Cykl powiązanych tematycznie publikacji

Integracja informacji międzyklasowej i wewnątrzklasowej w procesie redukcji atrybutów dla celów deskryptywnych

Autoreferat

Robert Susmaga

Instytut Informatyki
Politechnika Poznańska
Piotrowo 2
60–965 Poznań

Poznań, 21 kwietnia 2015 r.

Anielsko cierpliwym Najbliższym poświęcam

Autor

Prace wchodzące w skład cyklu powiązanych tematycznie publikacji

Czasopisma naukowe

[A01] R. Susmaga, „Reducts and Constructs in Attribute Reduction”, *Fundamenta Informaticae*, **61** (2), 2004, 159–181.
[lista A MNiSW(12.2014): 15pkt]
[IF(2004): 0,785]

[A02] R. Susmaga, „Reducts and Constructs in Classic and Dominance-based Rough Sets Approach”, *Information Sciences*, 271 (2014), 45–64.
[lista A MNiSW(12.2014): 45pkt]
[IF(2014): 3,893]

[A03] R. Susmaga, R. Słowiński, „Generation of Rough Sets Reducts and Constructs based on Inter-class and Intra-class Information”, *Fuzzy Sets and Systems*, (2014), dx.doi.org/10.1016/j.fss.2014.06.012.
[lista A MNiSW(12.2014): 40pkt]
[IF(2004): 1,880]
(udział pierwszego autora: 60%)

Materiały konferencyjne / monografie (w tym seria LNAI/LNCS)

[A04] R. Susmaga, „Reducts Versus Constructs: An Experimental Evaluation”, *Proceedings of the Workshop „Rough Sets in Knowledge Discovery”*, Warsaw, April 12–13, 2003, *Electronic Notes in Theoretical Computer Science*, **82** (4), Elsevier, 2003, 239-250.

[A05] R. Susmaga, „Tree-Like Parallelization of Reduct and Construct Computation”, [w] S. Tsumoto, R. Słowiński, H. J. Komorowski, J. W. Grzymala-Busse (red.), *Proceedings of the Rough Sets and Current Trends in Computing, 4th International Conference, RSCTC 2004*, Uppsala, Sweden, June 1–5, 2004, *Lecture Notes in Computer Science*, **3066**, Springer 2004, 455-464.

Wstęp

Reprezentowany w pracach prezentowanego cyklu publikacji obszar badań należy do nurtu informatyki zawiązanej z eksperymentalną analizą i eksploracją danych. Metody usprawniania procesów analizy i eksploracji w tym obszarze można w ogólności podzielić na szczegółowe oraz kompleksowe. Podczas gdy pierwsze mają na celu rozwijanie konkretnych procedur tego procesu w ustalonych z góry kierunkach, drugie służą raczej identyfikowaniu tych spośród potencjalnie niezmiernie licznych, istniejących kierunków rozwoju, które wydają się najbardziej obiecujące. Przydatne w tym miejscu okazuje się często integrowanie odmiennych technik i metod postępowania, które mogą okazywać się wzajemnie inspirujące i udoskonalające.

Prezentowany w niniejszym opracowaniu cykl powiązanych tematycznie publikacji wychodzi od wywodzącego się z teorii zbiorów przybliżonych pojęcia reduktu i oznaczającego podzbiór atrybutów, który jest minimalny, ale wystarczający do rozróżniania obiektów z różnych klas na poziomie gwarantowanym przez wszystkie atrybuty. Redukt może więc być traktowany jako selektor atrybutów, jednak o deskrypcyjnych raczej niż predykcyjnych, właściwościach. Wprowadzone następnie środowisko redukcji atrybutów wykorzystuje szersze niż we wcześniejszych podejściach zdyskontowanie informacji o przynależności obiektów do klas. Fakt ten pozwala na generowanie wyników na tyle ogólnych, że mogą być one potencjalnie skuteczniej wykorzystane, także we współpracy z metodami indukcyjnymi o dobrych właściwościach predykcyjnych (np. klasyfikatory obiektów). Tak uogólniony proces redukcji atrybutów może być dalej rozwijany przez proponowanie konkretnych algorytmów, np. algorytmów dokładnych, generujących rozwiązania wyczerpujące, niestety niezwykle kosztownych. Problemy związane z dużymi nakładami obliczeniowymi algorytmów dokładnych mogą być dalej pokonywane poprzez rozwój samych algorytmów w kierunkach nowych technik obliczeniowych, np. do postaci równoległej. Zrównoleglenie obliczeń pozwala na naturalne wykorzystanie szeroko rozpowszechnionych dziś systemów wieloprocesorowych i w wielu przypadkach przynosi zauważalne zyski.

Postępowaniem alternatywnym jest proponowanie algorytmów przybliżonych, kontrolowanych specjalizowanymi miarami w celu dostosowywania ich wyniku do konkretnych danych i oczekiwanych efektów. Z technicznego punktu widzenia, rozwiązania heurystyczne mogą być generowane szczególnie łatwo wtedy, gdy znane są algorytmy dokładne (ponieważ rozwiązanie heurystyczne stanowią ich zrelaksowane wersje). Skuteczne realizowanie procesu generowania rozwiązań przybliżonych nie gwarantuje jednak tego, że znalezione rozwiązania okażą się rozwiązaniami korzystnymi we wszystkich możliwych zastosowaniach. Dlatego właściwy dobór miar kontrolujących ten proces, w tym przypadku miar atrakcyjności atrybutów sterujących doborem tych atrybutów w procedurach heurystycznych, jest kluczowy z punktu widzenia przydatności ostatecznego wyniku (i ma na niego znaczący wpływ). Ich eksploracja stanowi więc kompleksowy poziom analizy danych, w odróżnieniu od rozwijania omówionych wcześniej aspektów algorytmicznych, zajmujących poziom szczegółowy.

Metody redukcji atrybutów dla danych nadzorowanych

Przedstawione poniżej metody są metodami indukcyjnymi przetwarzania nadzorowanych danych wielowymiarowych, przez które rozumie się wielowymiarowe opisy obiektów pochodzących z pewnego ustalonego zbioru obiektów. Wielowymiarowość opisów obiektów oznacza, że każdy pojedynczy obiekt opisany jest specyficznym wektorem deskryptorów odpowiadających atrybutom pochodzącym z ustalonego zbioru atrybutów. Z kolei fakt, że dane są nadzorowane oznacza, że obiekty zostały podzielone na odgórnie ustalone klasy, czyli zbiory skupiające obiekty o określonych właściwościach, wspólnych dla wszystkich obiektów w danej klasie (w sytuacji, gdy taki podział nie istnieje, dane noszą nazwę nienadzorowanych). Informacja o przynależności obiektu do danej klasy reprezentowana jest zwyczajowo w dodatkowym, specjalizowanym atrybucie, zwanym atrybutem decyzyjnym.

W kontekście statystycznej analizy danych, wektor wartości atrybutów opisujących jeden ustalony obiekt odpowiada obserwacji, a wektor wartości jednego ustalonego atrybutu opisujących wszystkie obiekty odpowiada zmiennej.

Dla tak rozumianych danych istotą przedstawianych zadań indukcyjnych jest w ogólności odkrywanie i budowanie modeli reprezentujących wielowymiarowe zależności między atrybutami (w innych kontekstach: zmiennymi, cechami), zgodne z opisami reprezentującymi zawarte w analizowanym zbiorze danych obiekty (w innych kontekstach: obserwacje, przykłady), z uwzględnieniem informacji o klasach, do których te obiekty należą. W szczególności, wobec faktu jednoznacznego wyrażania informacji klasowej przez atrybut decyzyjny, zadanie indukcyjne może być traktowane jako odkrywanie i budowanie modeli reprezentujących wielowymiarowe zależności między atrybutem decyzyjnym a pozostałymi atrybutami (wielowymiarowość w tym kontekście oznacza, że zależność dotyczy zbiorów atrybutów).

Tworzone modele powinny być dostosowane do analizowanej populacji, a ich tworzenie może być inspirowane różnymi celami i związanymi z nimi założeniami. W szczególności:

- deskrypcyjny cel analiz zakłada, że posiadany zbiór danych stanowi tę populację,
- predykcyjny cel analiz zakłada, że posiadany zbiór danych jest próbą tej populacji.

Najistotniejsze wymagania stawiane metodom generującym modele to przede wszystkim skuteczność i szybkość, budowanym przez nie modelom – jakość i łatwość interpretacji.

Jednoczesne spełnianie wszystkich z powyższych wymagań jest w ogólnym przypadku co najmniej bardzo trudne dla celów deskrypcyjnych i zasadniczo niemożliwe dla celów predykcyjnych. Jest to wynikiem zależności pomiędzy poszczególnymi wymaganiami, które idą zwyczajowo w wykluczających się parach, np. szybkość działania metody często obniża jej skuteczność, a wysokiej jakości tworzonego modelu nie zawsze towarzyszy jego zadowalająca łatwość interpretacji. Sytuację dodatkowo komplikują niejednoznaczności deskryptorów danych, które mogą być w ogólności zakłócone, niedokładne czy (w rozmaitym sensie) niespójne. W przypadku celów predykcyjnych analiz nieuniknionym zjawiskiem jest przeuczenie, którego kontrolowanie wymaga dalszych środków. W rezultacie zdecydowana

większość metod oferuje rozwiązania przybliżone, i jedynie nieliczne metody gwarantują rozwiązania optymalne pod wybranymi względami.

Metodą usprawniającą rozwiązywanie problemów o dużej złożoności jest ich dekompozycja na problemy składowe, które mogą być rozwiązywane mniejszym nakładem środków. W procedurze budowania modeli reprezentujących wielowymiarowe zależności między atrybutem decyzyjnym a pozostałymi atrybutami takim przydatnym problemem etapu pośredniego jest wartościowanie atrybutów i/lub ich podzbiorów (w innych kontekstach: selekcja cech). Wyniki tego wartościowania często stanowią wartość samą w sobie. W rezultacie jego zastosowania możliwe jest wprowadzenie podziału atrybutów na przydatne (wykorzystywane dalszych analizach) i nieprzydatne (ignorowane w dalszych analizach) [C04].

Oprócz oczywistych zysków wynikających z faktu, że atrybuty uznane za nieprzydatne nie muszą być dalej przechowywane ani pozyskiwane, co pozwala na zwolnienie zasobów komputerowych i zmniejszenie szeroko rozumianych kosztów akwizycji danych, wykorzystanie w dalszych analizach jedynie atrybutów uznanych za przydatne może znacząco przyczynić się do przyspieszenia obliczeń, a także do zwiększenia jakości budowanych przez nie modeli. Inne potencjalne redukcyjne analizy danych (nie uwzględniane w opisywanych pracach) obejmują np. ewaluację i redukcję obiektów.

Jednym z popularnych podziałów metod wartościowania/redukowania atrybutów jest podział [C07] na:

- metody typu „wrapper” (ang. „metody opakowujące”),
- metody typu „filter” (ang. „metody filtrujące”).

(poniżej wykorzystane będą angielskie terminy).

Model typu „wrapper” jest bardzo specyficzny – dla uprzednio wybranych: miary trafności klasyfikacji, klasyfikatora i procedury walidacyjnej, model ten znajduje w pracochłonnym eksperymencie zbiór atrybutów maksymalizujący wartość danej miary osiąganą przez dany klasyfikator w danym teście. W tym sensie uzyskiwany wynik jest wysoce „specjalizowany”, i oczywiście wysoce zależny od wybranych składowych eksperymentu (miary, klasyfikatora i procedury). W przypadku bardzo dużych zbiorów danych model może być dodatkowo niezwykle kosztowny obliczeniowo, a w przypadku bardzo małych może generować mało wiążące wyniki. Jak się też okazuje, oprócz maksymalnej zarejestrowanej wartości miary trafności klasyfikacji model ten nie dostarcza jawnie żadnych dodatkowych argumentów uzasadniających wygenerowany wynik. W rezultacie, wyniki generowane przez ten model są wprawdzie często bardzo skuteczne (w sensie wybranej miary), ale za to słabo interpretowalne i mało informatywne, co za tym idzie, mało podatne na uogólnienia.

W modelu „filter” zbiór atrybutów jest wyznaczany jako zbiór spełniający jak najlepiej określone z góry, wybrane warunki. Modele tego typu są więc właściwie środowiskami do testowania istniejących i proponowania nowych, mniej czy bardziej specjalizowanych warunków na przydatność atrybutów (i ich zbiorów). Jako takie, generują one dużo ogólniejsze wyniki, ponieważ dobry i skuteczny zestaw warunków może być stosowany do rozmaitych danych, także w zupełnie innych obszarach zastosowań. Jest to szczególnie przydatne w analizach, którym przyświecają cele

deskrypcyjne, gdzie często wymagana jest wysoka precyzja opisu rozważanej populacji obiektów. Wyniki tego modelu są też dość naturalnie interpretowalne (zgodnie z zastosowanymi warunkami).

Podsumowanie modeli typu „filter" i „wrapper" można wyrazić następująco:

- pierwszy odpowiada na pytanie „jakie...",
- drugi odpowiada na pytanie „które...",

„... atrybuty powinny być wartościowane pozytywnie (i zachowane), a które negatywnie (i usunięte)". Charakterystyki tych modeli są więc dość odmienne, co oznacza, że powinny być one właściwie dobierane do poszczególnych zastosowań. Oczywiście nie ma żadnych przeciwwskazań do jednoczesnego testowania czy nawet potencjalnego stosowania obu modeli.

Prezentowane poniżej metody redukcji atrybutów implementują zasadniczo rozwiązania charakterystyczne dla modelu typu „filter" (choć ich formy weryfikacji obejmują także procedury walidacyjne właściwe dla modelu „wrapper"). Wyniki tych metod cechują się więc przede wszystkim wysoką precyzją i łatwością interpretacji (wymóg celu deskrypcyjnego), i potencjalnie mniejszą skutecznością klasyfikacyjną (wymóg celu predykcyjnego). Do ich generowania zaprezentowano algorytmy dokładne jak i heurystyczne. W przypadku tych pierwszych przedstawiono i przebadano techniki usprawniające ich działanie (szczegółowy poziom analiz). W przypadku tych drugich zarysowano potencjalne techniki pozwalające na analizowanie specjalizowanych współczynników kontrolujących ich działanie (kompleksowy poziom analiz).

Metodyki Zbiorów Przybliżonych w eksploracji danych

Metodyka RSA (ang. Rough Sets Approach, RSA) opiera swoje analizy na ustalonej relacji, zwanej relacją bazową. W zależności od przyjętej relacji bazowej, wyróżnia się:

- klasyczną metodykę zbiorów przybliżonych [C09] (ang. Classic RSA, CRSA) – relacja bazowa: nierozróżnialność (ang. indiscernibility, IND),
- dominacyjną metodykę zbiorów przybliżonych [C10] (ang. Dominance-based RSA, DRSA) – relacja bazowa: dominacja (ang. dominance, DOM).

W przypadku CRSA, dla ustalonego zbioru atrybutów, relacja bazowa wyznacza na zbiorze obiektów klasy abstrakcji, nazywane zbiorami elementarnymi. Zbiory te stanowią najbardziej naturalne „granule", z których powinny składać się analizowane podzbiory obiektów. Zawieranie się w jakimś zbiorze obiektów tylko części którejś z „granul" jest traktowana jako forma niejednoznaczności (która w tej formie stanowi sedno badań opisywanej metodyki). Usunięcie tej niejednoznaczności z danego zbioru obiektów może polegać albo na usunięciu z niego albo dodaniu do niego części „granul", skutkując utworzeniem:

- takiego zbioru obiektów (podzbiór danego zbioru), z którego usunięto istniejące części niekompletnych „granul",
- takiego zbioru obiektów (nadzbiór danego zbioru), do którego dodano brakujące części niekompletnych „granul".

Zbiory powyższe, zawsze jednoznacznie określone dla każdego zbioru wyjściowego, nazywane są jego dolnym i górnym przybliżeniem i stanowią podstawę dalszych analiz tego zbioru (stanowią też źródło nazwy przypisywanej całej metodyce).

W szczególności dla danych nadzorowanych, po zastosowaniu powyższych koncepcji do analizy klas decyzyjnych, możliwe jest wprowadzenie pojęcia spójności danych – dane spójne w kategoriach ustalonego zbioru atrybutów to takie dane, w których dolne przybliżenia wszystkich klas decyzyjnych są równe górnym przybliżeniom tych klas.

W metodyce DRSA relacja bazowa DOM ma inne właściwości niż relacja bazowa IND w CRSA, co wynika z różnic w założeniach dotyczących analizowanych danych. Zgodnie z tymi założeniami, dziedziny atrybutów, zwanych w tym kontekście kryteriami, są uporządkowane preferencyjnie (dotyczy to także atrybutu decyzyjnego). W rezultacie, dla ustalonego zbioru kryteriów, relacja bazowa DOM wyznacza na zbiorze obiektów zbiory elementarne o innym charakterze („stożki” zamiast „granul”). Co ciekawe, metodyka DRSA stara się abstrahować od tego faktu i definiuje, w miarę możliwości, pojęcia analogiczne do pojęć wprowadzonych w CRSA, w szczególności (kluczowe w obu metodykach) dolne i górne przybliżenia, które także tutaj dzięki swojej jednoznaczności stanowią podstawę dalszych analiz.

W dalszych rozważaniach termin „atrybut” będzie używany w sensie szerokim, dotyczącym zarówno (klasycznych) atrybutów z CRSA, jak i kryteriów z DRSA, co ułatwia dalsze uogólnianie przedstawianych pojęć.

Dzięki analogicznemu zastosowaniu powyższych koncepcji do analizy klas decyzyjnych w przypadku danych nadzorowanych, możliwe jest wprowadzenie pojęcia spójności danych. W tym przypadku dane spójne w kategoriach ustalonego zbioru kryteriów to takie dane, w których dolne przybliżenia wszystkich kumulacji klas decyzyjnych są równe górnym przybliżeniom tych kumulacji.

Po wprowadzeniu (na dwa różne, ale analogiczne sposoby) pojęcia spójności, CRSA i DRSA definiują także analogiczne miary odstępstwa od sytuacji pełnej spójności, które tym samym stają się miarami poziomu spójności danych. Miary te charakteryzują się słabą monotonicznością (implikowaną przez słabą monotoniczność zastosowanych relacji bazowych) ze względu na uwzględniony zbiór atrybutów, co oznacza, że usunięcie dowolnego elementu z tego zbioru nie może zwiększyć poziomu spójności, a jedynie albo go zmniejszyć albo pozostawić na takim samym poziomie.

Powyższe fakty stwarzają naturalne podstawy do wprowadzenia takiego środowiska redukowania atrybutów, w którym jawnie zachowuje się wybrany z góry poziom spójności danych (np. poziom gwarantowany przez wszystkie atrybuty). Atrybuty wywierające zerowy (lub znikomy) wpływ na poziom spójności danych mogą być uznawane za nieprzydatne i usuwane z dalszych rozważań (redukowane). Pozostałe atrybuty stanowią tzw. redukt, czyli podzbiór atrybutów, który jest minimalny ze względu na zawieranie, ale zapewnia odpowiedni poziom spójności danych.

W praktyce sytuacja jest bardziej skomplikowana, ponieważ wzajemne interakcje (których istnienie ani skala nie są zwykle proste do przewidzenia) pomiędzy atrybutami powodują, że wpływ danego atrybutu na poziom spójności zależy od pozostałych, rozważanych w danej chwili atrybutów i jest w rezultacie tego zmienny. Oznacza to, że metody redukcji skazane są na badanie nie poszczególnych atrybutów, lecz ich zbiorów,

co implikuje zasadniczą zmianę złożoności problemu generowania rozwiązań (z wielomianowej na wykładniczą). Rezultatem tego są jednocześnie rozwijane zarówno dokładne jak i heurystyczne wersje algorytmów redukcji.

Od informacji międzyklasowej do wewnątrzklasowej

Istotą analizy danych nadzorowanych jest jawne wykorzystanie informacji międzyklasowej. W przedstawianych metodykach jest to oczywiście realizowane w rezultacie kontrolowania poziomu spójności danych poprzez aplikowanie odpowiednich relacji różnicujących do obiektów z różnych klas:

- DIS w CRSA,
- NDM w DRSA.

Na informacji międzyklasowej bazuje więc ogół implementowanych w tych metodykach analiz danych nadzorowanych, w tym algorytmy redukcji atrybutów (zarówno dokładne jak i przybliżone), których idea zasadza się na zapewnianiu, że:

- metodyka CRSA: obiekty z różnych klas nie są nierozróżnialne (w sensie relacji IND; w praktyce stosowana jest uwzględniająca negację powyższej relacja DIS),
- metodyka DRSA: obiekty z wyższych klas nie są dominowane przez obiekty z niższych klas (w sensie relacji DOM; w praktyce stosowana jest uwzględniająca negację powyższej relacja NDM).

Redukcja jest etapem zazwyczaj poprzedzającym metody indukcyjne właściwego uczenia się pojęć, które mogą być stosowane zarówno w celach deskrypcyjnych lub predykcyjnych (np. klasyfikatory obiektów w postaci drzew decyzyjnych lub zbiorów reguł). Metody te starają się odkryć wielowymiarowe (tzn. dotyczące wielu atrybutów) deskryptory obiektów, które różnicują obiekty z różnych klas i jednocześnie unifikują obiekty z tych samych klas.

Jednak gdy poprzedzająca metoda redukcji atrybutów posługuje się jedynie mechanizmem pozwalającym na różnicowanie obiektów należących do rożnych klas, to może ona (niejako nieumyślnie) usuwać atrybuty korzystne do unifikowania obiektów z tych samych klas. W ogólności bowiem atrybut słabo różnicujący obiekty z rożnych klas może dobrze unifikować obiekty z tych samych klas. Oznacza to, że usuwanie takich atrybutów może wiązać się z bezpowrotną (bo realizowaną już na etapie redukcji) utratą przydatnych informacji.

Z powyższych rozważań wynika, że korzystne dla procesu redukcji może być jawne uwzględnienie informacji wewnątrzklasowej [A01, A02, A04] (na równi z informacją międzyklasową). Pociąga to za sobą konieczność wprowadzenia dodatkowych relacji. Poprzednie relacje (natury różnicującej), czyli IND w CRSA i DOM w DRSA, nie spełniają swojej roli w odniesieniu do obiektów z tych samych klas (ich główna rola jest bowiem inna). Dlatego zdefiniowane zostały nowe relacje, których zadaniem jest unifikowanie (raczej niż różnicowanie) obiektów. Są to:

- SIM w CRSA [A01],
- PRX w DRSA [A02].

Podobnie jak relacje różnicujące, relacje unifikujące mogą być stosowane do wszystkich obiektów, także do obiektów z różnych klas. Głównym jednak ich przeznaczeniem jest zastosowanie do obiektów z tych samych klas (podczas gdy głównym przeznaczeniem relacji różnicujących jest zastosowanie do obiektów z różnych klas).

Wprowadzenie relacji unifikujących tworzy nowe środowisko, w którym poprzednie i nowe relacje spełniają wzajemnie dualne role [A02], pozwalając (np. w procesie redukcji atrybutów) na:

- różnicowanie obiektów z rożnych klas,
- unifikowanie obiektów z tych samych klas,

co gwarantuje jednoczesne wykorzystanie pełni informacji klasowej (tj. zarówno wewnątrzklasowej jak i międzyklasowej) o obiektach, i przekazuje tym samym więcej informacji wykorzystywanym na kolejnych etapach metodom indukcyjnym [A04].

Synteza informacji międzyklasowej i wewnątrzklasowej

W poprzednich środowiskach (tzn. w środowiskach, w których istniały tylko relacje rozróżniające, stosowane do obiektów z rożnych klas) metodyki CRSA techniki redukcji atrybutów mogły generować rozwiązania o nadmiernym poziomie rozróżnialności. Odpowiada to sytuacji zbytniego zbliżenia się do pewnego wyidealizowanego „punktu" charakteryzującego pełną rozróżnialność obiektów (wszystkie obiekty rozróżnialne). W nowym środowisku staje się jasne, że istnieją dwa takie „punkty" (bieguny), o przeciwległych charakterystykach, a konkretniej:

- biegun pełnej rozróżnialności,
- biegun pełnego podobieństwa.

Proces redukcji uwzględniający tylko informacje międzyklasowe (czyli rozróżnialność obiektów z różnych) klas może nadmiernie zbliżyć powstające rozwiązanie do pierwszego z biegunów. Oczywiście nie musi tak być w każdym wypadku, a to ze względu na fakt, że wymagana minimalność (ze względu na zawieranie) zbiorów atrybutów hamuje nadmierne rozróżnianie obiektów, natomiast sprzyja (choć niejawnie) ich upodabnianiu (dotyczy to jednak obiektów zarówno z tych samych, jak i z różnych klas). Niekorzystne konsekwencje tego typu mogą być zresztą także niwelowane w rezultacie działania uruchamianych w dalszej kolejności metod indukcyjnych (np. metod generowania reguł, które często są poprzedzane metodami redukcji atrybutów).

Wydaje się jednak, że widoki na najlepsze rezultaty ma jawne kontrolowanie tego zjawiska na jak najwcześniejszym etapie. Generowanie zredukowanych podzbiorów atrybutów w nowym środowisku (czyli takim, w którym istnieje zarówno relacja rozróżniająca, stosowana do obiektów z rożnych klas, jak i relacja upodabniająca, stosowana do obiektów z tych samych klas) ma więc dużo większe szanse powodzenia we współpracy z metodami indukcyjnymi [A02].

Co ciekawe, w nowym środowisku, oprócz dwóch przeciwległych biegunów można zidentyfikować także trzeci punkt charakterystyczny, stanowiący niejako „złoty środek". Punkt ten jest zasadniczo wyznaczony przez atrybut decyzyjny. Inaczej

mówiąc, „wypośrodkowany” (w powyższym sensie) zbiór atrybutów wyznacza w zbiorze wszystkich obiektów takie same relacje rozróżniania obiektów z różnych klas oraz upodabniania obiektów z tych samych klas jak atrybut decyzyjny. W wyidealizowanym przypadku, gdy jakiś atrybut jest równy (lub równoważny w sensie założeń przyjętych w danej metodyce) atrybutowi decyzyjnemu, zbiór złożony wyłącznie z tego atrybutu staje się idealnym, bo jednoelementowym (a tym samym minimalnym), zredukowanym zbiorem atrybutów. Jest to oczywiście cecha jedynie nowego środowiska, ponieważ w starych (jak już zaznaczono) możliwe było także mimowolne generowanie rozwiązań charakteryzujących się nadmiernie dużym poziomem rozróżniania obiektów (co w wielu przypadkach okazywało się niepotrzebnie, a nawet szkodliwe). W sytuacji, gdy idealnie „wypośrodkowany” zbiór atrybutów nie istnieje, środowisko pozwala na generowanie takich zbiorów atrybutów, które są do niego jak najbardziej podobne pod obiema wybranymi względami, czyli poziomem rozróżniania obiektów z rożnych klas oraz poziomem upodabniania obiektów z tej samej klasy.

Analogiczne rozumowanie można przedstawić dla metodyki DRSA [A03].

Rozwiązania przedstawionego typu mogą być przydatne w klasycznych kontekstach predykcyjnych, w których odkrywa się maksymalny poziom trafności przewidywania wartości atrybutu decyzyjnego na podstawie pozostałych atrybutów (klasyfikacja), jak i w innych, deskrypcyjnych, np. przy identyfikowaniu sytuacji, w których określa się górne i dolne granice „rozdzielczości” aktualnego atrybutu decyzyjnego i proponuje jego nowe postaci, np. poprzez:

- łączenie pewnych wartości w jedną wspólną,
- rozdzielanie pewnych wartości na wiele różnych,

co pozwala na poprawę poziomu „kompatybilności” atrybutu decyzyjnego z pozostałymi atrybutami.

Należy zwrócić uwagę, że jawne wykorzystanie pełni informacji klasowej na etapie redukcji atrybutów nie było dotychczas stosowane w metodykach RSA (jak i wielu innych metodach selekcji cech), ponieważ nie wykorzystywano w nich informacji wewnątrzklasowych. Podejścia takie wpisują się jednak w szeroko reprezentowany trend w analizie danych (nadzorowanych jak i nienadzorowanych) [C05]. Za dobry przykład może służyć w tym przypadku (nienadzorowana) metoda analizy skupień „k-średnich”, której celem jest proponowanie jak najlepszych skupień obiektów, przy czym o jakości tworzonych skupień decydują wskaźniki obliczane na podstawie dwóch następujących macierzy:

- macierzy odległości międzyskupieniowych,
- macierzy odległości wewnątrzskupieniowych.

Jakość wyników w tej metodzie jest (w ogólności) uznawana za większą im mniejsze są odległości pomiędzy obiektami należącymi do tych samych skupień (pierwsza macierz) a jednocześnie im większe są odległości pomiędzy obiektami należącymi do rożnych skupień (druga macierz). Reprezentowane przez te macierze źródła informacji międzyskupieniowych i wewnątrzskupieniowych są oczywiście nienadzorowanymi ekwiwalentami źródeł informacji nadzorowanych (czyli międzyklasowych i wewnątrzklasowych).

Eksploatacja informacji międzyklasowej i wewnątrzklasowej

Odpowiednikami powyższych dwóch zasadniczych rodzajów macierzy w opisywanych metodykach RSA są (zbiorcze) macierze:

- rozróżnialności / podobieństwa (w sensie odpowiednich relacji) w CRSA [A01],
- dominacji / bliskości (w sensie odpowiednich relacji) w DRSA [A02].

Każda z tych macierzy gromadzi informacje o zbiorach atrybutów gwarantujących zachodzenie odpowiednich relacji pomiędzy parami obiektów.

W metodyce CRSA (która zasadniczo zakłada dyskretność dziedzin atrybutów) relacją bazową jest relacja IND, która naturalnie indukuje tzw. „granule" obiektów [C09]. Natomiast relacje używane wraz z informacjami międzyklasowymi i wewnątrzklasowymi to: DIS (ang. discernibility), stosowana do informacji międzyklasowych oraz SIM (ang. similarity), stosowana do informacji wewnątrzklasowych [A01]. Relacja DIS stanowi proste dopełnienie relacji IND. Z kolei relacja SIM może być łatwo wywiedziona z relacji IND, ponieważ obie relacje pełnią podobna role – informują o podobieństwach pomiędzy obiektami, przy czym IND informuje o szczególnym rodzaju podobieństwa zupełnego, czyli o nierozróżnialności, a SIM – częściowego, czyli tolerancji (wprowadzenie relacji SIM było konieczne, ponieważ IND jest relacją zbyt mocną). IND stanowi relacje równoważności (zwrotna, symetryczna i przechodnia), a jej osłabiona wersja, czyli SIM, stanowi relacje tolerancji, (zwrotna i symetryczna), reprezentuje wiec rodzaj symetrycznego podobieństwa.

W przypadku metodyki DRSA (która zasadniczo zakłada uporządkowanie dziedzin atrybutów), relacja bazowa w tej metodyce jest relacją (słabej) dominacji DOM, która jest zwrotna, antysymetryczna i przechodnia [C10]. Jej dopełnienie, relacja NDM (ang. non-dominance), stosowana jest wraz z informacjami międzyklasowymi [A02]. Sytuacja jest jednak bardziej skomplikowana w przypadku informacji wewnątrzklasowych, co wynika z faktu, że w rezultacie uporządkowania dziedzin w DRSA relacja DOM naturalnie indukuje „stożki" obiektów, i jako taka nie może być w prosty sposób wykorzystana do reprezentowania podobieństwa. Aby to stało się możliwe, wprowadzona została relacja INS (ang. inseparability), z której w prosty sposób wywiedziono relacje (symetrycznego) podobieństwa PRX (ang. proximity) [A02].

Wprowadzona relacja INS jest wzorowana na relacji IND, w odróżnieniu jednak od niej nie wymaga jednak dyskretności dziedzin atrybutów i jest relacją kontekstową. W szczególnych warunkach jej właściwości naśladują właściwości relacji IND. Z kolei relacja PRX ma właściwości wzorowane na relacji SIM, i dlatego może być stosowana do eksploracji informacji wewnątrzklasowych.

Redukty i konstrukty w CRSA

CRSA wprowadza relację nierozróżnialności (IND) obiektów w kategoriach wybranych atrybutów i skupia się na pojęciu spójności danych w sensie tej relacji. W przypadku danych nadzorowanych spójność danych jest uznawana za naruszoną, gdy obiekty

pochodzące z różnych klas są nierozróżnialne w kategoriach wszystkich dostępnych atrybutów. Ponieważ przynależność dowolnych obiektów do różnych klas jest równoważna z rozróżnialnością tych obiektów na atrybucie decyzyjnym, idea spójności może być w całości wyrażona z użyciem relacji IND (noszącej wobec tego faktu nazwę relacji bazowej w tej metodyce).

Oprócz istnienia nieodzownego w tej sytuacji atrybutu decyzyjnego i niepustego zbioru pozostałych atrybutów opisujących niepusty zbiór obiektów, metodyka zakłada także możliwość ustalania relacji IND dla tych obiektów, co w praktyce oznacza dodatkowe założenie dyskretności dziedzin wszystkich atrybutów, w tym atrybutu decyzyjnego (dzięki procesowi dyskretyzacji możliwe jest także dodatkowe rozważanie atrybutów o oryginalnie ciągłych dziedzinach).

W celu wszechstronnego kontrolowania zjawiska niespójności, metodyka wprowadza współczynnik (tzw. jakość przybliżenia klasyfikacji obiektów) kwantyfikujący numerycznie poziom spójności danych w powyższym sensie.

Słaba monotoniczność relacji IND ze względu na rozważany zbiór atrybutów (usunięcie dowolnego atrybutu nie może zwiększyć jakości przybliżenia klasyfikacji obiektów, a jedynie albo ją zmniejszyć albo pozostawić na takim samym poziomie) zapewnia naturalną w tej metodyce metodę generowania zredukowanych podzbiorów atrybutów (czyli reduktów): redukt w CRSA jest podzbiorem atrybutów, który jest minimalny ze względu na zawieranie i zapewnia odpowiednio wysoką wartość współczynnika jakości przybliżenia [C09]. Ponieważ w CRSA poziom tej jakości jest bezpośrednio zależny tego, na ile odróżniane są od siebie rozróżnialne obiekty z różnych klas, redukt w tym kontekście może być równoważnie zdefiniowany jedynie z wykorzystaniem (fundamentalnego w tej metodyce) pojęcia rozróżnialności obiektów [C11]. Ze względu na wykorzystanie w definicji jedynie obiektów z różnych klas, redukt będzie od tej pory nazywany inter-reduktem.

Podsumowując: inter-redukt w CRSA to minimalny ze względu na zawieranie podzbiór atrybutów zapewniający rozróżnianie (w sensie relacji DIS) obiektów z różnych klas na poziomie gwarantowanym przez wszystkie atrybuty. Charakterystyczny dla tej metodyki jest fakt rozróżnialności także elementów (dyskretnej) dziedziny atrybutu decyzyjnego, dzięki czemu w powyższym sformułowaniu możliwe jest operowanie pojęciami „różnych” klas. Ze względu na symetrię macierzy rozróżnialności (bezpośrednio wynikającej z symetrii bazowej w CRSA relacji IND) możliwe jest wykorzystanie dowolnego z (dwóch identycznych) trójkątów macierzy

Wykorzystana w definicji inter-reduktu relacja DIS jest (słabo) monotoniczna ze względu na zbiór atrybutów, co oznacza, że dodanie atrybutu do tego zbioru nie może zmniejszyć (a jedynie albo zwiększyć albo pozostawić bez zmian) poziomu rozróżnialności, tzn. nie może zmniejszyć liczby obiektów rozróżnianych. Oznacza to jednocześnie, że sensowne jest poszukiwanie minimalnych ze względu na zawieranie podzbiorów atrybutów, które zapewniają odpowiedni poziom rozróżnialności obiektów z różnych klas.

Pojęciem zbudowanym w CRSA analogicznie do inter-reduktu, ale wykorzystującym w swej definicji jedynie obiekty z tych samych klas, jest intra-redukt [A04]: minimalny ze względu na zawieranie podzbiór atrybutów zapewniający upodabnianie (w sensie relacji SIM) obiektów z tych samych klas na poziomie

gwarantowanym przez wszystkie atrybuty. Relacja SIM jest zbudowana bezpośrednio z wykorzystaniem relacji IND (bazowej w CRSA).

Sensowność definicji intra-reduktu bazuje na fakcie, że wykorzystana w niej relacja SIM jest także (słabo) monotoniczna ze względu na zbiór atrybutów, co oznacza, że dodanie atrybutu do tego zbioru nie może zmniejszyć (a jedynie zwiększyć lub pozostawić na tym samym poziomie) liczby obiektów upodabnianych przez atrybuty z tego zbioru.

Inter-redukt i intra-redukt są więc w CRSA analogicznymi strukturami, różniącymi się przede wszystkim wykorzystaniem informacji o obiektach:

- inter-redukt wykorzystuje tylko informację międzyklasową (do eksploatacji której stosowana jest relacja DIS),
- intra-redukt wykorzystuje tylko informację wewnątrzklasową (do eksploatacji której stosowana jest relacja SIM).

Ponieważ każde z powyższych dwu pojęć wykorzystuje tylko jedną cześć dostępnej informacji o wszystkich obiektach, naturalnym kolejnym krokiem jest próba zdefiniowania pojęcia wykorzystującego całość tej informacji. Pojęciem takim jest konstrukt [A04]: minimalny ze względu na zawieranie podzbiór atrybutów zapewniający jednoczesne rozróżnianie obiektów z różnych klas na poziomie gwarantowanym przez wszystkie atrybuty oraz upodabnianie obiektów z tych samych klas na poziomie gwarantowanym przez wszystkie atrybuty.

W konstruktach CRSA zastosowana jest agregacja relacji (DIS i SIM), w której obowiązuje ścisły rozdział ich zastosowań [A01]:

- DIS dla obiektów z różnych klas,
- SIM dla obiektów z tych samych klas.

Należy podkreślić, że w metodyce CRSA oba wprowadzone pojęcia (intra-redukt i konstrukt) są interpretowalne w sposób analogiczny do oryginalnego inter-reduktu.

Redukty i konstrukty w DRSA

DRSA wprowadza relację dominacji obiektów w kategoriach wybranych atrybutów (DOM) i skupia się na pojęciu spójności danych w sensie tej relacji. W przypadku danych nadzorowanych spójność danych jest uznawana za naruszoną, gdy obiekty pochodzące z niższych klas dominują obiekty z klas wyższych w kategoriach wszystkich dostępnych atrybutów. Ponieważ przynależność dowolnych obiektów do niższych i wyższych klas jest równoważna z (silną) dominacją tych obiektów na atrybucie decyzyjnym, idea spójności może być wyrażona w całości z użyciem relacji DOM (noszącej wobec tego faktu nazwę relacji bazowej w tej metodyce).

Oprócz istnienia nieodzownego w tej sytuacji atrybutu decyzyjnego i niepustego zbioru pozostałych atrybutów opisujących niepusty zbiór obiektów, metodyka zakłada także możliwość obliczania relacji DOM dla tych obiektów, co w praktyce oznacza założenie uporządkowania preferencyjnego dziedzin wszystkich atrybutów, w tym atrybutu decyzyjnego (rozszerzenia metodyki pozwalają jednakże na dodatkowe rozważanie atrybutów o nieuporządkowanych dziedzinach).

Analogicznie do CRSA, metodyka wprowadza współczynnik (tzw. jakość przybliżenia klasyfikacji obiektów) kwalifikujący numerycznie poziom spójności danych w powyższym sensie. Wykorzystując ten współczynnik, redukt w DRSA może być zdefiniowany w sposób ogólny, jako minimalny ze względu na zawieranie podzbiór atrybutów, który zapewnia taki sam poziom spójności danych (wyrażany przez współczynnik jakości klasyfikacji) jak zbiór wszystkich atrybutów [C10]. Ponieważ jednak w tej metodyce poziom jakości klasyfikacji jest bezpośrednio zależny tego, na ile obiekty z niższych klas dominują obiekty z wyższych klas, redukt w tym kontekście może być równoważnie zdefiniowany jedynie z wykorzystaniem (fundamentalnego w tej metodyce) pojęcia dominowania się obiektów [C14]. Jak poprzednio, ze względu na wykorzystanie w definicji jedynie obiektów z różnych klas, redukt będzie od tej pory nazywany inter-reduktem.

Podsumowując: inter-redukt w DRSA to minimalny ze względu na zawieranie podzbiór atrybutów zapewniający niedominowanie (w sensie relacji NDM) obiektów z wyższych klas przez obiekty z niższych klas na poziomie gwarantowanym przez wszystkie atrybuty. Charakterystyczny dla tej metodyki jest fakt uporządkowania także elementów dziedziny atrybutu decyzyjnego, dzięki czemu w powyższym sformułowaniu możliwe jest operowanie pojęciami „niższych” i „wyższych” klas. Pozwala to na skupienie uwagi na jednym tylko z dwóch (w ogólności różnych) trójkątów macierzy dominacji (różność ta jest konsekwencją antysymetrii relacji DOM, na której bazuje DRSA).

Wykorzystana w definicji inter-reduktu relacja NDM jest (słabo) monotoniczna ze względu na zbiór atrybutów, co oznacza, że dodanie atrybutu do tego zbioru nie może zmniejszyć (a jedynie albo zwiększyć albo pozostawić bez zmian) poziomu niedominowania, tzn. nie może zmniejszyć liczby obiektów z wyższych klas dominowanych przez obiekty z niższych klas. Potwierdza to sensowność poszukiwania minimalnych ze względu na zawieranie podzbiorów atrybutów, które zapewniają odpowiedni poziom dominowania obiektów z niższych klas przez obiekty z wyższych klas.

Analogicznym do inter-reduktu, ale wykorzystującym w swej definicji jedynie obiekty z tych samych klas, jest w DRSA intra-redukt [A02]: minimalny ze względu na zawieranie podzbiór atrybutów zapewniający ujednolicanie obiektów z tych samych klas na poziomie gwarantowanym przez wszystkie atrybuty (ujednolicanie w sensie PRX). W metodyce DRSA relacja ujednolicania PRX nie jest jednak zbudowana z wykorzystaniem relacji DOM (bazowej w CRSA), co wynika bezpośrednio z braku odpowiednich właściwości tej drugiej. Zamiast tego, do budowy PRX wykorzystano relację kontekstową INS [A02] (relacja ta jest wzorowana na relacji IND z CRSA).

Sensowność definicji intra-reduktu bazuje na fakcie, że wykorzystana w niej relacja PRX jest także monotoniczna ze względu na zbiór atrybutów, co oznacza, że dodanie atrybutu do tego zbioru nie może zmniejszyć (a jedynie zwiększyć lub pozostawić na tym samym poziomie) liczby obiektów ujednolicanych przez atrybuty z tego zbioru.

Inter-redukt i intra-redukt są więc także w DRSA analogicznymi strukturami, różniącymi się przede wszystkim wykorzystaniem informacji o obiektach:

- inter-redukt wykorzystuje tylko informację międzyklasową (do eksploatacji której stosowana jest relacja NDM),

- intra-redukt wykorzystuje tylko informację wewnątrzklasową (do eksploatacji której stosowana jest relacja PRX).

Ponieważ każde z powyższych dwu pojęć wykorzystuje tylko jedną cześć dostępnej informacji o wszystkich obiektach (z naturalnym uwzględnieniem faktu wykorzystania tylko jednego trójkąta niesymetrycznej macierzy dominacji), naturalnym kolejnym krokiem jest próba zdefiniowania pojęcia wykorzystującego całość tej informacji. Pojęciem takim jest konstrukt [A02]: minimalny ze względu na zawieranie podzbiór atrybutów zapewniający jednoczesne niedominowanie obiektów z wyższych klas przez obiekty z niższych klas na poziomie gwarantowanym przez wszystkie atrybuty oraz ujednolicanie obiektów z tych samych klas na poziomie gwarantowanym przez wszystkie atrybuty.

W konstruktach DRSA zastosowana jest agregacja relacji (NDM i PRX), w której obowiązuje ścisły rozdział ich zastosowań [A02]:

- NDM dla obiektów z różnych klas,
- PRX dla obiektów z tych samych klas.

Należy podkreślić, że w metodyce DRSA wprowadzone pojęcia (intra-redukt i konstrukt) są interpretowalne w sposób analogiczny do oryginalnego inter-reduktu.

Fundamentalne cechy reduktów i konstruktów

Nietrudno dostrzec charakterystyczne, dopełniające się (zarówno w CRSA jak i w DRSA), cechy reduktów i konstruktów związane z przetwarzaniem informacji międzyklasowej (obiekty z różnych klas) oraz wewnątrzklasowej (obiekty z tych samych klas):

- inter-redukty:
  - różnicują obiekty z różnych klas,
  - (ignorują obiekty z tych samych klas),
- intra-redukty:
  - (ignorują obiekty z różnych klas)
  - unifikują obiekty z tych samych klas,
- konstrukty:
  - różnicują obiekty z różnych klas,
  - unifikują obiekty z tych samych klas,

przy czym:

- różnicowanie jest rozumiane w sensie relacji DIS w CRSA i relacji DOM w DRSA,
- unifikowanie jest rozumiane w sensie relacji SIM w CRSA i relacji PRX w DRSA.

W rezultacie inter-redukty i intra-redukty są więc analogicznymi pojęciami, które zasadniczo rozgranicza jedynie rodzaj przetwarzanej informacji (inter-redukty: informacja międzyklasowa, intra-redukty: informacja wewnątrzklasowa). Z kolei konstrukty stanowią formę agregacji obu rodzajów reduktów, i wykorzystuje się w nich pełnię informacji klasowej. Oczywiście w obu przypadkach, w zależności od metodyki,

do rodzaju przetwarzanej informacji dostosowuje się rodzaj stosowanej relacji. Dzięki zunifikowaniu ich definicji jednakże, wykorzystywane procedury i algorytmy są w znaczącej części identyczne [A03].

Omówione pojęcia są dokładnie tymi, które utożsamiają wspomniane wcześniej dwa bieguny (inter-redukt i intra-redukt) oraz leżący między nimi „złoty środek" (konstrukt). Dlatego gdy (w wyidealizowanym przypadku) jakiś atrybut jest równy (lub równoważny w sensie założeń przyjętych w danej metodyce) atrybutowi decyzyjnemu, zbiór złożony z tego atrybutu stanowi jednocześnie inter-redukt, intra-redukt i konstrukt. Dodatkowo, jest to jedyny możliwy przypadek jednoelementowego konstruktu, choć możliwe jest istnienie innych (tzn. nie będących równoważnymi atrybutowi decyzyjnemu) jednoelementowych inter-reduktów i intra-reduktów.

Ścisłe integrowanie informacji międzyklasowej i wewnątrzklasowej przez konstrukty implikue również istnienie właściwości wiążących istnienie inter-reduktów, intra-reduktów i konstruktów [A03].

Środowisko redukcji generujące konstrukty, czyli zredukowane podzbiory atrybutów, przy uzyskiwaniu których wykorzystuje się zarówno informacje międzyklasowe jak i wewnątrzklasowe, jest oczywiście uogólnieniem starszych środowisk, w których jawnie wykorzystywano wyłącznie informacje międzyklasowe. Warty podkreślenia jest jednakże fakt, że uogólnienie to jest „niekonfliktowe" z innymi uogólnieniami w tym sensie, że nie dokonuje ingerencji w elementy w nich zdefiniowane, a jedynie rozszerza je o nowe, kompatybilne z poprzednimi, elementy. W rezultacie możliwe jest adaptowanie i stosowanie innych, rozwijanych równolegle z opisywanymi (a także proponowanych wcześniej), uogólnień tych środowisk.

W metodyce CRSA, już na bardzo wczesnym etapie, popularności nabrało wiele uogólnień klasycznych reduktów (w obecnym sensie: inter-reduktów), np. redukty dynamiczne [C01], redukty częstościowe [C02], redukty modelu o zmiennej precyzji [C08] czy redukty oparte na entropii [C15], których celem było m. in. polepszenie właściwości predykcyjnych generowanych zbiorów atrybutów. Standardowa definicja inter-reduktu, wymagająca aby wszystkie obiekty rozróżniane przez zbiór wszystkich atrybutów były także rozróżniane także przez inter-redukt, została stworzona z myślą o jasnej interpretacji w kontekście deskrypcyjnym. Wymóg ten może jednak prowadzić do nadmiernego rozrostu podzbiorów wynikowych w rezultacie włączania do niego atrybutów zasadniczo mało przydatnych, choć niezbędnych do rozróżniania pewnej mało znaczącej liczby obiektów. W zastosowaniach predykcyjnych zabieg taki często okazuje się nadmiernie restrykcyjny, a zrelaksowanie odpowiedniego warunku, implementowane na różne sposoby w różnych uogólnieniach (pozwalających np. na zignorowaniu pewnej kontrolowanej liczby obiektów), polepsza zwykle istotne właściwości generowanego zbioru atrybutów.

W rezultacie wydaje się możliwe, że stosowane w tych uogólnieniach wskaźniki poziomu rozróżnialności obiektów, zdefiniowane z wykorzystaniem informacji międzyklasowej, można zaadaptować do oceniania poziomu upodabniania obiektów na podstawie informacji wewnątrzklasowej. Prowadzi to do zastosowania analogicznych uogólnień do intra-reduktów, a po prostym zintegrowaniu obu rodzaju wskaźników, do zastosowania ich do konstruktów.

Docelowo możliwe jest więc uogólnianie na wskazane sposoby nie tylko inter-reduktów, ale także intra-reduktów i konstruktów. Uwzględnienie informacji wewnątrzklasowych nie powinno być więc postrzegane jako uogólnienie alternatywne do innych możliwych uogólnień, lecz raczej jako operację do nich „ortogonalną" (a więc niezależną i „niekonfliktową"). Tak rozumiana „ortogonalność" rozważanych uogólnień pozwala na tworzenie ich skutecznych kombinacji, o potencjalnie bardziej korzystnych właściwościach, zarówno deskrypcyjnych jak i predykcyjnych (w zależności od konkretnego rodzaju integrowanych uogólnień). Teoretycznie więc nic nie stoi na przeszkodzie wydajnemu definiowaniu i wykorzystywaniu konstruktów dynamicznych, konstruktów częstościowych, itp.

Dalsze właściwości reduktów i konstruktów

Ilość i jakość informacji czerpanej z danych przez rozmaite procedury indukcyjne zależy w znacznej mierze od charakterystyki samych danych, w których znacznej zmienności podlegać mogą ich bardzo różne aspekty, m. in. konfiguracja liczby klas oraz liczności poszczególnych klas. Przy ustalonej liczbie obiektów, ta zmieniająca się konfiguracja posiada dwa wyraźne „punkty" (bieguny) o przeciwległych charakterystykach, z których jeden biegun stanowi sytuacja „wszystkie obiekty z tej samej klasy" (liczba klas równa jeden), a drugi sytuacja „każdy obiekt z innej klasy" (liczba klas równa liczbie obiektów). W praktyce dane charakteryzujące się wspomnianymi konfiguracjami „biegunowymi" należałoby traktować jak dane nienadzorowane: jest to oczywiste w przypadku sytuacji, gdy liczba klas jest równa jeden, ale dotyczy także (niejawnie) sytuacji, gdy liczba klas jest równa liczbie obiektów (w obu tych sytuacjach o obiektach z różnych klas, rozumianych jako zbiory, które „skupiają obiekty o określonych właściwościach, wspólnych dla wszystkich obiektów w danej klasie" można bowiem mówić jedynie ściśle teoretycznie).

W danych przypadku nadzorowanych wyróżnić można więc sytuacje, w których konfiguracja liczby klas oraz liczności poszczególnych klas odchyla się od „punktu" równowagi i zmierza w stronę jednego z biegunów (nie osiągając go jednakże) [A02]:

- sytuacja „wieloklasowa", w której liczba klas rośnie (potencjalnie zbliżając się do liczby obiektów), a liczności wszystkich klas są wyrównane,
- sytuacja „niezrównoważona", w której liczba klas maleje (potencjalnie zbliżając się do jeden), a liczność pewnej klasy dominuje licznością pozostałe.

Należy zauważyć, że podział ten jest inny (ogólniejszy) niż popularny podział zbiorów danych na „zrównoważone" i „niezrównoważone", przy tworzeniu którego brane są pod uwagę liczności poszczególnych klas. W omawianym przypadku główną zmieniającą się wielkością jest liczba klas, której zmiany implikują jednakże zmiany liczności klas (w szczególnym przypadku możliwe jest uwzględnienie poziomu niezrównoważenia liczności klas przy ustalonej z góry ich liczbie).

Charakterystyczne jest, że w rozważanych sytuacjach znacząco zmienia się względna ilość dostępnej informacji międzyklasowej i wewnątrzklasowej:

- sytuację „wieloklasową" charakteryzują:
  - nadmiar informacji międzyklasowej,

  - niedobór informacji wewnątrzklasowej,
- sytuację „niezrównoważoną” charakteryzują:
  - niedobór informacji międzyklasowej,
  - nadmiar informacji wewnątrzklasowej,

W poprzednich środowiskach, generujących inter-redukty (lub ich odpowiedniki wykorzystujące jawnie jedynie informacje międzyklasowe) mogło to odbijać się niekorzystnie na jakości wyników, ze względu na niewielką ilość faktycznie wykorzystywanej informacji. Analogiczna sytuacja dotyczyłaby zresztą także intra-reduktów (lub ich odpowiedników wykorzystujących jawnie jedynie informacje wewnątrzklasowe). Problem taki nie dotyczy jednakże konstruktów, ponieważ wykorzystują one pełnię dostępnej informacji, której sumaryczna ilość (przy ustalonej liczbie obiektów) pozostaje stała we wszystkich możliwych konfiguracjach liczby i liczności klas pomiędzy dwoma rozważanymi ekstremami.

Eksperymentalne badania klasyfikacyjne z użyciem rzeczywistych i syntetycznych danych ujawniają też dobre właściwości predykcyjne konstruktów, w szczególności ich przewagę nad inter-reduktami (wykorzystywanych zwyczajowo w roli zredukowanych zbiorów atrybutów) w sytuacji danych charakteryzujących się dużą liczbą przekłamań [A01, A04]. W omawianym przypadku rosnącej (w kontrolowany sposób) liczbie przekłamań w danych towarzyszył rosnący poziom błędu klasyfikowania dla konstruktów, mniejszy jednakże od (szybciej) rosnącego poziomu błędu dla inter-reduktów. Jednocześnie w eksperymentach klasyfikacyjnych wykorzystujących metody redukcji typowe dla zastosowań predykcyjnych, konstrukty (generowane podejściem heurystycznym) uzyskują bardzo zbliżone wyniki [A03].

Generowanie reduktów i konstruktów

Dzięki zunifikowanym definicjom generowanie reduktów i konstruktów może być dokonywane z użyciem niemal identycznych algorytmów (różnice polegają jedynie na adaptacjach pozwalających na identyfikowanie odpowiednich relacji między obiektami, które nie mają większego wpływu na złożoność obliczeniową). Oznacza to możliwość wykorzystania znanych i sprawdzonych algorytmów [C12] generowania reduktów, zarówno dokładnych, jak i heurystycznych. Najsprawniejsze z tych algorytmów wykorzystują macierze [A02]:

- metodyka CRSA
  - macierz rozróżnialności (w sensie relacji DIS), zdefiniowana dla obiektów z różnych klas,
  - macierz podobieństwa (w sensie relacji SIM), zdefiniowana dla obiektów z tych samych klas,
- metodyka DRSA
  - macierz niedominacji (w sensie relacji NDM), zdefiniowana dla obiektów z różnych klas,
  - macierz bliskości (w sensie relacji PRX), zdefiniowana dla obiektów z tych samych klas,

które dla każdej pary obiektów przechowują informacje o zbiorach atrybutów gwarantujących zachodzenie odpowiednich relacji pomiędzy tymi obiektami. Obie macierze stanowią (w ramach danej metodyki) rozłączne podmacierze całościowej macierzy obiekt-obiekt, przy czym pierwsza odpowiada informacji międzyklasowej, a druga informacji wewnątrzklasowej.

Macierze są w praktyce przetwarzane do postaci list, i w tej postaci biorą udział w generowaniu zredukowanych podzbiorów atrybutów. Ze względu na pokaźną redundancję swoich elementów, tworzone listy poddawane są wstępnej absorpcji, co nie wywiera wpływu na ostateczną postać wyników, prowadzi jednak do znacznego zredukowania liczby przetwarzanych elementów i, w rezultacie, przyspieszenia obliczeń [C12].

Kolejne kroki wszystkich algorytmów generowania podzbiorów atrybutów są od momentu utworzenia listy wspólne dla obu rodzajów reduktów, jak i dla konstruktów [A02, A03], ponieważ to, czy wynikiem danego algorytmu są inter-redukty, intra-redukty czy konstrukty zależy tylko i wyłącznie od tego, co zostało umieszczone na przetwarzanej liście:

- relacja DIS / NDM dla obiektów z różnych klas – inter-redukty,
- relacja SIM / PRX dla obiektów z tych samych klas – intra-redukty,
- relacja DIS / NDM dla obiektów z różnych klas oraz relacja SIM / PRX dla obiektów z tych samych klas – konstrukty.

Kierunki rozwoju algorytmów dokładnych

Algorytmy generujące wszystkie istniejące rozwiązania dokładne, charakteryzujące się największą złożonością obliczeniową (rosnącą wykładniczo wraz z całkowitą liczbą atrybutów w danych), mogą być znacząco usprawniane (pod względem czasowym) dzięki zastosowaniu zrównoleglania obliczeń, które polega na jednoczesnym wykorzystaniu wielu jednostek obliczeniowych (procesorów lub rdzeni procesorowych), w odróżnieniu od sekwencyjnego zastosowania jednej jednostki.

Skuteczna metoda zrównoleglenia obliczeń w tzw. modelu hierarchicznym polega na rekurencyjnej dekompozycji zadania obliczeniowego na podzadania, które są w pełni niezależne od siebie (co eliminuje konieczność jakiejkolwiek ich synchronizacji). Proponowany hierarchiczny model [A05] zrównoleglania jest bardziej elastyczny od tzw. modelu płaskiego [C13], w którym zadanie obliczeniowe jest dekomponowane jednokrotnie na ustaloną z góry liczbę podzadań. Model płaski gwarantuje wprawdzie utworzenie podzadań wykorzystujących wiele jednostek obliczeniowych, nie gwarantuje jednak wyrównanego obciążenia obliczeniowego tych podzadań, w rezultacie czego niektóre z nich mogą kończyć pracę dużo wcześniej niż inne, prowadząc do bardzo nierównomiernego wykorzystania przydzielonych jednostek (w ekstremalnym przypadku zdecydowana większość obliczeń realizowana jest przez tylko jedną jednostkę, co może zniwelować potencjalne zyski wynikające z zastosowania wielu jednostek).

W modelu hierarchicznym podział zadań na podzadania jest rekurencyjny, może więc być wykonywany wielokrotnie, także w późniejszych momentach czasu, w rezultacie czego podział ten będzie dotyczył niemal wyłącznie zadań bardziej

wymagających obliczeniowo (zadania niewymagające zakończą pracę wcześniej). Prowadzi to do lepszego wyrównywania obciążenia, może jednak angażować zbyt wielkie liczby jednostek obliczeniowych. Usprawniony pod tym względem jest tzw. kontrolowany model hierarchiczny, w którym podział zadań na podzadania pozostaje rekurencyjny, ale liczba powstających podzadań jest dodatkowo ograniczona od góry (co w szczególnym przypadku pozwala na stosowanie go w trybie sekwencyjnym).

Wyniki eksperymentów obliczeniowych mających na celu określenie praktycznych właściwości kontrolowanego modelu hierarchicznego (rozumianych jako wartości głównych parametrów kontrolujących jego działanie) i porównanie go z innymi modelami potwierdzają przydatność tego modelu i jego przewagę nad nimi.

Kierunki rozwoju algorytmów heurystycznych

Postępowaniem alternatywnym do rozwijania kosztownych obliczeniowo algorytmów generujących wszystkie istniejące rozwiązania dokładne jest proponowanie mniej precyzyjnych, a tym samym mniej kosztownych obliczeniowo algorytmów heurystycznych. W tym przypadku przez algorytmy heurystyczne rozumie się algorytmy generujące tylko jedno (dokładne) rozwiązanie. W pewnym sensie algorytmy te zachowują korzystne cechy algorytmów dokładnych, ich mniejszy potencjał wynika jednak z tego, że bardzo trudne (czy wręcz niemożliwe) jest zagwarantowanie znalezienia przez nie rozwiązania o zadanych z góry właściwościach. Dotyczy to zarówno prostych właściwości formalnych (np. minimalności zbioru pod względem liczebności elementów) jak i bardziej złożonych właściwości przedmiotowych (np. przydatność rozwiązania w kontekstach predykcyjnych), związanych z konkretnymi zastosowaniami tych rozwiązań.

Identyfikowanie rozwiązań o dobrych właściwościach przedmiotowych jest szczególnie przydatne, ponieważ dotyczy zarówno algorytmów heurystycznych, jak i dokładnych. Wynika to z faktu, że znalezienie rozwiązań charakteryzujących się dobrymi właściwościami przedmiotowymi może być trudne także wtedy, gdy znane są wszystkie rozwiązania dokładne. To z kolei wynika z faktu, że dane rozwiązanie (czyli inter-redukt, intra-redukt czy konstrukt), pomimo jasnej interpretacji deskryptywnej, może, ale nie musi posiadać innych, korzystnych w specyficznym zastosowaniu, właściwości przedmiotowych.

Głównym zadaniem algorytmów heurystycznych jest więc skuteczne kierowanie procesu generowania w stronę rozwiązań o pożądanych właściwościach. Jest to możliwe dzięki wykorzystaniu odpowiednich miar kontrolujących cały proces [C06], w szczególności właściwych miar oceny atrakcyjności atrybutów.

Miary atrakcyjności oceniają w ogólności jakość prawidłowości odkrytych w danych, np. w postaci reguł decyzyjnych. Zakładając dwuczęściową strukturę reguły decyzyjnej (część warunkowa i część decyzyjna), jej ocena w kategoriach miary może zostać utworzona następująco. Części reguły zachodzą, gdy następuje ich dopasowanie do pewnych obiektów, przy czym dopasowanie części warunkowej oznacza, że prawdziwa jest pewna przesłanka, a dopasowanie części decyzyjnej oznacza, że prawdziwa jest pewna konkluzja. Dla danego zbioru obiektów, możliwe jest jednoznaczne podsumowanie zależności pomiędzy wektorem wyrażającym prawdziwość przesłanek a wektorem wyrażającym prawdziwość konkluzji.

Podsumowanie to wyraża się macierzą rozmiaru 2×2 nieujemnych wartości całkowitych, które zliczają, ile razy wystąpiła dana kombinacja dopasowań. Charakteryzując skalarnie taką macierz z myślą opisania zależności między wspomnianymi wektorami, miara atrakcyjności generuje (pośrednią) ocenę reguły, która posłużyła do utworzenia tych wektorów.

Ponieważ jednak miary te zasadniczo oceniają zależność wektorów (niezależnie od faktu istnienia odpowiedniej reguły), mogą być także wykorzystane do ewaluowania atrakcyjności atrybutów (w kontekście atrybutu decyzyjnego), o ile rozważona zostanie zależność pomiędzy danym atrybutem a atrybutem decyzyjnym. W przypadku binarnych dziedzin obu atrybutów, macierz podsumowująca jest taka jak powyżej, dzięki czemu nie są wymagane żadne adaptacje oryginalnych miar.

Traktując zachodzenie bazowej w danej metodyce relacji (IND, DOM) dla pary obiektów jako przesłankę, a przynależność tych obiektów do tej samej klasy jako konkluzję, otrzymuje się, po rozważeniu wszystkich par obiektów, analogiczną macierz. W rezultacie powstaje możliwość oceniania wynikowej zależności przez miary atrakcyjności. Stanowi to przykład jednoczesnego wykorzystania informacji międzyklasowej i wewnątrzklasowej, i nawiązuje do definicji konstruktu, który musi zapewniać odpowiednio wysoki poziom różnicowania obiektów z różnych klas i, jednocześnie, odpowiednio wysoki poziom unifikowania obiektów z tych samych klas. Z tego powodu zasadne wydaje się testowanie możliwości różnych miar atrakcyjności w procesie kontrolowania algorytmów heurystycznych generowania konstruktów.

Rolę takich miar mogą odgrywać w szczególności miary konfirmacji Bayesa, oryginalnie stosowane do oceny atrakcyjności reguł indukowanych w zastosowaniach deskrypcyjnych. Specyfiką tych miar jest osiąganie przez nie wartości dodatnich, gdy wystąpieniu przesłanki towarzyszy zwiększenie szansy na wystąpienie konkluzji i ujemnych, gdy wystąpienie przesłanki towarzyszy zmniejszenie tej szansę (w pozostałych przypadkach wartości miar są zerowe) [C03].

Badania eksperymentalne z miarami opartymi na tego typu danych (np. różnica sum przekątnej i przeciwprzekątnej macierzy) ujawniają właściwe zachowanie się tych miar w kontrolowaniu procesu generowania konstruktów w CRSA [A03]. Generowane za ich pomocą konstrukty posiadają właściwości predykcyjne porównywalne z innymi, uznanymi metodami selekcji atrybutów. Dodatkowo, zachowują one łatwość interpretacji właściwą konstruktom, czyli gwarantowanie rozróżniania obiektów z różnych klas i, jednocześnie, upodabniania obiektów z tych samych klas.

Podsumowanie

Podstawowe elementy wkładu naukowego zawartego w prezentowanym cyklu powiązanych tematycznie publikacji przedstawiają się następująco:

- W kontekście metodyki CRSA oraz metodyki DRSA:
  - opisano i przeanalizowano oparty na informacji międzyklasowej istniejący paradygmat redukcji atrybutów i system wcześniejszych pojęć obejmujących odpowiednie relacje międzyobiektowe oraz ich wykorzystanie w definicji pojęcia inter-reduktu,

  - zidentyfikowano potrzebę wprowadzenia informacji wewnątrzklasowej i jej zintegrowania z informacją międzyklasową w procesie redukcji atrybutów [A01, A02],
  - wprowadzono odpowiednie relacje międzyobiektowe dla informacji wewnątrzklasowej oraz pojęcie intra-reduktu [A01, A02],
  - wprowadzono pojęcie konstruktu, agregującego międzyklasowe właściwości inter-reduktu i wewnątrzklasowe właściwości intra-reduktu [A01, A02],
  - zaprezentowano lokalne i globalne (w odniesieniu do klas) odmiany inter-reduktów, intra-reduktów i konstruktów oraz definicje rdzeni reduktowych i konstruktowych [A02].
- Wprowadzono zunifikowane ramy definicji reduktów i konstruktów [A03], dzięki którym możliwe jest:
  - homogeniczne definiowanie reduktów i konstruktów (zarówno w CRSA jak i DRSA),
  - skuteczne generowanie reduktów i konstruktów z użyciem wspólnych algorytmów, dokładnych jak i heurystycznych (zarówno w CRSA jak i DRSA).
- Przeanalizowano właściwości teoretyczne reduktów i konstruktów dla danych o różnych konfiguracjach klas [A02] oraz wzajemne zależności warunkujące ich istnienie [A03].
- Przebadano skuteczność predykcyjną konstruktów (niezależnie od ich właściwości deskrypcyjnych) jako selektorów atrybutów w eksperymentach klasyfikacyjnych z rzeczywistymi i syntetycznymi zbiorami danych [A01, A03].
- Przedstawiono algorytmy dokładne generowania zbiorów wszystkich reduktów i konstruktów [A05].
- Zaproponowano i przebadano eksperymentalnie metodyczne usprawnienia algorytmów dokładnych generowania reduktów i konstruktów wykorzystujące zrównoleglanie obliczeń [A05].
- Przedstawiono algorytmy heurystyczne generowania reduktów i konstruktów oraz nakreślono sposoby ich kontrolowania z wykorzystaniem fakultatywnych miar atrakcyjności atrybutów [A03].
- Zweryfikowano eksperymentalnie przydatność miary atrakcyjności w roli sterującej procesem selekcji atrybutów [A03].

Prace stanowiące pozostały dorobek naukowy

Czasopisma naukowe

[B01] K. Dembczyński, P. Gaweł, A. Jaszkiewicz, W. Kotłowski, M. Kubiak, R. Susmaga, P. Wesołek, A. Wojciechowski, P. Zielniewicz: „Community traffic: a technology for the next generation car navigation", *Control and Cybernetics*, **41** (4), 2012, 867-883.
[lista B MNiSW(12.2014): 10pkt]

[B02] M. Grzymisławski, K. Słowiński, J. Nowacki, R. Susmaga, „Wyjściowe dane kliniczne a występowanie powikłań w przebiegu marskości wątroby – analiza doświadczenia klinicznego z zastosowaniem metodyki zbiorów przybliżonych", *Gastroenterologia Polska*, **11** (1), 2004, 27-34.
[lista A MNiSW(12.2014): 07pkt]

[B03] R. Pindur, R. Susmaga, „Fast Rule Extraction with Binary-Coded Relations", *Intelligent Data Analysis*, **7** (1), 2003, 27–42.
[lista A MNiSW(12.2014): 15pkt]

[B04] R. Pindur, R. Susmaga, J. Stefanowski, „Aggregation of Decision Rules by Hyperplanes", *Fundamenta Informaticae*, **61** (2), 2004, 117–137.
[lista A MNiSW(12.2014): 15pkt]
[IF(2004): 0,785]

[B05] R. Susmaga, I. Masłowska, L. Budzyńska, "The Concept of Topological Information in Text Representation", *Foundations of Computing and Decision Sciences*, **36** (1), 2011, 57–78.
[lista A MNiSW(12.2014): 09pkt]

[B06] R. Susmaga, I. Szczęch, „Can Interestingness Measures Be Usefully Visualized?", *International Journal of Applied Mathematics and Computer Science*, **25** (2), 2015.
[lista A MNiSW(12.2014): 25pkt]
[IF(2013): 1,39]

[B07] R. Susmaga, I. Szczęch: „Visualization Support for the Analysis of Properties of Interestingness Measures", *Bulletin of the Polish Academy of Sciences: Technical Sciences*, (accepted for publication).
[lista A MNiSW(12.2014): 25pkt]
[IF(2013): 1,00]

Materiały konferencyjne / monografie (w tym seria LNAI/LNCS) / inne

[B08] J. Błaszczyński, R. Słowiński, R. Susmaga: „Rule-based Estimation of Attribute Relevance", *Lecture Notes in Computer Science*, **6954**, 2011, 36–44.

[B09] L. Budzyńska, J. Jelonek, E. Łukasik, A. Naganowski, R. Słowiński, R. Susmaga, „Knowledge Discovery from Digital Media Objects Using Preference

Semantics", [w] P. Hobson, E. Izquierdo, Y. Kompatsiaris, N.E. O'Connor (red.), *Knowledge-Based Media Analysis for Self-Adaptive and Agile Multimedia Technology. Proceedings of EWIMT 2004*, Queen Mary University, London, 2004, 15–23.

[B10] L. Budzyńska, J. Jelonek, E. Łukasik, R. Słowiński, R. Susmaga, „Multistimulus ranking versus pairwise comparison in assessing quality of musical instruments sounds", *Proceedings of the 118th Convention of Audio Engineering Society, AES'118*, May 28–31, Barcelona, Spain, 2005.

[B11] K. Dembczyński, R. Pindur, R. Susmaga, „Dominance-Based Rough Sets Classifier without Induction of Decision Rules", *Proceedings of the Workshop „Rough Sets in Knowledge Discovery"*, Warsaw, April 12–13, 2003, *Electronic Notes in Theoretical Computer Science*, **82** (4), Elsevier, 2003, 84–95.

[B12] K. Dembczyński, R. Pindur, R. Susmaga, „Generation of Exhaustive Set of Rules within Dominance-based Rough Set Approach", *Proceedings of the Workshop „Rough Sets in Knowledge Discovery"*, Warsaw, April 12–13, 2003, *Electronic Notes in Theoretical Computer Science*, **82** (4), Elsevier, 2003, 96–107.

[B13] M. Flinkman, W. Michalowski, S. Nilsson, R. Słowiński, R. Susmaga, S. Wilk: „Identifying important attributes for the Siberian forest management using rough sets analysis", [w] T. Trzaskalik, J. Michnik (red.), *Multiple Objective and Goal Programming: Recent Developments*, Springer-Verlag, Heidelberg, Germany, 2002, 272–282.

[B14] P. Gaweł, K. Dembczyński, W. Kotłowski, M. Kubiak, R. Susmaga, P. Wesołek, P. Zielniewicz, A. Jaszkiewicz: „Community Traffic: A Technology for the Next Generation Car Navigation", [w] Pechenizkiy, M., Wojciechowski, M. (red.), *Proceedings of the 16th East-European Conference on Advances in Databases and Information Systems (ADBIS 2012), New Trends in Databases and Information Systems: Advances in Intelligent Systems and Computing*, **185**, Springer-Verlag, 2013, 339–348.

[B15] P. Gaweł, K. Dembczyński, R. Susmaga, P. Wesołek, P. Zielniewicz, A. Jaszkiewicz: „Adapting Travel Time Estimates to Current Traffic Conditions", [w] M. Pechenizkiy, M. Wojciechowski, (red.), *Proceedings of the 16th East-European Conference on Advances in Databases and Information Systems (ADBIS 2012), New Trends in Databases and Information Systems: Advances in Intelligent Systems and Computing*, **185** , Springer-Verlag, 2013, 339–348.

[B16] J. Jelonek, R. Słowiński, R. Susmaga, „Sequential Construction of Features Based on Genetically Transformed Data", [w] Kay Chen Tan *et al.* (red.), *Recent Advances in Simulated Evolution and Learning*, World Scientific, August 2004, 623–642.

[B17] M. Libera, R. Susmaga, „Zastosowanie metod uczenia maszynowego w problematyce powierzchniowej trwałości zmęczeniowej", *IX Kongres eksploatacji urządzeń technicznych: materiały konferencyjne*, Krynica, Sekcja

Podstaw Eksploatacji Komitetu Budowy Maszyn PAN, Instytut Technologii Eksploatacji w Radomiu, **1**, 2001, 147–153.

[B18] E. Łukasik, R. Susmaga, „Classification Experiments with Plosive Consonants", *Proceedings of Artificial Intelligence Methods (AI-METH'03)*, 5–7 November, Gliwice, 2003, 79–80.

[B19] E. Łukasik, R. Susmaga, „Nienadzorowane metody uczenia maszynowego w wizualizacji cech dźwięku skrzypiec", *Materiały Konferencyjne 50. Otwartego Seminarium z Akustyki (OSA)*, Szczyrk, 2003, 329–334.

[B20] E. Łukasik, R. Susmaga, „Phoneme, Gender And Speaker Variability Visualization in Voiceless Stop Consonants", *Proceedings of Signal Processing 2003*, November, Poznań, 2003, 35–40.

[B21] E. Łukasik, R. Susmaga, „Unsupervised Machine Learning Methods In Timbral Violin Characteristics Visualization", *Proceedings of Stockholm Music Acoustics Conference (SMAC'03)*, 6–9 August, Stockholm, Sweden, 2003, 83–86.

[B22] R. Susmaga, "Confusion Matrix Visualization", [w] M. A. Kłopotek, S. T. Wierzchoń, K. Trojanowski (red.), *Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM'04 Conference*, Zakopane, Poland, May 17–20, 2004, *Advances in Soft Computing*, Springer, 2004, 107–116.

[B23] R. Susmaga, „Inter-class and Intra-class Reducts" (extended abstract), [w] Z. Suraj, (red.), *Proceedings of the 6th International Conference on Soft Computing and Distributed Processing (SCDP'02)*, Rzeszów (2002), 74–77.

[B24] R. Susmaga, I. Szczęch, „Selected Group-Theoretic Aspects of Confirmation Measure Symmetries", *Research Report RA–019/13*, Poznań University of Technology, 2013.

[B25] R. Susmaga, I. Szczęch: „Statistical Significance of Bayesian Confirmation Measures", *Research Report RA–010/12*, Poznań University of Technology, 2012.

[B26] R. Susmaga, I. Szczęch: „The Property of χ^2_{01}-Concordance for Bayesian Confirmation Measures", [w] V. Torra *et al*. (red.): Modeling Decisions for Artificial Intelligence, *MDAI 2013*, Barcelona, Spain, November 20–22, 2013, *Lecture Notes in Artificial Intelligence*, **8234**, 2013, 226–236.

[B27] R. Susmaga, I. Szczęch: „Visual-Based Detection of Properties of Confirmation Measures", [w] Andreasen, T., Christiansen, H., Cubero Talavera J.C., Raś, Z.W. (red.), *Foundations of Intelligent Systems. Proceedings of the 21st International Symposium on Methodologies for Intelligent Systems, ISMIS 2014*, Roskilde, Denmark, June 25–27, 2014, *Lecture Notes in Computing Science*, **8502**, Springer (2014), 133–143.

[B28] R. Susmaga, I. Szczęch: „Visualization of interestingness measures", [w]: „Proceedings of the 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics", *Fundacja UAM*, Poznań (2013), 95–99.

[B29] P. Wesołek, P. Gaweł, A. Jaszkiewicz, K. Dembczynski, R. Susmaga, P. Zielniewicz: „Oszacowanie straty czasu w podróżach samochodowych w Warszawie wynikającej z nadmiernego natężenia ruchu”, *Raport Badawczy RB–001/12*, Politechnika Poznańska, 2012.

Publikacje nie wchodzące w skład prezentowanego cyklu poruszają różne aspekty teoretyczne oraz praktyczne analizy danych, często bardzo blisko związane z problemami poruszanymi w artykułach wchodzących w skład cyklu, i dotyczą:

- propozycji algorytmów / koncepcji: [B03, B04, B11, B12, B23, B24, B25, B26],
- propozycji metod selekcji i konstrukcji atrybutów w różnych środowiskach: [B05, B08, B16],
- propozycji technik wizualizacji danych wielowymiarowych: [B06, B07, B22, B27, B28],
- zastosowań klasycznej / dominacyjnej teorii zbiorów przybliżonych: [B02, B09, B10, B13, B17],
- zastosowań nadzorowanych / nienadzorowanych metod uczenia maszynowego: [B18, B19, B20, B21],
- zastosowań metod regresji liniowej / nieliniowej: [B01, B14, B15, B29].

Miary konfirmacji jako miary atrakcyjności

W szczególności, prace [B06, B07, B08, B24, B25, B26, B27, B28] stanowią zbiór publikacji, których tematyka jest zbieżna z tematyką przedstawianą w prezentowanym cyklu publikacji. Prace te skupiają się na różnych aspektach miar atrakcyjności, w tym miarach konfirmacji Bayesa, które okazują się przydatne w algorytmach redukcji atrybutów oraz procedurach oceniania atrybutów, zarówno w kontekstach deskrypcyjnych jak i predykcyjnych. Poziom analiz jest jednakże wyższy (kompleksowy raczej niż szczegółowy) niż w tamtych metodach (przykładowe wykorzystanie miary atrakcyjności w kontrolowaniu heurystycznego algorytmu generowania konstruktów jest zademonstrowane w pracy [A03] cyklu publikacji).

Skuteczne stosowanie miar w ocenianiu przydatności atrybutów do wybranych celów wymaga jednakże wcześniejszego przeanalizowania tych miar pod względem ich podstawowych właściwości. To z kolei może wymagać utworzenia metod wspomagających tę analizę, np. w postaci technik wizualizacyjnych, które pozwalają na łatwe pozyskanie informacji o zachowaniu się analizowanych miar w różnych obszarach ich wielowymiarowych dziedzin. Oprócz kontrolowania algorytmów heurystycznych, miary atrakcyjności (konkretniej: miary konfirmacji Bayesa) mogą być wykorzystane w ocenianiu przydatności atrybutów na potrzeby eksperymentów klasyfikacyjnych.

Przykładową grupą miar o szeroko znanych i korzystnych właściwościach teoretycznych są miary konfirmacji Bayesa. Miary te, stosowane oryginalnie do oceny atrakcyjności reguł decyzyjnych indukowanych w zastosowaniach deskrypcyjnych, mogą być z równym powodzeniem aplikowane do oceny atrybutów oraz ich zbiorów generowanych zarówno w celach deskrypcyjnych jak i predykcyjnych. Ocena ta może

być ponadto dokonywana bezpośrednio lub pośrednio, z użyciem dodatkowej metody indukcyjnej w postaci zewnętrznego klasyfikatora, np. zbioru reguł decyzyjnych. Traktując fakt wykorzystania danego atrybutu w regułach jako przesłankę, a fakt osiągnięcia wysokiej trafności klasyfikowania jako konkluzję, otrzymuje się (potencjalną) zależność przesłanka-konkluzja, która podlega wartościowaniu przez miary konfirmacji. Adaptacja ta pozwala na stosowanie miar konfirmacji jako przydatnego narzędzia w problemie oceniania atrybutów, stanowiącego naturalny etap pośredni w ogólnym procesie redukcji.

Analizy miar atrakcyjności

W pracach [B24, B25, B26] przedstawiono analizy wybranych właściwości miar atrakcyjności. Właściwości porządkują obszerny zbiór wszystkich dostępnych miar w grupy miar, które są podobne pod względem swoich różnych cech i zachowań, a tym samym ułatwiają dobór najlepszych możliwych miar do konkretnych zastosowań. Przedstawione analizy dotyczą konkretnie miar konfirmacji Bayesa (stanowiących typowe przykłady miar atrakcyjności).

Przedmiotem prac [B25, B26] są statystyczne właściwości miar konfirmacji Bayesa, w tym zgodność ze statystycznie istotną zależnością pomiędzy występowaniem przesłanki i konkluzji. W sytuacjach, w których analizowane dane mogą być obarczone błędami obserwacyjnymi, może dochodzić do bardzo niepożądanych sytuacji, np. takich, w których rozważana miara konfirmacji zaniża lub zawyża generowane oceny. Prace definiują współczynnik (tzw. konkordancję) oceniający zależność pomiędzy występowaniem przesłanki a konkluzji w danych eksperymentalnych, a następnie wprowadzono sposób kwantyfikowania poziomu zgodności tego współczynnika z analizowaną miarą. Odstępstwa od tej zgodności mogą być dalej analizowane i interpretowane w kategoriach ryzyka: miara konfirmacji, która zaniża oceny jest nazywana miarą wykazującą awersję do ryzyka, a miara, która zawyża oceny jest nazywana miarą wykazującą skłonność do ryzyka.

Z kolei celem pracy [B24] jest demonstracja pochodzących z teorii grup aspektów symetrii miar konfirmacji Bayesa. Po wprowadzeniu niezbędnych elementów teorii grup, przede wszystkim grupy diedralnej D_4, wraz z jej najbardziej przydatnymi właściwościami, praca przedstawia miary konfirmacji i ich podstawowe właściwości, z uwzględnieniem właściwości symetrii oraz podziału tych symetrii na symetrie korzystne (dla danej miary) i symetrie niekorzystne. Następnie praca prezentuje interpretacje symetrii w kategoriach teorii grup, wykazując, że wszystkie symetrie można traktować jako permutacje, a ich kombinacje jako złożenia permutacji. Rozpatrywany zbiór wszystkich symetrii jest interpretowany jako grupa izomorficzna z przedstawioną grupą a podzbiory korzystnych symetrii jako podgrupy właściwe tej grupy. Ostatecznie przedstawiono warunki pozwalające na weryfikowanie spójności podziału symetrii na korzystne i niekorzystne oraz wyniki analizy kombinatorycznej wszystkich możliwych podziałów, które są spójne z przyjętymi warunkami.

Metody wspomagające analizy miar atrakcyjności

W pracach [B06, B07, B27, B28] zaproponowano i zweryfikowano wizualne techniki analizy miar atrakcyjności, których empiryczny charakter jest dobrym uzupełnieniem

rozmaitych rozważań teoretycznych. Wizualizacja pozwala na łatwe pozyskanie informacji o zachowaniu się analizowanych miar we wszystkich obszarach ich dziedzin, usprawniając znacząco proces doboru miar do poszczególnych zastosowań, np. do oceniania atrybutów oraz ich zbiorów. Wprowadzone techniki zaprezentowano w zastosowaniu do miar konfirmacji Bayesa, stanowiących typowe przykłady miar atrakcyjności (choć technika wizualizacyjna może być aplikowana do dużo szerszej klasy współczynników, np. popularnych współczynników oceny klasyfikatorów).

Wprowadzona w pracach [B06, B28] technika wizualizacji wykorzystuje fakt posiadania przez rozważane miary wspólnej, czterowymiarowej dziedziny, elementy której są reprezentowane przez nieujemne macierze 2×2 o sumie dodatniej. Dzięki odpowiednim przeskalowaniu tej sumy możliwe jest reprezentowanie relatywnej wersji oryginalnej, czterowymiarowej dziedziny miar w barycentrycznym układzie współrzędnych (na bazie czworościanu), reprezentacja którego możliwa jest w przestrzeni trójwymiarowej. Ponieważ wartość dodatkowej funkcji od czterech argumentów dziedziny, np. dowolnej miary atrakcyjności, może być wizualizowana jako kolor odpowiadającego mu punktu, wizualizacja miary w całej swej dziedzinie sprowadza się do wizualizacji kolorowego czworościanu. W celach ilustracyjnych dokonano analizy szeregu miar, demonstrując możliwości prezentowanej techniki wizualizacji w uwidacznianiu różnorodnych charakterystycznych cech (np. ekstremów, zer, obszarów wzrostu / spadku, itp.) pojedynczych miar, agregacji par miar (np. różnic między nimi) oraz agregacji grup miar (np. wariancji w tych grupach).

Prace [B07, B27] demonstrują jak opisywany system wizualizacji może być zastosowany w szczególności do wizualnej analizy miar atrakcyjności pod względem ich wybranych właściwości. Podejście wizualne jest użyteczną alternatywą dla czasochłonnych analiz teoretycznych, z uwagi na możliwość sprawnego i prostego identyfikowania faktu spełniania albo niespełniania (a także ewentualnie spełniania w zakresie ograniczonym / przybliżonym) wybranych właściwości przez analizowane miary. W szczególności skuteczna jest wizualizacja ważnych właściwości miar konfirmacji Bayesa, co pozwala na łatwą identyfikację tych właściwości nawet u złożonych miar, w tym u miar typu hierarchicznego (czyli miar konfirmacji zdefiniowanych jako monotoniczne agregacje innych miar konfirmacji).

Zastosowania miar atrakcyjności

Adaptacja miar konfirmacji Bayesa do problemu oceniania atrakcyjności atrybutów w konkretnych celach predykcyjnych (klasyfikacja obiektów, dane nadzorowane) wymaga, na wzór innych miar oceny atrakcyjności, sprecyzowania (dla danego atrybutu) odpowiedniej przesłanki i odpowiedniej konkluzji. W najprostszych sytuacjach może to oznaczać badanie indywidualnej zależności pomiędzy atrybutem ocenianym a atrybutem decyzyjnym, jednak uzyskana tak ocena indywidualna dotyczy zasadniczo pojedynczych atrybutów i w ogólności nie jest kompatybilna z oceną ich wieloelementowych zbiorów (ze względu na ignorowanie potencjalnych wzajemnych interakcji pomiędzy atrybutami występującymi w zbiorach).

Z kolei ocena zbiorów atrybutów prawidłowo uwzględniająca potencjalne interakcje pomiędzy atrybutami może być implementowana w sytuacji, w której jako przesłankę i konkluzję traktuje się wykorzystanie danego atrybutu przez wygenerowany klasyfikator

(np. zbiór reguł) i poprawne klasyfikowanie obiektów przez ten klasyfikator [B08]. Oznacza to, że wartość oceny danego atrybutu rośnie wraz z faktem częstego wykorzystania tego konkretnego atrybutu w klasyfikatorze (np. w przypadku klasyfikatora regułowego jest to jego występowanie w warunkach reguł) charakteryzujących się wysoką trafnością. Skuteczność tak uzyskanej oceny może być niezależnie oceniana w eksperymentach klasyfikacyjnych, w których w testach walidacyjnych szacuje się przydatność zbiorów złożonych z najwyżej ocenionych atrybutów. Doświadczenia z wybraną miarą konfirmacji wykazują, że atrybuty o najwyższych konfirmacjach według tej miary okazują się bardzo przydatne w konstruowaniu zbiorów atrybutów zapewniających wysokie skuteczności klasyfikacji.

Podsumowanie tematyki miar atrakcyjności

W ramach publikacji poświęconych miarom atrakcyjności:

- Przeanalizowano wybrane właściwości miar konfirmacji Bayesa.
- Zaproponowano system wizualizacji miar atrakcyjności pozwalający na empiryczne analizowanie ich różnych cech i właściwości oraz przedstawiono wykorzystanie tego systemu do przeprowadzenia analiz wybranych miar konfirmacji Bayesa.
- Zademonstrowano wykorzystanie miar konfirmacji Bayesa w problemie oceniania przydatności atrybutów w problemach klasyfikacyjnych.

Referencje

[C01] J. Bazan, A. Skowron, P. Synak, „Dynamic Reducts as a Tool for Extracting Laws from Decisions Tables”, *LNCS*, **869**, 1994, 346–355.

[C02] M. Borkowski, D. Ślęzak, „Application of Discernibility Tables to Calculation of Approximate Frequency Based Reducts”, *LNAI* **2005**, Springer-Verlag, 2001, 123–130.

[C03] R. Carnap, *Logical Foundations of Probability* (2nd ed.), University of Chicago Press, 1962.

[C04] M. Dash, H. Liu, „Feature selection for classification”, *Intelligent Data Analysis*, **1** (3), 1997.

[C05] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, John Wiley, New York, 2001.

[C06] R. Hilderman, H. Hamilton, *Knowledge Discovery and Measures of Interest*, Kluwer Academic Publishers, 2001.

[C07] R. Kohavi, G.H. John, „Wrappers for Feature Selection” *Artificial Intelligence*, **97**, 1997, 273–324.

[C08] M. Kryszkiewicz, „Maintenance of reducts in the variable precision rough set model”, [w] T.Y. Lin, N. Cercone (eds), *Rough Sets and Data Mining – Analysis of Imperfect Data*, Kluwer Academic Publishers, Boston, 1997, 355–372.

[C09] Z. Pawlak, *Rough Sets, Theoretical Aspects of Reasoning About Data*, Kluwer Academic Publishers, Dordrecht, 1991.

[C10] Z. Pawlak, R. Słowiński, „Rough set approach to multi-attribute decision analysis”, *European Journal of Operational Research*, **72** (1994), 443–459.

[C11] A. Skowron, C. Rauszer, „The discernibility matrices and functions in information systems”, [w] R. Słowiński (ed.), *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory*, Kluwer Academic Publishers, Dordrecht, 1992, 331–362.

[C12] R. Susmaga, „New test for inclusion minimality in reduct generation”, *Foundations of Computing and Decision Sciences*, **25** (2), 2000, 121–146.

[C13] R. Susmaga, „Parallel computation of reducts”, *LNAI* **1424**, Springer-Verlag, 1998, 450–457.

[C14] R. Susmaga, R. Słowiński, S. Greco, B. Matarazzo, „Computation of Reducts and Rules in Dominance-Based Rough Sets Approach”, *Control and Cybernetics*, **29** (4) 2000, 969–988.

[C15] D. Ślęzak, „Approximate Entropy Reducts”, *Fundamenta Informaticae*, **53**, 2002, 365–390.

Dodatek

Zestawienie wybranych skrótów / terminów

- Skróty nazw metodyk analizy danych:
  - RSA (ang. Rough Sets Approach) – metodyka zbiorów przybliżonych,
  - CRSA (ang. Classic RSA) – klasyczna metodyka zbiorów przybliżonych,
  - DRSA (ang. Dominance-based RSA) – dominacyjna metodyka zbiorów przybliżonych.
- Skróty dotyczące relacji:

  CRSA:
  - IND (ang. indiscernibility) – dosł. nierozróżnialność (relacja bazowa w CRSA),
  - DIS (ang. discernibility) – dosł. rozróżnialność (relacja dopełniająca relacji IND),
  - SIM (ang. similarity) – dosł. podobieństwo (relacja wywiedziona z relacji IND).

  DRSA:
  - DOM (ang. dominance) – dosł. dominacja (relacja bazowa w DRSA),
  - NDM (ang. non-dominance) – dosł. nie-dominacja (relacja dopełniająca relacji DOM),
  - INS (ang. inseparability) – dosł. nieseparowalność (relacja pomocnicza w DRSA),
  - PRX (ang. proximity) – dosł. bliskość (relacja wywiedziona z relacji INS).
- Terminy dotyczące relacji:

  W ramach CRSA / DRSA:
  - termin określający obiekty spełniające relację IND / DIS: (*nie*)*rozróżnianie* (CRSA),

- termin określający obiekty spełniające relację SIM: *upodabnianie* (CRSA),
- termin określający obiekty spełniające relację NDM / DOM: *(nie)dominowanie* (DRSA),
- termin określający obiekty spełniające relację PRX: *ujednolicanie* (DRSA).

Poza ramami CRSA / DRSA:

- termin stanowiący uogólnienie rozróżniania i dominowania: *różnicowanie*,
- termin stanowiący uogólnienie upodabniania i ujednolicania: *unifikowanie*.

Zestawienie zbiorów prac

- Prace autora (w tym współtworzone) opublikowane po uzyskaniu stopnia doktora:
  - wchodzące w skład cyklu powiązanych tematycznie publikacji: [A01–A03],
  - stanowiące pozostały dorobek naukowy: [B01–B29].
- Prace cytowane: [C01–C15]
  - prace autora (w tym współtworzone) opublikowane przed uzyskaniem stopnia doktora),
  - prace innych autorów.

R. Susmaga

.....

Robert Susmaga