View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Repozytorium UW

Algorytmy uczenia się relacji podobieństwa z wielowymiarowych zbiorów danych

(Autoreferat rozprawy doktorskiej)

Andrzej Janusz

Streszczenie

Pojęcie podobieństwa pełni istotną rolę w dziedzinach uczenia maszynowego i sztucznej inteligencji. Jest ono powszechnie wykorzystywane w zadaniach dotyczących nadzorowanej klasyfikacji, grupowania, wykrywania nietypowych obiektów oraz planowania [2, 24, 38]. Ponadto w dziedzinach takich jak wyszukiwanie informacji (ang. information retrieval) lub wnioskowanie na podstawie przykładów (ang. case-based reasoning) pojęcie podobieństwa jest kluczowe ze względu na jego obecność na wszystkich etapach wyciągania wniosków [1]. Jednakże samo podobieństwo jest pojęciem niezwykle złożonym i wymyka się próbom ścisłego zdefiniowania. Stopień podobieństwa między dwoma obiektami może być różny w zależności od kontekstu w jakim się go rozpatruje. W praktyce trudno jest nawet ocenić jakość otrzymanych stopni podobieństwa bez odwołania się do zadania, któremu mają służyć. Z tego właśnie powodu modele oceniające podobieństwo powinny być wyuczane na podstawie danych, specjalnie na potrzeby realizacji konkretnego zadania.

W niniejszej rozprawie opisano model podobieństwa zwany Regułowym Modelem Podobieństwa (ang. Rule-Based Similarity) oraz zaproponowano algorytm tworzenia tego modelu na podstawie danych. Wykorzystuje on elementy teorii zbiorów przybliżonych [27] do konstruowania funkcji podobieństwa pozwalającej aproksymować podobieństwo w zadanym kontekście. Konstrukcja ta rozpoczyna się od wykrywania zbiorów wysokopoziomowych cech obiektów. Mogą być one interpretowane jako istotne aspekty podobieństwa. Mając zdefiniowane tego typu cechy możliwe jest wykorzystanie idei modelu kontrastu cech Tversky'ego [37] (ang. feature contrast model) do budowy precyzjnej oraz zgodnej z obserwacjami psychologów funkcji podobieństwa dla rozważanego problemu. Dodatkowo, niniejsza rozprawa zawiera opis dwóch rozszerzeń Regułowego Modelu Podobieństwa przystosowanych do działania na danych o bardzo wielu atrybutach. Starają się one włączyć do modelu szerszy zakres aspektów podobieństwa. W pierwszym z nich odbywa się to poprzez konstruowanie wielu zbiorów cech z reduktów decyzyjnych. Aby zapewnić ich zróżnicowanie, zaproponowano algorytm łączący heurystykę zachłanną z elementami losowymi. Podejście to jest szczególnie wskazane dla zadań związanych z problemem małej liczby obiektów i dużej liczby cech (ang. the few-objects-many-attributes problem), np. analizy danych mikromacierzowych. Podobny pomysł może być również wykorzystany w dziedzinie analizy tekstów. Realizowany jest on przez drugie z proponowanych rozszerzeń modelu. Łączy ono metodę semantycznego indeksowania z algorytmem obliczania bireduktów informacyjnych, aby reprezentować teksty dobrze zdefiniowanymi pojęciami.

Funkcja podobieństwa zaproponowanego modelu może być wykorzystana do klasyfikacji nowych obiektów oraz do łączenia dokumentów tekstowych w semantycznie spójne grupy. Eksperymenty, których wyniki opisano w rozprawie, dowodzą, że zaproponowane modele mogą skutecznie konkurować nawet z powszechnie uznanymi rozwiązaniami.

Słowa kluczowe: Regułowy Model Podobieństwa, Nauka Podobieństwa, Teoria Zbiorów Przybliżonych, Model Kontrasu Cech Tversky-ego, Rozumowanie Oparte na Przykładach, Ekstrakcja Cech

1 Opis problemu

Umiejętność identyfikacji podobnych obiektów pełni fundamentalną rolę w procesie podejmowania decyzji i uczenia się [28, 29, 36]. Fakt ten zauważony został przez wielu znakomitych naukowców, którzy poświęcili swe badania odkrywaniu własności podobieństwa i projektowaniu modeli pozwalających automatycznie mierzyć stopień podobieństwa między zadanymi obiektami [7, 8, 9, 37]. Niestety ze względu na złożoność i subiektywny charakter tego pojęcia, jak dotąd nikomu nie udało się podać jego ścisłej definicji. Mimo to podobieństwo jest wykorzystywane przez liczne algorytmy uczenia maszynowego w zastosowaniach takich jak nadzorowana klasyfikacja, grupowanie oraz identyfikacja nietypowych obiektów [1, 25, 35]. Niestety, ze względu na trudności związane z manualnym wyborem modelu podobieństwa dla danych, skuteczność algorytmów wykorzystujących podobieństwo bywa często ograniczona. Problem ten jest szczególnie widoczny, gdy dane dotyczące interesujących obiektów posiadają bardzo dużą liczbę atrybutów [4].

Tematem rozprawy jest problem uczenia się, w jaki sposób oceniać, czy w danym kontekście dwa wskazane obiekty są do siebie podobne. W tym celu wykorzystywano już bardzo wiele modeli, które miały łączyć intuicyjne własności podobieństwa postulowane przez psychologów z wydajnością i dokładnością w zastosowaniach. Duża część z nich bazowała na różnego rodzaju metrykach odległości. W podejściu tym obiekty traktowane są jak punkty w przestrzeni metrycznej zdefiniowanej przez ich cechy, a podobieństwo między obiektami jest nierosnącą funkcją ich odległości. Obiekty uznaje się za podobne, jeśli są dostatecznie blisko w tak zadanej przestrzeni [4, 37]. Modele tego typu można często poprawić, przypisując wagi do atrybutów w celu wyrażenia ich istotności dla modelu. Optymalizacja takich wag skutkuje lepszym dopasowaniem do danych. Algorytmy, które temu służą, można zatem traktować jako przykłady metod uczenia się podobieństwa z danych. Były one tematem wielu badań, np. [5, 24, 33, 38, 40].

Z jednej strony modele tego typu mogą się wydawać zgodne z intuicją – obiekty posiadające zbliżone wartości na atrybutach powinny być podobne. Jednakże z drugiej strony psychologowie tacy jak Amos Tversky empirycznie udowodnili, że w niektórych kontekstach podobieństwo nie posiada własności narzucanych przez miary odległości, takich jak symetria, czy też zachowywanie własności trójkąta [7, 37]. Sytuacja ta zdarza się szczególnie często, gdy rozpatruje się obiekty o bardzo dużej złożoności, które częstokroć są opisane dużą liczbą cech. Dzieje się tak, ponieważ złożone obiekty mogą być do siebie podobne w jednym aspekcie a jednocześnie niepodobne w innych. Zatem aby zdecydować, które z możliwych aspektów podobieństwa są ważniejsze, konieczna jest dodatkowa wiedza na temat kontekstu [8, 9, 37].

Ponadto zależności między lokalnym (tj. ograniczonym jedynie do pewnych cech lub aspektów) a globalnym podobieństwem mogą być wysoce nieliniowe i, aby je uchwycić, konieczne jest przejście na poziom bardziej ogólnych charakterystyk obiektów. Ze względu na potencjalnie nieograniczoną liczbę tego typu cech w praktyce często niemożliwym jest, aby konstrukcją wysokopoziomowych charakterystyk dla rozpatrywanego zbioru danych zajmowali się eksperci. Również z tego powodu wysokopoziomowe cechy oraz metody ich agregacji powinny być automatycznie wyuczane na podstawie dostępnych danych. Oczywiście, tak jak i w innych zagadnieniach uczenia maszynowego, algorytmy uczenia się podobieństwa powinny równoważyć złożoność oraz skuteczność [25, 35]. Tworzenie za bardzo skomplikowanego modelu może być zbyt kosztowne obliczeniowo, aby można było wykorzystać taki model do rozwiązywania rzeczywistych problemów. Ponadto model taki mógłby zbytnio dopasowywać się do danych wykorzystywanych do nauki, co często prowadzi do słabych wyników dla nowych obiektów.

W rozprawie omówiono problem uczenia się oceniania podobieństwa w kontekście

narzuconym przez wykonywane zadanie (np. klasyfikacja, grupowanie, itp.). Na podstawie obserwacji psychologów zajmujących się badaniem podobieństwa sformułowano wymagania dotyczące budowy oraz pożądanych własności modelu. Opisano również wyzwania związane z tym problemem oraz propozycje praktycznych rozwiązań. Na koniec sprawdzono skuteczność zaproponowanych metod przez liczne eksperymenty na prawdziwych zbiorach danych. Główne rezultaty rozprawy można zatem podzielić na cztery kategorie:

1. Analiza i porównanie własności znanych modeli podobieństwa z punktu widzenia metod analizy danych i sztucznej inteligencji.

2. Propozycja ogólnego modelu podobieństwa (nazwanego Regułowym Modelem Podobieństwa) wraz z charakterystyką jego wybranych własności.

3. Propozycja trzech algorytmów konstruowania modelu Regułowego Modelu Podobieństwa na podstawie różnych typów danych.

4. Implementacja oraz empiryczna ewaluacja jakości zaproponowanych algorytmów na licznych zbiorach danych o różnej charakterystyce i w kontekście różnych zadań.

Większość z opisywanych cząstkowych wyników była już prezentowana na międzynarodowych konferencjach i warsztatach. Zostały one opublikowane w ponad dwudziestu artykułach, które ukazały się w recenzowanych materiałach konferencyjnych oraz w renomowanych czasopismach. Publikacje autora rozprawy dotyczące budowy oraz zastosowań Regułowego Modelu Podobieństwa to m.in. [10, 11, 12, 14, 15, 17, 20]. Na kształt prezentowanego modelu duży wpływ miały również inne kierunki badań prowadzonych przez autora. Dotyczyły one takich zagadnień jak problem selekcji cech, czy też uczenie się przy użyciu zespołów klasyfikatorów [13, 16, 18, 19, 21, 23, 31, 39]. Ponadto nienadzorowana wersja zaproponowanego modelu podobieństwa była zainspirowana doświadczeniami autora z pracy nad metodami wyszukiwania informacji i semantycznym indeksowaniem dokumentów tekstowych prowadzonej w ramach projektu SYNAT [22, 32, 34] (jedynie wybrane publikacje). Dwie spośród publikacji autora zostały nagrodzone na międzynarodowych konferencjach. Praca [22] otrzymała tytuł Najlepszej Pracy konferencji *RSCTC'2012*, a za pracę [19] autor otrzymał Nagrodę im. Zdzisława Pawlaka, przyznawaną podczas konferencji *FedCSIS'2012*. Ponadto prowadzone badania zaowocowały sukcesami na międzynarodowych konkursach analizy danych, między innymi pierwszym miejscem podczas *Australasian Data Mining 2009 Analytic Challenge: Ensembling* [13, 16] oraz trzecim miejscem w konkursie *IEEE ICDM Contest: TomTom Traffic Prediction for Intelligent GPS Navigation – the GPS task*. Autor współorganizował również konkurs *RSCTC'2010 Discovery Challenge: Mining DNA Microarray Data for Medical Diagnosis and Treatment* [39] oraz pełnił rolę przewodniczącego komitetu organizacyjnego konkursu *JRS'2012 Data Mining Competition: Topical Classification of Biomedical Research Papers* [18].

2 Główne wyniki rozprawy

Model podobieństwa, będący najważniejszym wynikiem rozprawy, ma swoje źródło w teorii zbiorów przybliżonych (ang. rough set theory) [27]. Aby lepiej przedstawić jego budowę i działanie, należy odwołać się do kilku podstawowych pojęć z tej dziedziny. W teorii zbiorów przybliżonych dostępna wiedza o danym zbiorze obiektów z uniwersum Ω jest reprezentowana poprzez *system informacyjny* $\mathbb{S} = (U, A)$, gdzie $U \subseteq \Omega$ jest skończonym niepustym zbiorem obiektów, a A jest skończonym niepustym zbiorem atrybutów (cech) tychże obiektów. Jeśli dodatkowo w zbiorze cech systemu informacyjnego jest wyróżniony

jeden lub więcej atrybutów, wedle których dzieli się obiekty na klasy, to system taki nazywamy decyzyjnym. Wyróżnione atrybuty systemu decyzyjnego nazywamy atrybutami decyzyjnymi lub w skrócie decyzjami. System taki oznaczamy poprzez $\mathbb{S}_d = (U, A \cup \{d\})$.

Mając dany system informacyjny $\mathbb{S}$, możemy przybliżać pojęcia odpowiadające dowolnym podzbiorom jego obiektów przy użyciu zbiorów opisanych za pomocą formuł języka logiki decyzyjnej L_A, określonych na wartościach atrybutów. W szczególności, jeśli interesujące nas pojęcie[1] odpowiadające zbiorowi obiektów $X \subset U$, to powiemy, że *zbiorem przybliżonym* dla tego pojęcia jest para $(\underline{X}, \overline{X})$, gdzie

$$\begin{aligned} \underline{X} &= \{u \in U : [u]_A \subseteq X\}, \\ \overline{X} &= \{u \in U : [u]_A \cap X \neq \emptyset\}. \end{aligned}$$

W powyższej formule $[u]_A = \{u' \in U : \forall_{a \in A}\ a(u') = a(u)\}$. Zbiory takie nazywamy klasami nierozróżnialności w $\mathbb{S}$, a zbiory $\underline{X}$ i $\overline{X}$ nazywamy odpowiednio dolnym i górnym przybliżeniem X.

Zbiory przybliżone można również wykorzystywać do aproksymowania dowolnych relacji. Niezmiernie istotny jest przy tym odpowiedni wybór przestrzeni aproksymacji [30]. Przestrzeń, która generuje zbyt małe klasy nierozróżnialności, może pozwalać na bardzo dokładne przybliżanie pojęć lub relacji na dostępnym zbiorze obiektów, lecz przybliżenie to może okazać się mylne dla nowych obiektów. Z tego powodu przy wyborze przestrzeni aproksymacji stosuje się Regułę Minimalnego Opisu (ang. Minimum Description Length). Z realizacją tej heurystycznej reguły w kontekście zbiorów przybliżonych związane jest pojęcie *reduktu decyzyjnego*. Reduktem decyzyjnym nazwiemy minimalny[2] podzbiór atrybutów $DR \subseteq A$, który nie łączy ze sobą klas nierozróżnialności obiektów posiadających różne decyzje, o ile były one rozróżnialne w pełnym systemie decyzyjnym. Możliwe jest również zdefiniowanie reduktu dla systemów informacyjnych bez wyróżnionej decyzji jako podzbioru atrybutów, który nie zmniejsza liczby nierozróżnionych obiektów w stosunku do pełnego zbioru cech. Redukty takie nazywamy informacyjnymi.

2.1 Analiza i porównanie własności znanych modeli podobieństwa

Podobieństwo można traktować jako binarną relację τ pomiędzy obiektami z rozważanego uniwersum Ω. Wiele badań empirycznych prowadzonych przez psychologów i kognitywistów pokazało, że percepcja podobnych obiektów u ludzi jest istotnie zależna od czynników zewnętrznych, takich jak dostępna informacja, wcześniejsze doświadczenia życiowe badanych oraz kontekst [8, 37]. Ten ostatni czynnik jest szczególnie ważny. Przykładowym kontekstem dla oceny podobieństwa może być atrybut decyzyjny rozważanych obiektów. W kontekście tym możemy wyróżnić pewną przydatną własność relacji podobieństwa – jeśli dwa obiekty są podobne to muszą należeć do tej samej klasy decyzyjnej. Własność tę możemy wykorzystać w czasie nauki podobieństwa dla zadanego systemu decyzyjnego.

Do przybliżania relacji podobieństwa bardzo często wykorzystuje się specjalne funkcje zwane funkcjami lub miarami podobieństwa. W rozprawie podjęto próbę formalnego opisu własności, jakie powinna posiadać funkcja podobieństwa pasująca do zadanej relacji:

Definicja 1 (Właściwa funkcja podobieństwa)**.**
Niech τ będzie relacja podobieństwa pomiędzy obiektami z uniwersum Ω, $U \subseteq \Omega$ będzie zbiorem znanych obiektów a $Sim : U \times \Omega \to \mathbb{R}$ będzie funkcją. Dodatkowo dla dowolnego $\lambda \in \mathbb{R}$ określmy $\tau_{(\lambda)}^{Sim} = \{(u_1, u_2) \in U \times U : Sim(u_1, u_2) \geqslant \lambda\}$. Funkcję Sim nazwiemy właściwą funkcją podobieństwa dla τ na zbiorze obiektów U wtedy i tylko wtedy, gdy istnieją $\epsilon_1, \epsilon_2 \in \mathbb{R}$, $\epsilon_1 > \epsilon_2$, takie, że spełnione są oba poniższe warunki:

[1] W teorii zbiorów przybliżonych pojęcia utożsamia się ze zbiorami obiektów, które do nich należą.
[2] Oznacza to, że żaden z właściwych podzbiorów reduktu nie spełnia tej definicji.

1. $|\tau^{Sim}_{(\epsilon_1)}| > 0$ *oraz* $\tau^{Sim}_{(\epsilon_1)}$ *zawiera się w* τ*,*

2. $|(U \times U) \setminus \tau^{Sim}_{(\epsilon_2)}| > 0$ *oraz* $\tau^{Sim}_{(\epsilon_2)}$ *jest nadzbiorem* τ*.*

Każdy ze zbiorów $\tau^{Sim}_{(\lambda)}$ dla właściwej funkcji podobieństwa dla relacji τ może być traktowany jako pewne przybliżenie tej relacji wewnątrz zbioru $U \times U$. Pierwszy z warunków w Definicji 1 wymaga, by począwszy od pewnej liczby ϵ_1 wszystkie pary należące do przybliżeń definiowanych przez właściwą funkcję podobieństwa były prawdziwie w tej relacji. Własność ta implikuje, że w kontekście klasyfikacji i dla dostatecznie dużych λ obiekty z każdej pary w $\tau^{Sim}_{(\lambda)}$ posiadają tę samą wartość decyzji. Drugi warunek mówi, że istnieje pewna wartość graniczna ϵ_2, że dla $\lambda \leqslant \epsilon_2$ wszystkie pary podobnych obiektów są zawarte w zbiorze $\tau^{Sim}_{(\lambda)}$. Warunki te mogą być odczytywane jako pewna analogia do pojęcia zbioru przybliżonego, jako że zbiory $\tau^{Sim}_{(\epsilon_1)}$ i $\tau^{Sim}_{(\epsilon_1)}$ mogą być traktowane jako, odpowiednio, dolne i górne przybliżenie relacji podobieństwa τ.

Metody analizy danych i sztucznej inteligencji wykorzystują bardzo wiele funkcji mierzących podobieństwo. Podstawową i zarazem najczęściej używaną klasą miar podobieństwa są funkcje bazujące na metrykach odległości (ang. distance-based similarity functions). Rozprawa zawiera obszerny przegląd tego typu miar wraz ze wskazaniem ich najciekawszych własności oraz zastosowań. We wszystkich modelach opartych na metrykach odległości podobieństwo jest nierosnącą funkcją odległości pomiędzy wektorowymi reprezentacjami porównywanych obiektów. Ich cechą wspólną jest to, że generowane przez nie przybliżenia relacji podobieństwa posiadają szereg własności narzuconych przez metrykę. Na przykład dowolne przybliżenie relacji podobieństwa bazujące na funkcji odległości posiada własność symetrii, co jest sprzeczne z wynikami badań psychologów zajmujących się percepcją [8, 37]. Ponadto, jako że modele bazujące na odległości niejednokrotnie wybiera się w sposób niezależny od danych, nie są one w stanie brać pod uwagę jakiegokolwiek kontekstu ani różnic w istotności poszczególnych aspektów podobieństwa. Problem ten można częściowo rozwiązać poprzez rozdzielenie ewaluacji podobieństwa na poziomie lokalnym oraz globalnym (ang. the local-global principle).

Przykładem innego podejścia do zagadnienia mierzenia podobieństwa jest model kontrastu cech Tversky'ego [37]. W modelu tym obiekty nie są reprezentowane przez wektory wartości atrybutów, lecz przez zbiory ich wysokopoziomowych, niejednokrotnie abstrakcyjnych cech. Jako równanie modelu Tversky zaproponował:

$$Sim(x, y) = \theta f(X \cap Y) - \Big(\alpha f(Y \setminus X) + \beta f(X \setminus Y)\Big),$$

gdzie zbiory X i Y są binarnymi charakterystykami obiektów, f jest skalą interwałową (ang. an interval scale), a nieujemne stałe θ, α, β są parametrami modelu.

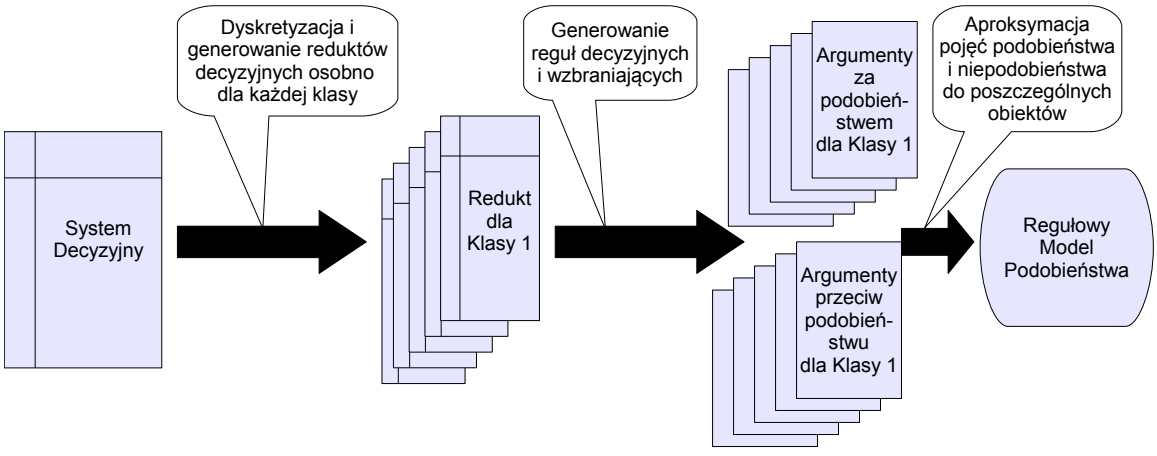
Wybór odpowiednich cech pozwala modelowi Tversky'ego uchwycić kontekst, w którym oceniane jest podobieństwo. Dodatkowo, w zależności od wartości parametrów θ, α, β przybliżenia podobieństwa uzyskane przy pomocy modelu kontrastu cech mogą posiadać różne własności, np. jeśli $\alpha \neq \beta$, to wynikowa relacja nie będzie symetryczna. W praktyce trudno jest jednak dobrze określić zbiór możliwych wysokopoziomowych cech obiektów, co istotnie ogranicza zastosowania tego modelu. Jedną z głównych motywacji Regułowego Modelu Podobieństwa jest chęć przezwyciężenia tego problemu.

Rozprawa zawiera również opis hierarchicznego modelu podobieństwa, w którym poszczególne aspekty podobieństwa połączone są w sieć [2, 10]. Struktura taka, zwana ontologią podobieństwa, wyznaczana jest dla konkretnych zastosowań przez ekspertów. W modelu tym, dla określenia stopni podobieństwa między parami obiektów, konieczne jest wyznaczenie osobnych klasyfikatorów dla każdego z aspektów. Wejściem dla klasyfikatorów odpowiadających aspektom znajdującym się wyżej w hierarchii są wyniki klasyfikatorów z niższych poziomów. Takie wielopoziomowe podejście do uczenia się podobieństwa jest

wyjątkowo elastyczne i pozwala automatycznie wyuczać się złożonej agregacji lokalnych podobieństw. Wadą tego podejścia jest złożoność obliczeniowa oraz potrzeba dużego zaangażowania ze strony ekspertów podczas manualnego etykietowania danych do nauki.

2.2 Proponowany model uczenia się podobieństwa z danych

Motywacją dla Regułowego Modelu Podobieństwa są obserwacje psychologów, którzy zauważyli, że relacja podobieństwa może posiadać inne własności niż modele bazujące na metrykach. W pracach takich jak [7, 37] podważono zasadność wszystkich własności cechujących podejście oparte o funkcje odległości, wliczając w to nawet zwrotność czy symetrię. Zauważono również, że z jednej strony podobieństwo powinno się oceniać na podstawie wysokopoziomowych cech, lecz z drugiej strony informacje o cechach tego typu rzadko są dostępne bezpośrednio w zbiorach danych [2, 9]. Z tego powodu konstrukcja modelu podobieństwa zaproponowanego w rozprawie zakłada etap automatycznego wykrywania wysokopoziomowych charakterystyk obiektów opisanych w systemie informacyjnym. Ważne jest aby proces ten odbywał się zgodnie z kontekstem dla oceny podobieństwa. Istotne cechy obiektów traktuje się jako rodzaj argumentów za lub przeciw podobieństwu porównywanych obiektów. Funkcja podobieństwa Regułowego Modelu Podobieństwa agreguje te argumenty w sposób analogiczny do modelu kontrastu cech Tversky'ego. Jednakże w przypadku zaproponowanego modelu wagi poszczególnych rodzajów argumentów nie muszą być nadawane przez eksperta, lecz są określane bezpośrednio na podstawie danych, z uwzględnieniem wpływu innych obiektów na kontekst dla oceny podobieństwa zadanej pary.



Rysunek 1: Schemat budowy Regułowego Modelu Podobieństwa.

Rysunek 1 przedstawia schemat budowy Regułowego Modelu Podobieństwa dla przypadku, w którym podobieństwo oceniane jest w kontekście problemu decyzyjnego. Wysokopoziomowe cechy obiektów w proponowanym modelu są wtedy definiowane przez lewe strony reguł decyzyjnych (ang. decision rules) i wzbraniających (ang. inhibitory rules). Reguły te generowane są z reduktów decyzyjnych wyliczonych z danych, osobno dla każdej klasy decyzyjnej. Zbiory cech wyznaczone przez reguły decyzyjne mogą stanowić argumenty za podobieństwem dwóch obiektów, a te wyznaczone przez reguły wzbraniające świadczą przeciwko podobieństwu. Dla i-tej klasy decyzyjnej będą one oznaczane przez $F^{+}_{(i)}$ i $F^{-}_{(i)}$, gdzie:

$$F^+_{(i)} = \Big\{\phi : \Big(\phi \to (d = i)\Big) \in RuleSet_i\Big\},$$
$$F^-_{(i)} = \Big\{\phi : \Big(\phi \to \neg(d = i)\Big) \in RuleSet_i\Big\},$$

gdzie ϕ to formuła języka L_A odpowiadające lewej stronie pewnej reguły. Zbiór odpowiadający znaczeniu ϕ w U oznaczać będziemy przez $\phi(U)$, a fakt posiadania cechy ϕ przez obiekt u odnotujemy poprzez $u \vDash \phi$.

W Regułowym Modelu Podobieństwa przybliżanie relacji podobieństwa odbywa się poprzez aproksymację pojęcia *bycia podobnym* do poszczególnych obiektów z danych. Z punktu widzenia teorii zbiorów przybliżonych tego typu pojęcia są dobrze określone. Przybliżenie pojęcia *bycia podobnym do obiektu* u można zdefiniować jako zbiór tych obiektów z U, które posiadają przynajmniej jedna cechę z $F^+_{(i)}$:

$$SIM_{(i)}(u) = \bigcup_{\phi \in F^+_{(i)} \wedge u \vDash \phi} \phi(U)$$

Jeśli reguły wykorzystane do wyznaczania zbioru F^+_i były pewne, to zbiór $SIM_{(i)}(u)$ odpowiada on dolnemu przybliżeniu tego pojęcia. Analogicznie możemy zdefiniować przybliżenie pojęcia *niepodobieństwa do* u:

$$DIS^0_{(i)}(u) = \bigcup_{\phi \in F^-_{(i)} \wedge u \nvDash \phi} \phi(U)$$

Dla wygody zdefiniujmy również zbiór obiektów posiadających przynajmniej jedną cechę z F^-_i, która jest wspólna z u:

$$DIS^1_{(i)}(u) = \bigcup_{\phi \in F^-_{(i)} \wedge u \vDash \phi} \phi(U)$$

Dodatkowo niech $SIM(u) = SIM_{d(u)}(u)$ oraz $DIS(u) = DIS^0_{d(u)}(u)$.

W Regułowym Modelu Podobieństwa stopień podobieństwa obiektu u_1 do u_2 mierzy się sprawdzając, na ile u_2 pasuje do pojęć *bycia podobnym* i *bycia niepodobnym* do u_1. Wykorzystuje się w tym celu dwie funkcje:

$$Similarity(u_1, u_2) = \frac{\Big|SIM(u_1) \cap SIM_{d(u_1)}(u_2)\Big|}{|SIM(u_1)| + C_{sim}},$$

$$Dissimilarity(u_1, u_2) = \frac{\Big|DIS(u_1) \cap DIS^1_{d(u_1)}(u_2)\Big|}{|DIS(u_1)| + C_{dis}}.$$

Dodatnie stałe C_{sim} oraz C_{dis} można traktować jako parametry modelu. W ogólnym przypadku są one konieczne by uniknąć dzielenia przez zero, lecz jeśli założymy, że wszystkie reguły wykorzystane przy konstrukcji modelu są pewne oraz pokrywają wszystkie obiekty z U, funkcje *Similarity* i *Dissimilarity* będą dobrze określone nawet dla $C_{sim} = C_{dis} = 0$.

Funkcja podobieństwa proponowanego modelu agreguje wartości funkcji *Similarity* i *Dissimilarity* dla danej pary obiektów. Można ją wyrazić jako:

$$Sim_{RBS}(u_1, u_2) = F\Big(Similarity(u_1, u_2), Dissimilarity(u_1, u_2)\Big) \tag{1}$$

gdzie $F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ jest dowolną funkcją monotonicznie rosnącą względem pierwszego argumentu (wartości *Similarity*) oraz monotonicznie malejącą ze względu na drugi argument (wartość *Dissimilarity*). W rozprawie omówiono szereg własności tak skonstruowanej funkcji podobieństwa oraz pokazano, że przy pewnych założeniach dotyczących reguł wykorzystywanych przy konstrukcji modelu funkcja Sim_{RBS} posiada własność z definicji 1 dla relacji podobieństwa w kontekście klasyfikacji.

Budowa Regułowego Modelu Podobieństwa często wymaga konstrukcji reduktu decyzyjnego ze zbiorów danych zawierających atrybuty numeryczne. W takim przypadku pojęcie reduktu musi zostać przedefiniowane. Najczęściej rozumie się je jako zbiór atrybutów wraz z cięciami zadanymi na zbiorach ich wartości tak, by po zamienieniu ich na przedziały nowo powstałe atrybuty symboliczne spełniały klasyczną definicję reduktu. Jednym z wyników rozprawy jest efektywny algorytm bazujący na [26], który pozwala generować tego typu redukty.

Algorytm tworzenia proponowanego modelu został zmodyfikowany w celu dopasowania go do problemów wymagających analizy danych wielowymiarowych (np. powyżej 1000 atrybutów). Powstały w ten sposób dwa rozszerzenia pierwotnego modelu. Pierwsze z nich zaprojektowano w celu uczenia się podobieństwa w kontekście klasyfikacji z danych, w których liczba atrybutów może kilkaset razy przekraczać liczbę dostępnych obiektów. Podejście to wykorzystuje pojęcie dynamicznego reduktu decyzyjnego [3] do konstruowania wielu zróżnicowanych zbiorów wysokopoziomowych cech reprezentujących różne aspekty podobieństwa. Lokalne podobieństwa względem poszczególnych aspektów mogą być agregowane w sposób analogiczny do agregacji drzew decyzyjnych w algorytmie Lasów Losowych (ang. Random Forest):

$$Sim_{DRBS}(u_1, u_2) = \frac{1}{N} \cdot \sum_{j=1}^{N} \Big(Sim^{(j)}_{RBS}(u_1, u_2) \Big), \tag{2}$$

W rozprawie zawarto propozycję algorytmu generowania wielu zróżnicowanych dynamicznych reduktów decyzyjnych.

Drugie rozszerzenie dla Regułowego Modelu Podobieństwa ma na celu umożliwienie uczenia się podobieństwa dokumentów tekstowych w kontekście wyznaczonym przez ich semantykę. Aby to umożliwić, dokumenty reprezentowane są przez zbiory pojęć związanych z ich tematyką. Tworzy się je przy pomocy algorytmu łączącego metodę semantycznego indeksowania [6, 22] oraz nowatorską technikę konstruowania bireduktów informacyjnych [20, 31].

Definicja 2 (Biredukt informacyjny)**.**
Niech $\mathbb{S} = (U, A)$ będzie systemem informacyjnym. Parę (B, X), gdzie $B \subseteq A$ a $X \subseteq U$, nazwiemy bireduktem informacyjnym wtedy i tylko wtedy, gdy B rozróżnia wszystkie pary obiektów z X i są spełnione następujące warunki:

1. *Nie istnieje żaden podzbiór $C \subsetneq B$, który rozróżnia wszystkie obiekty w X.*
2. *Nie istnieje żaden nadzbiór $Y \supsetneq X$, którego wszystkie obiekty są rozróżnione przez atrybuty z B.*

Wykorzystanie bireduktów informacyjnych sprawia, że na proces oceniania podobieństwa w proponowanym modelu można patrzeć jak na rodzaj interakcji pomiędzy agentami. Każdy z wygenerowanych bireduktów może odpowiadać jednemu agentowi, posiadającemu unikatowe doświadczenie (zbiór znanych przykładów) oraz własne preferencje (zbiór cech, przez pryzmat których agent patrzy na dane). W ten intuicyjny sposób można uchwycić różne punkty widzenia na semantykę porównywanych tekstów. W rozprawie przedstawiono efektywny algorytm generowania bireduktów informacyjnych.

2.3 Empiryczna ewaluacja zaproponowanych rozwiązań

Wszystkie algorytmy zaproponowane w rozprawie zostały gruntownie przetestowane w serii eksperymentów na różnych zbiorach danych. W testach z podobieństwem w kontekście klasyfikacji obiektów wykorzystano zarówno referencyjne tablice danych z repozytorium UCI (ang. UC Irvine Machine Learning Repository: http://archive.ics.uci.edu/ml/) jak i rzeczywiste zbiory mikromacierzy pobrane z repozytorium ArrayExpress

(http://www.ebi.ac.uk/arrayexpress). Eksperymenty z nienadzorowanym modelem uczenia się podobieństwa tekstów zostały przeprowadzone na zbiorze artykułów naukowych z dziedziny biomedycyny uzyskanych z repozytorium PubMed Central (http://www.ncbi.nlm.nih.gov/pmc).

Wszystkie eksperymenty opisane w rozprawie zostały przeprowadzone w środowisku Systemu R (http://www.r-project.org/). W czasie testów wykorzystano biblioteki *apriori*, *class*, *cluster*, *e1071*, *kknn*, *parallel* oraz *randomForest*. Dla wydajności część kodu została napisana bezpośrednio w języku C i wywoływana z poziomu R poprzez interfejs *.C*. Skrypty oraz zbiory danych pozwalające powtórzyć najważniejsze eksperymenty opisane w rozprawie są dostępne na życzenie.

Jakość testowanych modeli w kontekście klasyfikacji była weryfikowana poprzez ewaluację skuteczności algorytmu 1-NN, działającego z odpowiadającymi im funkcjami podobieństwa. Jako miarę skuteczności wykorzystano średnią dokładność (ang. accuracy) oraz średnią zrównoważoną dokładność (ang balanced accuracy). Skuteczność tę określano na podstawie wielokrotnej weryfikacji krzyżowej (ang. cross-validation) z podziałem na pięć lub dziesięć zbiorów (w zależności od zbioru danych). Istotność różnic w wynikach weryfikowano za pomocą testu statystycznego.

Pierwotną wersję Regułowego Modelu Podobieństwa porównano m.in. z metodą uczenia się podobieństwa wykorzystującą algorytm genetyczny do wyznaczania wag lokalnych podobieństw oraz z modelem łączącym podejście bazujące na metryce odległości z technikami selekcji istotnych atrybutów. O ile w testach na standardowych zbiorach danych z UCI duże różnice w wynikach (na korzyść Regułowego Modelu Podobieństwa) zanotowano jedynie dla zbiorów danych zawierających atrybuty symboliczne, to w wynikach dla wielowymiarowych danych mikromacierzowych widać było wyraźną przewagę zaproponowanego modelu.

Zaproponowane rozszerzenie przeznaczone dla danych wielowymiarowych zdecydowanie przewyższyło jakością wszystkie inne z testowanych modeli. Dodatkowo eksperymenty na 11 zbiorach danych mikromacierzowych pokazały, że model ten może z sukcesem konkurować nawet z takimi algorytmami klasyfikacji jak Lasy Losowe (ang. Random Forest), czy SVM.

Ewaluacja modelu podobieństwa tekstów polegała na sprawdzeniu jego przydatności w zadaniu nienadzorowanego grupowania artykułów z dziedziny biomedycyny. Do każdego dokumentu w wykorzystanym korpusie przypisano pojęcia (tematy główne) z ontologii MeSH (ang. Medical Subject Headings, http://www.nlm.nih.gov/mesh/) z wykorzystaniem własnej implementacji algorytmu ESA (ang. Explicit Semantic Analysis) [6]. Zaproponowany model porównano z klasycznym podejściem bazującym na mierze kosinusowej oraz dwoma innymi miarami podobieństwa tekstów. Jakość wyników grupowania oceniana była zarówno miarami wewnętrznymi [35], jak i poprzez sprawdzanie spójności tematycznej otrzymanych grup dokumentów. Wyniki zaproponowanego modelu dowiodły jego przydatności w ocenianiu podobieństwa tematycznego tekstów.

3 Podsumowanie

W rozprawie poruszony został problem uczenia się relacji podobieństwa z danych przechowywanych w systemach informacyjnych. Relacja ta z jednej strony powinna posiadać naturalne własności podobieństwa i odzwierciedlać ludzką percepcje [7, 8, 9, 37], a z drugiej powinna cechować się przydatnością w rozwiązywaniu praktycznych problemów związanych z analizą danych. Szczególną uwagę poświęcono sytuacji, w której porównywane obiekty posiadają bardzo dużą liczbę atrybutów. W takim przypadku często zawodzą najpowszechniej wykorzystywane modele bazujące na metrykach odległości [4, 37].

Zaproponowany został model podobieństwa, w którym tak jak w modelu kontrastu

cech Twersky'ego wektorowa reprezentacja obiektów została zamieniona na reprezentację przez zbiory wysokopoziomowych cech. W modelu tym, nazwanym Regułowym Modelem Podobieństwa, ocena stopnia podobieństwa między parą obiektów zależy nie tylko od samych wartości atrybutów, lecz także od kontekstu w jakim dokonuje się porównania. Kontekst ten zależy zarówno od zadania jakiemu służy ocena podobieństwa, jak i od innych obiektów opisanych w dostępnym w zbiorze danych. Ta szczególna cecha proponowanego modelu sprawia, że pozostaje on w zgodzie z obserwacjami poczynionymi przez wielu psychologów zajmujących się tym problemem [7, 8, 9, 37].

Regułowy Model Podobieństwa wykorzystuje pojęcia z teorii zbiorów przybliżonych zaproponowanej przez Zdzisława Pawlaka [27]. Na proces uczenia się podobieństwa w tym modelu można spojrzeć jak na konstruowanie przestrzeni przybliżeń, która jest odpowiednia dla przybliżania pojęcia *bycia podobnym* do obiektów dostępnych w systemie informacyjnym, w z góry zadanym kontekście. Podczas budowy modelu korzysta się z technik wyboru istotnych atrybutów, które od lat rozwijane są w ramach teorii zbiorów przybliżonych. Ponadto funkcja podobieństwa modelu skonstruowana jest przez analogię do *funkcji niepewności* oraz *funkcji zawierania*, znanych z klasycznej teorii Prof. Pawlaka [30].

W rozprawie opisane są trzy algorytmy budowania Regułowego Modelu Podobieństwa, przystosowane do działania z trzema różnymi typami zbiorów danych. Pierwszy z nich jest odpowiedni dla uczenia się podobieństwa w kontekście klasyfikacji z systemów decyzyjnych posiadających standardową liczbę atrybutów (nie większą niż kilkadziesiąt). Kolejny algorytm sprawdza się również dla danych o bardzo dużej liczbie atrybutów (tj. większej niż 1000). Ostatni z zaproponowanych algorytmów ma zastosowanie przy uczeniu się podobieństwa tekstów w kontekście ich znaczenia. Wszystkie z zaproponowanych rozwiązań zostały gruntownie przetestowane w serii eksperymentów na różnych zbiorach danych. Ich wyniki pokazują, że Regułowy Model Podobieństwa stanowi dobrą alternatywę dla klasycznych modeli. Dzięki jego intuicyjnym własnościom oraz skuteczności nie tylko jest on w stanie ułatwić uzyskiwanie dobrych wyników w zadaniach związanych z wspomaganiem podejmowania decyzji czy klasyfikacją, lecz może także zaoferować ekspertom zrozumiałe dla nich wyjaśnienia dokonywanych wyborów poprzez wskazanie motywujących je przykładów.

Literatura

[1] A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *Artificial Intelligence Communications*, 7(1):39–59, 1994.

[2] J. Bazan, P. Kruczek, S. Bazan-Socha, A. Skowron, and J. J. Pietrzyk. Automatic planning of treatment of infants with respiratory failure through rough set modeling. In *Proceedings of RSCTC 2006*, volume 4259 of *Lecture Notes in Artificial Intelligence*, pages 418–427, Berlin, 2006. Springer. see also the extended version in Fundamenta Informaticae 85, 2008.

[3] J. G. Bazan, A. Skowron, and P. Synak. Dynamic reducts as a tool for extracting laws from decisions tables. In *ISMIS '94: Proceedings of the 8th International Symposium on Methodologies for Intelligent Systems*, pages 346–355, London, UK, 1994. Springer-Verlag.

[4] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is ńearest neighbor"meaningful? In *In International Conference on Database Theory*, pages 217–235, 1999.

[5] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 539–546, Washington, DC, USA, 2005. IEEE Computer Society.

[6] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of The Twentieth International Joint Conference for Artificial Intelligence*, pages 1606–1611, Hyderabad, India, 2007.

[7] I. Gati and A. Tversky. Studies of similarity. In E. Rosch and B. Lloyd, editors, *Cognition and Categorization*, pages 81–99. L. Erlbaum Associates, Hillsdale, N.J., 1978.

[8] R. Goldstone, D. Medin, and D. Gentner. Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology*, 23:222–262, 1991.

[9] U. Hahn and N. Chater. Understanding similarity: A joint project for psychology, case based reasoning, and law. *Artificial Intelligence Review*, 12:393–427, 1998.

[10] A. Janusz. Similarity relation in classification problems. In C.-C. Chan, J. W. Grzymala-Busse, and W. P. Ziarko, editors, *Proceedings of RSCTC 2008*, volume 5306 of *Lecture Notes in Artificial Intelligence*, pages 211–222, Heidelberg, 2008. Springer.

[11] A. Janusz. Learning a Rule-Based Similarity: A comparison with the Genetic Approach. In *Proceedings of the Workshop on Concurrency, Specification and Programming (CS&P 2009), Kraków-Przegorzały, Poland, 28-30 September 2009*, volume 1, pages 241–252, 2009.

[12] A. Janusz. Rule-based similarity for classification. In *Proceedings of the WI/IAT 2009 Workshops, 15-18 September 2009, Milan, Italy*, pages 449–452, Los Alamitos, CA, USA, 2009. IEEE Computer Society.

[13] A. Janusz. Combining multiple classification or regression models using genetic algorithms. In M. S. Szczuka et al., editor, *Proceedings of RSCTC 2010*, volume 6086 of *Lecture Notes in Artificial Intelligence*, pages 130–137, Heidelberg, 2010. Springer.

[14] A. Janusz. Discovering rules-based similarity in microarray data. In *Proceedings of International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2010)*, Lecture Notes in Artificial Intelligence, pages 49–58, Berlin, Heidelberg, 2010. Springer-Verlag.

[15] A. Janusz. Utilization of dynamic reducts to improve performance of the rule-based similarity model for highly-dimensional data. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and International Conference on Intelligent Agent Technology - Workshops*, pages 432–435. IEEE, 2010.

[16] A. Janusz. Combining multiple predictive models using genetic algorithms. *Intelligent Data Analysis*, 16(5):763–776, 2012.

[17] A. Janusz. Dynamic Rule-Based Similarity model for DNA microarray data. *Lecture Notes in Computer Science Transactions on Rough Sets XV*, 7255:1–25, 2012.

[18] A. Janusz, H. S. Nguyen, D. Ślęzak, S. Stawicki, and A. Krasuski. JRS'2012 Data Mining Competition: Topical Classification of Biomedical Research Papers. In J.T. Yao et al., editor, *Proceedings of the 8th International Conference on Rough Sets and Current Trends in Computing (RSCTC 2012), Chengdu, China, August 17-20, 2012*, volume 7413 of *Lecture Notes in Artificial Intelligence*, pages 417–426. Springer, Heidelberg, 2012.

[19] A. Janusz and D. Ślęzak. Utilization of attribute clustering methods for scalable computation of reducts from high-dimensional data. In M. Ganzha, L. A. Maciaszek, and M. Paprzycki, editors, *Federated Conference on Computer Science and Information Systems - FedCSIS 2012, Wrocław, Poland, 9-12 September 2012, Proceedings*, pages 295–302, 2012.

[20] A. Janusz, D. Ślęzak, and H. S. Nguyen. Unsupervised similarity learning from textual data. *Fundamenta Informaticae*, 119(3).

[21] A. Janusz and S. Stawicki. Applications of approximate reducts to the feature selection problem. In *Proceedings of International Conference on Rough Sets and Knowledge Technology (RSKT)*, volume 6954 of *Lecture Notes in Artificial Intelligence*, pages 45–50. Springer, 2011.

[22] A. Janusz, W. Świeboda, A. Krasuski, and H. S. Nguyen. Interactive document indexing method based on explicit semantic analysis. In J.T. Yao et al., editor, *Proceedings of the 8th International Conference on Rough Sets and Current Trends in Computing (RSCTC 2012), Chengdu, China, August 17-20, 2012*, volume 7413 of *Lecture Notes in Artificial Intelligence*, pages 156–165. Springer, Heidelberg, 2012.

[23] K. Kurach, K. Pawłowski, Ł. Romaszko, M. Tatjewski, A. Janusz, and H. S. Nguyen. An ensemble approach to multi-label classification of textual data. In S. Zhou, S. Zhang, and G. Karypis, editors, *Advanced Data Mining and Applications*, volume 7713 of *Lecture Notes in Computer Science*, pages 306–317. Springer Berlin Heidelberg, 2012.

[24] M. Martín-Merino and J. Las Rivas. Improving k-nn for human cancer classification using the gene expression profiles. In *IDA '09: Proceedings of the 8th International Symposium on Intelligent Data Analysis*, pages 107–118, Berlin, Heidelberg, 2009. Springer-Verlag.

[25] T. M. Mitchell. *Machine Learning*. McGraw Hill series in computer science. McGraw-Hill, 1997.

[26] H. S. Nguyen. On efficient handling of continuous attributes in large data bases. *Fundamenta Informaticae*, 48(1):61–81, 2001.

[27] Z. Pawlak and A. Skowron. Rudiments of rough sets. *Information Sciences*, 177(1):3–27, 2007.

[28] S. Pinker. *How the mind works*. W. W. Norton, 1998.

[29] R. C. Schank. *Dynamic Memory: A Theory of Learning in Computers and People*. Cambridge University Press, New York, 1982.

[30] A. Skowron and J. Stepaniuk. Tolerance approximation spaces. *Fundamenta Informaticae*, 27(2/3):245–253, 1996.

[31] D. Ślęzak and A. Janusz. Ensembles of bireducts: Towards robust classification and simple representation. In T.-H. Kim, H. Adeli, D. Ślęzak, F. E. Sandnes, X. Song, K.-I. Chung, and K. P. Arnett, editors, *Future Generation Information Technology - Third International Conference, FGIT 2011, Jeju Island, Korea. Proceedings*, volume 7105 of *Lecture Notes in Computer Science*, pages 64–77. Springer, 2011.

[32] D. Ślęzak, A. Janusz, W. Świeboda, H. S. Nguyen, J. G. Bazan, and A. Skowron. Semantic analytics of PubMed content. In A. Holzinger and K.-M. Simonic, editors, *Information Quality in e-Health - 7th Conference of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2011, Graz, Austria. Proceedings*, volume 7058 of *Lecture Notes in Computer Science*, pages 63–74. Springer, 2011.

[33] A. Stahl and T. Gabel. Using evolution programs to learn local similarity measures. In *In Proceedings of the Fifth International Conference on Case-Based Reasoning*, pages 537–551. Springer, 2003.

[34] M. S. Szczuka, A. Janusz, and K. Herba. Clustering of rough set related documents with use of knowledge from DBpedia. In J. Yao, S. Ramanna, G. Wang, and Z. Suraj, editors, *Rough Sets and Knowledge Technology*, volume 6954 of *Lecture Notes in Computer Science*, pages 394–403. Springer Berlin/Heidelberg, 2011.

[35] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, Boston, 2006.

[36] P. Thagard. *Mind: Introduction to Cognitive Science*, chapter 10. MIT Press, Cambridge, Massachusetts, segunda edition, 2005.

[37] A. Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977.

[38] A. Wojna. *Analogy-based reasoning in classifier construction*. PhD thesis, Warsaw University, Faculty of Mathematics, Informatics and Mechanics, 2004.

[39] M. Wojnarski, A. Janusz, H. S. Nguyen, J. Bazan, C. Luo, Z. Chen, F. Hu, G. Wang, L. Guan, H. Luo, J. Gao, Y. Shen, V. Nikulin, T.-H. Huang, G. J. McLachlan, M. Bošnjak, and D. Gamberger. RSCTC'2010 discovery challenge: Mining DNA microarray data for medical diagnosis and treatment. In M. S. Szczuka et al., editor, *Proceedings of RSCTC 2010*, volume 6086 of *Lecture Notes in Artificial Intelligence*, pages 4–19, Heidelberg, 2010. Springer.

[40] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell. Distance metric learning with application to clustering with side-information. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada*, pages 505–512. MIT Press, 2002.