STUDIA INFORMATICA 2012
Volume 33 Number 2A (105)

Agnieszka NOWAK-BRZEZIŃSKA
Uniwersytet Śląski, Instytut Informatyki

EKSPLORACJA ODCHYLEŃ W REGUŁOWYCH BAZACH WIEDZY

Streszczenie. Artykuł przedstawia problematykę wykrywania odchyleń w regułowych bazach wiedzy. Reguły nietypowe, uznawane tu za odchylenia, powinny być przedmiotem analiz ekspertów i inżynierów wiedzy, gdyż mogą wpływać na efektywność wnioskowania w systemach wspomagania decyzji. Autorka prezentuje różne podejścia w znajdowaniu odchyleń w regułach. W artykule ujęto także wykonane eksperymenty wraz z interpretacją wyników.

Słowa kluczowe: odchylenia, analiza skupień, regułowe bazy wiedzy, efektywność procesów wnioskowania

MINING OUTLIERS IN RULE KNOWLEDGE BASES

Summary. The paper presents the problem of outlier detection in the rule knowledge bases. Unusual (rare) rules, regarded here as the deviation, should be the subject of analysis experts and knowledge engineers because they can influence the efficiency of inference in decision support systems. The author presents a different approach in finding outliers in the rules. The experiments with their results are also presented in the paper.

Keywords: outliers, cluster analysis, rules knowledge bases, efficiency

Wprowadzenie

W ostatnich latach w dziedzinie eksploracji danych szczególnego znaczenia nabrały metody wykrywania odchyleń w danych. Wynika to z faktu znajdowania coraz to nowych zastosowań tego zagadnienia, tj. w: wykrywaniu defraudacji, nietypowych objawach chorobowych, nieautoryzowanych włamaniach do serwerów, wykrywaniu wadliwych serii produkcyjnych czy chociażby znajdowaniu nowości w tekstach (np. nowych artykułów na dany temat). Okazuje się, iż wykrywając w danych pewne nieregularności, udaje nam się posiąść sporo – istotnej dla efektywnego działania – wiedzy. Niewykrycie odchyleń w danych

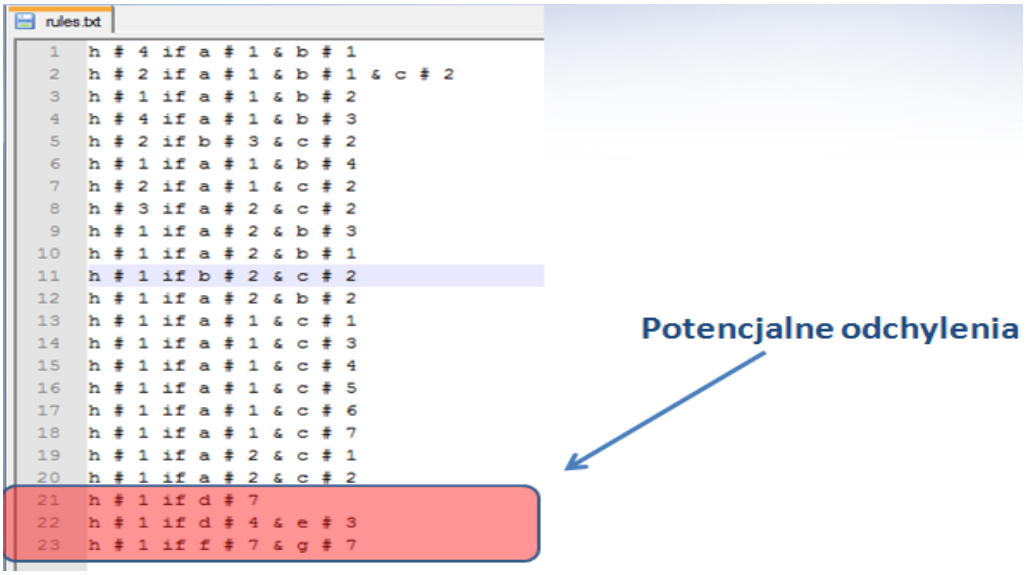
w istotny może sposób wpłynąć na błędy w procesach decyzyjnych. Wiele metod eksploracji wiedzy (ang. *data-mining*) nie pozwala na użyteczną analizę danych bez wcześniejszego wykrycia ewentualnych odchyleń czy błędów.

Obecnie w literaturze możemy spotkać zarówno metody typowo statystyczne, jak i oparte na badaniu gęstości danych czy np. metody wywodzące się z analizy skupień. Zaletą tych ostatnich jest fakt, iż wykrywanie odchyleń w ramach algorytmów grupowania jest jakby elementem składowym tych algorytmów: obiekty niedające się włączyć do żadnego skupienia możemy traktować jako odchylenia.

1.1. Odchylenia w regułowych bazach wiedzy

Odchyleniem w kontekście regułowych baz wiedzy będą reguły niepasujące do reszty wiedzy (reguł) zapisanej w bazie wiedzy. Reguły takie są regułami rzadkimi i powinny stanowić przedmiot bardziej wnikliwych badań w dziedzinie, zwłaszcza ze strony ekspertów, tak aby w przyszłości w miarę możliwości rozszerzać bazę wiedzy w dotąd niezbadanym obszarze. Wykrycie odchyleń w regułach pozwoli na optymalizację wnioskowania w regułowych bazach wiedzy, a przez to zwiększy efektywność systemów wspomagania decyzji.

Rozważania omówimy na przykładzie bazy wiedzy złożonej z 23 reguł prostych o zmiennej liczbie warunków. Rysunek 1 prezentuje bazę, w której na pierwszy rzut oka widać, iż 3 ostatnie reguły znacząco różnią się od reszty reguł w bazie wiedzy, gdyż w części warunkowej składają się na nie zupełnie inne atrybuty.



Rys. 1. Regułowa baza wiedzy z potencjalnymi odchyleniami
Fig. 1. Rule knowledge base with potential outliers

1.2. Po co wykrywać odchylenia w regułach?

Może być wiele przyczyn powstawania odchyleń w regułowej bazie wiedzy. Takie rzadkie (nietypowe) reguły mogą reprezentować przypadki wyjątkowe i specyficzne, ale i reguły błędne, powstałe w wyniku nieprzemyślanej modyfikacji bazy wiedzy. Szybkość ich wykrycia niewątpliwie wpływa na jakość decyzji, a więc na skuteczność danego systemu wspomagania decyzji. Znalezione reguły bądź grupy reguł odstających powinny wzbudzić zainteresowanie inżyniera wiedzy. Będzie on mógł, w kontakcie z ekspertem dziedzinowym, uzupełniać bazę wiedzy w obszarze, który uznano za odchylenie. W kontekście złożonych baz wiedzy, o których autorka wspomina w pracach [3, 4, 5], w dalszej części niniejszej pracy prezentuje zaś jedynie ogólne wyjaśnienia, wykrycie odchyleń w regułowych bazach wiedzy przed grupowaniem reguł pozwoli zwiększyć jakość tworzonych skupień reguł. Wiąże się z tym także lepsza separowalność skupień, ich jednorodność, a więc i lepsza jakość formowanych reprezentantów tak tworzonych grup reguł. To z kolei z pewnością zwiększy efektywność wnioskowania w złożonych bazach wiedzy.

2. Czynniki wpływające na efektywność systemów wspomagania decyzji

W pracach [3, 5] autorka zaprezentowała model hierarchicznej bazy wiedzy ze skupieniami reguł decyzyjnych, gdzie oprócz szczegółowego omówienia proponowanej koncepcji złożonych baz wiedzy, z przedstawieniem modelu reprezentacji tak złożonej wiedzy, przedstawiła także metody optymalizacji procesów wnioskowania dokonywanych w tak złożonych bazach wiedzy. Przez *złożone bazy wiedzy* autorka rozumie bazy wiedzy złożone nie tylko pod względem dużej liczby reguł, lecz także skomplikowanej struktury wewnętrznej takich reguł. Zakłada się bowiem, że reguła jest *złożona* wtedy, gdy części warunkowe jednej reguły stanowią część decyzyjną innej reguły w bazie wiedzy. Ponieważ sam rozmiar bazy wiedzy w znaczący sposób wpływa na czas wnioskowania, zdają się być potrzebne takie modyfikacje istniejącej struktury dziedzinowej bazy wiedzy, które ograniczą liczbę reguł analizowaną w procesie wnioskowania w celu znalezienia reguł do uaktywnienia. W pracy [3] autorka przedstawiła model tzw. złożonej bazy wiedzy (ang. *composited knowledge base*), w którym zbiór reguł jest prezentowany w postaci struktury skupień reguł podobnych do siebie bądź w części warunkowej, bądź w decyzyjnej. Zastosowanie analizy skupień (ang. *cluster analysis*) w tym przypadku – jako metody grupowania obiektów o podobnych własnościach – wydaje się oczywiste [1, 2]. Spośród wielu możliwych do zastosowania algorytmów analizy skupień algorytmy aglomeracyjne z grupy algorytmów hierarchicznych pozwalają na naturalne łączenie w grupy tych obiektów (reguł), które są do siebie najbardziej podobne w da-

nym kroku algorytmu. Utworzona hierarchia skupień dodatkowo mówi nam, jak bardzo podobne są do siebie połączone elementy (o tym decyduje poziom w hierarchii). Dla takich typów struktur znane są efektywne techniki ich przeszukiwania, pozwalające w znaczący sposób ograniczyć złożoność obliczeniową.

3. Metody wykrywania odchyleń w danych

W literaturze przedmiotu znane są już dziesiątki metod wykrywania obiektów odstających (ang. *outlier*). Wszystkie zakładają (zgodnie z definicją Hawkinsa [9]), że *odchyleniem jest obiekt tak bardzo odstający od reszty obserwacji, że istnieje podejrzenie, iż wygenerował go odmienny mechanizm*.

Wychodząc od takiego właśnie założenia, wszystkie dostępne rozwiązania, niezależnie od tego, czy opierają się na metodach typowo statystycznych, matematycznych czy z pogranicza eksploracji wiedzy, szukają danych istotnie różniących się od reszty elementów w zbiorze. Wśród metod wykrywania odchyleń wyróżniamy metody oparte na:

- rozkładzie danych (ang. *distribution-based*),
- odległości danych (ang. *distance-based*),
- gęstości (ang. *density-based*),
- grupowaniu (ang. *clustering-based*) [6, 7, 8].

Dla każdej z grup zostały wybrane metody najbardziej reprezentatywne, najczęściej stosowane w praktyce i krótko opisane w kolejnych podrozdziałach. Na ich przykładzie zostaną omówione zalety i wady takiego podejścia do znajdowania odchyleń w danych.

3.1. Metody statystyczne

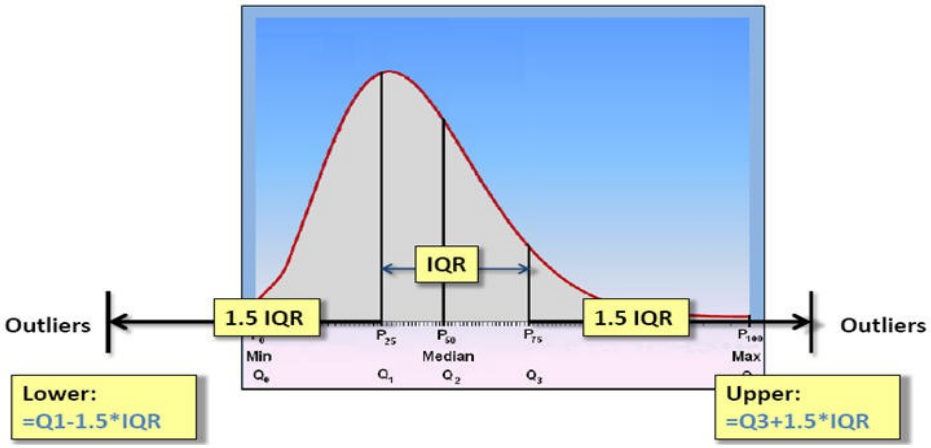
Statystyczne podejście do wykrywania odchyleń w danych nadaje się raczej dla danych prezentowanych w przestrzeni o małej liczbie wymiarów. W grupie metod statystycznych dominują dwie następujące techniki:

- wykrywanie odchyleń z wartości średniej i odchylenia standardowego,
- wykrywanie odchyleń na podstawie rozstępu międzykwartylowego.

Pierwsza metoda polega na tym, iż za odchylenie można uznać dane, które wykraczają poza przedział: $< średnia \pm 2 * odchylenie\_std >$. Metoda ta opiera się na założeniach tzw. testu Grubbsa (ang. *maximum normed residual test*, 1969 r.), gdzie wyszukiwanie odchyleń w danych jest procesem iteracyjnym tak długo, dopóki istnieją jakiekolwiek odchylenia.

Druga z metod polega na tym, iż za odchylenia uważa się dane, które nie spełniają ram przedziału wyznaczonego na podstawie kwartyli: pierwszego i trzeciego, oraz na podstawie

tzw. rozstępu międzykwartylowego. Konkretniej, mając dla danego zbioru wyznaczone wartości pierwszego (Q1) oraz trzeciego (Q3) kwartyla, wyznaczamy wartości tzw. rozstępu międzykwartylowego (IQR), jako różnicę między Q3 i Q1. Jeśli za minimalną wartość uznamy wartość nie mniejszą niż: $\min = Q1 - 1{,}5 * IQR$, a za wartość maksymalną wartość nie większą niż: $\max = Q3 + 1{,}5 * IQR$, wówczas za odchylenie uznamy każdą taką wartość, która nie mieści się w przedziale $< \min, \max >$. Wizualizacja tego podejścia została zaprezentowana na rysunku 2.

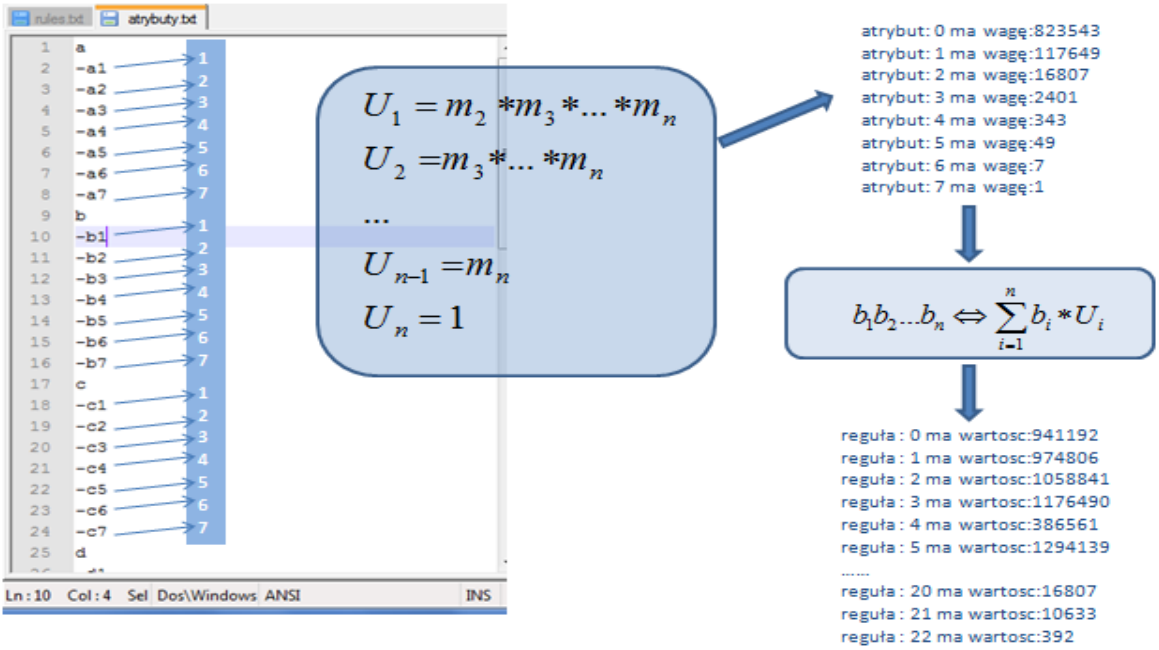


Rys. 2. Wykrywanie odchyleń metodą rozstępu międzykwartylowego
Fig. 2. Outlier detection based on inter quartile method

Oczywiście podane tutaj wartości: 2 dla pierwszej metody oraz 1,5 dla drugiego podejścia są wartościami umownymi, najczęściej spotykanymi w publikacjach z tego zakresu. Prawdziwy problem stanowić mogą dane wielowymiarowe i dane o różnych typach, znacznie częściej jakościowe niż ilościowe, gdy mamy do czynienia z wiedzą z dowolnej dziedziny. Aby móc zastosować metody opierające się na średniej i odchyleniu standardowym czy na kwartylach, musimy mieć jedną wartość określającą cały obiekt zamiast osobno wartości dla każdej cechy opisującej dany obiekt. Możemy dokonać pewnego rodzaju digitalizacji dzięki metodzie zaproponowanej przez Pawlaka [10] dla systemów wyszukiwania informacji. Aby przekształcić każdą informację wielowymiarową (a więc wektor) w jedną wartość numeryczną, i to tak by nie zaburzyć relacji między obiektami w rzeczywistości, dokonujemy kodowania wartości wszystkich atrybutów z osobna: numerujemy każdą dwójkę (atrybut, wartość) w każdym atrybucie od 1 do k (0 odpowiada sytuacji, gdy atrybut nie wystąpił w danej regule). W ten sposób każdej regule (części warunkowej) $(a_1 , v_1)*(a_2 , v_2) *,...,*(a_n , v_n)$ będzie odpowiadał ciąg liczb $b_1, b_2,...,b_n$. Następnie każdej regule przypisujemy unikalną wartość (będziemy ją dalej nazywać wartością reguły) za pomocą następującej formuły:

$$b_1 b_2 \ldots b_n \Leftrightarrow \sum_{i=1}^{n} b_i * U_i ,$$

w której wartości U_i obliczamy następująco: $U_1 = m_2 * m_3 * ... * m_n$, $U_2 = m_3 * ... * m_n$, ..., $U_{n-1} = m_n$ oraz $U_n = 1$, gdzie $m_i = cardV_{A_i}$. Wartości te będziemy nazywać wagami atrybutów. Przykład prezentuje rysunek 3.



Rys. 3. Wyznaczanie wartości dla każdej reguły
Fig. 3. Calculating values for rules

Jeśli weźmiemy pod uwagę regułę 1 z bazy wiedzy: h # 4 if a # 1 & b # 1, to – uwzględniając wartości każdego atrybutu – za pomocą proponowanej tu formuły możemy wyznaczyć wartość dla danej reguły. Jako że atrybut *a* ma wagę 823543, *b* zaś ma wagę 117649, to suma wartości wag tych dwóch atrybutów, tworzących część warunkową reguły pierwszej, wynosi 941192. Z kolei gdy weźmiemy do rozważań regułę ostatnią o postaci: h # 1 if f # 7 & g # 7, to suma wartości wag atrybutów warunkowych *f* i *g* z ich wartościami wyniesie tutaj jedynie 392 i będzie to wartość tak odstająca od reguł stanowiących resztę bazy wiedzy, że można będzie uznać tę regułę za odchylenie (stosując proste statystyczne założenie, że to, co np. nie mieści się w przedziale wartości średniej ± odchylenie standardowe, jest potencjalnym odchyleniem w danych).

3.2. Metody oparte na głębokości (odległości)

Podstawowa idea polega tu na pomiarze odległości (np. euklidesowej) *d* danego obiektu od jego *k*-sąsiadów. Różnice istnieją jedynie w metodzie wyznaczania odchyleń na podstawie wyznaczonej w ten sam sposób odległości [6, 7, 8]:

- Metoda Ramaswamy: *Dla każdego obiektu wyznaczamy odległość d jako odległość od pozostałych obiektów w jego k-sąsiedztwie. Pierwsze M obiektów o maksymalnej odległości uznajemy za odchylenia*.

- Metoda Angiulli i Pizzutti: *Dla każdego obiektu wyznaczamy odległość d jako odległość od pozostałych obiektów w jego k-sąsiedztwie. Następnie wyznaczamy sumę odległości obiektu od obiektów w jego k-sąsiedztwie. Pierwsze M obiektów o maksymalnej sumie odległości uznajemy za odchylenia*.

3.3. Metody oparte na gęstości

W literaturze najbardziej popularną metodą wykrywania odchyleń w danych wydaje się być metoda LOF (ang. *Local Outlier Factor*) Breuniga [11]. Analizujemy gęstość obiektów, w przestrzeni wielowymiarowej. Mała gęstość świadczy o niewielkim podobieństwie obiektów a zatem raczej sugeruje, iż dane do siebie nie pasują. Badamy tzw. lokalne sąsiedztwo obiektów i dla każdego z obiektów wyznaczamy tzw. wskaźnik LOF. Wysoka wartość LOF oznacza odchylenia w danych (i małą gęstość obiektów w sąsiedztwie obiektu analizowanego). LOF w porównaniu z innymi metodami nie jest podejściem binarnym: nie dokonuje jednoznacznej klasyfikacji obiektów jako tych, które na pewno są (albo na pewno nie są) odchyleniami. Zaletą LOF w stosunku do metod opartych na pomiarze odległości jest fakt, że algorytm LOF potrafi wykryć odchylenia, których nie da się wykryć, mierząc odległość obiektów względem siebie. Wówczas bowiem może zaistnieć sytuacja, w której znajdziemy dwie takie same odległości, jednak ewidentnie w przypadku jednej z nich będzie mowa o jednej tak dużej odległości obiektu od reszty obiektów o dużej gęstości, w drugim zaś przypadku o większej liczbie takich odległości, sugerujących mniej gęste skupienie. Wówczas odległość z pierwszego przypadku powinniśmy uznać za odchylenie, a z drugiego już nie. W takich sytuacjach algorytmy mierzące odległość się nie sprawdzą, a pomocne będą metody opierające się na pomiarze gęstości obiektów w zbiorze. Jeśli przez *q* określimy obiekt, dla którego mierzymy wskaźnik LOF, przez *p* zaś każdy kolejny obiekt z jego sąsiedztwa ($N_{k\text{-distance(q)}}$), wówczas algorytm LOF przebiega zgodnie z następującymi krokami:

1. Dla każdego obiektu w zbiorze wyznaczamy odległości *d* od jego *k*-najbliższych sąsiadów (*k-distance*) i określamy sąsiedztwo tego obiektu ($N_{k\text{-distance}(q)}(q)$):
$$N_{k-distance(q)}(q) = \{p \in R \setminus \{q\} \mid d(q,p) \le k - distance(q)\}.$$
2. Wyznaczamy tzw. odległość osiągalną (*reach-dist*) każdego obiektu od obiektów w jego sąsiedztwie (k): $reach - dist_k(q,p) = \max\{k - dis\tan ce(p), d(q,p)\}$.
3. Wyznaczamy tzw. lokalną osiągalną gęstość (lrd) każdego obiektu jako odwrotność średniej z osiągalnych odległości, opierającej się na sąsiedztwie tego obiektu:
$$lrd_k = 1/\left(\frac{\sum_{p \in N_{k-dis\tan ce(q)}} reach - dist_k(p,q)}{|N_{k-dis\tan ce(q)}q|}\right).$$

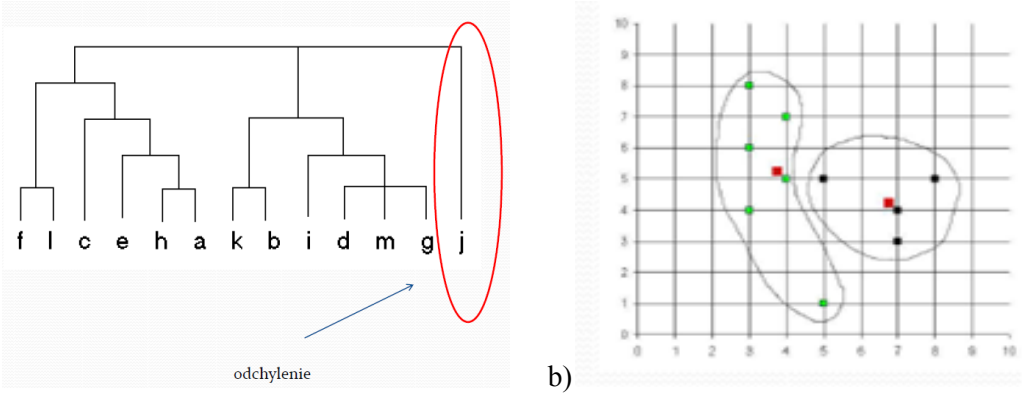
4. Obliczamy dla każdego obiektu jego wskaźnik LOF(q):

$$LOF_k(q) = \frac{\sum_{p \in N_{k-distance(q)}} \frac{lrd_k(p)}{lrd_k(q)}}{|N_{k-distance(q)}|}.$$

Póki wartość LOF dla danego obiektu jest mniejsza od 1, nie uznajemy obiektu za odchylenie.

3.4. Metody oparte na analizie skupień

Najbardziej interesujące, z punktu widzenia naukowych zainteresowań autorki niniejszego artykułu [4, 5], są metody wykrywające odchylenia w danych związane z grupowaniem danych. W literaturze znaleźć można wiele możliwych rozwiązań w zależności od wybranej metody grupowania: hierarchicznej bądź *k*-optymalizacyjnej. Wszystkie podejścia zakładają, że odchyleniem są obiekty niedające się włączyć do żadnej grupy bądź np. mało liczne grupy. Użycie metod analizy skupień do wykrywania odchyleń wydaje się być bardzo intuicyjnym podejściem. Znanych jest wiele algorytmów grupowania, które w wyniku pozwalają wykryć nie tylko skupienia, lecz także odchylenia od tych skupień. Zwłaszcza hierarchiczne algorytmy grupowania (np. AHC) (rys. 4a) bardzo łatwo radzą sobie z wykryciem odchyleń jako tych obiektów, które są dołączane do skupień w ostatnich etapach. Mając zbudowane skupienia, wystarczy wyznaczyć odległość obiektu od najbliższego skupienia oraz odległość między obiektem a reprezentantem najbliższego skupienia, by określić rozbieżność między indywidualnymi przypadkami obiektów a obiektami tworzącymi grupy (podobnymi). Potem rozbieżności takiej miary można używać do wykrywania odchyleń w danych. Są jednak i takie algorytmy grupowania (metody *k*-optymalizacyjne, rys. 4b), które nie są odporne na obiekty odległe i wtedy rezultaty grupowania mogą być znacznie obciążone właśnie istniejącymi odchyleniami. Innymi słowy, gdyby tych odchyleń nie było, inny byłby podział obiektów, inni reprezentacji skupień, a także mniejsze byłyby koszty obliczeniowe algorytmu (liczba iteracji, parametry kosztów grupowania).



a) b)

Rys. 4. Metody grupowania: a) metody hierarchiczne; b) metody k-optymalizacyjne
Fig. 4. Clustering methods: a) hierarchical methods; b) non-hierarchical clustering methods

Metody wykrywania odchyleń w danych na podstawie uprzedniego grupowania tychże danych są odpowiedzią na problem z dużymi zbiorami danych, dla których konieczność wyznaczania odległości każdego obiektu od pozostałych obiektów staje się nieefektywna zarówno czasowo, jak i pamięciowo.

Główne założenie: *dane normalne to takie które, należą do dużych i gęstych skupień, podczas gdy odchylenia nie należą do żadnych skupień bądź tworzą bardzo małe skupienia*.

W literaturze można znaleźć wiele podejść do wykrywania odchyleń w grupach obiektów. Najpopularniejsze są następujące metody:

- dane, które nie należą do żadnego skupienia, są odchyleniami,
- dane tworzące bardzo małe skupienia także są odchyleniami,
- dane należące do skupień o małej gęstości także należy uznać za odchylenia.

O ile w przypadku algorytmów hierarchicznych odchyleniami są obiekty dołączane do istniejących skupień na ostatnim etapie, a więc obiekty o niskim podobieństwie do grup powstałych we wcześniejszych krokach, o tyle w przypadku algorytmów k-optymalizacyjnych (niehierarchicznych), jak *k*-means czy *k*-medoid (omawiane już w wielu pracach z zakresu analizy skupień), niewykryte przed grupowaniem obiekty, będące odchyleniami, w znaczący sposób wpływają na jakość tworzonych skupień oraz zaburzają reprezentantów tak tworzonych grup. Dlatego konieczne wydaje się być opracowanie metod wykrywania odchyleń jeszcze przed procesem grupowania. Po pierwsze może to pozwolić na skrócenie czasu formowania grup, a po drugie pozwoli na lepszą separowalność nie tylko samych skupień, lecz także ich reprezentantów.

4. Wykrywanie odchyleń w regułowych bazach wiedzy

Gdy baza wiedzy ma strukturę skupień reguł podobnych do siebie, w procesie wnioskowania wystarczy przejrzeć tylko grupę o największym stopniu podobieństwa ze zbiorem podanych faktów. Czas wnioskowania zostaje znacznie zmniejszony (kilkudziesięciokrotnie), nie tracąc na efektywności (kompletności i dokładności) wnioskowania [5]. Tak jednak będzie tylko w przypadku, gdy grupy reguł będą faktycznie jednorodne i reprezentanci reguł będą jednoznacznie określać zawartość skupień. Zatem gdy skupienia będą tworzone przed wykryciem obiektów, będących odchyleniami od reszty, to założenie nie będzie spełnione. Tylko gdy zapewnimy odpowiednią separowalność grup i ich reprezentantów, pozwolimy, by w tak złożonych bazach wiedzy efektywność wnioskowania była na odpowiednim poziomie (łatwiej będzie znaleźć relewantne skupienie) [4, 5].

4.1. Podejście statystyczne

Opierając się na wartościach reguł wyznaczonych metodą prof. Pawlaka, prezentowaną w rozdziale 3.1 niniejszego artykułu, odchyleniami będziemy nazywać reguły, których wartości wykraczają poza uznany za normę przedział: $< średnia \pm 2 * odchylenie\_std >$ bądź $< Q1 - 1.5 * IQR; Q3 + 1.5 * IQR >$.

4.2. Podejście oparte na pomiarze odległości

Reguły, których odległość (np. euklidesowa) od pozostałych reguł w bazie wiedzy jest zbyt duża, będziemy uznawać za potencjalne odchylenia.

4.3. Podejście oparte na pomiarze gęstości

Gdy oprzemy się na pomiarze odległości euklidesowej, interesuje nas jedynie odległość danego obiektu, ale w ramach jego sąsiedztwa, nie zaś w całym zbiorze reguł. To my określamy, co rozumiemy przez sąsiedztwo (k – liczba sąsiadów). Wówczas gęstość mierzymy jako odległość reguły od reguł będących najbardziej do niej podobnych. Wysoka gęstość świadczy o spójności reguł stanowiących sąsiedztwo; analogicznie – mała gęstość świadczy o niepodobieństwie reguł, a więc o wykryciu reguły nietypowej (rzadkiej, odstającej od pozostałych).

4.4. Podejście oparte na grupowaniu

W przyszłości analizie poddane zostaną oba omawiane wcześniej przypadki: wykrywanie odchyleń przed grupowaniem oraz rezultat grupowania. W pierwszym przypadku do wykrycia odchyleń jeszcze przed grupowaniem zostanie zastosowane podejście statystyczne, tzn. mierząc w pierwszym kroku algorytmu aglomeracji reguł odległości (podobieństwo) wszystkich reguł od pozostałych, znajdziemy te, których odległości wykraczają poza zakres <średnia±odchylenie_std> bądź poza wartość Q3+1,5*IQR. Wykryte w ten sposób reguły odstające będą stanowić przedmiot dalszego zainteresowania inżyniera wiedzy, zbiór reguł niebędących odchyleniami będzie zaś podlegał grupowaniu. W drugim przypadku proces grupowania będzie zatrzymany, w momencie gdy podobieństwo tworzonych grup (bądź odległości grup podlegających grupowaniu) będzie zbyt niskie. Aby właściwie określić moment zatrzymania grupowania, a tym samym moment, w którym wszystkie dotąd niezwiązane reguły zostaną uznane za odchylenia, proponuje się wyznaczenie progu podobieństwa procentowo. Omawiany tu próg podobieństwa będzie wyznaczony w następujący sposób: *gdy w bieżącym kroku grupowania podobieństwo grup jest już mniejsze niż średnie podobieństwo*

w poprzednim kroku grupowania minus k% (gdzie właściwy poziom procentowy, np. 5% czy 10%, określać będzie użytkownik) proces grupowania powinien zostać przerwany.

5. Eksperymenty

Przebieg eksperymentów oraz analiza ich wyników zostaną przedstawione dla pojedynczej bazy wiedzy, składającej się z 23 reguł o zmiennej liczbie warunków w regułach (prezentowanej w rozdziale 1.1). Jednak należy mieć na uwadze, iż w trakcie badań przeanalizowano znacznie więcej przykładowych regałowych baz wiedzy, ale ich pełne przedstawienie znacznie wykroczyłoby poza ramy niniejszego artykułu. Celem eksperymentów jest znalezienie w bazie wiedzy reguł nietypowych (odchyleń). W sumie przeanalizowano 23 przypadki metod omówionych w rozdziale 4 (statystycznych, bazujących na odległości czy na gęstości). Implementacja i analiza metod opisanych w rozdziale 4.4, a więc opartych na analizie skupień, zostaną omówione w kolejnych pracach autorki. Wśród wykonanych eksperymentów znalazły się:

- metoda „ze średniej i odchylenia standardowego" (dla różnych wartości odchylenia standardowego = 1, 1,5, 2, 3),
- metoda „z rozstępu międzykwartylowego" (dla różnych wartości IQR = 1, 1,5, 2, 3),
- metoda oparta na pomiarze odległości (dla różnych wartości parametru *pct*, określającego procent obiektów stanowiących sąsiedztwo = {5%, 10%, 15%, 20%}),
- metoda LOF (dla różnych wartości k = 2, 3, 4, 5 oraz k równej pierwiastkowi z liczby reguł w bazie wiedzy),
- metoda Ramaswamy (dla różnych wartości k = 2, 3,4),
- metoda Angiulli i Pizzutti (dla różnych wartości k = 2, 3, 4).

Wyniki wykonanych eksperymentów przedstawia tabela 1.

Tabela 1

Czasy wnioskowania dla klasycznego podejścia oraz częściowych reguł decyzyjnych

Reguła	A1	B1	A2	B2	A3	B3	A4	B4	C1	C2	C3	D1	D2	D3	E1	E2	E3	E4	E5	F1	F2	F3	F4
0																+	+					+	+
1																+					+	+	+
2															+	+		+	+				+
3															+		+	+	+				+
4	+	+													+	+	+	+	+			+	+
5															+			+	+			+	+
6																			+				
7	+	+													+		+	+					
8	+	+	+	+		+									+	+	+	+	+				+
9	+	+		+											+	+		+	+				+
10	+	+		+												+	+		+			+	+
11	+	+		+												+	+						
12															+		+	+					+

Tabela 1

Czasy wnioskowania dla klasycznego podejścia oraz częściowych reguł decyzyjnych

13															+	+	+	+					+
14																			+				+
15																			+				+
16																			+				+
17									+					+	+	+		+	+			+	+
18	+	+													+			+	+				
19	+	+																					
20	+	+	+	+		+			+	+	+	+	+	+	+	+	+	+	+		+	+	+
21	+	+	+	+		+			+	+		+	+	+	+	+	+	+	+		+	+	+
22	+	+	+	+		+			+	+	+	+	+	+	+	+	+	+	+		+	+	+

Oznaczenia: R – reguła, A1 – odchylenie standardowe = 1, A2 – odchylenie standardowe = 1,5, A3 – odchylenie standardowe = 2, A4 – odchylenie standardowe = 3, B1 – rozstęp międzykwartylowy = 1, B2 – rozstęp międzykwartylowy = 1,5, B3 – rozstęp międzykwartylowy = 2, B4 – rozstęp międzykwartylowy = 3, C1 – Ramaswamy (k = 4), C2 – Ramaswamy (k = 3), C3 – Ramaswamy (k = 2), D1 – Angiulli i Pizzuti (k = 2), D2 – Angiulli i Pizzuti (k = 3), D3 – Angiulli i Pizzuti (k = 4), E1 – LOF (k = pierwiastek(liczba_reguł)), E2 – LOF (k = 2), E3 – LOF (k = 3), E4 – LOF (k = 4), E5 – LOF (k = 5), F1 – Pct(5%), F2 – Pct(10%), F3 – Pct(15%), F4 – Pct(20%), „+” oznacza, że dana reguła została uznana za odchylenie.

Analizując tabelę 1, dla każdej reguły możemy łatwo określić, jak często została uznana za potencjalne odchylenie. Wyniki dla każdej reguły przedstawiają się następująco: dla reguły „0” częstość bycia odchyleniem to 4, co stanowi 17,39%. Odpowiednio dla pozostałych reguł częstość wynosi: „1”: (4, 17,39%), „2”: (5, 21,74%), „3”: (5, 21,74%), „4”: (9, 39,13%), „5”: (5, 21,74%), „6”: (1, 4,35%), „7”: (5, 21,74%), „8”: (11, 47,83%), „9”: (8, 34,78%), „10”: (8, 34,78%), „11”: (5, 21,74%), „12”: (4, 17,39%), „13”: (5, 21,74%), „14”: (2, 8,70%), „15”: (2, 8,70%), „16”: (2, 8,70%), „17”: (8, 34,78%), „18”: (5, 21,74%), „19”: (2, 8,70%), „20”: (19, 82,61%), „21”: (18, 78,26%), „22”: (19, 82,61%). Zatem jeśli jako odchylenia chcemy traktować reguły, które zostały uznane za odchylenia przez min. 80% przeanalizowanych przypadków, odchyleniami będą tylko 3 reguły: o numerach odpowiednio 20, 21 oraz 22. Jeśli próg obniżymy do 40%, to oprócz tych trzech reguł (20, 21 oraz 22) za odchylenia uznamy także reguły 4 oraz 8.

6. Podsumowanie

W dużych zbiorach danych identyfikacja przypadków rzadkich bądź wyjątków (niespotykanych wzorców) staje się coraz bardziej potrzebna. Odchyleniem jest każdy obiekt, który nie pasuje do reszty obiektów w zbiorze.

Problem wykrycia odchyleń w danych w ostatnim czasie przyjął zupełnie inny kierunek w dziedzinie *data mining*. Kiedyś był jednym z elementów przygotowania danych do analizy (przypadki niepasujące do ogółu wyrzucano ze zbioru), do elementów wartych dalszej eks-

ploracji w celu wykrycia dotąd nieznanych zależności. Jednak dziś strategia uległa zdecydowanej zmianie: usuwanie danych uznanych za odchylenia naraża nas na utratę istotnej wiedzy z systemu. Zatem jedynym rozsądnym rozwiązaniem wydaje się być wykrywanie odchyleń przed eksploracją wiedzy w badanych zbiorze. Jednak problem wykrycia odchyleń w zbiorach wielowymiarowych jest niezwykle trudny, gdyż wymaga uwzględnienia wielu aspektów, takich jak: różne typy danych, brakujące wartości w danych, dane błędne, dane unikatowe. Różnorodność problemów związanych z zagadnieniem odchyleń w danych znalazła odzwierciedlenie w ogromnej liczbie prac z tego zakresu, które mają już nawet swój odrębny termin *outlier mining*.

Wczesne wykrycie odchyleń pozwoli na wzbogacenie wiedzy systemu o rzadką, aczkolwiek cenną wiedzę ekspertów, zwłaszcza w regułowych bazach wiedzy. Dlatego właśnie problem ten będzie dalej analizowany, a wyniki wykonanych badań zostaną zaprezentowane w kolejnych pracach z tego zakresu.

BIBLIOGRAFIA

1. Kaufman L., Rousseeuw P. J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley Sons, New York 1990.
2. Koronacki J., Ćwik J.: Statystyczne systemy uczące się. WNT, Warszawa 2005.
3. Nowak A.: Złożone bazy wiedzy: struktura i procesy wnioskowania. Rozprawa doktorska, Uniwersytet Śląski, Instytut Informatyki, Katowice 2009.
4. Nowak-Brzezińska A., Wakulicz-Deja A.: Wybór miary podobieństwa a efektywność grupowania reguł w złożonych bazach wiedzy. Studia Informatica, Vol. 31, No. 2A(89), Wydawnictwo Politechniki Śląskiej, Gliwice 2010, s. 189÷202.
5. Nowak-Brzezińska A.: Eksploracja wiedzy a efektywność systemów wspomagania decyzji. Studia Informatica, Vol. 32, No. 2A(96), Wydawnictwo Politechniki Śląskiej, Gliwice 2011, s. 403÷416.
6. Pearson R. K.: Mining imperfect data – dealing with contamination and incomplete records. SIAM, I-X, 1-305, 2005.
7. Seo S.: A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets. Univeristy of Pittsburgh, 2006.
8. Cherednichenko S.: Outlier Detection in Clustering. Master's Thesis, University of Joensuu, Department of Computer Science, 2005.
9. Hawkins D.: Identification of Outliers. Chapman and Hall, 1980.
10. Pawlak Z., Wiktor M.: Information storage and retrieval system – mathematical foundations. Computation Center Polish Academy of Sciences (CC PAS), Warsaw, Poland 1974.

11. Breunig M. M. et.al.: LOF: Identifying Density-Based Local Outliers. KDD 2000.

Wpłynęło do Redakcji 15 stycznia 2012 r.

Abstract

The paper presents the problem of outlier detection in the rule knowledge bases. Unusual (rare) rules, regarded here as the deviation, should be the subject of analysis of experts and knowledge engineers because they can influence the efficiency of inference in decision support systems.

The problem of detecting outliers in the data has recently adopted a completely different direction in the field of data mining. Firstly it was one of the elements of data preprocessing in data mining (finding missing values, outliers in data and removing them, data normalization). Today, however, it has a strong change of strategy: removing objects considered as outliers exposes us to significant loss of knowledge of the system. Extremely difficult is a problem with finding outliers in multidimensional datasets: data different type, missing values, deviation in data etc. The variety of problems associated with the issue of outlier detection in the data reflected in the huge number of works in this field – it is so called process of "outlier mining". In large data sets to identify rare cases or exceptions (unusual pattern) becomes more and more necessary.

Especially in knowledge bases area - early detection of outliers in rules will contribute to the knowledge system of a rare but valuable knowledge of experts. That is why this issue will be further analyzed and the results of research will be presented in subsequent work in this area.

The author presents a different approach in finding outliers in the rules: based on distribution, on distance, density or clustering. For each of them the most often used methods are presented with their advantages and disadvantages. The experiments with their results are also presented in the paper.

Adres

Agnieszka NOWAK-BRZEZIŃSKA Uniwersytet Śląski, Instytut Informatyki, Wydział Informatyki i Nauki o Materiałach, ul. Będzińska 39, 41-200 Sosnowiec, Polska, agnieszka.nowak@us.edu.pl.