa pojęcia: wsparcia oraz ufności. Oba będą często gościły w niniejszej pracy, dlatego

zostaną krótko przedstawione.

**Wsparcie** (ang. *support*) dla zbioru atrybutów $X$ to liczba (lub procent) wszystkich transakcji bazy $D$, które zawierają atrybuty $X$. Oznaczane jest poprzez $sup(X)$.

Prostymi słowami można stwierdzić, że jest to miara, która określa jak często w bazie danych pojawiają się cechy ze zbioru $X$. Patrząc odwrotnie na powyższą definicję należy wprowadzić pojęcie wspierania:

Transakcja $T_j$ wspiera zdarzenie $X$, wtedy kiedy zawiera wszystkie atrybuty zawarte w $X$ (może zawierać ich więcej niż posiada $X$).

W tej pracy zostało założone iż dziedziną funkcji wsparcia jest zbiór $\langle 0, 1\rangle$. Zatem miara ta wyznacza „ułamek" wszystkich transakcji. Terminu *wsparcie* używa się najczęściej nie w odniesieniu do zbioru atrybutów lecz w nawiązaniu do reguły asocjacyjnej:

Wsparcie reguły asocjacyjnej $A \Rightarrow B$ ($sup(A \Rightarrow B)$) to wsparcie zbioru $A \cup B$ ($sup(A \cup B)$).

Ta miara wskazuje jak silna jest przedstawiona reguła w konkretnym zbiorze danych. Im większa jest wartość omawianej funkcji tym częściej zbiór (na którym oparta jest reguła) pojawia się w bazie. W tym miejscu należy zaznaczyć, że to użytkownik lub badacz decyduje jak „silnych" reguł potrzebuje. W większości algorytmów taka decyzja jest respektowana poprzez ustalenie minimalnej wartości wsparcia. Celem jest nie generowanie reguł, dla których wsparcie znajduje się poniżej uzgodnionego progu.

Drugim niezbędnym pojęciem przy eksploracji jest **ufność** (lub **wiarygodność**, ang. *confidence*).

Ufność reguły asocjacyjnej $X \Rightarrow Y$ to część liczby transakcji które wspierają $Y$ wśród tych które wspierają $X$. Oznaczana jest przez $conf(X \Rightarrow Y)$

$$\mathrm{conf}(X \Rightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X)}$$

Sam problem wydobywania reguł jest zazwyczaj opisywany właśnie za pomocą tych dwóch miar poprzez podanie minimalnych ich wartości. To znaczy, że pożądane są tylko te asocjacje, których wsparcie jest wyższe niż minimalne wsparcie, a ufność jest większa niż minimalna ufność. Te wartości graniczne są oznaczane odpowiednio *minSup* oraz *minConf* (z ang. minimumSupport oraz minimumConfidence). Oba ograniczenia są ustawiane przez użytkownika, w celu sterowania procesem eksploracji.

Z przedstawionymi powyżej definicjami związane jest jeszcze jedno pojęcie. Każda reguła asocjacyjna jest stworzona w oparciu o pewien **zbiór częsty** (ang. frequent itemset). Jest to zbiór atrybutów których wsparcie jest

większe niż minimalne wsparcie minSup. Innymi słowy, problem eksploracji reguł można sprowadzić najpierw do zadania wyszukania zbiorów częstych. Następnie każdy z nich jest wykorzystywany do budowy kilku implikacji.

Wszystkie odsłonięte tutaj pojęcia będą wykorzystywane w kolejnych rozdziałach tej pracy. Natomiast ich dokładniejsze przedstawienie znajduje się w pracy [1]

# 3 Ilościowe reguły asocjacyjne

Znakomita większość algorytmów odkrywających zbiory częste oparta jest na założeniu akceptacji baz danych o wyłącznie binarnych wartościach. Jest to prawdopodobnie spadek z przeszłości po popularnej analizie koszyka zakupowego (ang. Market Basket Analysis). Kiedyś stosowane były w większości dane binarne. Z kolei obecnie częściej zbierane są informacje o ciągłych dziedzinach. Przykładem mogą być wyniki badań nad kwiatem o wdzięcznej nazwie Kosaciec, które zaowocowały legendarną już bazą danych Iris[2]. Wiedza jakie relacje ukryte są pomiędzy atrybutami numerycznymi jest niezwykle przydatna. W tym celu buduje się ilościowe reguły asocjacyjne (ang. quantitative association rules)[3]. Mogą one opisywać implikację danych o dowolnych wartościach (w tym numerycznych). W literaturze można spotkać następujące konwencje zapisu:

$$A = [a_1, a_2] \wedge B = [b_1, b_2] \Rightarrow C = [c_1, c_2]$$

lub

$$A : [a_1, a_2] \wedge B : [b_1, b_2] \Rightarrow C : [c_1, c_2]$$

Formalnie: $a_1 \leq A \leq a_2$ oraz $b_1 \leq B \leq b_2$ oraz $c_1 \leq C \leq c_2$

Wszystkie zapisy służą pokazaniu, iż tym razem nie tylko stwierdzone jest, że atrybuty A, B oraz C biorą udział w implikacji, ale dodatkowo podane są przedziały wartości jakie wchodzą w jej skład. To jest jedyna różnica w stosunku do reguł binarnych, poza tym wszystkie definicje oraz miary (wsparcie, ufność) pozostają dalej uznawane.

W powyższych zapisach zastosowano przedziałowy selektor danych. To znaczy, że wybrane wartości należały do konkretnego przedziału (np. [a,b]). Istnieją też inne rodzaje selektorów (deskryptorów) takie jak większościowy

[2]C.L. Blake and C.J. Merz. UCI repository of machine learning databases. University of California, 1998 (https://archive.ics.uci.edu/ml/datasets/Iris)

[3]Termin ten wprowadza artykuł [4]

oraz równościowy. Pierwszy z nich do selekcji danych używa operatorów porównania $<$ lub $>$ albo ich słabszych odpowiedników $\leq$ lub $\geq$. Ostatni deskryptor, oznaczany znakiem $=$ wybiera konkretne wartości. Wszystkie selektory są stosowane identycznie jak ich matematyczne odpowiedniki. Do atrybutów numerycznych można stosować wszystkie podejścia. Mimo to w dalszej części pracy wykorzystywany jest wyłącznie zapis przedziałowy.

Warto przyjrzeć się pojedynczemu składnikowi reguły: $X = [x_1, x_2]$. Zostało założone iż rozpatrywany przedział jest domknięty z obu stron. Zasadność takiego warunku można zauważyć na prostym przykładzie. Mając zbiór $Z = \{1, 147, 2, 151, 148, 3\}$ można bezspornie podzielić go na dwa podzbiory: $Z_1 = \{1, 2, 3\}$ oraz $Z_2 = \{147, 148, 151\}$. Brzegowe wartości obu zbiorów (uporządkowanych) mogą utworzyć granicę przedziałów dzielących dziedzinę danego zbioru: [1, 3] i [147, 151]. Jak widać, nie jest możliwe stworzenie ciągłego podziału (tak by suma przedziałów tworzyła ciągły przedział), ponieważ zbiór Z nie daje żadnych informacji na temat luki (3,147). Jakiekolwiek założenia mogłyby być sprzeczne z rzeczywistością. Dystans ten nie może zostać zniwelowany, a oba przedziały muszą być domknięte z obu stron aby dobrze opisywać zbiory $Z_1$ i $Z_2$. Ten fakt dobrze wpisuje się w idee całej eksploracji reguł, której zadaniem jest odkrywanie asocjacji już istniejących w danych, bez dokonywania żadnych założeń na temat próbek, które mogą pojawić się w przyszłości. Istnieją oczywiście również dziedziny eksploracji danych, które mają odmienną filozofię, jak choćby klasyfikacja.

Podział większych zbiorów wartości ciągłych (rzeczywistych, całkowitych, itp.) nie jest już tak oczywisty jak na powyższym prostym przykładzie. Wtedy zmiana granic wpływa na wsparcie tworzonego przedziału, co jest kluczowe w eksploracji reguł. To znaczy, że dysponując pewną regułą ilościową można zbudować regułę silniejszą jeśli zwiększy się przedział jednego ze składników ilościowych. Niestety czym silniejsza jest asocjacja tym mniej może być interesująca dla użytkownika, ponieważ może być zbyt ogólna i dobrze znana.
Tu nieśmiało zarysował się problem, któremu czoła chce stawić ta praca. Ogólnie i w przybliżeniu zadanie polega na takim doborze zakresów poszczególnych atrybutów by uzyskać możliwie dobrą regułę. Choć nie ma formalnej definicji dla ..dobrej reguły" to intuicyjnie może być ona pojmowana jako ta pomiędzy regułą bardzo silną (lecz jednocześnie pospolitą i nieciekawą), a bardzo słabą (przez co być może incydentalną i osobliwą, choć zapewne intrygującą).

| TID | wiek | kolor oczu | wynik testu |
|---|---|---|---|
| 1 | 9 | brązowy | 1 |
| 1 | 48 | zielony | 1 |
| 3 | 4 | niebieski | 0 |
| 4 | 51 | brązowy | 1 |
| 5 | 5 | niebieski | 0 |
| 6 | 49 | zielony | 0 |
| 7 | 7 | niebieski | 1 |
| 8 | 5 | niebieski | 1 |
| 9 | 51 | brązowy | 1 |

Tablica 1: Przykładowe dane dotyczące oczu.

### 3.1 Przykład

Dla zaprezentowania teorii wprowadzonej do tej pory, przedstawiona zostanie prosta reguła ilościowa. W tym celu stworzony został niewielki zbiór danych. Zawiera on wiek, kolor oczu oraz wynik pewnego testu dla dziewięciu podmiotów. Pierwszy atrybut jest numeryczny, drugi nominalny, ostatni binarny. Dane zawiera tabela 1

Obserwując uważnie w szczególności atrybut o wartościach całkowitych można zauważyć iż w zbiorze kryje się kilka reguł asocjacyjnych. Dla przykładu można przyjąć, że poszukiwane asocjacje powinny mieć wsparcie większe lub równe $\frac{4}{9}$ całego zbioru (to znaczy, że zbiór częsty będzie wspierany przez co najmniej 4 transakcje).

Analizując atrybut *wiek* widać naturalny wręcz podział na dwie kategorie: osobników młodych $\langle 4, 9\rangle$ oraz bardziej doświadczonych $\langle 48, 51\rangle$. Taki podział idealnie wpisuje się w wymaganie minimalnego wsparcia, ponieważ oba podzbiory go spełniają. Teraz wprost manifestuje swoją obecność reguła:

$$\underbrace{wiek : [4-9]}_{wsparcie=\frac{5}{9}} \Rightarrow \underbrace{kolor\_oczu = niebieski}_{wsparcie=\frac{4}{9}}$$

Pod obiema stronami implikacji zapisane są wsparcia poszczególnych podzbiorów. Natomiast cała reguła ma wsparcie $\frac{4}{9}$ co w przybliżeniu daje 44% i ufność $\frac{4}{5}$ czyli 80%.

## 4 Analiza skupień

Analiza skupień (ang. cluster analysis) lub inaczej grupowanie (ang. data clustering) to rozwiązanie problemu odkrywania struktury grup w kolekcji

obiektów nieoznaczonych żadnymi etykietami klas. Jest metodą klasyfikacji bez nadzoru (ang. unsupervised learning). To znaczy, że wyników podziału na zbiory nie można porównać z rozwiązaniem „referencyjnym", ponieważ takie nie istnieje dla metod bez nadzoru. Dodatkowo ten fakt powoduje niejednoznaczność grupowania oraz brak obiektywnej i globalnej miary jego poprawności.

**Zdefiniowanie problemu:**

Celem analizy skupień jest taki podział obiektów, by te znajdujące się wewnątrz grup były maksymalnie podobne do siebie, natomiast podobieństwo pomiędzy tymi, które przynależą do różnych skupisk powinno być minimalne. Innymi słowy, optymalizując powyższe warunki, poszukiwane są spójne podobszary danych.

Wygląda na to, że idea i kierunek działania dla rozwiązania tego problemu jest znany. Niemniej jednak w praktyce zastosowanie opisanej optymalizacji jest trudne. Sam problem grupowania jest uznawany za NP-zupełny (odniesienie w [12]). W dodatku jego przynależność do metod uczenia bez nadzoru sprawia iż problematyczne jest zdefiniowanie kompletnego celu końcowego. A co za tym idzie, niełatwo znajduje się warunki zakończenia działania dla algorytmów rozwiązujących to zadanie.

## 4.1 Miara podobieństwa

Pod pojęciem grupy[4] kryje się więc zbiór obiektów, które są „podobne", wobec tego mają one porównywalne właściwości. W takiej formie jest to dość mgliste i niejednoznaczne wyjaśnienie. Dlatego w każdym konkretnym przypadku stosowania analizy skupień konieczne jest ścisłe zdefiniowanie miary podobieństwa (lub adekwatnie niepodobieństwa) rozpatrywanych obiektów.

Konieczność powołania takiej definicji może budzić zaniepokojenie w momencie napotkania danych o charakterze kategorycznym (podrozdział 2.1). Są to wartości ze skończonej przestrzeni, dla których nie sposób wyznaczyć funkcję odległości. Jednakże nawet takie warunki nie budują bariery nie do pokonania, czego dowodem są wyniki prac [9] oraz [8] - rozdział „Distance Measures".

Zadanie wytyczenia funkcji podobieństwa przyjmuje formę bardziej przejrzystą i niebudzącą tylu wątpliwości jeśli grupowaniu podlegają tylko obiekty opisane właściwościami o wartościach ciągłych (numerycznych). Takie założenie uprawnia do wyboru jednej z wielu funkcji odległości i zbudowania w oparciu o nią adekwatnej funkcji podobieństwa. Praca [6] przedstawia odległość Euklidesową jako najczęściej stosowaną w takim przypadku.

[4]Pojęcia: grupa, skupisko, zbiór, są tutaj używane wymiennie.

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Jest to uzasadnione dwoma czynnikami. Pierwszy to fakt iż miara ta jest bardzo dobrze znana i powszechnie używana. Drugi, że dzięki temu jest intuicyjna albo sprawia wrażenie właśnie takiej. Powyższa funkcja stanowi szczególny przypadek miary Minkowskiego, która z kolei opisuje odległości w przestrzeniach o większej liczbie wymiarów:

$$L_m(\mathbf{x}, \mathbf{y}) = \left( \sum_{i=1}^{n} |x_i - y_i|^m \right)^{1/m}$$

Nadzwyczaj uciążliwą wadą tej ostatniej miary jest bardzo szybka utrata intuicyjności wraz ze wzrostem wymiarowości przestrzeni. O ile dla czterowymiarowych danych można odważyć się na pewne wnioski płynące z analizy odległości pomiędzy obiektami, o tyle już dla podwojonej ich liczby czyli ośmiu wymiarów jest to karkołomne wyzwanie. Należy zauważyć, że pomimo iż nikogo nie przerażają wektory o kilkunastu elementach to jednak przestrzenie wielowymiarowe mogą być kłopotliwe w interpretacji. To naturalne, że człowiek traci orientację i intuicję „matematyczną" po wyjściu poza przestrzeń którą zna od dziecka (trójwymiar). Dlatego w większości przypadków, funkcja podobieństwa jest po prostu bezdyskusyjnie przyjmowana jako funkcja odległości. W praktyce rozważania są ograniczane do rozwiązania następnego kłopotu.

## 4.2 Normalizacja

Kolejnym problemem (tym razem niezależnym od liczby wymiarów) jest dominacja składników o „szerokich" dziedzinach. Innymi słowy takich, dla których zakres wartości jest większy niż w innych składnikach przestrzeni. Na całą miarę Minkowskiego wpływ ma suma „odległości" na poszczególnych wymiarach (atrybutach). Najlepiej, żeby wszystkie składniki miały wartości zbliżone do siebie (chociażby pod względem rzędu wielkości). W przeciwnym przypadku nawet pojedynczy składnik może zdominować zachowanie całej sumy. Można zauważyć to zjawisko dysponując nawet tylko dwoma atrybutami, z których jeden przyjmuje wartości z bardzo małego zakresu, a drugi rozciąga się na całą, bardzo szeroką dziedzinę. Można spostrzec, że w ramach drugiego wymiaru różnica odległości dowolnych dwóch punktów będzie zawsze dużo większa niż dla pierwszego. Tym samym więc ten drugi ma większy

wpływ na wartość całej funkcji odległości. Zatem niewielkie zmiany położenia w ramach „szerokiego" atrybutu spowodują stosunkowo duże zmiany wartości pełnej miary Minkowskiego. Takie zachowanie obniża intuicyjność oraz uniemożliwia ustalenie warunków dla których można mówić o obiektach podobnych do siebie (co często czyni się na podstawie wyznaczenia progu dla odległości). Lekiem na wyżej wymienione dolegliwości może być **normalizacja** danych oddzielnie względem poszczególnych atrybutów. Zazwyczaj używane jest skalowanie przy wykorzystaniu zakresu lub wariancji [6]. W celu rozwiązania tego problemu adoptuje się również rozwiązania ze statystyki, w tym standaryzację:

> **Standaryzacja** (ang. *standard score*) jest sposobem normalizacji zmiennej losowej. Po tym procesie zmienna posiada zerową wartość oczekiwaną ($\mu = 0$) oraz wariancję równą jeden ($\sigma = 1$). Najpopularniejsza jest *standaryzacja Z*, którą opisuje wzór:
>
> $$z = \frac{x - \mu}{\sigma}$$

**Podsumowanie**

Istnieją też inne miary odległości wykorzystywane w analizie skupień, jak chociażby nieskomplikowana odległość Manhattan czy ambitna miara Mahalanobis. Jednakże w ramach tej pracy badane będą tylko i wyłącznie dane o charakterze liczb rzeczywistych oraz najczęściej jednowymiarowej dziedzinie, stąd prosty wniosek, że wystarczająca jest miara Minkowskiego.

Podsumowując, analizie skupień mogą podlegać dowolne obiekty pod warunkiem, że istnieje dla nich miara podobieństwa. Mimo to najczęściej pod pojęciem *obiektu* kryje się uporządkowany wektor cech (atrybutów) o wymiarze d. Dzięki temu można rozpatrywać go jako *punkt* w $d-$wymiarowej przestrzeni. W takim kontekście można mówić nawet o kształcie grup: wklęsłe, wypukłe, albo bardziej drobiazgowo: koliste, prostokątne, itp. Niektóre algorytmy jako swoje właściwości mają również specyfikowany kształt zbiorów wynikowych[5].

Daleko idący i atrakcyjny przegląd dziedziny analizy skupień został przedstawiony w pracy [6].

## 4.3 Rodzaje analizy skupień

Współczesna informatyka dostarcza szerokiego wachlarzu rozwiązań problemu analizy skupień. Każde z nich działa na swój sposób i ma wyjątkowe właściwości. Ta różnorodność jest cechą pozytywną ze względu na fakt, że nie

[5]Przykładem jest algorytm dzielenia według siatki, który produkuje wyłącznie prostokątne podobszary.

wszystkie zadania grupowania są takie same. Oczywiste jest, iż w praktyce stosuje się różne miary podobieństwa. Ale dodatkowo stawiane są różnorodne wymagania odnośnie właściwości jakie mają spełniać nowo powstałe grupy. I chociażby z tego ostatniego powodu można podzielić wszystkie metody analizy skupień na kilka kategorii, oto lista pojęć jakimi są one identyfikowane:

### 4.3.1 Właściwości grup wynikowych

To jakie cechy będą mieć przedziały po analizie skupień zależy od zastosowanego algorytmu. Natomiast, to jakie są dostępne opcje prezentuje poniższa lista:

1. **równe szerokości przedziałów** (ang. *equi-width*) – każdy interwał ma w przybliżeniu równą średnicę[6].

2. **równe głębokości** (ang. *equi-depth*) – uzyskane zbiory zawierają w przybliżeniu tyle samo elementów. Nazwa angielska i jej polskie tłumaczenie najprawdopodobniej są zainspirowane grupowaniem hierarchicznym, w którym to poziom głębokości definiuje również w pewnym sensie liczność grup.

3. **jednolite przedziały** (ang. *homogeneity-based bins*) – rozmiar grupy jest tak dobierany, by rozkład jej elementów był możliwie jednolity. Taką kategorię opisuje praca [12].

4. **grupowanie rozmyte** (ang. *fuzzy clustering*) – otrzymane grupy mogą mieć parami niepuste części wspólne. W potocznym znaczeniu - częściowo nakładają się na siebie. Kluczowe znaczenie ma dobór ograniczeń na wspólny podzbiór, czyli odpowiedź na pytanie jak bardzo grupy mogą się przenikać ze sobą. Drugim pytaniem jest, czy dopuszczalne jest zawieranie zbiorów (być może przydatne w pewnych specyficznych zastosowaniach).

   Formalnie: każdy obiekt zbioru może z pewnym prawdopodobieństwem należeć do każdego z powstałych przedziałów. Dla przykładu: dysponując zbiorem grup $\{G_1, G_2, G_3\}$, element x, dzięki swoim cechom, z prawdopodobieństwem $0,78$ powinien przynależeć do zbioru $G_1$, natomiast z $0,12$ do $G_2$ oraz z $0,1$ do $G_3$. Stosując grupowanie rozmyte nie dokonuje się rozstrzygnięcia do której z grup należy dany obiekt. Jest on pojmowany jako członek wszystkich grup z dodatkową informacją na temat „poziomu" tego członkostwa.

[6] Przez średnicę rozumie się tutaj maksymalną odległość pomiędzy dwoma dowolnymi elementami zbioru.

5. **grupowanie „twarde”** (ang. *hard clustering*) – każdy z analizowanych obiektów może należeć tylko i wyłącznie do jednej z grup. Dlatego utworzone zbiory nie posiadają części wspólnych. Jest to tradycyjne podejście do tematyki analizy skupień, a przez to jest najczęściej stosowane. Jego przeciwieństwem są metody rozmyte.

6. **grupowanie „naturalne”** (ang. *natural clustering*) - to koncepcja wydobycia ze zbioru takich skupisk, jakie tam się naturalnie znajdują. Sensem tego podejścia jest jak najmniejszy wpływ sztucznie i manualnie wybranych parametrów na końcowy podział. Do tej kategorii można zaliczać metody gęstościowe, jak np. algorytm DBSCAN, czy metody jakościowe, jak np. algorytm Quality Threshold.

Zapoznanie się z taką kategoryzacją analizy skupień pozwala na optymalny dobór metody, a później konkretnego algorytmu dla przedstawionego rzeczywistego problemu. Użytkownik musi wskazać dokładnie jakich cech oczekuje. Należy jednak wspomnieć o tym, że istnieje możliwość mieszania powyżej przedstawionych opcji. Jest dozwolone zbudowanie takiego algorytmu, który będzie rozpatrywać zarówno kryterium równej szerokości tworzonego zbioru, ale również będzie czuły na ilość zawartych w nim obiektów lub ich rozkład. Wszystko zależy od warunków jakie są stawiane przed procesem grupowania.

### 4.3.2 Metody podziału

Znając już warunki jakie musi spełniać pożądane grupowanie, należy wybrać jeszcze jedną z wielu procedur grupujących[7]. Ponownie decyzję można oprzeć o kilka ogólnych kategorii. Procedury analizy skupień dzielą się na:

1. **hierarchiczne** (ang. *hierarchical*) – grupy są łączone ze sobą w celu utworzenia hierarchii – dwie mniejsze grupy mogą tworzyć ze sobą większą (pod warunkiem, że są do siebie „podobne”). Mówiąc inaczej, cały zbiór zawiera kilka mniejszych zbiorów, z kolei te zawierają w sobie jeszcze mniejsze i tak dalej. Tutaj należy wyróżnić dwa podejścia:

   (a) **wstępujące** – każdy obiekt początkowo traktowany jest jako grupa, następnie iteracyjnie grupy są łączone w pary. Taki proces kończy się zazwyczaj po połączeniu dwóch ostatnich grup w cały zbiór, albo po osiągnięciu założonego wcześniej celu dotyczącego właściwości podziału. W jednym i w drugim przypadku produktem końcowym jest hierarchia grup (struktura zawierania).

[7] Obecne rozważania nadal nie nawiązują do żadnych konkretnych algorytmów.

(b) **zstępujące** – metoda rozpoczyna działanie od całego zbioru i dzieli go na mniejsze fragmenty, te nowo powstałe są rozcinane w podobny sposób rekurencyjnie. W tym przypadku warunek stopu też podlega ustaleniu. Absolutny koniec to uzyskanie wszystkich jednoelementowych zbiorów. W wyniku działania otrzymywana jest hierarchia grup (struktura dzielenia). Lecz tym razem, brak kompletnej struktury, to brak pełnej wiedzy na temat podobieństwa „małych" zbiorów, łącznie z jednoelementowymi.

Jeśli dowolna miara odległości zostanie wybrana jako funkcja podobieństwa, to dla obu powyższych podejść należy dokonać wyboru sposobu obliczania odległości pomiędzy grupami. Innymi słowy, zbiór obiektów jako taki nie stanowi pojedynczego punktu w przestrzeni, co powoduje niejednoznaczności przy obliczaniu funkcji odległości. Może być ona ustalona pomiędzy „środkami"[8] zbiorów. Ale również możliwe jest zastosowanie odległości maksymalnej, czy minimalnej. Zagadnienie to nazywane jest *metodami wiązania*, a więcej na ten temat w [10] w rozdziale Analiza skupień.

Algorytmy hierarchiczne to jedne z niewielu, które oprócz wyboru funkcji podobieństwa nie wymuszają dodatkowych parametrów. Natomiast na samą funkcję nie ma nałożonych żadnych ograniczeń ani wymagań (oprócz wartości zwracanej, która ma być miara określającą jak bardzo jeden obiekt jest podobny do drugiego). Budowana jest wtedy całą hierarchia, którą można analizować nawet wizualnie na dendrogramach ([RYSUNEK]), czy diagramach venna ([RYSUNEK]).

2. **metody oparte na podziałach** (ang. *partitional clustering*)[9] To metody stosujące pojedynczy i rozłączny podział całego zbioru zamiast hierarchii podziałów. Biorąc pod uwagę ich cechy można wyznaczyć pewne kategorie i są to kolejno procedury:

   (a) **deterministyczne** albo **niedeterministyczne** – niektóre algorytmy grupujące uznawane są za niedeterministyczne, ponieważ ich działanie opiera się w jakimś stopniu o „losowość". Przykładem może być algorytm k-means, który losowo ustala położenie pierwotne środków grup.

[8] To tak zwana metoda środka ciężkości zbioru. Można go wyznaczyć poprzez średnią (zwykłą lub ważoną) wartości we wszystkich punktach zbioru.

[9] Definicja zaczerpnięta z encyklopedii uczenia maszynowego [11] - hasło *partitional clustering*

(b) **minimalizujące błąd wynikowy** (ang. *Error Minimization Algorithms* [8]) – tak właściwie jest to raczej idea stojąca za niektórymi konkretnymi algorytmami. Polega ona na minimalizacji pewnego kryterium błędu, najczęściej opisanego funkcją odległości. Najbardziej znany jest błąd kwadratowy (ang. , *squared error alorithms* [6]), a konkretnie błąd średniokwadratowy (ang. MSE - *Mean Squared Error*), czy suma kwadratów błędów (ang. *Sum of Squared Error*). Należy więc ustalić jakie odległości są rozpatrywane – najczęściej badana jest odległość punktów do środków grup ich zawierających. Cały proces polega wówczas na globalnej optymalizacji funkcji błędu.

Dobrą ilustracją tej procedury jest algorytm k-means, który początkowo losuje pakiet środków przyszłych grup, np. $\{S_1, S_2, S_3\}$. Niestety to jest jego wada, że liczbę i położenie tych środków należy ustalić apriori. Celem jest takie przemieszczenie środków względem obiektów aby zminimalizować błąd (np. średniokwadratowy sumy odległości punktów od środków). Rozwinięcie zagadnienia znajduje się w pracy [8] w podrozdziale „Partitioning Methods".

(c) **oparte na gęstościach** (ang. *density-based algorithms*) wykonują swoją pracę analizując gęstość rozłożenia punktów zbioru w przestrzeni. Tym razem, każdy obiekt z zadanego zbioru musi być opisany wektorem liczb (odpowiadającym kolejnym jego cechom, atrybutom obiektu). Tak stworzony wektor jest traktowany jako punkt w przestrzeni (liczba wymiarów odpowiada rozmiarom wektora). Tworzone grupy cechują się ustaloną i jednolitą wewnętrzną koncentracją. Natomiast dzięki kontrastowi gęstości można wydzielać kolejne zbiory. Punkty należące do obszarów o niskim nasyceniu są traktowane jako szum (zakłócenia lub próbki odstające od zbioru) i nie podlegają grupowaniu. Przykładem implementacji tej metody jest algorytm DBSCAN[10]

Algorytmy zbudowane w oparciu o tą idee mają spektakularne zalety. Po pierwsze i co najważniejsze nie wymagają dogłębnej znajomości dziedziny analizowanych danych (ani liczby ani położenia środków). Pozwalają także na wykrycie grup o bardzo dowolnym, a nawet wyrafinowanym kształcie. Te dwa argumenty już sprawiają, że można podchodzić do nich bardzo entuzjastycznie.

[10]Ester M. i in.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Proc. of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)

Niemniej jednak takie atuty mają też swoje skutki uboczne. Problemem bywa ustalenie poziomu koncentracji wewnętrznej grup oraz poziomu gęstości poniżej której punkty traktowane są jako szum. Procedura jest w stanie działać bardzo dobrze dla danych w których skupiska punktów są dobrze odizolowane, co niestety nie zawsze zachodzi w świecie rzeczywistym.

(d) **oparte na grafach** (ang. *graph-theoretic clustering* lub *graph-based clustering*) – w zarysie polegają na rozpięciu grafu o węzłach w punktach zbioru. Najbardziej znane algorytmy wykorzystują minimalne drzewo rozpinające (ang. *Minimal Spanning Tree*). Artykuł [8] – rozdział „Graph-theoretic clustering".

(e) Istnieje jeszcze wiele innych podejść do metody grupowania opartej na podziałach. Są to między innymi procedury wykorzystujące sieci neuronowe ([8] – „Neural networks"), czy nawet techniki ewolucyjne. W nielicznych sytuacjach można również zastosować podział według narzuconej „siatki" (ang. *grid-based methods* [8]).

Jednakże dotychczas przedstawione metody oraz procedury stanowią wystarczającą bazę dla wyboru rozwiązania najlepiej spełniającego cel niniejszej pracy.

## 4.4 Rozważania na temat właściwości grupowania

Dysponując całym wachlarzem algorytmów grupujących, należy wybrać właściwości jakie są oczekiwane względem zbiorów końcowych. Pewną wskazówką może być klasyfikacja tych algorytmów naszkicowana w rozdziale 4.3 Rodzaje analizy skupień.

**Odniesienie do właściwości podziałów** (rozdz. 4.3.1)

Analiza skupień w przypadku tej pracy nie może ograniczyć się do grupowania ani o równych szerokościach ani o zbliżonej liczności. To specyfika i naturalny rozkład danych w bazie powinien narzucić rozpiętość i rozmiary poszczególnym grupom. Z drugiej strony zostaną wprowadzone zdroworozsądkowe ograniczenia oparte właśnie na tych parametrach. Wiadomo, że szerokość grupy określa poniekąd jej „jakość". Natomiast liczba elementów będzie niezmiernie ważna dla późniejszego odkrywania reguł.

Tworzenie pakietu rozłącznych grup to podejście tradycyjne i dobrze zakorzenione w intuicji wielu ludzi. Niemniej jednak warto wykorzystać potencjał grupowania rozmytego. Ta kwestia rozstrzygnie się w kolejnych rozdziałach.

**Odniesienie do metod podziałów** (rozdz. 4.3.2)

W kontekście dostępnych ogólnych schematów postępowania, pierwsze

pytanie na jakie trzeba odpowiedzieć, to czy hierarchia grupowania jest potrzebna?

Zdecydowanie wymagany jest płaski podział, który można następnie potraktować jako atrybut nominalny dla reguł asocjacyjnych. Dlatego ani hierarchia grup, ani dokładność tego procesu nie są przydatne. Po przedstawieniu reguły, np. $wiek : [5 - 10] \Rightarrow wzrost = niski$ nie jest już istotne czy przedział $[5 - 10]$ został zbudowany w oparciu o dwa czy trzy inne przedziały. Nie jest też ważne jak grupują się obiekty pojedyncze. Natomiast z punktu widzenia metody zapewniania prywatności, grupowanie hierarchiczne z zachowywaniem hierarchii byłoby nie do przyjęcia.

W następstwie powyższych wniosków, algorytmu należy szukać wśród metod opartych na podziałach „płaskich". A w tych ramach, należy odpowiedzieć na pytanie, czy zachowanie losowe jest w tym zastosowaniu akceptowalne?

Grupowanie jest elementem transformacji, która jak było sugerowane wcześniej powoduje przekształcenie całej bazy (nawet fizycznie, z jednego pliku na drugi). Teoretycznie nie wymusza to deterministyczności całego procesu. Raz przekształcona baza może służyć do wielokrotnego poszukiwania reguł. Ostatecznie cały proces można powtórzyć kilka razy przed właściwą eksploracją reguł, żeby uniknąć negatywnego skutku „niefortunnej" sytuacji w momencie losowania.

Nie istnieje niestety możliwość dobrego zastosowania metody gęstościowej dla późniejszej eksploracji reguł asocjacyjnych. Dzieje się tak z kilku powodów, po pierwsze, jak zwykle problematyczne bywa ustalenie wartości wszystkich parametrów, z poziomem gęstości na czele. Dodatkowo w oryginalnych algorytmach nie ma możliwości sterowania wielkością grupy ani zakresem jej wartości. Ostatnim problemem, jest fakt iż ta metoda pozwala na wyodrębnienie podobszarów, które są „kontrastowe" między sobą względem gęstości. Tymczasem dla eksploracji reguł cenny jest nawet podział przestrzeni spójnej. Oczywiście podział końcowy powinien opierać się na skupiskach możliwie odrębnych, ale jeśli zajdzie potrzeba rozcięcia zwartego skupiska na kilka części, to też powinno być wykonane.

Ani algorytmy grafowe ani bezpośrednie zastosowanie sieci neuronowych nie są możliwe w przypadku baz przeznaczonych do eksploracji. Głównym powodem jest zazwyczaj duży rozmiar używanych kolekcji. Jednocześnie złożoność pamięciowa tych rozwiązań jest delikatnie mówiąc godziwa, co przyczynia się do tego iż ich zastosowanie staje się mało zachęcające.

To krótkie podsumowanie uświadamia możliwości w wyborze algorytmu analizy skupień z zastosowaniem dla budowana reguł ilościowych. Pomijając wykluczone podejścia, w rozwiązaniu można zastosować metody oparte na podziałach w tym szczególnie deterministyczne, ale również te z elementami niedeterministycznymi. Dopuszczalne są metody z szeroko pojętej kategorii

minimalizujących błąd wynikowy.

# 5 Zbiór wartości binarnych

Po wstępie teoretycznym musi pojawić się uzasadnienie, dlaczego tak, zdawałoby się, odmienne dziedziny eksploracji danych są przedstawiane w tej pracy razem. Łączy ich cel – odkrycie interesujących ilościowych reguł asocjacyjnych. Podążając za pracą [4] wyróżnia się dwa podstawowe typy reguł asocjacyjnych: binarne (ang. *boolean association rules*) oraz ilościowe (ang. *quantitative association rules*). Jak już zostało wspomniane wcześniej, do odkrywania asocjacji binarnych istnieje szereg algorytmów, natomiast w drugim przypadku jest ich niewiele oraz bywają mało efektywne. Stąd już w 1996 roku powstał pomysł, żeby zbiór danych formatu numerycznego przekształcić w kolekcję pozycji $\{0, 1\}$. Tym samym problem wyszukiwania ilościowych reguł staje się zadaniem odkrycia binarnych reguł asocjacyjnych. Ten rozdział zawiera przegląd możliwości zamiany pojedynczych atrybutów na binarne.

## 5.1 Transformacja na wartości binarne

W odniesieniu do **atrybutów kategorycznych** sposób przekształcenia jest oczywisty. Dla przykładu niech obiekty będą charakteryzowane przez *„atrybut1”*, który przyjmuje $n$ wartości. Każdy element tej cechy powinien stworzyć nowy atrybut, klasycznie pod tytułem *„atrybut1:wartośćK”*. Tym razem jest on już dwuwartościowy, znów klasycznie, przyjmuje się dziedzinę $\{0, 1\}$. Jedynka oznacza, że pierwotna cecha *„atrybut1”* miała w danej transakcji wartość *„wartośćK”*, zero wstawiane jest w przeciwnym przypadku.

W odniesieniu do **kolekcji typu numerycznego** (np. liczb rzeczywistych) droga do zbioru dwuwartościowego nie jest tak bezdyskusyjna. Nie istnieje jednoznaczne i najlepsze rozwiązanie. Być może istnieją problemy, które będą wymagać w tym miejscu odwzorowania każdej unikalnej wartości w zbiór $\{0, 1\}$. Ale w znakomitej większości przypadków taka procedura doprowadzi do uzyskania ogromnej i bardzo rzadkiej macierzy.

Większość pozycji literatury dotykającej problemu reguł ilościowych sugeruje iż transformacja atrybutów numerycznych powinna odbyć się poprzez dyskretyzację ich dziedziny wartości. Następnie tak podzielony atrybut można potraktować jako kategoryczny i przekształcić w dwuwartościowy tak jak to jest opisane wyżej.

## 5.2 Dyskretyzacja

Jest to procedura przetwarzająca informacje ciągłe w dyskretne (źródło: [6]). Formalnie, szczególnie w matematyce, pojęcie to dotyczy procesu przekształcania modeli ciągłych w dyskretne ich odpowiedniki. W topologii dotyczy przekształcania przestrzeni spójnych w przestrzenie dyskretne. Przestrzeń spójna intuicyjnie składa się z „jednego kawałka", natomiast dyskretna opiera się o punkty, które są niejako „oddzielone" od siebie.

W kontekście eksploracji danych, do dyspozycji jest kolekcja wartości konkretnego atrybutu ciągłego. Sam atrybut można potraktować jako pewną funkcję. Jej przeciwdziedzina tworzy przestrzeń spójną. Dlatego proces transformacji na dane binarne z tego punktu widzenia wygląda na zadanie dyskretyzacji przeciwdziedziny (po prostu w celu zamiany jej na topologiczną przestrzeń dyskretną). Uzyskany podział powinien pokryć całą przestrzeń. Niemniej jednak zazwyczaj o przeciwdziedzinie wiadomo niewiele. Znany jest wyłącznie skończony zbiór próbek tej przestrzeni. Dlatego w praktyce, mniej lub bardziej słusznie, mówi się o dyskretyzacji kolekcji próbek i w tym celu wykorzystuje się metody analizy skupień.

## 5.3 Właściwości dyskretyzacji

Sam proces grupowania niesie za sobą pewne konsekwencje. Przede wszystkich spowoduje częściową utratę informacji, a jednocześnie ustanowi pewien poziom prywatności danych pierwotnych. Ma to praktyczne zastosowania, a jedno wynika z tego co już zostało wspomniane wcześniej: podstawowym i najprostszym sposobem zapewnienia prywatności danych indywidualnych jest ich podział na przedziały. Natomiast utrata konkretnych wartości ma kilka zastosowań ogólnych. Są to między innymi próby uproszczenia danych, na przykład w celu ich prezentacji wizualnej. Innym przykładem jest chęć przygotowania wstępnej analizy lub próby szybkiego wyodrębnienia ogólnej wiedzy.

## 5.4 Przykład transformacji

W celu zaprezentowania działania transformacji atrybutów różnego rodzaju, użyty zostanie zbiór danych zawarty w tabeli 1 omawiany już w rozdziale 3.1. Zawiera on zarówno wartości całkowite jak i nominalne. Minimalne wsparcie niech tym razem wynosi 40%. Pierwszy atrybut (kolumna wiek) zostanie poddany analizie skupień, czego wynikiem będzie utworzenie dwóch przedziałów $\langle 4-9\rangle$ oraz $\langle 48-51\rangle$. Wartości w dotychczasowych rekordach zostaną zamienione na wspomniane przedziały. Ten krok prezentuje tabela 2, która

| **TID** | **wiek** | **kolor oczu** | **wynik testu** |
|---|---|---|---|
| 1 | $\langle 4-9 \rangle$ | brązowy | 1 |
| 1 | $\langle 48-51 \rangle$ | zielony | 1 |
| 3 | $\langle 4-9 \rangle$ | niebieski | 0 |
| 4 | $\langle 48-51 \rangle$ | brązowy | 1 |
| 5 | $\langle 4-9 \rangle$ | niebieski | 0 |
| 6 | $\langle 48-51 \rangle$ | zielony | 0 |
| 7 | $\langle 4-9 \rangle$ | niebieski | 1 |
| 8 | $\langle 4-9 \rangle$ | niebieski | 1 |
| 9 | $\langle 48-51 \rangle$ | brązowy | 1 |

Tablica 2: Wyniki dyskretyzacji atrybutu wiek z tabeli 1.

| TID | wiek: $\langle 4-9 \rangle$ | wiek: $\langle 48-51 \rangle$ | oczy: niebieski | oczy: brązowy | oczy: zielony | wynik testu |
|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| 2 | 0 | 1 | 0 | 0 | 1 | 1 |
| 3 | 1 | 0 | 1 | 0 | 0 | 0 |
| 4 | 0 | 1 | 0 | 1 | 0 | 1 |
| 5 | 1 | 0 | 1 | 0 | 0 | 0 |
| 6 | 0 | 1 | 0 | 0 | 1 | 0 |
| 7 | 1 | 0 | 1 | 0 | 0 | 1 |
| 8 | 1 | 0 | 1 | 0 | 0 | 1 |
| 9 | 0 | 1 | 0 | 1 | 0 | 1 |

Tablica 3: Wyniki przekształcenia danych z tabeli 1 do postaci binarnej.

jest już bazą danych wyłącznie kategorycznych. W tym momencie należy zastosować binaryzację. Dla przykładu, drugi atrybut jest przekształcony na tyle kolumn ile zawiera unikalnych wartości, ostatni pozostanie bez zmian. Wynik prezentuje tabela 3 zawierająca binarną bazę danych.

# 6 Analiza skupień, a wyszukiwanie reguł asocjacyjnych

Jak zostało zaznaczone już nawet we wstępie, w celu odkrycia ilościowych reguł asocjacyjnych zostanie przeprowadzone grupowanie atrybutów numerycznych. W tym celu należy uściślić właściwości tego procesu i wymagania odnośnie wynikowych grup. Sam proces eksploracji podzielony jest na dwa

etapy. Pierwszy to znalezienie zbiorów częstych, drugi to budowa reguł z odkrytych zbiorów.

## 6.1 W kontekście wyszukiwania zbiorów częstych

Rozważania warto rozpocząć od przypomnienia, że zbiór częsty to taki zbiór atrybutów bazy danych, który w całym zbiorze transakcji ma wsparcie większe niż minSup. Ogólnie znane jest też twierdzenie, które mówi:

Każdy podzbiór zbioru częstego jest zbiorem częstym. To oznacza, że żaden zbiór nieczęsty nie może stać się częstym poprzez dodanie do niego nowych atrybutów. A wniosek ostateczny brzmi:

Jeśli pojedynczy atrybut jest nieczęsty, to nigdy nie wejdzie w skład reguły asocjacyjnej.

Z ostatniego stwierdzenia można wysnuć już prosty postulat: cechę numeryczną należy tak przekształcać, by nowo powstałe atrybuty były częste. Innymi słowy utworzenie grupy która nie będzie częsta nie przyniesie korzyści z punktu widzenia wyszukiwania reguł. Z tego powodu na grupowanie należy nałożyć pierwsze kryterium – powinno znać i respektować minimalne wsparcie ustalane przed procesem eksploracji. A konkretniej:

Wynikowe grupy powinny zawierać co najmniej tyle elementów ile wynosi iloczyn minimalnego wsparcia i liczby wszystkich transakcji w bazie:

$min\_rozmiar\_grupy = minSup * liczba\_transakcji$

Pod warunkiem, że minSup jest dane jako ułamek, a nie jako konkretna liczba próbek.

Ten sam problem został zaprezentowany w pracy [4] pod nazwą problemu „MinSup" i zdefiniowany nieco inaczej. Tam obecna definicja mówi, że jeśli liczba przedziałów będzie duża to wsparcie pojedynczego przedziału będzie małe. A wtedy niektóre reguły zawierające tą cechę mogą zostać nie odkryte z powodu braku minimalnego wsparcia.

Bardzo trudno jest uzasadnić sens powoływania przedziałów tworzących atrybuty, które nie wezmą udziału w żadnej regule asocjacyjnej. Z punktu widzenia problemu „MinSup" im szersze zakresy przedziałów tym lepiej.

## 6.2 W kontekście powoływania reguł asocjacyjnych

Żeby z dowolnego zbioru częstego utworzyć regułę asocjacyjną, należy podzielić go na dwie części i jedną potraktować jako lewą stronę implikacji, a drugą jako prawą. Jeśli taka reguła spełnia warunek minimalnej ufności to może być spokojnie zaprezentowana jako jeden z wyników całego procesu eksploracyjnego. W przeciwnym przypadku zostaje odrzucona. Jak to odnosi się do procesu analizy skupień? Otóż, niski poziom wiarygodności pojawia

się wtedy, kiedy wsparcie poprzednika implikacji jest duże w stosunku do wsparcia całej reguły (czyli tych transakcji, które wspierają poprzednik i następnik jednocześnie). Jeśli w poprzedniku znajduje się atrybut utworzony z pierwotnie numerycznego, to zwiększając szerokość przedziału możliwe jest pogorszenie wsparcia całej reguły. A w konsekwencji reguła może zostać odrzucona z powodu nie spełnienia kryterium minimalnej wiarygodności (minConf). Innymi słowy, im mniejsze przedziały, tym lepiej z punktu widzenia poziomu ufności.

Dla przykładu warto rozważyć regułę $A \Rightarrow B$ która ma wsparcie $w$. Dodatkowo niech zarówno $A$ jak i $B$ oddzielnie też mają wsparcia $w$. To oznacza, że reguła ma ufność na poziomie 100%. Jeśli jednak zwiększy się zakres $A$, a więc jednocześnie zacznie go wspierać więcej transakcji, to poziom ufności całej reguły spadnie. Przykładowo, jeśli wsparcie $A$ wzrośnie dwukrotnie to poziom wiarygodności zmaleje aż do 50%.

Rozwiązywanie problemów minSup i minConf działa antagonistycznie. Poprawa jednego może pogorszyć sytuację drugiego. Dlatego właśnie tak trudno jest ustalić optymalny rozmiar grup. Na szczęście wiele zbiorów danych zawiera w sobie naturalne skupiska, które spełniają dolne ograniczenie.

# 7 Zapewnianie prywatności

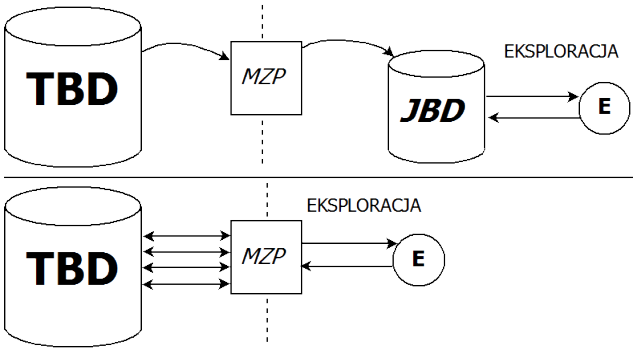
To dziedzina eksploracji danych obejmująca zbiór metod zapewniania prywatności podczas procesów wydobywania wiedzy. Prywatność rozumiana jest w tym przypadku jako częściowe ukrycie informacji, które zdaniem ich właściciela są zbyt wrażliwe by je ujawniać. Jednocześnie użyte metody ukrywania umożliwiają budowanie modeli w oparciu o całościowy rozkład wartości.

## 7.1 Sposoby rozproszenia danych

Cały proces zapewniania prywatności jest techniką wspierającą eksplorację danych, które nie znajdują się w posiadaniu badaczy. Baza na której będzie budowany model, może znajdować się w wielu miejscach. W literaturze spotyka się wyszczególnienie trzech typów takiego rozproszenia.

1. **Dane podzielone poziomo** (ang. *horizontally partitioned data*)
   Każda ze stron posiada tylko część rekordów, ale wszystkie fragmenty opisane są tymi samymi atrybutami. Są to więc te same informacje, ale na temat różnych obiektów.

2. **Dane podzielone pionowo** (ang. *vertically partitioned data*)
   W różnych miejscach znajdują się różne informacje, ale o tych samych




Rysunek 1: Modele udostępniania danych po zastosowaniu metod prywatyzacji. a) Nieinteraktywne (szkic na górze) b) Interaktywne (szkic na dole) Poszczególne bloki: TBD to poufna baza danych, MZP zawiera moduł zapewniania prywatności, JBD to baza jawnie udostępniona, E symbolizuje eksploratora

rekordach. To znaczy, że fizycznie ten sam obiekt jest opisywany przez inne atrybuty w zależności od konkretnej bazy.

3. **Dane położone centralnie** (scentralizowane, ang. *centralised data*) Tak określa się przypadek, kiedy wszystkie wymagane dane znajdują się w jednym miejscu, jednej bazie. Dzięki temu nie ma potrzeby komunikacji w celu uzyskania pełnych informacji.

## 7.2 Modele źródeł danych

Źródła danych, które już uległy metodom zapewniania prywatności można podzielić na dwa typy:

1. **Nieinteraktywne** (ang. *non-interactive model*) – cała baza danych podlega pewnemu procesowi zapewnienia prywatności, po czym zostaje udostępniona w całości. Metodę prezentuje rysunek 1a.

2. **Interaktywne** (ang. *interactive model*) – poszczególne rekordy są zdobywane poprzez zadawanie pytań do oryginalnej bazy. Odpowiedzi są tak dostosowywane by zapewnić odpowiedni poziom prywatności i nie ujawnić informacji poufnych. Szkic tego procesu zawiera rysunek 1b.

## 7.3 Poziomy prywatności

Poufność można uzyskać na kilku poziomach abstrakcji danych:

1. **poziom indywidualny** (ang. *individual level of privacy preserving*)
   Właściciel danych nie zezwala na ujawnienie poszczególnych rekordów. Dozwolone jest budowanie dowolnych modeli w oparciu o całość danych, ale pod warunkiem, że nie będzie możliwe sprawdzenie rzeczywistych wartości w poszczególnych atrybutach. Typowym przykładem może być chęć ochrony wynagrodzenia w tabeli zawierającej dane pracowników. Szansa na powiązanie realnej pensji z konkretnym pracownikiem byłaby naruszeniem jego prywatności.

2. **poziom zagregowany** (ang. *aggregate level of privacy preserving*)
   Instytucja dzieląca się pewną kolekcją próbek może mieć świadomość jaka wiedza jest w nich zawarta. Jednocześnie może nie być zainteresowana ujawnianiem jej w całości. Odnosząc się do samych reguł można to ująć jako niechęć do odtajniania konkretnych implikacji. Najczęstszym sposobem zachowania prywatności w takich przypadkach jest odpowiednie zmniejszenie wsparcia dla wybranych reguł.

## 7.4 Założenia i idee

Krótki wstęp teoretyczny przedstawiony powyżej służy wyjaśnieniu ustanowionych dla tej pracy założeń. Po pierwsze rozwiązanie końcowe będzie pracować wyłącznie na bazie scentralizowanej. Uzasadnieniem jest fakt, że każdy z typów rozproszenia danych można bez większego trudu przekształcić do bazy przechowywanej centralnie (chociażby za pośrednictwem zaufanej instytucji).

Sam proces eksploracji będzie wykorzystywać model nieinteraktywny, a więc ujawnioną w całości bazę (po poddaniu jej procesom zapewniania prywatności). Natomiast jedynym celem jaki jest postawiony to zapewnienie prywatności na poziomie indywidualnych danych.

**Idea:** Ujawnienie zagregowanych informacji bez wyjawiania indywidualnych wartości poszczególnych rekordów. A w konsekwencji oparcie reguł o te zagregowane dane.

**Wytyczna:** Prywatność indywidualnych informacji opiera się na niepewności co do ich oryginalnych wartości. Im trudniej jest przewidzieć jakie one były tym większa jest prywatność.

## 7.5 Kategorie zapewniania prywatności

Wszystkie dostępne techniki zapewniania prywatności można zgrupować w kilka ogólnych kategorii:

1. **Blokowanie** (ang. *blocking*)
   Jest to najprostszy z dostępnych sposobów, stosowany w sytuacjach

kiedy poufność jest krytyczna w stosunku do niewielkiej ilości informacji. Polega na całkowitym utajnieniu wybranych wartości poprzez zastąpienie ich *symbolem braku danych* (ang. *missing value*), niekiedy nazywanym wartością *NULL*. Oczywistym skutkiem jest niemożność wykorzystania takich informacji do budowy żadnego modelu.

2. **Próbkowanie** (ang. *sampling*)
   Równie proste rozwiązanie, ale nie zapewniające prywatności wszystkich indywidualnych informacji. Wartość analizowanego atrybutu nie jest udostępniana w całości lecz w postaci losowo wybranego podzbioru. W ten sposób można ominąć rekordy, które zawierają najbardziej wrażliwe informacje.

3. **Zamiana wartości** (ang. *swapping*)
   Metoda polega na zamianie miejscami wartości dla danego atrybutu. Powstaje nowy zbiór z wymieszanymi wartościami, ale nadal nie zmienionymi. Pozwala to budować modele opisujące zagregowane informacje, ponieważ całościowy rozkład wartości nie jest zmieniany. Takie przeniesie danych może być wadą, gdyż tracone jest powiązanie pomiędzy atrybutami na poziomie poszczególnych rekordów.

4. **Zakłócanie wartości** (ang. *value distortion*)
   Jedna z najpopularniejszych i najbardziej uniwersalnych metod. Polega na modyfikacji każdego pojedynczego elementu przy pomocy liczby losowanej z dobrze znanego rozkładu. Ostatecznie ujawniane są wartości zmodyfikowane i parametry rozkładu modyfikującego. Możliwe jest dzięki temu odtworzenie rozkładu pierwotnych danych ale bez perspektyw na poznanie oryginalnych wartości pojedynczych rekordów.

5. **Agregacja** (ang. aggregation)
   Wymieniana jest tutaj na ostatniej pozycji lecz w ramach tej pracy jest najistotniejsza. Dane, które są niechętnie ujawniane należy podzielić na fragmenty. Innymi słowy, znając dziedzinę całego zbioru, można dokonać jej podziału na przedziały. Następnie dla każdego elementu ze zbioru wystarczy udostępnić wyłącznie przedział do którego należy. Dzięki temu oryginalne wartości zostają niejawne, a badanie całego zbioru nadal jest możliwe (oczywiście godząc się z pewną utratą informacji).

Nie należy zapominać o możliwości łączenia różnych metod z powyższej listy kategorii w zależności od potrzeb.

# 8 Rozważania na temat prywatności

Po rozważaniach na temat analizy skupień w kontekście odkrywania reguł ilościowych przyszedł czas na dyskusję na temat bezpieczeństwa i prywatności danych podczas tego procesu.

Celem badacza jest odkrycie wszystkich reguł ilościowych ukrytych w danych. Przykładowo niech ze zbioru danych wynika następująca implikacja:

$$Wynagrodzenie : [10000, 14000] \Rightarrow dom = 2$$

Jedyne czego pragnie badacz to zdobycie takich reguł. Nie interesuje go który rekord zawierał akurat tak duże wynagrodzenie jak 14 tysięcy. Oraz dodatkowo nie interesuje go jaki jest rozkład wartości na przedstawionym przedziale [10 000, 14 000], bo widocznie mniejszy przedział nie zawierał tylu elementów, by przekroczyć próg minimalnego wsparcia.

## 8.1 Idealne reguły ilościowe

Żeby zdobyć teoretycznie wszystkie (dobre i interesujące) reguły przy zadanym progu wsparcia można założyć, że właściciel danych dysponuje idealną metodą eksploracji reguł asocjacyjnych. Działa ona bez zarzutu i zwraca tylko najlepsze rezultaty. Dysponując wynikami takiej utopijnej procedury można zauważyć, że atrybut numeryczny pojawia się w regułach wyłącznie w postaci przedziałów. Gdyby koszt zakupu tak idealnych rezultatów był za duży, to przebiegły badacz mógłby poprosić o pierwotny zbiór danych ze zagregowanymi atrybutami ciągłymi według uzyskanych z idealnych reguł przedziałów. Taką sprzedaż można określić mianem zachowującej prywatność na poziomie indywidualnych wartości. Jednocześnie nabywający byłby w stanie za pomocą metod binarnej eksploracji danych [3] odkryć te same reguły co właściciel za pomocą utopijnej procedury.

Cała ta fikcyjna sytuacja służy wyciągnięciu podstawowego wniosku:

Odpowiednio zagregowane dane numeryczne są wystarczającym źródłem do powołania reguł ilościowych. Co ważniejsze, sama agregacja jest jednocześnie bardzo dobrym sposobem zapewnienia prywatności (rozdział Kategorie zapewniania prywatności podpunkt 5).

Innymi słowy, dokonując rozsądnego podziału rozwiązuje się dwa problemy niemal jednocześnie. Jedynym kłopotem jest niestety brak jasnej definicji optymalnego grupowania dla ustalonego zbioru danych, o czym była już mowa wcześniej.

Mianem agregacji można określić również dyskretyzację. Dlatego dotychczasowe rozważania na temat analizy skupień wpisują się niejako w rozwiązanie problemu zachowania prywatności. Z niewielkim wyjątkiem. Otóż podział

idealny dla eksploracji reguł ilościowych, nie musi być dobry z punktu widzenia zapewnienia prywatności na poziomie indywidualnym. Ten problem zostanie poruszony później.

Istnieją oczywiście też sposoby, które pozwalają, na zapewnienie prywatności bez konieczności agregacji wartości. Są nimi metody zakłócania oryginalnych wartości liczbowych. Przez zakłócenie rozumiane jest dodanie do każdej próbki losowej wartości z pewnego rozkładu. Są to bardzo skuteczne metody, ale w niniejszym zastosowaniu nie muszą być używane, bo agregacja i tak jest tutaj niezbędna. Więcej o metodach zakłócających w pracy [5].

## 8.2 Miara prywatności

W tej pracy wykorzystywany będzie sposób pomiaru stopnia prywatności wprowadzony w [5]. Wskazuje on dokładność z jaką można oszacować pierwotną wartość atrybutu na podstawie zmodyfikowanych danych.

> *„Jeżeli z dokładnością c% można stwierdzić, ze oryginalna wartość atrybutu leży w przedziale $\langle x_1, x_2 \rangle$, wówczas długość przedziału, czyli $x_2 - x_1$, określa wielkość prywatności przy poziomie c%."*
>
> źródło: [1, 5]

Dla przykładu, jeśli atrybut wykazuje prywatność 7 dla poziomu pewności 100 %, to z taką właśnie pewnością na podstawie zmodyfikowanej wartości, oryginalna wartość znajduje się w przedziale o długości 7. W przypadku dyskretyzacji, po modyfikacji otrzymuje się wyłącznie pakiet przedziałów do których należały pierwotne wartości. Oznacza to więc, że z pewnością 100% można wyłącznie stwierdzić, że oryginalna wartość jest gdzieś wewnątrz danego przedziału. Stąd prywatność dyskretyzacji jest równa szerokości wytworzonych zakresów dla poziomu 100%. Porównanie agregacji z metodą zakłócania wartości pokazuje tabela 4.

Oznaczenie $W$ to długość przedziałów zastosowana przy dyskretyzacji. Jest tu niejawne założenie, że są one równych szerokości. W celu ujednolicenia miary, zakłada się, że jeśli szerokości grup nie są takie same, to dla obliczenia miary prywatności stosuje się ich średnią arytmetyczną. Parametr $\sigma$ to odchylenie standardowe rozkładu normalnego, którym rozpraszane są wartości oryginalne w przypadku stosowania metody zakłócania wartości (rozdział Kategorie zapewniania prywatności podpunkt 4).

Jak widać, zakłócenie rozkładem normalnym ma większe możliwości zwiększenia prywatności – wystarczy zwiększyć odchylenie standardowe. Natomiast w przypadku dyskretyzacji należałoby poszerzyć przedziały co wpłynie na większą utratę informacji i niewątpliwie wpłynie negatywnie na budowane z tych danych modele.

| Poziom pewności | 50 % | 95 % | 99,9 % |
|---|---|---|---|
| Dyskretyzacja | $0,5 * W$ | $0,95 * W$ | $0,999 * W$ |
| Zakłócanie rozkładem normalnym | $1,34 * \sigma$ | $3,92 * \sigma$ | $6,58 * \sigma$ |

Tablica 4: Wartości miary prywatności dla różnych poziomów pewności.

### 8.3 Wartości binarne

Agregacja wartości dla atrybutu ciągłego oraz wykonana transformacja (rozdział *Transformacja na wartości binarne*) sprawiają, że uzyskiwany jest zbiór cech włącznie binarnych. Stąd można dojść do wniosku, że jeśli prywatność zapewniona przez dyskretyzację nie jest wystarczająca to można zastosować metody zachowywania prywatności działające na cechach dwuwartościowych.

**Losowe zakłócanie atrybutów binarnych** (ang. *randomisation-based method for distorting binary attributes*)

W przypadku zbiorów dwuwartościowych modyfikacja przebiega w następujący sposób:

Każda oryginalna wartość (0 albo 1) jest pozostawiana bez zmian z prawdopodobieństwem $p$ albo zamieniana na przeciwną z prawdopodobieństwem $1 - p$. Każdy atrybut może być rozproszony z innym prawdopodobieństwem. W takim przypadku badacz otrzymuje bazę zmodyfikowaną oraz wartość $p$ [1].

## 9 Quality Threshold Clustering

Quality Threshold Clustering (w skrócie QT Clustering albo QTC) to algorytm grupowania pierwotnie stworzony do analizy genów [13][11]. Jego priorytetem jest zapewnienie odpowiedniego poziomu jakości dla tworzonych grup. Nie wymaga specyfikowania potencjalnej liczby skupisk jakie wystąpią w bazie danych. Wręcz przeciwnie, pozwala na odkrycie ich naturalnej liczby. Wielkością grup wynikowych steruje parametr maksymalnej średnicy skupiska. Podstawowa idea opiera się na znalezieniu grup kandydujących na podstawie każdego punktu ze zbioru. Każda z nich musi spełniać wymaganie jakościowe i co ważniejsze każda z nich jest budowana w oparciu o pełny zbiór danych (jeszcze niepogrupowanych). Ostatecznie spośród wszystkich kandydatów wybierany jest najlepszy. Elementy w nim zawarte są usuwane z grupowanego zbioru (jako już przydzielone), a cały proces jest powtarzany. Cała procedura jest zapisana w pseudokodzie poniżej – Algorytm 1

[11] Informacje są zaczerpnięte również z encyklopedii uczenia maszynowego [11] (hasło: „Quality Threshold Clustering")

**Algorytm 1** Procedura grupowania Quality Threshold Clustering

**funkcja** QTCLUSTERING(G, d)
  **if** $|G| \leq 1$ **then return** G
  **end if**
  **for all** $i \in G$ **do**
    zbiór $A_i \leftarrow \{i\}$ ▷ $A_i$ jest i-tym kandydatem
    **while** $A_i \neq G$ **do**
      znajdź $j \in (G - A_i)$ dla którego ŚREDNICA$(A_i \cup j)$ jest minimalna
      **if** ŚREDNICA$(A_i \cup j) <$ d **then**
        $A_i \leftarrow A_i \cup \{j\}$
      **else**
        **break while**
      **end if**
    **end while**
  **end for**
  zbiór $C \leftarrow$ NAJWIĘKSZY_ZBIÓR_Z( $A_1, A_2, A_3, \ldots, A_{|G|}$ )
  **return** { $C$, QTCLUSTERING$(G - C, d)$ }
**end funkcja**

## 9.1 Właściwości

W oryginalnym algorytmie mianowanie najlepszego kandydata opiera się na wyborze największego pod względem liczby elementów. Niemniej jednak modyfikacja tego warunku może być wskazana w zależności od zastosowania algorytmu. Idea niezależnego generowania kandydatów z całej dostępnej puli próbek sprawia, że ta metoda ma wiele unikalnych zalet. Przede wszystkim jest niezależna od kolejności budowania grup ani występowania danych. Dodatkowo nie zawiera żadnego elementu losowego. To wszystko sprawia, że jest to jeden z niewielu, w pełni deterministyczny algorytm grupujący. Uruchamiając go wielokrotnie, zawsze jest pewność tego samego wyniku. Algorytm gwarantuje też, że wszystkie grupy wynikowe będą spełniały przedstawione wymaganie jakościowe (w tym przypadku maksymalną średnicę). Na dodatek wybierane są skupiska w kolejności od najlepszego (np. największego) do najsłabszego (np. najmniejszego). Daje to gwarancję, że najbardziej istotna struktura danych zostanie odkryta poprawnie.

Najważniejszą zaletą algorytmu jest fakt odkrywa on naturalne skupiska w zadanym zbiorze z dokładnością do ustawionej jakości. Oznacza to, że możliwe jest poznanie faktycznego obrazu struktury wartości, niesfałszowanej i niczym nie wymuszonej. Jednocześnie parametr sterujący jest dość intuicyjny dla badaczy. Dla kontrastu wystarczy wspomnieć np. konieczność ustalenia

poziomu gęstości dla algorytmu DBSCAN (rozdz. Metody podziału). Tutaj ustalenie średnicy może w pierwszym momencie nie być łatwe, ale szybkie zapoznanie z dziedziną wartości (chociażby zakresu min-max) wystarczy by zapanować nad sytuacją.

Algorytm ten ma delikatne podobieństwo do grupowanie hierarchicznego z kompletnym łączeniem (ang. *complete linkage hierarchical clustering*), ale uśredniając produkuje zdecydowanie większe grupy. Jak zaznaczają jego autorzy: L. Heyer, S. Kuglyak oraz S. Yooseph [13], lokalne decyzje podczas budowy kandydatów nie mają dużego wpływu na końcowy wynik. Jedynie grupa w danym kroku najsilniejsza jest istotna dla całej analizy skupień. Autorzy przypuszczają, że metoda jest mniej wrażliwa na niewielkie zmiany w danych, niż grupowanie hierarchiczne. Co akurat było istotną zaletą w przypadku ówczesnego zastosowania z powodu konieczności filtrowania i usuwania niektórych próbek genów.

## 9.2 Parametr

Algorytm QT wymaga zdefiniowania jakości grup wynikowych w postaci ograniczenia na ich średnicę. Celem grupowania w tym przypadku ma być dostarczenie zbiorów, które będą miały liczność przynajmniej na granicy minimalnego wsparcia. Zakłada się więc, że przed analiza skupień znane są parametry dalszej eksploracji (w tym minSup). Według wymagań, analizie będą poddawane poszczególne atrybuty, czyli dane jednowymiarowe. Konieczne jest krok ich wstępnego przetworzenia, aby poznać dziedzinę wartości. W tym przypadku wystarczająca jest znajomość wartości minimalnej (niech będzie oznaczona przez atrMin) oraz maksymalnej (atrMax). Pierwszym podejściem do rozwiązania tego zadania jest dość naiwne założenie iż dane mają w przybliżeniu rozkład jednostajny. To uprawnia to następującego wyboru:

$$prog = (max - min) \cdot minSup$$

Gdzie *minSup* to wartość minimalnego wsparcia dla reguł asocjacyjnych zdefiniowana w postaci ułamka.

## 9.3 Algorytm QT w kontekście reguł ilościowych

W przypadku prowadzenia grupowania dla wyszukiwania reguł ilościowych niezwykle istotne jest by, nie zostało ono przeprowadzone w sposób sztuczny, lecz by było kierowane danymi. Celem jest odkrycie realnie pojawiających się skupisk, dzięki czemu odkryte reguły będą maksymalnie interesujące. Ten argument przeważa na korzyść algorytmu QT. Niestety wykorzystując go

wprost przy grupowaniu zbiorów danych dla eksploracji reguł ilościowych napotyka się kilka przeszkód.

**Brak gwarancji minSup**

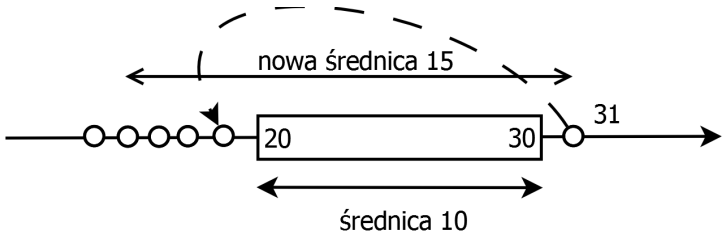
Zbiory występujące w realnym świecie są nieregularne. Dlatego nie można zakładać parametru średnicy w oparciu o jednostajność zbioru. Nawet pomimo intuicyjnie poprawnego odkrycia zawartych w nim skupisk pojawia się dodatkowy problem. Z właściwości algorytmu wynika gwarancja, że każdy kolejny wyznaczony przedział będzie mniej liczny. Monotoniczność tego procesu sprawia, że w pewnym momencie rozmiar grupy spada poniżej minimalnego wsparcia. Dlatego zbiory końcowe są praktycznie nieprzydatne z punktu widzenia eksploracji reguł asocjacyjnych.

**Zwiększanie progu jakościowego**

Pierwszym przychodzącym na myśl rozwiązaniem jest zwiększanie maksymalnej rozpiętości średnicy zbioru od razu w przypadku, kiedy uzyskany przedział przestanie zaspokajać minimalne wsparcie. Teoretycznie ponowne wyszukanie grupy z większym parametrem progu powinno zakończyć się znalezieniem grupy, której rozmiar będzie co najmniej większy od poprzednio niezadowalającego. Ten intuicyjny wniosek został zweryfikowany w praktyce i niestety nie jest on prawdziwy.

Pierwszy błąd logiczny to fakt, że progu nie można zwiększyć ani trochę. Żeby to uzasadnić należy wyobrazić sobie następującą sytuację. Początkowy próg jakości ustawiony na 10. Minimalne wsparcie to 5 obiektów (wyjątkowo jednostką są elementy, a nie ułamek całości). W zbiorze pojawia się ostatni przedział o rozpiętości 10 - niech to będzie zakres $\langle 20; 30\rangle$. Kolejny przedział jest odrzucany bo zawiera mniej niż pięć elementów. Niech maksymalna średnica zostanie zwiększona do 15. Jeden ze zbiorów kandydujących będzie tworzony z początkowego punktu 31. Niestety w „polu widzenia" tego punktu są również punkty po drugiej stronie utworzonego już przedziału $\langle 20; 30\rangle$. Istnieje więc ryzyko, że po modyfikacji parametru będą tworzyć się „przedziały nad przedziałami" (zawierając się w sobie). Całą sytuację prezentuje rysunek 2.

Rozwiązanie końcowe musi unikać powyższego zagrożenia, ale istnieje jeszcze jeden problem. Zaimplementowany oryginalny algorytm QT nie jest w stanie wytworzyć wyłącznie zadowalających rozmiarem przedziałów z powodu „problemu skrawków", który jest opisany dalej. Ten problem dotyczy wszystkich metod, które dokonują grupowania iteracyjnie.




Rysunek 2: Szkic wskazujący zagrożenie zwiększania parametru algorytmu QT w trakcie działania.

# 10 Problem skrawków

Skrawek (ang. *snippet*) to według słownika języka polskiego „resztka, pozostałość po cięciu". W grupowaniu algorytmem QT jest to grupa wartości, która nie może uzyskać żądanego minimalnego wsparcia z powodu sąsiadujących z nią utworzonych już wcześniej grup.

**Przykład**

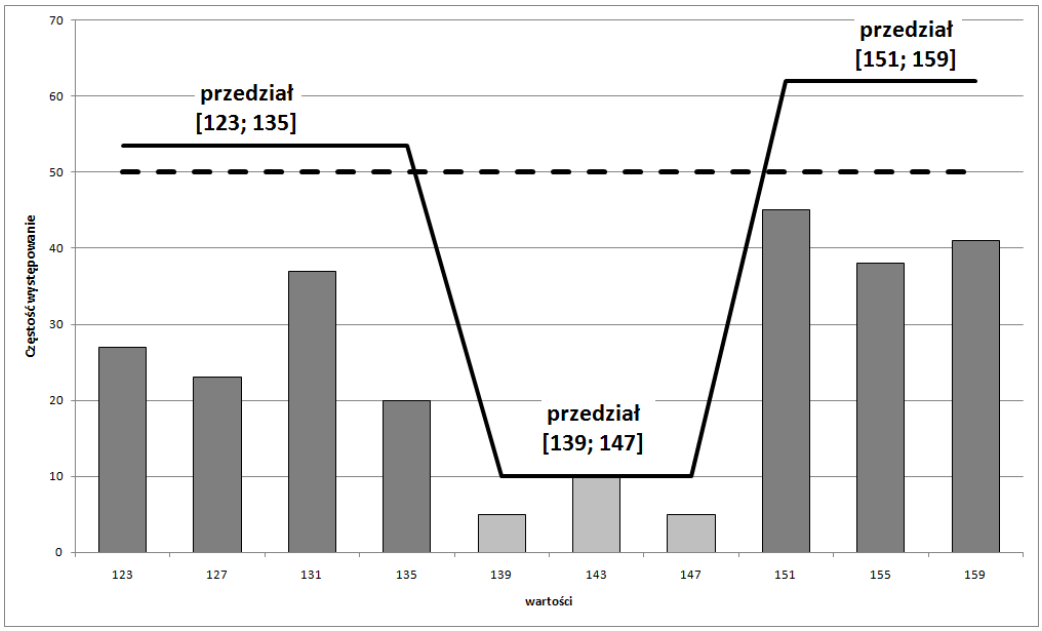
Niech w wyniku działania algorytmu wyszukującego grupy w sposób iteracyjny powstaną dwa przedziały $G_1 = \langle a_1, a_2 \rangle$ oraz $G_2 = \langle a_3, a_4 \rangle$, przy czym $a_2 < a_3$. Jeśli dokładnie pomiędzy nimi znajduje się zbiór obiektów $S$ to jest tak jakby odizolowany od punktów bazy danych przez sąsiedztwo $G_1$ oraz $G_2$. Zbiór $S$ można nazwać skrawkiem, wtedy jeśli jego wsparcie będzie niższe od zakładanego minimum, a więc dalszy jego podział będzie daremny. Przykład zbioru tego typu można zaobserwować na rysunku 3. Przedział $[139, 147]$ jest otoczony przez zbiory silne, ale sam nie spełnia wymogu wsparcia.

## 10.1 Rozciąganie przedziałów

Jedną z najprostszych metod radzenia sobie z problemem opisanym wyżej jest polityka rozciągania przedziałów. W przypadku napotkania skrawka może on być włączony do jednego z dwóch swoich sąsiadów. Istnieją dwa przypadki tego procesu:

1. Przedział rozpatrywany ma tylko jednego sąsiada - taka sytuacja zdarza się „na końcach" dziedziny. Czyli dla wartości najwyższych i najniższych. Jak wynika z przeprowadzonych eksperymentów, skrajne elementy dziedziny prawie zawsze tworzą skrawki. Jest to uzasadnione chociażby tym, że to właśnie ekstrema dziedziny wartości zawierają najczęściej próbki „fałszywe" i odstające od typowych obiektów.

Przedział znajdujący się w takiej konfiguracji musi być przyłączony do swo-




Rysunek 3: Histogram danych (wykres kolumnowy). Przykład skrawka na przedziale [139;147]. Wykres linią ciągłą przedstawia podział na grupy (płaskie fragmenty) z jednoczesną łączną sumą obiektów w grupie (podzieloną przez dwa dla przejrzystości wykresu). Wykres linią przerywaną to minimalne wsparcie (również podzielone przez dwa). Jak widać, tylko środkowy przedział nie przekracza minimalnego wsparcia.

jego sąsiada. Innymi słowy, znaleziona wcześniej grupa zostanie tak rozciągnięta, aby obejmować też dany niewielki zbiór.

2. Przedział rozpatrywany leży pomiędzy dwoma oznaczonymi już grupami. Ten przypadek jest omawiany od samego początku i przedstawiony na rysunku 3. Teraz rozwiązanie nie jest aż tak oczywiste jak w podpunkcie wyżej. Rozsądnie jest podzielić rozpatrywany przedział na dwa obszary, tak żeby rozciągnąć obie sąsiadujące grupy. Należy w tym przypadku oprzeć się o pewną heurystykę, ponieważ błędnym rozwiązaniem byłby podział na dwie równoliczne połowy. Wydaje się, że najlepiej sprawuje się rozpatrywanie bliskości. Niech każdy nieprzydzielony punkt trafi do tej grupy, do której odległość jest najmniejsza. Ze wszystkich znanych metod wiązania [10] skuteczna będzie metoda pojedynczego wiązania. W jej ramach odległość pomiędzy dwoma grupami jest równa dystansowi pomiędzy dwoma najbliższymi punktami. W tym przypadku będzie to odległość rozpatrywanego punktu do granicy przedziału. To wiązanie może być zastosowane tylko pod warunkiem, że granica grupy zostanie rozciągnięta dopiero po rozdzieleniu wszystkich elementów skrawka.

Równie dobrym wiązaniem może być metoda środków - obliczanie dystansu do środka grupy. Chociaż to rozwiązanie ma w wyjątkowych okolicznościach gorsze werdykty niż nakazywałaby intuicja, to jednak jest stabilniejsze od wiązania pojedynczego. Stabilność jest skutkiem wysokiej bezwładności środka grupy.

## 10.2 Nakładanie przedziałów

Całkowicie oryginalnym wyjściem z impasu wprowadzonego przez skrawki może być odpowiednie zaadoptowanie techniki grupowania z nakładaniem (ang. overlapping clustering). Ogólnie technika ta pozwala na nierozłączne dzielenie dziedziny. Innymi słowy, przedziały mogą się ze sobą „zazębiać". Takie podejście można wprost wykorzystać w przypadku problematycznych skrawków. Wystarczy rozszerzyć zbyt mały przedział wykorzystując elementy, które już były użyte. Zapis przedziałów nakładających umożliwia tylko algorytm analizy skupień, który od razu dokonuje binaryzacji atrybutów. Oryginalna wartość atrybutu ciągłego będzie należeć do dwóch przedziałów jednocześnie. Rekord ją zawierający, po transformacji będzie wspierać dwa binarne atrybuty zamiast jednego.

Nakładanie przedziałów ma ważną zaletę: Nie narusza struktury odkrytych wcześniej grup, które bardzo silnie odpowiadają skupiskom naturalnie zawartych w zbiorze. Mimo to, przedziały którym pierwotnie brakowało trochę wsparcia mogą też wziąć udział w procesie eksploracji i wpłynąć na budowę nieoczekiwanych reguł.

Jedynym minusem metody, jest konieczność unikania całkowitego zawierania zbiorów. Powołanie dwóch takich atrybutów mogłoby doprowadzić do powstania reguły między nimi, co byłoby sprzeczne z logiką reguł asocjacyjnych.

# Literatura


[1] Piotr Andruszkiewicz, *Privacy Preserving Classification and Association Rules Mining over Centralised Data*, Warsaw, 2011

[2] Piotr Andruszkiewicz, *Privacy Preserving Classification for Continuous and Nominal Attributes*, Intelligent Information Systems, 2008

[3] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami, *Mining association rules between sets of items in large databases.* In Peter Buneman and Sushil Jajodia, editors, SIGMOD Conference, pages 207–216. ACM Press, 1993

[4] Rakesh Agrawal, Ramakrishnan Srikant, *Mining Quantitative Association Rules in Large Relational Tables*, SIGMOD Conference, ACM Press, 1996

[5] Ramakrishnan Srikant Rakesh Agrawal. Privacy-preserving data mining. In Proc. of the ACM SIGMOD Conference on Management of Data, pages 439–450. ACM Press, May 2000.

[6] Jain, A.K., Murty M.N., and Flynn P.J. (1999): Data Clustering: A Review, ACM Computing Surveys, Vol 31, No. 3, 264-323. http://www.cs.rutgers.edu/ mlittman/courses/lightai03/jain99data.pdf

[7] Ying Yang, Geoffrey I. Webb, Xindong Wu, Discretization Methods, Springer, 2005

[8] Lior Rokach, *A survey of Clustering Algorithms*, Data Mining and Knowledge Discovery Handbook, Springer, Londyn, 2010

[9] Shyam Boriah, Varun Chandola, Vipin Kumar, *Similarity Measures for Categorical Data: A Comparative Evaluation*, 8 SDM SIAM, Atlanta, 2008

[10] StatSoft (2006). *Elektroniczny Podręcznik Statystyki PL*, Krakow, WEB: http://www.statsoft.pl/textbook/stathome.html.

[11] Encyclopedia of Machine Learning, Springer, 2010

[12] Brian Lenty, Arun Swamix, Jennifer Widom, *Clustering Association Rules*

[13] Heyer, L., Kruglyak, S., Yooseph, S. (1999). Exploring expression data: Identification and analysis of coexpressed genes. Genome Research, 9, 1106–1115

[14] Kazimierz Krzysztofek, Marek Szczepański, *Zrozumieć rozwój - od społeczeństw tradycyjnych do informacyjnych*, strony 186–189, Wydawnictwo Uniwersytetu Śląskiego, Katowice, 2002

[15] Justyna Berezowska, Michał Huet, *Społeczeństwo informacyjne w Polsce. Wyniki badań statystycznych z lat 2009-2013*, strony 20–22, Główny Urząd Statystyczny, Szczecin, 2014