STUDIA INFORMATICA 2013
Volume 34 Number 2A (111)

Agnieszka NOWAK-BRZEZIŃSKA
Instytut Informatyki, Uniwersytet Śląski

# WYKRYWANIE REGUŁ NIETYPOWYCH – METODY OPARTE NA ANALIZIE SKUPIEŃ

**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 strukturze skupień reguł. W artykule ujęto także wykonane eksperymenty wraz z interpretacją wyników.

**Słowa kluczowe**: obserwacje odstające, analiza skupień, regułowe bazy wiedzy, efektywność

# MINING OUTLIERS IN RULE KNOWLEDGE BASES – CLUSTERING BASED METHODS

**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 structure of rules' clusters. The experiments with their results are also presented in the paper.

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

## 1. Wprowadzenie

W ostatnich latach szczególnego znaczenia w dziedzinie eksploracji danych nabrały metody wykrywania odchyleń w danych. Wykrywanie defraudacji, nietypowe objawy chorobowe, nieautoryzowane włamania do serwerów to tylko niektóre z przykładów zastosowań wykrywania odchyleń. W pracy [5] przedstawiono metody typowo statystyczne oraz oparte na

badaniu gęstości danych, pozwalające wykrywać reguły nietypowe w regułowych bazach wiedzy. Niniejszy artykuł omawia metody wywodzące się z analizy skupień. Okazuje się, że obiekty nietypowe mogą znacznie zniekształcać tworzone skupienia, a przez to – utrudniać analizę danych. Jeśli analizowanymi obserwacjami będą reguły w bazach wiedzy, powiemy, że *nietypowe* będą reguły, które nie tworzą żadnego skupienia, bądź reguły tworzące skupienia mało liczne. Za nietypowe uznane będą także reguły, które znacząco wpływają na postać reprezentanta całego skupienia reguł bądź pogorszenie jakości skupień. Reguły takie – jako rzadkie – mogą 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.

## 2. Definicja reguł nietypowych

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. Nietypowe są nie tylko reguły sprzeczne (dla tych samych przesłanek – różne konkluzje) lecz także reguły w mało licznych skupieniach (nie udało się ich skleić z większymi grupami, mają nietypowe atrybuty bądź rzadkie wartości atrybutów). Nietypowe są też reguły *wpływowe* (wpływają istotnie (negatywnie) na jakość skupień bądź na postać reprezentanta grupy). W analizie regresji wyróżnia się 3 typy obserwacji nietypowych: odchylenia regresyjne (duże wartości rezyduów) – obserwacje odstające (ang. *outlier*), obserwacje wysokiej dźwigni (ang. *high leverage*) oraz tzw. obserwacje *wpływowe* (ang. *influential*). Popularnymi miarami wpływu poszczególnych obserwacji na równanie regresji są: DFFITS (ang. *Difference in FITs*), odległość Cooka czy DFBETAS (ang. *Difference in BETAs*) [3].

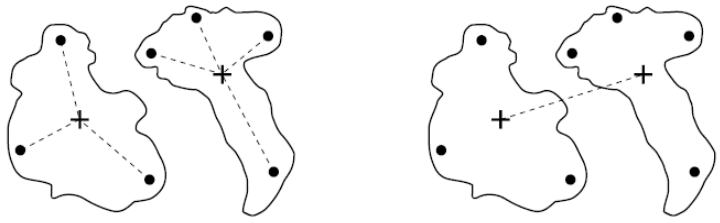
Wśród metod wykrywania odchyleń w danych wyróżnia się następujące metody:

1) oparte na rozkładzie danych (*distribution-based*),
2) oparte na odległości danych (*distance-based*),
3) oparte na gęstości (*density-based*),
4) oparte na grupowaniu (*clustering-based*) [6-10].

W pracy [5] przedstawiono wyniki analiz trzech pierwszych grup metod wraz z opisem wykonanych eksperymentów i interpretacją uzyskanych wyników. Ze względu na zaintere-

sowanie autorki tematyką analizy skupień oraz efektywnością wnioskowania w systemach wspomagania decyzji jako cel badań ustanowiono opracowanie metod wykrywania reguł nietypowych w strukturze skupień obiektów (w tym przypadku skupień reguł w bazie wiedzy). Wykrycie odchyleń jeszcze przed procesem generowania skupień pozwoli na skrócenie czasu formowania grup, a co ważniejsze – na lepszą separowalność nie tylko samych skupień lecz także ich reprezentantów. Lepsza separowalność grup i ich reprezentantów niewątpliwie wpłynie pozytywnie na efektywność wnioskowania w złożonych bazach wiedzy (łatwiej znaleźć relewantne skupienie). Te dwa czynniki stanowią kryteria przy opracowaniu optymalnej metody wykrywania nietypowych reguł. Wcześniejsze prace autorki [4, 5] pozwoliły zauważyć, że w algorytmach niehierarchicznych odchylenia (niestety) wpływają na jakość tworzonych grup oraz na reprezentantów, w algorytmach hierarchicznych zaś odchyleniami są najczęściej mało liczne grupy (jednoelementowe). Można też powiedzieć, że obserwacja typowa jest w skupieniu z innymi, podczas gdy nietypowa nie należy do żadnego skupienia. Ponadto typowe obserwacje leżą bliżej najbliższego centroidu, podczas gdy obserwacje nietypowe – daleko od najbliższego centroidu. Wreszcie możemy powiedzieć, że typowe obserwacje należą do dużych skupień, podczas gdy nietypowe – do małych. Pozostaje więc pytanie: jak wykrywać odpowiednio spójne i separowalne skupienia reguł?

Wśród wielu miar oceny jakości skupień dwie wydają się szczególnie przydatne: spójność i separowalność. Spójność (ang. *cohesion*) mówi o tym, jak blisko centrum grupy są obiekty, separacja zaś (ang. *separation*) o tym, jak wyraźne, jak bardzo odseparowane są grupy [1]. Można też użyć miary całkowitej odległości obiektów w grupie (ang. *total distance*), która bada, jak bardzo obiekty są odległe w grupie względem siebie. Spójność rozumiana jest tutaj jako suma podobieństw pomiędzy reprezentantem (centroidem, medoidem itp.) a obiektami wewnątrz grupy. Separacja oznacza odległość pomiędzy reprezentantami grup. Rys. nr 1 obrazuje spójność (lewa strona) oraz separowalność (prawa strona).



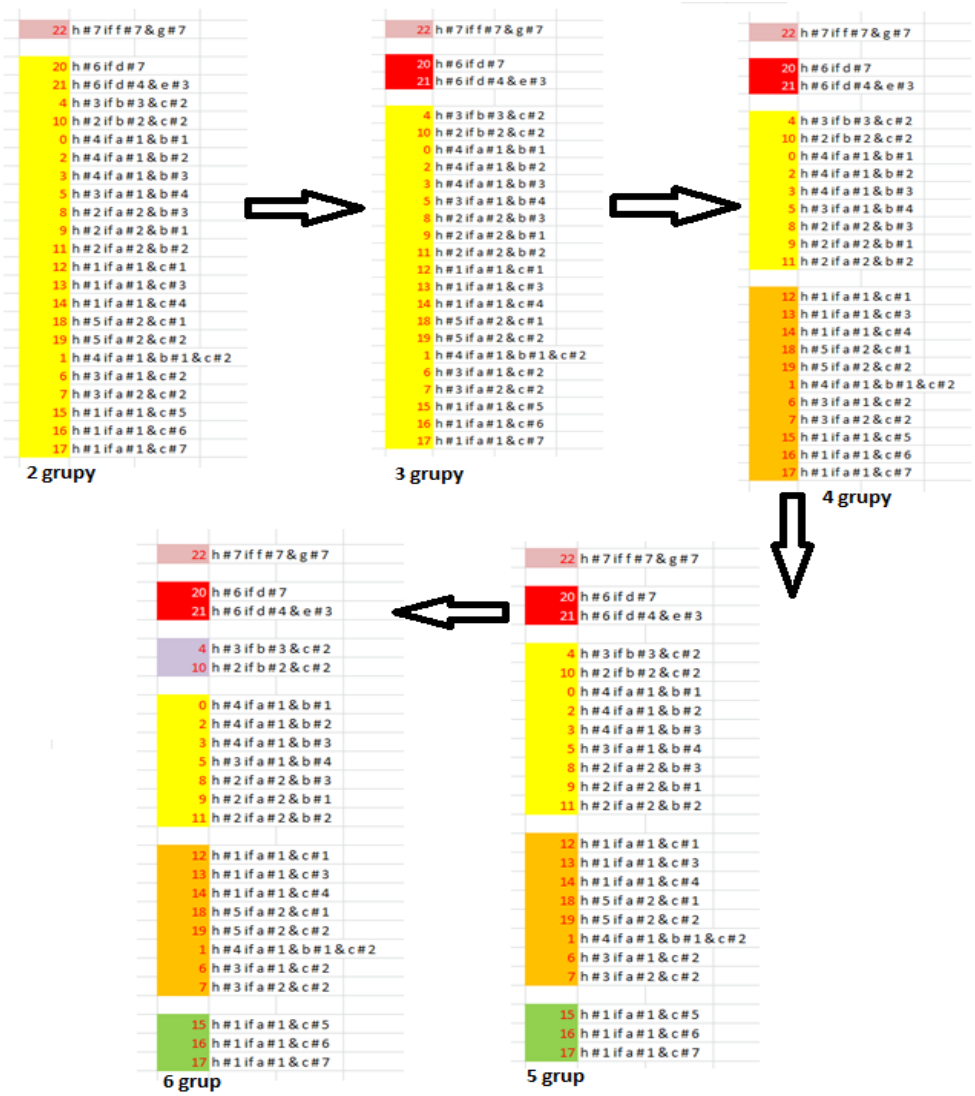
Rys. 1. Spójność (z lewej) i separacja (z prawej)
Fig. 1. Cohesion (at left) and separation (at right)

Miary oceny jakości skupień, o której tu mowa, mogą być wyrażone następująco:

- spójność $Cohesion = \sum_i \sum_{x \in c_i} d(x, m_i)^2$ ,
- separacja $Separation = \sum_i |C_i| d(m, m_i)^2$ ,

gdzie: $|C_i|$ – liczba elementów w $i$-tym skupieniu, $m_i$ – reprezentant $i$-tego skupienia, $m$ –reprezentant wszystkich skupień; $d(x, m_i)$ mierzy odległość obserwacji $x$ od reprezentanta $i$-tej grupy, $d(m, m_i)$ zaś mierzy odległość reprezentanta $i$-tej grupy od reprezentantów wszystkich grup.

W eksperymentach zrealizowanych na potrzeby niniejszego artykułu odległość mierzono odległością euklidesową. Model reprezentacji złożonych baz wiedzy przedstawiono m.in. w pracy [4], gdzie uwzględniono analizę wykorzystania różnych metryk podobieństwa oraz odległości do reprezentacji reguł w złożonych bazach wiedzy.




Rys. 2. Podział reguł na grupy (dla k=1, 2, …, 6)
Fig. 2. Clustering results (for k=1, 2, …, 6)

Kolejnym aspektem, który wiąże się z oceną jakości skupień, jest określenie optymalnej liczby grup. Można zauważyć, że im mniejsza jest wartość spójności, tym lepsze jest skupienie. Im większa jest wartość separacji, tym lepsze są skupienia, lepszy podział obiektów. Szukanie optymalnej liczby skupień przeprowadzono przez sprawdzenie wartości wyżej wy-

mienionych miar dla kolejnych wariantów liczby skupień (k=2, 3, 4, …). Okazało się, że optymalne wyniki dla analizowanej bazy wiedzy uzyskano przy liczbie skupień równej 6.

Wyniki podziału bazy wiedzy na grupy (dla k=1, 2, …, 6) przedstawiono na rys. 2.

## 3. Trójstopniowy proces wykrywania reguł nietypowych

W artykule proponuje się trójstopniowy proces wykrywania odchyleń w regułowych bazach wiedzy – AHCOB (Agglomerative Hierarchical Clustering Outlier Based).

### 3.1. Wykrywanie reguł sprzecznych

W pierwszym etapie następuje wykrycie reguł sprzecznych. Powiemy, że reguły r1 i r2 są sprzeczne, jeśli te same części przesłankowe tych reguł nie pociągają za sobą identycznych konkluzji. Wykorzystując kryterium podobieństwa, powiemy, że jeżeli podobieństwo części przesłankowych dwóch reguł r1 i r2 wynosi 1 (czyli mówi, że reguły są identyczne, jeśli patrzymy na część przesłankową), a podobieństwo części decyzyjnych jest mniejsze od jedności, wówczas reguły te uznamy za sprzeczne. Tak więc nietypowe (sprzeczne) będą reguły, dla których $sim(r1_C \text{ AND } r2_C)=1$, a $sim(r1_D \text{ AND } r2_D)<1$.

| | |
|---|---|
| 0 | h # 4 if a # 1 & b # 1 |
| 1 | h # 4 if a # 1 & b # 1 & c # 2 |
| 2 | h # 4 if a # 1 & b # 2 |
| 3 | h # 4 if a # 1 & b # 3 |
| 4 | h # 3 if b # 3 & c # 2 |
| 5 | h # 3 if a # 1 & b # 4 |
| 6 | h # 3 if a # 1 & c # 2 |
| 7 | h # 3 if a # 2 & c # 2 |
| 8 | h # 2 if a # 2 & b # 3 |
| 9 | h # 2 if a # 2 & b # 1 |
| 10 | h # 2 if b # 2 & c # 2 |
| 11 | h # 2 if a # 2 & b # 2 |
| 12 | h # 1 if a # 1 & c # 1 |
| 13 | h # 1 if a # 1 & c # 3 |
| 14 | h # 1 if a # 1 & c # 4 |
| 15 | h # 1 if a # 1 & c # 5 |
| 16 | h # 1 if a # 1 & c # 6 |
| 17 | h # 1 if a # 1 & c # 7 |
| 18 | h # 5 if a # 2 & c # 1 |
| 19 | h # 5 if a # 2 & c # 2 |
| 20 | h # 6 if d # 7 |
| 21 | h # 6 if d # 4 & e # 3 |
| 22 | h # 7 if f # 7 & g # 7 |

Rys. 3. Przykładowa regułowa baza wiedzy
Fig. 3. Example of rule knowledge base

W drugim etapie następuje wykrycie skupień jednoelementowych. Takie skupienia najwidoczniej mocno różniły się od pozostałych skupień, skoro nie udało się takiego skupienia włączyć do innych większych grup i uznano je za nietypowe. W trzecim etapie wykrywane są reguły wpływowe. Nietypowość tych reguł polega na tym, że ich obecność w strukturze sku-

pień znacznie wpływa na pogorszenie jakości utworzonych skupień i/lub na formę reprezentanta grupy.

Rozważania omówimy na przykładzie bazy wiedzy złożonej z 23 reguł prostych o zmiennej liczbie warunków (rys. 3).

Zakładając, że optymalną strukturą tej bazy wiedzy jest 6 skupień (przedstawionych na rys. 2), każdy z trzech etapów jest wykonywany już w ramach utworzonych skupień. I tak w pierwszej grupie reguły 7 oraz 19 uznane będą za sprzeczne.

### 3.2. Wykrywanie skupień mało licznych

W drugim etapie, gdzie następuje wykrywanie reguł nietypowych jako tych, które tworzą mało liczne skupienia, istotne jest określenie sposobów wykrycia takich właśnie skupień. Możliwych jest wiele rozwiązań:

- wyznaczenie rankingu odległości reguł od reprezentantów skupień. Za nietypowe uznaje się wówczas te reguły, których odległość od reprezentanta skupienia jest największa bądź mieści się np. w 10 % największych odległości;
- wyznaczenie liczności skupień, i jako nietypowych oznaczenie skupień jednoelementowych bądź mało licznych (np. 1% wszystkich reguł);
- na podstawie odległości reguły od reprezentanta wyznaczenie prawdopodobieństwa bycia członkiem grupy i uznanie reguł, których prawdopodobieństwo jest mniejsze niż założony próg, za nietypowe;
- znalezienie maksymalnej odległości obiektu od skupienia $d_{max} = \max_i \{\|x_i - C_i\|\}$ – a następnie wyznaczenie dla każdej reguły współczynnika jej odchylenia od grupy jako $o_i = \frac{\{\|x_i - C_i\|\}}{d_{max}}$. Wówczas jeśli $o_i$ > T dla *i*-tej reguły, to uznajemy ją za odchylenie: $Out <- Out \cup \{x_i\}$.

Spośród dostępnych metod w pierwszych eksperymentach użyto metody wykrywania skupień mało licznych przez określenie minimalnej (w %) liczności skupienia. Proponowany algorytm przedstawia się następująco:

```
Wejście:
   N – reguł w BW, K – skupień reguł
Wyjście:
   Out – zbiór odchyleń,
   Begin
   Out<-ϕ
   Wyznacz odległości między danymi
   Utwórz K skupień
   Dla każdego skupienia sprawdź: Czy sizeof(Ki) < t ?
   Jeśli tak, to: Out<-Out ∪{r ∈ Ki}
   End;
```

Ten etap dla analizowanej bazy wiedzy wykaże grupę drugą jako najmniej liczną (jednoelementową), a więc regułę 22.

### 3.3. Wykrywanie reguł wpływowych

W ostatnim etapie w ramach każdej grupy wykrywane są tzw. reguły *wpływowe*. Reguła jest *nietypowa i wpływowa* dla danej grupy, jeśli współczynniki jakości skupienia znacznie się zmienią (poprawią) po usunięciu tej reguły ze skupienia bądź gdy zmienia się znacząco reprezentant skupienia.

Algorytm wykrywania reguł nietypowych (wpływowych) przedstawia się następująco:

```
Wejście:
   K={K1,K2,..,Kk}  – zbiór skupień reguł
Wyjście:
   Out – zbiór reguł uznanych za wpływowe
   Begin
   Out<-ϕ
   For i=1 to K do
   Dla każdej j-tej reguły w skupieniu Ki sprawdź, jaka jest wartość parametru
   oceny skupień z j-tą regułą i bez niej. Jeśli różnica jest istotna, to jest
   to potencjalne odchylenie.
   Out<-Out ∪{rj ∈ Ki}
   End;
```

Wykrywanie wpływu poszczególnych reguł może zatem zwrócić inne rezultaty w zależności od parametru, na który reguła miałaby wpływać. Takimi parametrami będą: postać reprezentanta skupienia, wartość spójności i wartość separacji. I tak okazało się, że:

- regułami wpływowymi przy wzięciu pod uwagę postaci reprezentanta są reguły 20, 21 (z grupy 3) oraz 4 (z grupy 6),
- regułami wpływowymi przy wzięciu pod uwagę spójności grup są reguły 20,21 (z grupy 3) oraz 14 (z grupy 1),
- regułami wpływowymi przy wzięciu pod uwagę separacji są reguły 1, 6 i 12 (wszystkie z grupy 1). Wyniki te zostały zaprezentowane w tabeli 1 w następnym rozdziale.

## 4. Eksperymenty

W niniejszym rozdziale przedstawiono wyniki przeprowadzonych eksperymentów wraz z interpretacją wyników. Tabela 1 przedstawia reguły nietypowe w zależności od przyjętego kryterium oceny nietypowości (reguły sprzeczne, mało liczne skupienia czy reguły wpływowe); są one oznaczone symbolem „+”.

W ramach eksperymentów badano, jak czuły i wrażliwy na obiekty nietypowe jest proponowany algorytm AHCOB. W tym celu wyznaczono znane z kontroli jakości klasyfikatorów

miary czułości i specyficzności [2]. Jeżeli przez TN (ang. *True negative*) określimy liczbę reguł faktycznie typowych i przez algorytm określonych jako typowe, przez FP (ang. *False positive*) zaś liczbę reguł faktycznie typowych, choć przez system określonych jako nietypowe, wówczas specyficzność (ang. *Specificity*) będzie można mierzyć jako:

$$Specifity = \frac{TN}{FP + TN}.$$

Tabela 1

Reguły nietypowe – wyniki dla algorytmu AHCOB

| Reguła | Reguły sprzeczne | Skupienia małe | Reguły wpływowe | | |
|---|---|---|---|---|---|
| | | | reprezentant | spójność | separacja |
| **0** | | | | | |
| **1** | | | | | + |
| **2** | | | | | |
| **3** | | | | | |
| **4** | | | + | | |
| **5** | | | | | |
| **6** | | | | | + |
| **7** | + | | | | |
| **8** | | | | | |
| **9** | | | | | |
| **10** | | | | | |
| **11** | | | | | |
| **12** | | | | | + |
| **13** | | | | | |
| **14** | | | | + | |
| **15** | | | | | |
| **16** | | | | | |
| **17** | | | | | |
| **18** | | | | | |
| **19** | + | | | | |
| **20** | | | + | + | |
| **21** | | | + | + | |
| **22** | | + | | | |

Jeżeli zaś przez TP (ang. *True positive*) określimy liczbę reguł faktycznie nietypowych i jako takie określonych przez algorytm, przez FN (ang. *False negative*) zaś liczbę reguł faktycznie nietypowych, choć przez system sklasyfikowanych jako typowe, wówczas czułość (ang. *Sensitivity*) będziemy mierzyć jako:

$$Sensitivity = \frac{TP}{TP + FN}.$$

Wyniki przeprowadzonych w tym celu eksperymentów przedstawia tabela 2.

Czułość i specyficzność przy wykrywaniu reguł sprzecznych były maksymalne (100%), gdyż system podobnie jak ekspert dziedzinowy określił reguły 7 i 19 jako sprzeczne (i takie faktycznie były). Podobnie metody wykrywające skupienia mało liczne okazały się w pełni czułe i specyficzne (reguła 22 została uznana za nietypową, bo nią faktycznie była).

W przypadku wykrywania reguł wpływowych tylko metoda badająca wpływ na spójność grup okazała się maksymalnie czuła i specyficzna (wykryła reguły 14 oraz 20 i 21 i uznała je za wpływowe, jeśli chodzi o wartość spójności, bo faktycznie po ich usunięciu znacznie poprawiała się spójność w grupach). Parametry czułości i specyficzności uległy pogorszeniu w przypadku metod wykrywania reguł wpływowych ze względu na zmiany w reprezentancie grupy, separację czy całkowitą odległości reguł w skupieniach. Czułość metody wykrywającej reguły wpływowe przy wzięciu pod uwagę wpływu na reprezentanta wyniosła 67%, gdyż spośród trzech reguł faktycznie wpływowych (20, 21 oraz 1) algorytm AHCOB wykrył jedynie dwie: 20 i 21. W przypadku metody wykrywania reguł wpływających na separację czułość wyniosła jedynie 33%, gdyż spośród trzech reguł faktycznie wpływowych system wykrył jedynie jedną (regułę 1). Z kolei reguły 6 i 12, które były typowe, system błędnie wskazał jako nietypowe, co pogorszyło specyficzność (czasami nazywaną wrażliwością) metody.

Tabela 2

Czułość i specyficzność algorytmu AHCOB

| | Czułość (ang. *Sensitivity*) | Specyficzność (ang. *Specificity*) |
|---|---|---|
| Wykrywanie reguł sprzecznych | 100% | 100% |
| Wykrywanie skupień mało licznych | 100% | 100% |
| Wykrywanie reguł wpływowych (wpływ na postać reprezentanta) | 67% | 95% |
| Wykrywanie reguł wpływowych (wpływ na spójność) | 100% | 100% |
| Wykrywanie reguł wpływowych (wpływ na separację) | 33% | 90% |

Z wykrywaniem reguł nietypowych wiążą się dwa dość istotne problemy, określone w literaturze jako *masking* i *swamping*. *Masking*, nazywany maskowaniem odchyleń, polega na pominięciu „mniejszych" odchyleń w danych, s*wamping* zaś – na mylnym uznaniu obserwacji za nietypową, podczas gdy była ona jak najbardziej typowa. Nietypowe reguły w bazach wiedzy mogą przez *masking* i *swamping* rodzić błędy dwojakiego rodzaju:

- *błąd I rodzaju* – gdy nie znajdziemy reguły, która była nietypowa, mimo że taka była w bazie wiedzy,
- *błąd II rodzaju* – gdy regułę typową uznamy za nietypową.

Dużo bardziej istotny jest, zdaniem autorki, błąd I rodzaju, gdyż nie wykrywając pewnych faktycznie nietypowych reguł, możemy pozbawić ekspertów istotnej informacji o tym, jakie zależności wykryte w danych są szczególnie ciekawe i przez to warte dalszej eksploracji. Szczęśliwie przedstawione wyniki pozwalają uznać algorytm AHCOB za wystarczająco wrażliwy na reguły nietypowe. Minimum 90% reguł nietypowych było poprawnie sklasyfikowanych. Czułość tylko w jednym przypadku była niższa niż 50% (tylko w przypadku analizy separacji), co oznacza, że dużo więcej reguł uznano za nietypowe, niż to miało miejsce w rzeczywistości.

## 5. Podsumowanie

Eksploracja odchyleń (tzw. *outlier mining*) stała się w ostatnim czasie niezwykle ważnym elementem odkrywania wiedzy z danych. W regułowych bazach wiedzy wykrycie nietypowych reguł z pewnością będzie ogromnym wsparciem dla ekspertów i inżynierów wiedzy (pomoc w weryfikacji kompletności i spójności bazy wiedzy). W kontekście metod grupowania dużych zbiorów reguł wykrywanie reguł nietypowych może wpłynąć na strukturę skupień reguł, na metodę tworzenia reprezentantów. Gdy baza wiedzy jest stosunkowo jednorodna, może dochodzić do uznania za nietypowe reguł, które wcale nie są nietypowe (*swamping*). Wszystko zależy od: rozkładu danych (spójności reguł), metody tworzenia reprezentanta, metody tworzenia nowych skupień (SL,CL,AL) oraz dodatkowych parametrów wykrywania odchyleń. Dokładna analiza wszystkich możliwych czynników wpływających na zwiększenie efektywności wykrywania reguł nietypowych będzie podstawą przyszłych badań autorki.

**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. Koronacki J., Mielniczuk J.: Statystyka dla studentów kierunków technicznych i przyrodniczych. WNT, Warszawa 2009.
4. Nowak A.: Złożone bazy wiedzy: struktura i procesy wnioskowania. Rozprawa doktorska, Uniwersytet Śląski, Wydział Informatyki i Nauki o Materiałach, Katowice 2009.
5. Nowak-Brzezińska A.: Eksploracja odchyleń w regułowych bazach wiedzy. Studia Informatica, Vol. 33, No. 2A (105), Gliwice 2012, s. 479÷492.
6. Pearson R. K.: Mining imperfect data – dealing with contamination and incomplete records. SIAM, I-X, 1-305, 2005.

7. Breunig M., Kriegel H.-P., Ng R. T., Sander J.: LOF: Identifying Density-Based Local Outliers. KDD, 2000.
8. Cherednichenko S.: Outlier Detection in Clustering. University of Joensuu, Department of Computer Science, Master's Thesis, 2005.
9. Hawkins D.: Identification of Outliers. Chapman and Hall, 1980.
10. Loureiro A., Torgo L., Soares C.: Outlier Detection Using Clustering Methods: a data cleaning application. Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector, Bonn, Germany 2004.

Wpłynęło do Redakcji 8 stycznia 2013 r.

**Abstract**

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

The problem of discovering outliers in data has recently gained a completely different meaning in the field of data mining. At first, it was a part of the preprocessing process in data mining, but at present it is believed that removing outliers can lead to a significant loss of knowledge which potentially could be discovered. In large data sets identifying rare cases (unusual patterns) becomes more and more necessary. Especially, in the case of knowledge bases, early detection of unusual rules will eventually lead to extracting novel and useful knowledge for domain experts. That is why this issue will be further analyzed and the results of the research will be presented in subsequent work in this area.

In a previous paper [5] the author presented different approaches for finding outliers from a set of rules: based on distribution, distance, density or clustering and discussed the results from conducted experiments. This article presents a method of finding rare rules in knowledge bases by generating clusters of rules. Besides the information about advantages and disadvantages of the presented algorithm (called AHCOB) the experiments as well as the interpretation of achieved results are also included.

**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.