STUDIA INFORMATICA 2016
Volume 37 Number 1 (123)

Agnieszka NOWAK-BRZEZIŃSKA, Tomasz RYBOTYCKI
Uniwersytet Śląski, Instytut Informatyki

EKSPLORACJA MEDYCZNYCH REGUŁOWYCH BAZ WIEDZY

Streszczenie. Celem pracy jest eksploracja (grupowanie i wizualizacja) medycznych regułowych baz wiedzy. W artykule opisano narzędzie CluVis, zaimplementowane przez autorów, pozwalające analizować (grupować przy użyciu hierarchicznej analizy skupień) reguły i wizualizować (przy użyciu tzw. map prostokątów) ich skupienia. W ramach eksperymentów przeanalizowano wpływ miar podobieństwa wewnątrz- i międzygrupowego, metod wizualizacji, a także miary jakości skupień na wyniki eksploracji (wykrycie tendencji, nietypowości w danych).

Słowa kluczowe: medyczne regułowe bazy wiedzy, analiza skupień, wizualizacja, miary jakości grupowania

EXPLORATION OF MEDICAL RULE-BASED KNOWLEDGE BASES

Summary. In this work the topic of applying clustering as a knowledge extraction method from real-world medical data is discussed. The authors propose hierarchical clustering method and visualization techniques for knowledge base representation in the context of medical knowledge bases for which data mining methods are successfully employed and may resolve different problems. What is more, the authors analyze the impact of different clustering parameters on the result of searching through such structure.

Keywords: medical rule-based knowledge bases, cluster analysis, visualization, validity index

1. Wprowadzenie

W ostatnich latach systemy ekspertowe przestały być wyłącznie domeną naukowców i laboratoriów naukowych zajmujących się badaniami w dziedzinie sztucznej inteligencji. Możliwości zastosowań tej nowoczesnej technologii informatycznej obejmują wiele dziedzin

nauki i wiedzy: od medycyny poprzez geologię, technikę aż do zastosowań w dziedzinie wspomagania podejmowania decyzji gospodarczych i finansowych. Motorem wzrostu wielu specjalności medycznych jest m.in. działalność naukowo-badawcza ukierunkowana na wytworzenie i skomercjalizowanie specjalnych programów komputerowych, które mają pomagać lekarzowi w wyborze optymalnego rodzaju operacji [13]. Wspomaganie lekarzy w identyfikacji choroby i terapii z pewnością pozwoli udoskonalić procedury medyczne na każdym etapie. Tworzy się więc coraz częściej tzw. medyczne bazy wiedzy (w formie łańcuchów przyczynowo-skutkowych). Jednym z pionierskich systemów ekspertowych jest system MYCIN, przechowujący wiedzę (około 500 reguł, w postaci zbiorów par atrybut-wartość zapisanych jako klauzule Horna) na temat różnych infekcji krwi oraz zapalenia opon mózgowo-rdzeniowych. Powstał on, ponieważ czas oczekiwania na wyniki badań laboratoryjnych bakterii będących przyczyną choroby był zbyt długi i lekarze często podejmowali decyzje o leczeniu pacjentów bez czekania na wyniki, co drastycznie zmniejszało celność diagnozy. Na podstawie pytań, dotyczących objawów choroby i wyników części badań laboratoryjnych, system potrafił rozpoznać, czy pacjent jest chory i określić, czy choroba jest wywołana przez bakterie, po czym na podstawie dostępnych danych zaproponować optymalną terapię lekową.

Niestety, dzisiaj rozwój tego typu systemów wciąż jest powolny, choć wydawać by się mogło, że sporo czynników argumentuje potrzebę ich tworzenia. Wśród najważniejszych da się wymienić np. koszty tworzenia i wdrażania tego typu aplikacji, które w dłuższym okresie czasu są znacznie tańsze i pomogłyby w rozwiązywaniu problemów wymagających najbardziej specjalistycznej (najdroższej) wiedzy. Co równie istotne, brak jest ekspertów w wielu dziedzinach, przez co tym bardziej powinno się widzieć cel w tworzeniu systemów, które byłyby w stanie zachować wiedzę specjalistów i dysponować nią w szerszym zakresie (w tym samym czasie udostępniać ją wielu odbiorcom). Nie bez znaczenia jest i to, że tego typu systemy po prostu pracują efektywniej niż człowiek (nie męczą się, są bardziej niezawodne niż ludzie). Są także bardziej obiektywne i dokładne. Zdarza się, że ogrom wiedzy w danej dziedzinie może sprawić problem z jej interpretacją nawet ekspertom (np. by w rozsądnym/krótkim czasie potrafił przeanalizować trudny przypadek chorobowy i podjąć optymalną decyzję o dalszej terapii), stąd nadzieje pokładane w użyciu narzędzi komputerowych do usprawnienia procesu analizy danych i podejmowania decyzji. Należy jednak zauważyć, że duża ilość informacji może być nie tylko trudna w analizie dla narzędzia komputerowego, ale też w ich opisie czy charakterystyce. Lekarze korzystający z takiego medycznego systemu wspomagania decyzji mogą także mieć trudność w interpretacji wiedzy przekazanej przez ów system, zwłaszcza gdy wiedza jest złożona, przedstawia wiele zależności i ukrytych powiązań między danymi. Co więcej, w takich danych często daje się odkryć pewne nietypowości lub regularności (pewne powtarzalne grupy), dlatego prócz technik wykrywania różnych zależności w danych, kolejnym potrzebnym krokiem będzie wyposażanie takich aplikacji

w narzędzia wizualizacji złożonych danych, tak by proces interpretacji przez lekarza (a zatem i proces ostatecznego podejmowania decyzji co do leczenia pacjenta) był jak najkrótszy.

Wiele zależy od typu analizowanych danych, a reguły nie są łatwym obiektem analiz, gdyż mogą być złożone z wielu atrybutów różnych typów. Dodatkowo można sobie np. wyobrazić tak skrajne przypadki baz regułowych, w których reguły są całkowicie separowalne, nie mając żadnej przesłanki/konkluzji wspólnej. Reguły minimalne (krótkie) dodatkowo utrudniają przetwarzanie, gdyż zbyt ogólny zapis nie pozwala na eksplorację zbyt wielu powiązań między takimi danymi. Narzędzia/metody automatycznie generujące reguły z danych (np. RSES [1]) zwykle produkują bardzo duże liczby reguł (reguły są redundantne, mało optymalne). Ich duża liczba utrudnia szybką i poprawną jakościowo analizę. Reguły tworzą bazę wiedzy, która w procesie wnioskowania (proces tworzenia decyzji/rady przez system ekspertowy) jest analizowana w sposób liniowy, co oznacza, że im więcej jest reguł do analizy, tym dłużej trwa cały proces wnioskowania , a co za tym idzie tym podjęcie decyzji przez użytkownika takiego systemu (lekarza) i przekazanie decyzji np. pacjentowi jest opóźnione. Autorzy podjęli zatem próbę stworzenia narzędzia, które najpierw, znanymi i bardzo efektywnymi technikami eksploracji wiedzy, grupuje reguły podobne do siebie w skupienia (nadając każdemu skupieniu adekwatny opis – reprezentanta), a następnie wizualizuje utworzone grupy, pozwalając tym samym na ich szybszą analizę przez lekarzy-specjalistów. Dzięki temu lekarz widząc w danych pewne regularności, może dużo szybciej podjąć pewne decyzje co do dalszej terapii pacjenta, bądź widząc, że ma do czynienia z przypadkami nietypowymi, również odpowiednio zareagować (np. robiąc dodatkowe badania). Z tego względu autorzy zaprojektowali i stworzyli narzędzie CluVis, pozwalające analizować i wizualizować otrzymane złożone struktury danych, jakimi są regułowe bazy wiedzy. Szczególnej analizie poddano medyczne bazy wiedzy. Uzyskanie rzeczywistych baz medycznych jest bardzo trudne, więc skorzystano z darmowych repozytoriów danych medycznych i automatycznie przy użyciu narzędzia RSES wygenerowano z nich reguły. Następnie porównano wyniki, które dla nich uzyskano, z tymi otrzymanymi dla rzeczywistej bazy. W ramach badań przeprowadzono grupowanie reguł (algorytmami hierarchicznej analizy skupień), testując przy tym różne miary podobieństwa wewnątrz- i międzygrupowego, oceniono jakość utworzonych skupień oraz zwizualizowano je.

Praca składa się z 6 rozdziałów: rozpoczynając (rozdział 1) od wprowadzenia i przechodząc do opisu hierarchicznej analizy skupień wraz z opisem rozważanych miar podobieństwa i jakości skupień (rozdział 2). Rozdział 3 poświęcono metodom wizualizacji skupień, zaś kolejny (4) opisowi narzędzia CluVis. Eksperymenty ujęto w rozdziale 5. Ostatni rozdział stanowi podsumowanie.

2. Hierarchiczna analiza skupień

Grupowanie jako jedna z metod pozyskiwania wiedzy, a tym samym eksploracji danych, jest ściśle uwarunkowane źródłem danych oraz oczekiwaną postacią rezultatów. Zasadniczo metody hierarchiczne (wykorzystane w tej pracy) dzieli się na dwie grupy, ze względu na sposób tworzenia struktury wynikowej. Są to metody aglomeracyjne (na których skupili się autorzy) oraz metody deglomeracyjne (podziałowe). Aglomeracyjne metody hierarchicznej analizy skupień tworzą dla zbioru danych hierarchię klasyfikacji, zaczynając od takiego podziału, w którym każdy obiekt stanowi samodzielne skupienie, a kończąc na podziale, w którym wszystkie obiekty należą do jednego skupienia.

2.1. Ogólny hierarchiczny aglomeracyjny algorytm grupowania

Ogólny hierarchiczny aglomeracyjny algorytm grupowania [4] (wykorzystujący podobieństwo obiektów) można przedstawić w postaci pseudokodu:

Wejście: *n* obiektów, wartość progowa podobieństwa s_k oraz *warunek stopu*
Wyjście: struktura reprezentująca sekwencję grupowania obiektów
1. *umieść każdy obiekt w osobnym skupieniu*
2. *skonstruuj macierz podobieństwa skupień dla wszystkich par grup*
3. *dla zadanej wartości podobieństwa* s_k*:*
4. ***repeat***
5. *utwórz graf skupień, w którym każda para skupień o podobieństwie większym lub równym* s_k *jest połączona krawędzią*
6. *ewentualnie zmodyfikuj zadaną wartość podobieństwa* s_k
7. ***until*** *wszystkie grafy utworzą graf spójny* ***or*** *warunek stopu*
8. ***return*** *utworzona struktura*

Głównym problemem omawianego algorytmu jest określenie sensownej wartości podobieństwa, przy której skupienia powinny zostać połączone. W praktyce znacznie częściej stosuje się modyfikacje tego algorytmu, zgodnie z którymi grupowanie przebiega tylko do momentu, w którym podobieństwo łączonych podgrup jest wciąż odpowiednio wysokie bądź do momentu uzyskania określonej liczby grup. Bardzo ważnymi parametrami grupowania są tzw. miary podobieństwa wewnątrz- i międzygrupowego. Mowa tu zarówno o miarach użytych do wyszukiwania dwóch najbardziej podobnych do siebie reguł (zastosowano miary: Gowera [3], SMC [5], W SMC [5]) oraz miarach łączenia skupień (SL, CoL, CL, AL) [4], w których chodzi o to, by łącząc dwa mniejsze skupienia w jedno odpowiednio określić podobieństwo nowo powstałego skupienia do pozostałych skupień w strukturze. Wybierając miary podobieństwa reguł, kierowano się założeniem, by miara pozwalała na pomiar podobieństwa danych wielotypowych (ilościowych, jakościowych). Jedną z miar dobrze się do tego nadających jest miara Gowera:

$$GowerSim(O_i, O_j) = \frac{\sum_{k=1}^{n} s_{ijk} w_{ijk}}{\sum_{k=1}^{n} w_{ijk}}, \quad (1)$$

gdzie O_i, O_j to analizowane reguły, a w odpowiada *wadze* (zwykle $w = 1$) badanego atrybutu, natomiast czynnik s jest zależny od typu atrybutu. Gdy atrybut jest cechą symboliczną, wówczas $s = 1$, gdy wartości atrybutów są równe, zaś 0 w p.p. Gdy atrybut ma wartości na skali ilościowej i x_i, x_j są wartościami k-tego atrybutu dla reguł O_i, O_j zaś R_k jest różnicą między najwyższą i najniższą wartością dla k-tego atrybutu, wówczas s liczymy z formuły:

$$s_{ijk} = 1 - \frac{x_{ik} - x_{jk}}{R_k}. \quad (2)$$

Drugą miarą podobieństwa, rozważaną w niniejszej pracy, jest miara SMC (ang. Simple Matching Coefficient), nieuwzględniająca typu analizowanych danych (zwróci wartość 1, gdy porównywane reguły mają tę samą wartość dla danego atrybutu ($A_i \cap A_j \neq \phi$) i 0 w p.p.):

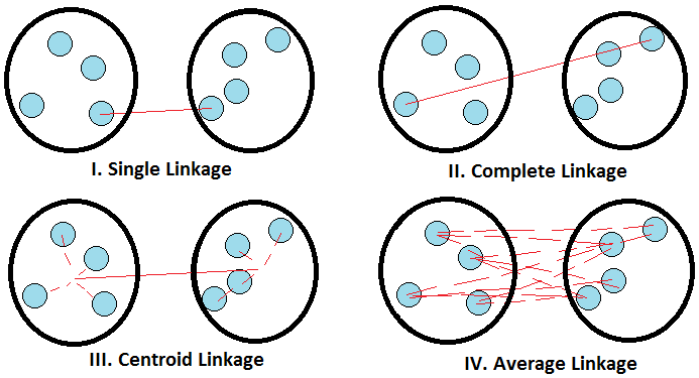
$$simpleSim(O_i, O_j) = \overline{\overline{A_i \cap A_j}}, \quad (3)$$

gdzie: O_i, O_j to analizowane reguły, a A_i, A_j to odpowiadające im zbiory par atrybut-wartość.

Ostatnią analizowaną miarą podobieństwa jest miara W SMC (Weighted SMC):

$$weightedSim(O_i, O_j) = \frac{\overline{\overline{A_i \cap A_j}}}{\overline{\overline{A_i \cup A_j}}}, \quad (4)$$

(oznaczenia tożsame z (3)) będąca modyfikacją SMC w tym sensie, że podobieństwo to jest dodatkowo dzielone przez liczbę atrybutów opisujących reguły, by nie faworyzować długich reguł. Spośród miar łączenia skupień w niniejszej pracy analizie poddano cztery najbardziej popularne: SL (Single Linkage), CoL (Complete Linkage), AL (Average Linkage) oraz CL (Centroid Linkage) – wszystkie wymienione zostały przedstawione na rysunku 1.



Rys. 1. Miary łączenia skupień
Fig. 1. Inter-cluster similarity measures

2.2. Miary jakości skupień

Czasami w odniesieniu do grupowania używa się terminu *uczenie nienadzorowane*, co oznacza, że utworzone skupienia nie są weryfikowane pod względem *jakości* na żadnym etapie procesu grupowania, gdzie przez jakość skupienia rozumie się jego skłonność do bycia wewnętrznie minimalnie oraz zewnętrznie maksymalnie zróżnicowanym [11]. Jedną z najbardziej popularnych metod określania jakości wygenerowanych grup jest obliczanie tzw. indeksu Dunna zaprezentowanego w pracy [2]. Oryginalny indeks Dunna korzysta jednak nie z podobieństwa obiektów, a z ich odległości, dlatego w niniejszej pracy zastosowano jego zmodyfikowaną wersję [6] przedstawioną za pomocą wzoru (5):

$$MDI_m = \frac{\max\left(\delta\left(c_i, c_j\right)\right)}{\min \Delta_k}, \tag{5}$$

gdzie δ jest miarą podobieństwa międzygrupowego, C_i, C_j są skupieniami, $i \neq j$, m jest liczbą skupień, a *min* Δk określa minimalne podobieństwo wewnątrzgrupowe. Ze względu na występowanie funkcji *min* w mianowniku ułamka istnieje możliwość, że mianownik będzie równy zeru (gdy wewnątrz grupy występują dwie, skrajnie niepodobne reguły), wtedy przyjmuje się *MDI = 0*. Decyzję tę argumentuje się faktem, iż poszukiwana jest liczba skupień *m*, dla której wartość *MDI* jest maksymalna, a wartość *0* jest skrajnie minimalną wartością.

Do celów porównawczych zaimplementowano także modyfikację innej miary jakości skupień – *MDBI* (ang. Modified Davies-Bouldin Index) – wyrażoną formułą:

$$MDBI_m = \frac{\sum_{i=0,\, j=i+1}^{m} \frac{Comp(c_i) + Comp(c_j)}{\delta\left(c_i, c_j\right)}}{m}, \tag{6}$$

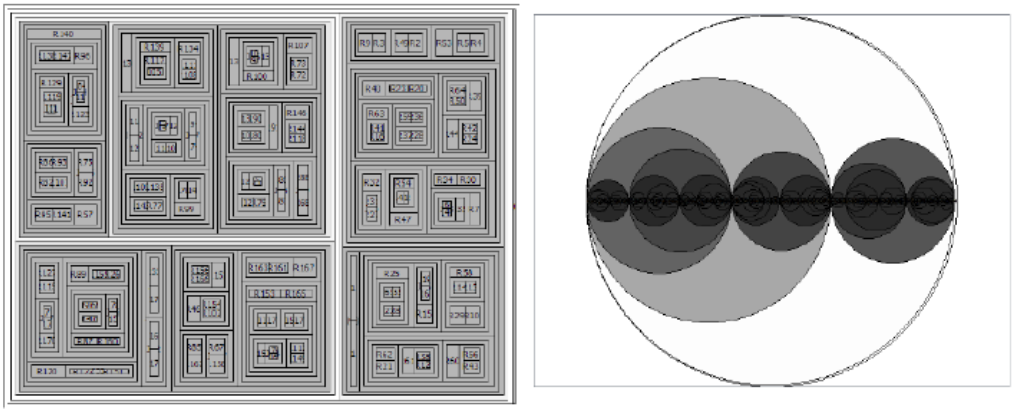
gdzie: δ oznacza podobieństwo międzygrupowe, C_i oraz C_j skupienia, *Comp* ścisłość (ang. *compactness*) skupienia, a *m* liczbę skupień. W pracy przyjęto, że jeśli $\delta(C_i, C_j) = 0$, to wtedy cały składnik sumy także jest równy zeru. *Comp* dla *k*-tej grupy złożonej z *N* reguł to podobieństwo jej reprezentanta Rep_k do każdej reguły O_i w obrębie tej grupy i można je wyrazić wzorem (7):

$$Comp\left(C_k\right) = \sum_{i=1}^{N} \delta\left(Rep_k, O_i\right). \tag{7}$$

3. Metody wizualizacji skupień

Istnieje wiele różnych metod wizualizacji skupień, różniących się od siebie m.in. kształtem obiektów wizualizacji czy ich rozmieszczeniem. Dendrogram jest najbardziej popularną techniką wizualizacji skupień spotykaną w literaturze, jednak gdy danych do wizualizacji jest

dużo, metoda ta zawodzi. Lepszym rozwiązaniem jest użycie tzw. map prostokątów (ang. *treemaps*) [7, 8]. Na rysunku 2 przedstawiono przykładowe wizualizacje wygenerowane w ramach pracy za pomocą zaimplementowanych algorytmów.



Rys. 2. Klasyczna i kolista mapy prostokątów przedstawiające to samo grupowanie
Fig. 2. Rectangular treemap and Circular treemap representing same grouping

W przypadku klasycznych map prostokątów obszar roboczy dzielony jest rekurencyjnie na prostokąty. Rozmiar prostokątów zazwyczaj tożsamy jest z rozmiarem wizualizowanego skupienia (liczbą obiektów w skupieniu). Klasyczna mapa prostokątów ma wiele wariacji ze względu na sposób rozmieszczenia obiektów wizualizacji (najbardziej popularne: slice-and-dice/squarified). Slice-and-dice dzieli obszar roboczy w zależności od tego, czy jest szerszy czy dłuższy, na podłużne prostokąty – odpowiednio pionowe lub poziome. Technika ta ma jednak jedną zasadniczą wadę, a mianowicie ma tendencję do tworzenia cienkich prostokątów, które negatywnie wpływają na czytelność wizualizacji. Drugi z algorytmów wizualizacji (tzw. kolista mapa prostokątów (Circular Treemap)) jest wariantem klasycznej mapy prostokątów, w której jako obiekty wizualizacji wykorzystane są koła. Diagram taki nie wykorzystuje efektywnie dostępnego obszaru roboczego. Duży wpływ na ilość wolnego miejsca na ekranie ma liczba oraz wielkość wizualizowanych skupień. Biorąc pod uwagę wady obu metod, dobrze jest wykorzystywać obie podczas eksploracji BW.

4. CluVis

Przegląd literatury wykazał, że nie ma takich aplikacji, które pozwalałyby grupować, a potem wizualizować reguły w dziedzinowych bazach wiedzy [12]. O ile istnieją narzędzia grupowania czy wizualizacji jako odrębne oprogramowania (i często trudno połączyć ich możliwości) dla typowych tablicowo zapisanych zbiorów danych (rekordów w bazie danych, np. danych pacjentów) [9], o tyle dla danych w postaci reguł tego typu narzędzi nie znaleziono. CluVis [6] jest aplikacją umożliwiającą grupowanie reguł za pomocą hierarchicznego

aglomeracyjnego algorytmu grupowania (AHC) i wizualizację wygenerowanych skupień za pomocą dwóch technik graficznej prezentacji skupień: klasycznej metody map prostokątów (z rozkładem slice-and-dice) oraz jej kolistej odmiany. Interaktywność jest kluczową cechą niniejszej aplikacji, ponieważ ułatwia użytkownikowi analizę wyników wizualizacji, np. przez podświetlanie badanego skupienia (rys. 3) lub umożliwienie natychmiastowego przejścia w głąb hierarchii wybranej grupy.

Rys. 3. Funkcja podświetlenia skupień w narzędziu CluVis
Fig. 3. Cluster highlighting function in CluVis

CluVis pozwala na analizę i wizualizację grup reguł zawartych w regułowych bazach wiedzy wygenerowanych przez system RSES [1]. Sama aplikacja została napisana w języku C++ z użyciem zestawu bibliotek graficznych Qt w wersji 5.4. Powodem, dla którego zdecydowano się na wykorzystanie tego języka, jest przede wszystkim fakt, że programy napisane w nim są z reguły szybsze niż te pisane w innych językach obiektowych, a do tego kod wynikowy języka C++ jest bardzo wydajny. Dodatkowo język ten jest niezwykle plastyczny i popularny, co przyczyniło się do powstania wielu przydatnych bibliotek przeznaczonych wyłącznie dla tego języka (w tym także bibliotek graficznych). Jednym z zestawów takich bibliotek jest Qt, w skład którego wchodzą dodatkowo narzędzia, takie jak m.in. Qt Creator (zintegrowane środowisko programistyczne, w którym napisana została aplikacja CluVis). CluVis pozwala zachować utworzoną wizualizację w pliku graficznym (PNG) oraz dodatkowo generuje raport z grupowania (w pliku tekstowym TXT bądź w XML traktowanym jak arkusz kalkulacyjny XLS). Z raportu możemy wówczas odczytać najlepsze jakościowo grupowanie, uwzględniając wiele istotnych informacji na temat utworzonych grup reguł. CluVis jest narzędziem open source (jego kod dostępny jest pod adresem https://github.com/Tomev/CluVis).

5. Eksperymenty

Głównym celem pracy było przedstawienie możliwości wspomagania interpretacji medycznych regułowych baz wiedzy przez połączenie funkcji grupowania (na bazie podobieństwa) dużych zbiorów reguł (opisujących wiedzę z danej dziedziny, np. medycyny) i wizualizacji uzyskanych grup reguł. Wyniki grupowania mogą się mocno różnić w zależności od tego, wedle jakich parametrów nastąpiło tworzenie grup. Skupiono się jedynie na optymalnych (pod względem kryteriów oceny jakości skupień – indeksów MDI i MDBI opisanych w rozdziale 2) podziałach reguł. Eksperymentom poddano 6 zbiorów danych w ramach repozytorium UCI ML Repository (Spect, Soybean, Echo, Hepatitis, Post op., Pima), a następnie porównano wyniki z tymi uzyskanymi dla rzeczywistej bazy danych dotyczącej także medycyny (guz Krukenberga). Tabela 1 przedstawia krótkie charakterystyki testowanych zbiorów danych (liczbę reguł oraz atrybutów baz).

Tabela 1

Ogólna charakterystyka testowanych baz wiedzy

BW	Spect	Soybean	Echo	Hepatitis	Post op.	Pima	Krukenberg
# atrybutów	23	36	13	20	9	9	23
# reguł	67	63	63	35	46	457	200

Przy użyciu narzędzia CluVis pogrupowano reguły (testując różne parametry grupowania) i zwizualizowano ich skupienia. Celem stało się m.in. sprawdzenie, czy któreś z miar/metod mają tendencję do tworzenia np. dużych skupień, albo długich (szczegółowych) opisów grup. Równie ciekawe miało być sprawdzenie, czy pewne miary podobieństwa wewnątrz- bądź międzygrupowego są typowe dla tworzenia małych skupień, bądź sprzyjają tworzeniu grup słabej jakości (złe wartości dla indeksów MDI/MDBI). Wyniki analizy użycia różnych miar podobieństwa wewnątrz- i międzygrupowego przedstawia tabela 2.

Tabela 2

Różnice w jakości skupień dla miar podobieństwa wewnątrz- i międzygrupowego

	Podobieństwo wewnątrzgrupowe				Podobieństwo międzygrupowe				
	G	SMC	W_SMC	P	SL	CoL	AL	CL	P
# małych skupień	4,17± 7,53	6,72± 8,16	4,07±5,03 (0,00-32,00)	Ns	8,15± 9,29	2,45± 3,31	4,17± 4,16	5,17± 8,57	P< 0,001
MDI	1,20± 0,85	2,38± 2,47	2,05±2,03 (0,00-10,06)	P< 0,001	2,86± 3,14	0,927± 0,149	2,16± 1,46	1,53± 1,25	P< 0,001
MDBI	122,03± 152,38	338,88± 1058,03	222,85± 351,66	Ns	66,16± 91,08	193,67± 217,85	505,90± 1241,44	145,93± 146,11	P< 0,001

Ns – różnica nie jest istotna statystycznie dla p = 0,05. W tabeli przedstawiono wartość średnią ± SD (odchylenie standardowe).

Tabela przedstawia analizę różnic statystycznych (ANOVA [14]) dla miar jakości skupień w kontekście użytych miar łączenia skupień. Można zauważyć, że jedynie stosując indeks MDI, różnice są istotne statystycznie. Najniższe wartości MDI dostarcza miara Gowera,

najwyższe SMC. Na poziomie istotności p <0,001 powiemy, że najmniej małych skupień generuje metoda CoL (łączenia skupień), zaś najwięcej – SL. Najwyższe wartości MDBI uzyskamy dla metody AL, najniższe dla SL. Te same trendy zauważono dla rzeczywistego zbioru guza Krukenberga.

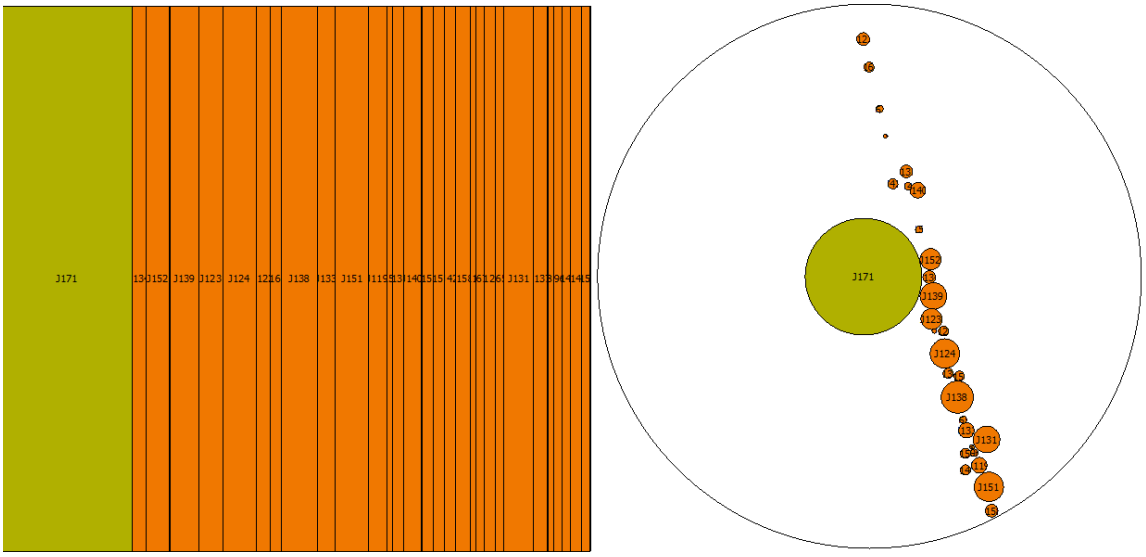
Kolejnym etapem było porównanie miar podobieństw wewnątrz- i międzygrupowych pod kątem liczby reguł w grupie, a także długości reprezentanta skupienia, składającego się z każdego atrybutu występującego w skupieniu (wyniki prezentuje tabela 3).

Tabela 3

Rozmiar grupy/reprezentanta grupy

	Podobieństwo wewnątrzgrupowe				Podobieństwo międzygrupowe				
	G	SMC	W_SMC	P	SL	CoL	AL	CL	P
% reguł w grupie	8,56± 16,24	8,59± 20,60	8,55± 18,01	Ns	8,61± 22,54	8,54± 14,05	8,64± 18,34	8,48± 17,53	Ns
Długość reprezentanta	10,79± 7,8	6,87± 5,69	7,64± 5,71	P< 0,001	7,79± 7,28	9,34± 6,61	8,77± 6,36	7,78± 6,37	P< 0,001

Można zauważyć, że długość reprezentanta skupień mocno zależy od wybranej metody wewnątrz- bądź międzygrupowego podobieństwa. Średnia długość najkrótszego reprezentanta skupienia reguł tworzona jest dla miary SMC, najdłuższego dla miary Gowera. Najkrótszy reprezentant zawiera tylko dwie przesłanki w opisie grupy, podczas gdy najdłuższy aż 35 deskryptorów (dwójek atrybut-wartość). Te same analizy wykonano osobno dla rzeczywistego zbioru danych Krukenberga (baza wiedzy zawiera 200 reguł, opisanych 23 atrybutami typu wiek, lokalizacja guza, typ zabiegu, informacja, czy pacjent żyje). Potwierdziły się tendencje opisane powyżej. Ma to o tyle istotne znaczenie, że można analizując rzeczywiste zbiory danych regulować uzyskane wyniki grupowania i wizualizacji (z opisami grup) przez użycie odpowiednich parametrów grupowania i wizualizacji.



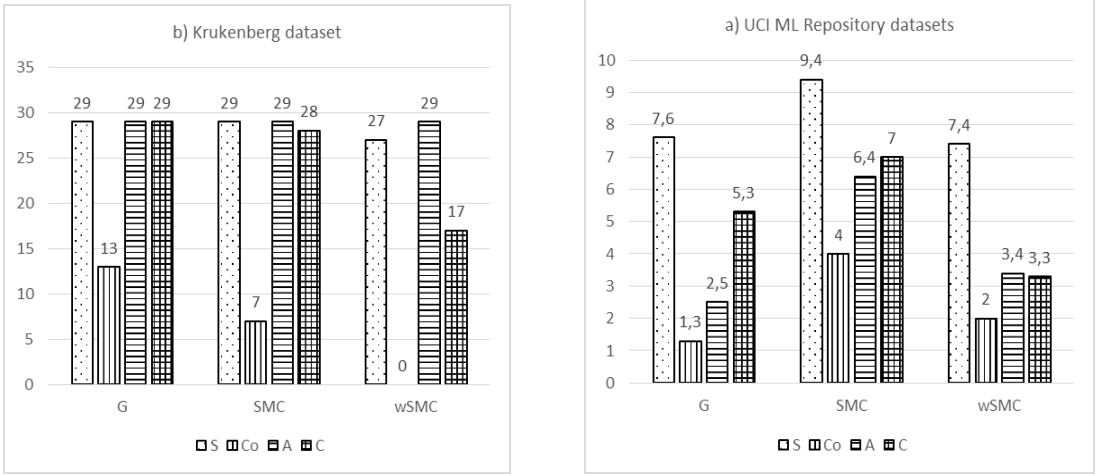
Rys. 4. Wizualizacja dla zbioru Krukenberga
Fig. 4. The visualization of Krukenberg set

Rysunek 4 przedstawia ciekawy przypadek wizualizacji skupień dla tejże bazy, gdzie widać 23 małe skupienia i 7 dużych grup reguł.

Grupowanie to nie ma reguł niepogrupowanych (co jest charakterystyczne dla grupowania metodą Complete Link). Największe skupienie zawiera 43 reguły. Jego reprezentant ma postać (*BORRMAN = rozlanyNaciek*) & (*WAGA = spadek<10%*) &…& (*sch_I = 5Fu + LV3dni*) & (*wiek = poniżej_50*) & (*BOOST = nie*) => (*stan = ZYJE*) i oznacza, że te 43 reguły opisują przypadki pacjentów, dla których odpowiednie atrybuty miały najczęściej takie wartości (liczona jest moda dla atrybutów kategorycznych i średnia arytmetyczna dla atrybutów liczbowych przy tworzeniu reprezentanta grupy).

Grupowanie reguł w skupienia może dostarczyć wielu możliwych podziałów, w zależności od tego, jakich parametrów grupowania użyto. Nie każdy z tych podziałów jest optymalny. Celem badań stała się analiza liczby pojedynczych grup (są to szczególnie w medycynie przypadki warte identyfikacji) w zależności od użytych parametrów grupowania. Wśród parametrów mogących mieć wpływ na jakość grupowania są różne miary podobieństwa reguł (G, SMC, W SMC), miary łączenia skupień (SL, CoL, AL, CL), ale i miary tworzenia reprezentantów grup.

Rysunek 5 przedstawia wyniki uzyskane dla różnych miar podobieństwa (G-Gower, S – SMC, W – W SMC) i różnych miar łączenia skupień (S – SL, Co – CoL, A – AL., C – CL).



Rys. 5. Liczba małych skupień reguł
Fig. 5. The number of unclustered rules

Można zauważyć, że dla miary CoL liczba małych skupień jest kilkakrotnie mniejsza niż dla reszty miar, co oznacza, że generuje ona dużo większe skupienia i redukuje liczbę odchyleń (prawdopodobnie jedno jest konsekwencją drugiego). W dalszych badaniach warto będzie sprawdzić także wiele innych czynników, np. zbadać, na co mają wpływ wewnątrz- i międzygrupowe miary podobieństwa.

6. Podsumowanie

Metody ekstrakcji wiedzy stały się bardzo popularne w ostatnim czasie, zwłaszcza w kontekście znajdowania w nich jakichś potencjalnych zależności, anomalii czy trendów [10]. Celem pracy było pokazanie funkcjonalności technik grupowania danych i ich wizualizacji w zadaniu eksploracji dużych dziedzinowych baz wiedzy. O ile sporo jest prac poświęconych grupowaniu i wizualizacji dużych (typowych, tablicowych) zbiorów danych, o tyle danych tak złożonych jak reguły w dziedzinowych bazach wiedzy dotąd w pracach naukowych nie grupowano i nie wizualizowano jednocześnie. Opisano zarówno algorytm AHC do grupowania hierarchicznego, jak i algorytmy wizualizacji skupień. Samo grupowanie nie wystarczy, gdy chcemy efektywnie eksplorować regułowe bazy wiedzy. Proces grupowania doposażony w wizualizację pozwala na łatwiejsze odkrywanie prawdziwej struktury z dziedziny, identyfikując jakieś przypadki nietypowe bądź właśnie regularności/trendy w danych. Taka dwutorowa eksploracja złożonych danych ma szanse dostarczyć spodziewanych efektów z punktu widzenie ekspertów dziedzinowych (np. lekarzy), którzy dużo sprawniej będą mogli realizować swoje zadania.

Eksperymenty wykonane w ramach niniejszej pracy pozwoliły wyciągnąć następujące wnioski. Po pierwsze, użycie różnych miar podobieństwa między- bądź wewnątrzgrupowego wpływa na rozmiar skupień, ich strukturę wewnętrzną, a także liczbę skupień. Metoda pojedynczego łączenia (SL) ma tendencję do tzw. łańcuchowania, przez co często uzyskana struktura to sporo małych skupień, których nie udało się połączyć w grupy większe. Z kolei metoda najdalszego łączenia sprawia, że uzyskujemy niewiele dużych (licznych) skupień, nie pozwalając wykrywać zbyt wielu odchyleń w danych.

Odrębnym celem pracy było stworzenie uniwersalnego narzędzia jak CluVis, które pozwoliłoby jednocześnie grupować i wizualizować regułowe bazy wiedzy.

Liczba medycznych systemów ekspertowych stale wzrasta i dzięki postępowi w takich dziedzinach, jak akwizycja wiedzy, rozumowanie oparte na wiedzy czy integracja systemów w środowiskach klinicznych, jest szansa, że postęp ten i efektywność tego typu rozwiązań będzie coraz większy. Dla lekarzy, którzy mieliby korzystać z takich narzędzi, kluczową kwestią jest dostarczenie im środków, które wspomogą specjalistów w podjętych przez nich decyzjach (pozwolą na interpretację dużych zasobów wiedzy i/lub zidentyfikują nietypowe przypadki chorobowe), dlatego warto prowadzić dalsze badania w tym kierunku.

Praca jest częścią projektu „Eksploracja regułowych baz wiedzy” finansowanego w ramach środków Narodowego Centrum Nauki (NCN: 2011/03/D/ST6/03027).

BIBLIOGRAFIA

1. Bazan J.G., Szczuka M.S., Wroblewski J.: A new version of rough set exploration system. Rough Sets and Current Trends in Computing, Springer-Verlag, Berlin 2002, s. 397÷404.
2. Dunn J.C.: Well separated clusters and optimal fuzzy partitions. Journal of Cybernetica, Vol. 4, 1974, s. 95÷104.
3. Gower J.C.: A general coefficient of similarity and some of its properties. Biometrics, Vol. 27, International Biometric Society, Washington 1971, s. 857÷871.
4. Morzy T.: Eksploracja danych. Metody i algorytmy. Wydawnictwo Naukowe PWN, Warszawa 2013.
5. Nowak-Brzezińska A., Jach T.: Wnioskowanie w systemach z wiedzą niepewną. Studia Informatica, Wydawnictwo Politechniki Śląskiej, Gliwice 2011.
6. Rybotycki T.: Wizualizacja struktur hierarchicznych dla regułowych baz wiedzy. Sosnowiec 2015.
7. Shneiderman B.: Tree visualization with tree-maps: 2-d space-filling approach. Transactions on Graphics (TOG), Association for Computing Machinery, New York 1992.
8. Wetzel K.: Pebbles – using circular treemaps to visualize disk usage. 2004.
9. Nowak-Brzezińska A., Xięski T.: Exploratory clustering and visualization. 18th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, KES 2014, Gdynia, Poland 2014, s. 1082÷1091.
10. Han J., Kamber M., Pei J.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., 2011.
11. Kovács F., Legány C., Babos A.: Cluster Validity Measurement Techniques. AIKED'06 Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, ISBN:111-2222-33-9, 2006, s. 388÷393.
12. Nowak-Brzezińska A., Rybotycki T.: Visualization of medical rule-based knowledge bases. Journal of Medical Informatics & Technologies, Vol. 24, 2015, s. 91÷98.
13. Treptow A.: Medyczna strefa Schengen nadzieją dla szpitali. Puls Biznesu, nr 123, 2012.
14. Stanisz A.: Przystępny kurs statystyki z zastosowaniem Statistica PL na przykładach z medycyny. Tom 1-3, Statsoft Polska, Kraków 2006.

Abstract

In this work the topic of applying clustering as a knowledge extraction method from real-world data is discussed. The authors propose a hierarchical clustering and treemap visualization technique for knowledge base representation in the context of medical knowledge bases, for which data mining techniques are successfully employed and may resolve different problems. What is more, the authors analyze the impact of different clustering parameters (inter and intra-cluster similarity measures) on the result of searching through such a structure. Particular attention was also given to the problem of clusters visualization. Authors review selected two-dimensional approaches, stating their advantages and drawbacks in the context of representing complex cluster structures.

Adresy

Agnieszka NOWAK-BRZEZIŃSKA: Silesian University, Institute of Computer Science, ul. Będzińska 39, 41-200 Sosnowiec, Poland, agnieszka.nowak@us.edu.pl.
Tomasz RYBOTYCKI: rybotyckitomasz@gmail.com.