INSTYTUT BADAŃ SYSTEMOWYCH
POLSKIEJ AKADEMII NAUK

ANALIZA SYSTEMOWA W FINANSACH I ZARZĄDZANIU

Wybrane problemy
Tom 11

Pod redakcją
Jerzego HOŁUBCA

Warszawa 2009

INSTYTUT BADAŃ SYSTEMOWYCH
POLSKIEJ AKADEMII NAUK

ANALIZA SYSTEMOWA W FINANSACH I ZARZĄDZANIU

Wybrane problemy
Tom 11

Pod redakcją
Jerzego HOŁUBCA

Warszawa 2009

Wykaz opiniodawców artykułów zamieszczonych w niniejszym tomie:

prof. dr hab. inż. Jerzy HOŁUBIEC
dr inż. Lech KRUŚ
doc. dr hab. inż. Wiesław KRAJEWSKI
doc. dr hab. Jacek MALINOWSKI
dr inż. Edward MICHALEWSKI
prof. dr Adam SKOREK
dr hab. Ryszard SMARZEWSKI
prof. dr hab. inż. Andrzej STRASZAK
dr Dominik ŚLĘZAK
prof. dr hab. inż. Stanisław WALUKIEWICZ
doc. dr hab. Sławomir ZADROŻNY

ISBN 9788389475220

Druk: Zakład Poligraficzny Jerzy Kosiński, Warszawa

OSTRE I ROZMYTE GRUPOWANIE DANYCH ORAZ JEGO ZASTOSOWANIA

Magdalena Laskowska
Studia Doktoranckie IBS PAN

It's in the nature of man to split the set of objects that function in his life. He's not able to comprehend the magnitude of information that he must process, without grouping them together.

There are many ways of grouping. Good example is rough and fuzzy clustering. Discussing them and their practical appliance in database systems is what this article will be dedicated to.

Wprowadzenie

Codziennie dociera do nas olbrzymia ilość informacji. Tylko część z nich jest dla nas ważna. Człowiek bez problemu potrafi wydobyć z obrazu (pewnego zbioru danych) interesujące go obiekty. Obiektami mogą być przedmioty na zdjęciu, postacie, nawet ich emocje, czy schorzenia. Jedno spojrzenie wystarczy, by podzielić obiekty np. według nastroju, wzrostu czy wieku.

1. Grupowanie danych

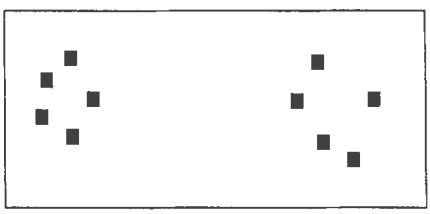
Grupowanie danych jest to proces mający na celu podział zbioru danych na pewną liczbę grup tak, aby były zachowane dwie cechy tego podziału:

- *homogeniczność w grupach*, tzn. dane w obrębie danej grupy powinny być jak najbardziej podobne do siebie,
- *heterogeniczność pomiędzy grupami*, tzn. dane należące do różnych grup powinny być jak najbardziej różne od siebie[1].

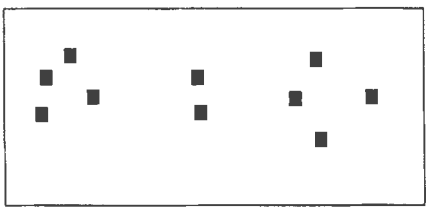
Rozważmy rysunki 1 i 2. Chcemy dokonać podziału zbiorów przedstawionych na tych rysunkach na dwa podzbiory biorąc pod uwagę sąsiedztwo elementów. Zatem chcemy podzielić zbiór na podzbiory, by elementy należące do jednego

[1] Zob. Rutkowski L. (2005): Metody i techniki sztucznej inteligencji, Wydawnictwo Naukowe PWN, Warszawa, 300s..

podzbioru były sobie „bliskie" (cecha homogeniczności w grupie), zaś elementy zaliczone do drugiego podzbioru były od siebie „odległe" (cecha heterogeniczności).



Rys. 1. Przykładowy zbiór danych



Rys.2. Przykładowy zbiór danych

Dla rysunku 1 podział na dwa podzbiory nie wymaga znajomości miar odległości i jest intuicyjny. Jednak w przypadku rysunku 2 problem podziału na dwie grupy nie jest już tak oczywisty, bo nie wiadomo do której grupy zaliczyć dwa „środkowe" elementy. Ile jest wszystkich możliwych podzbiorów dla tego zbioru danych? Liczba sposobów, na jakie możemy przeprowadzić podział M obiektów na c grup, wynosi[2]:

$$\frac{1}{c!}\sum_{i=1}^{c}\binom{c}{i}(-1)^{(c-i)}i^{M} \qquad (0.1)$$

Dla zbioru 10-elementowego przedstawionego na rysunkach 1 i 2 mamy 511 możliwości podziału na dwie grupy. Przy podziale na trzy grupy mamy już 9.330 możliwości. Widzimy, że przeglądanie wszystkich możliwości jest zadaniem trudnym, dlatego konieczne są metody optymalnego podziału.

W celu znalezienia optymalnego podziału stosuje się często miary odległości. Zmierzenie odległości pomiędzy obiektami ułatwia podział zbioru na podzbiory. Jeśli dla obiektów z rysunku 2 zmierzono by odległość, to obiekty, które wydają się mieścić na środku, znajdowałyby się bliżej którejś z grup i wtedy tam, by je przyporządkowano.

Najczęściej stosowaną miarą odległości jest *miara euklidesowa*, czyli geometryczna odległości między dwoma punktami w przestrzeni **X**. Rozważmy dwa punkty $x_d = (x_{d1}, x_{d2}, \ldots, x_{dn})$ oraz $y_i = (y_{i1}, y_{i2}, \ldots, y_{in})$, takie że $x_d, y_i \in X$. *Odległość euklidesową* między tymi dwoma punktami definiujemy następująco:

[2] Jest to tzw. Liczba Stirlinga II rodzaju.

$$d(x_d, y_i) = \sqrt{\sum_{k=1}^{n} (x_{dk} - y_{ik})^2} \tag{0.2}$$

Oczywiście im mniejsza jest ta odległość, tym bardziej obiekty są homogeniczne względem siebie.

Poza odległością euklidesową, często stosuje się odległość przeciętną oraz odległość miejską[3]. W przypadku zmiennych binarnych miara ta jest określana miarą Hamminga[4] i służy do określenia ilości bitów, o które różnią się dwa ciągi bitów. W zależności od modelowanego problemu należy dobrać tak miarę, aby jak najlepiej odzwierciedlić rzeczywistość.

Po obliczeniu odległości pomiędzy każdymi dwoma obicktami ze zbioru, przyporządkowujemy je do grup tak, aby odległość względem siebie była możliwie jak najmniejsza.

Istnieje wiele rodzajów grupowania mających szereg różnych zastosowań. Najpopularniejsze z nich to grupowanie ostre oraz rozmyte. Rutkowski L.(2005) wprowadza również trzeci rodzaj – grupowanie posybilistyczne, jako modyfikację grupowania rozmytego. Ostre grupowanie wymusza na obiekcie całkowitą przynależność do grupy lub brak tej przynależności. Dokonujemy podziału na c grup A_i, gdzie $i=1,...,c$ zachowując przy tym warunki (1.3)-(1.5).

$$\bigcup_{i=1}^{c} A_i = \mathrm{X}, \tag{0.3}$$

$$A_i \cap A_j = \varnothing,\ 1 \le \mathrm{i} \ne \mathrm{j} \le \mathrm{c}, \tag{0.4}$$

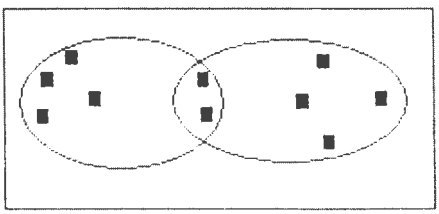
$$\varnothing \subset \mathrm{A_i} \subset \mathrm{X}, 1 \le i \le c. \tag{0.5}$$

Zbiór wszystkich grup tworzy zbiór oryginalny, sprzed operacji grupowania. Grupy są rozłączne, tzn. nie zawierają tych samych obiektów, żadna z nich nie jest pusta, jak również nie zawiera całego zbioru danych.

Dla danych z rysunku 1 przykład ostrego grupowania przedstawiono na rysunku 3. Niestety w przypadku danych z rysunku 2 dokonanie podziału na dwie grupy nie jest tak jednoznaczne. Obiekty znajdujące się na „środku" można albo nie przyporządkowywać do żadnej z grup albo przyporządkować je do obu jednocześnie (patrz rysunek 3). W przypadku obu tych rozwiązań nic są zachowane

[3] Inne nazwy tej miary to miara taksówkowa lub wielkomiejska. Jest to również szczególny przypadek metryki Minkowskiego.
[4] Inna nazwa to odległość Hamminga.

wszystkie warunki (1.3)-(1.5), zatem grupowanie ostre nie jest dobrym rozwiązaniem modelowanego problemu.



Rys. 3. Przykład ostrego grupowania danych

Najczęściej rozpatrywany zbiór nie pozwala na podział danych tak jednoznaczny, ponieważ obszary występowania grup mogą zachodzić na siebie. Naturalnym rozszerzeniem podziału ostrego jest podział rozmyty. Opiera się on na założeniu, że obiekty mogą należeć do wielu grup z różnymi stopniami przynależności[5].

Zarówno dla podziałów ostrych i rozmytych stosuje się macierz podziału **U** o wymiarach $c \times M$, która zawiera stopnie przynależności μ_{ik} k-tej danej x_k do i-tej grupy, gdzie $k = 1,...,M$, $i = 1,...,c$.

Niech $X = \{x_1,...,x_M\}$ będzie zbiorem skończonym, zaś c liczbą całkowitą określającą ilość grup, $2 \le c \le M$. *Przestrzeń ostrego podziału* zbioru **X** definiujemy następująco:

$$Z_1 = \left\{ U \in R^{c \times M} \middle| \mu_{ik} \in \{0,1\}, \forall i,k : \sum_{i=1}^{c} \mu_{ik} = 1, \forall k : 0 < \sum_{k=1}^{M} \mu_{ik} < M, \forall i \right\}. \tag{0.6}$$

Zaś *przestrzeń rozmytego podziału* zbioru **X** definiujemy następująco:

$$Z_2 = \left\{ U \in R^{c \times M} \middle| \mu_{ik} \in [0,1], \forall i,k : \sum_{i=1}^{c} \mu_{ik} = 1, \forall k : 0 < \sum_{k=1}^{M} \mu_{ik} < M, \forall i \right\}. \tag{0.7}$$

[5] Stopień przynależności jest to liczba rzeczywista z przedziału $[0,1]$.

Przestrzeń ostrego oraz rozmytego podziału zbioru X nie dopuszcza istnienia pustych grup ani zawierających wszystkie dane ze zbioru X. Różnica pomiędzy przestrzenią ostrą a rozmytą jest taka, że w podziale ostrym obiekt należy wyłącznie do jednej grupy, a w podziale rozmytym obiekt może jednocześnie należeć do wszystkich grup z pewnym stopniem przynależności, ale suma wszystkich stopni przynależności musi być równa 1.

Dla danych przedstawionych na rysunku 2 ostry podział na dwie grupy (c=2) może być reprezentowany przez macierz U:

$$U = \begin{bmatrix} 1111110000 \\ 0000001111 \end{bmatrix}.$$

Pierwszy wiersz reprezentuje pierwszą grupę, zaś drugi grupę drugą. Przynależność do grupy wynosi 0, gdy obiekt nie należy do zbioru, zaś 1, gdy obiekt należy do zbioru. Obiekty, które intuicyjnie nie należą do żadnej z grup muszą być do niej na sztywno przypisane.

Dla danych przedstawionych na rysunku 2 lepszy jest podział rozmyty, gdyż intuicyjnie „środkowe elcmenty" należą zarówno do pierwszej, jak i drugiej grupy. Dla danych przedstawionych na rysunku 2 rozmyty podział na dwie grupy (c=2) może być reprezentowany przez macierz U:

$$U = \begin{bmatrix} 0.98 & 0.97 & 0.95 & 0.9 & 0.55 & 0.42 & 0.3 & 0.2 & 0.1 & 0.01 \\ 0.02 & 0.03 & 0.05 & 0.1 & 0.45 & 0.58 & 0.7 & 0.8 & 0.9 & 0.99 \end{bmatrix}$$

Podział ten wydaje się lepszy od podziału ostrego. Jednak nie jest wolny od wad. Możemy mu zarzucić, że elementy, które intuicyjnie nie należą do żadnej z grup, muszą mieć taką wartość funkcji przynależności, aby sumowała się ona do jedynki dla wszystkich grup. To ograniczenie nie daje nam dowolności w grupowaniu. Podział posybilistyczny znosi ten wymóg i pozwala na bardzo małą wartość funkcji przynależności, co w praktyce oznacza brak przynależności. Więcej informacji na ten temat można znaleźć w pracy Rutkowskigo L(2005).

Inne metody grupowania danych można znaleźć w pracy Krzyśko M., Wołyński W., Górecki T., Skorzybut M. (2008). Szczególnie interesującym przypadkiem jest sam algorytm k-średnich (algorytm tworzenia grupowania ostrego) oraz jego rozmyte rozszerzenia. Romesburg H. (2004) przedstawił szczegółowe omówienie algorytmu k-śrenich oraz innych algorytmów grupujących.

2. Zastosowanie ostrego i rozmytego grupowania danych w wydobywaniu informacji w bazach danych

Język SQL jest doskonałym przykładem wykorzystania grupowania danych ze względu na określone kryteria. We współczesnych firmach, a także korporacjach użytkownicy korzystają z grupowania danych, ale zazwyczaj nie ograniczają się tylko do tego procesu. Rozważmy jako przykład bazę danych firmy transportowej (schemat przedstawia rysunek 4). Baza ta przechowuje informacje o klientach, pracownikach, samochodach oraz zamówieniach. Dodatkowo są informacje o fakturach, trasach i towarach.

Chcąc wydobyć nazwiska pracowników, którzy mieszkają w Warszawie, a ich wynagrodzenie jest większe od średniego wynagrodzenia w omawianej firmie transportowej mamy do czynienia z grupowaniem danych.

select NAZWISKO_PR from PRACOWNIK

where MIASTO='Warszawa'

group by NAZWISKO_PR, WYNAGRODZENIE_PR

having WYNAGRODZENIE_PR>(select avg(WYNAGRODZENIE_PR) from PRACOWNIK);

Podzieliliśmy pracowników na dwie grupy w sposób ostry. Macierz U liczy dwa wiersze i tysiąc kolumn, a wygląda następująco:

$$U = \begin{bmatrix} 10\ldots000 \\ 01\ldots111 \end{bmatrix}$$

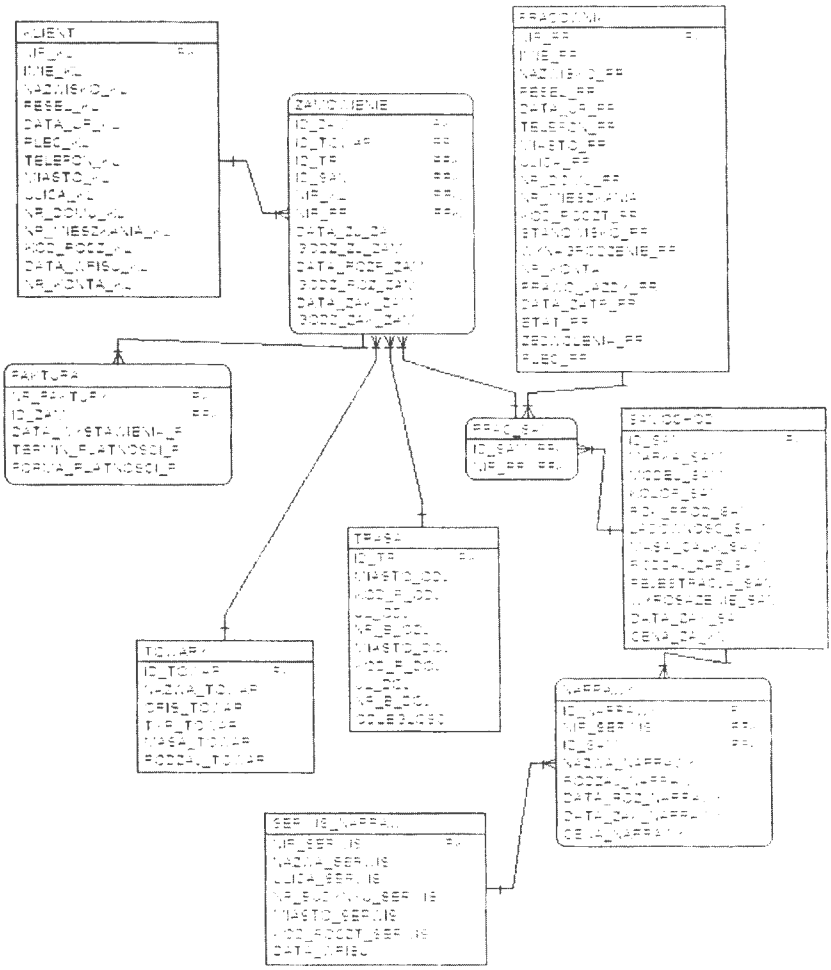
Chcąc wydobyć informacje o pracownikach, którzy mieszkają w Warszawie, a ich wynagrodzenie jest zbliżone do średniego wynagrodzenia w firmie musimy dokonać grupowania w sposób rozmyty.

select NAZWISKO_PR from PRACOWNIK

where MIASTO='Warszawa'

group by NAZWISKO_PR, WYNAGRODZENIE_PR

having WYNAGRODZENIE_PR≈(select avg(WYNAGRODZENIE_PR) from PRACOWNIK);



Rys. 4. Schemat bazy danych firmy transportowej

Oczywiście w klasycznym języku SQL oraz jego standardowych odmianach T_SQL, P-SQL nie ma operatora ≈. Rozwiązaniem tego problemu może być dodanie do tabeli *PRACOWNIK* jednego atrybutu o nazwie przynależność do zbioru rozmytego $\mu_{avg}(wynagrodzenie)$. Jako wartość tego atrybutu przypisujemy stopień zgodności wynagrodzenia pracownika ze średnim wynagrodzeniem w całej firmie. Wtedy zapytanie będzie uzupełnione o dodatkowy warunek - wymagany poziom wartości atrybutu przynależności do zbioru rozmyte-

go$\left(\mu_{avg}(wynagrodzenie) > c, c \in [0,1]\right)$. Macierz U dla tego przypadku ma wymiary takie same, jak w przypadku grupowania ostrego i wygląda następująco:

$$U = \begin{bmatrix} 0.92 & 0.85 & \ldots & 0.3 & 0.1 & 0 \\ 0.08 & 0.15 & \ldots & 0.7 & 0.9 & 1 \end{bmatrix}$$

Niestety takie rozwiązanie generuje nowe problemy, które są związane uzupełnieniem wartości dodanego atrybutu. Oczywiście można za pomocą języka SQL wyliczać wartość tego atrybutu, a następnie uzupełniać kolumnę danymi. Jednak w przypadku dużej ilości rekordów operacja ta nie jest możliwa w czasie rzeczywistym. Innym rozwiązaniem tego problemu jest lepsze przygotowanie danych tak, aby nie była konieczna modyfikacja języka SQL.

Podsumowanie

Grupowanie danych do niezwykle ważny proces, który znajduje szereg zastosowań, m.in. w bazach danych. W życiu codziennym banki wykorzystują grupowanie do dzielenia klientów na klasy ryzyka kredytowego, firmy ubezpieczeniowe sprzedając autocasco grupują klientów według wieku. Inne ciekawe przykłady zastosowania ostrego grupowania danych przedstawia Szwabowski J., Deszcz J. (2005). Dotyczą one głównie sektora budowlanego. Więcej zastosowań można znaleźć w pracy Lenkiewicz S. (2007).

Grupowanie rozmyte jest najczęściej stosowane w grupowaniu pikseli w obrazach, a także kompresji danych. Inny ciekawy przykład zastosowania grupowania rozmytego przedstawia Podstawny J., Matysiak B. (2007).

W niniejszym artykule zajęłam się ostrym oraz rozmytym grupowaniem danych oraz jego zastosowaniami. Grupowanie danych jest związane w sposób trwały z bazami danych. Zarówno ostre, jak i rozmyte grupowanie znajduje szereg zastosowań przy podziale zbioru danych na podzbiory. Warto zwrócić uwagę, że danymi mogą być ludzie, ich cechy charakteru oraz wyglądu, a także obrazy, poszczególne piksele, czy tkanki mózgowe. Bez grupowania ostrego i rozmytego nie byłby możliwy podział tych danych na podzbiory z zachowaniem homogeniczności w grupach i heterogeniczności pomiędzy grupami.

Literatura

[1]. Rutkowski L. (2005): *Metody i techniki sztucznej inteligencji.* Wydawnictwo Naukowe PWN, Warszawa.

[2]. Szwabowski J., Deszcz J. (2005): *Metody wielokryterialnej analizy porównawczej. Podstawy teoretyczne i przykłady zastosowań w budownictwie.* Wydawnictwo Politechniki Śląskiej, Gliwice.

[3]. Krzyśko M., Wołyński W., Górecki T., Skorzybut M. (2008): *Systemy uczące się.* Wydawnictwo Naukowo-Techniczne, Warszawa, 346-357.

[4]. Romesburg H. (2004): *Cluster Analysis for Researchers.* Lulu Press, North Karolina.

[5]. Lenkiewicz S. (2007): *Analiza skupień i obszary jej zastosowań.* W. Hołubiec J. (Red.) Analiza systemowa w finansach i zarządzaniu. Wybrane problemy. IBS PAN, Warszawa, T. 9, 112-123.

[6]. Podstawny J., Matysiak B. (2007): *Zastosowanie wartości lingwistycznych w automatycznym module transakcji zapytań rozmytych na zapytania* dokładne. Wydawnictwo Komunikacji i Łączności, Warszawa.

ISBN 9788389475220