Metody zbiorów przybliżonych w uczeniu się podobieństwa z wielowymiarowych zbiorów danych

Andrzej Janusz

WMIM, Uniwersytet Warszawski
ul. Banacha 2, 02-097 Warszawa, Polska

andrzejanusz@gmail.com

13.06.2013

Dlaczego właśnie podobieństwo?

Myślenie...

i formowanie pojęć

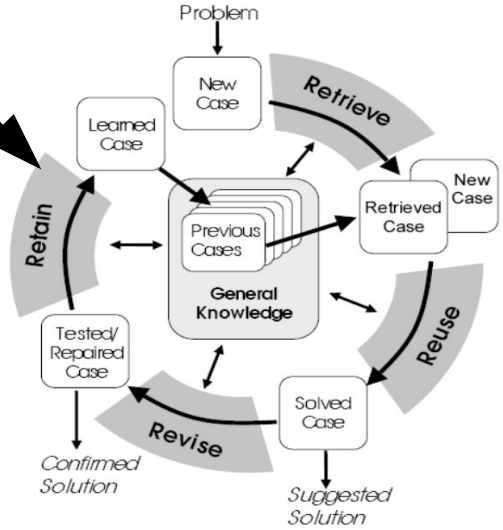
Podejmowanie decyzji

sampl eID	AFFX-3_at	3322_i_at	4969_s_at	...	22095_s_at	22379_at	Diagnosis
GSM1	4.010	12.434	32.443	...	1.665	12.44	3
GSM2	5.314	43.765	5.763	...	3.567	7.645	2
GSM3	3.275	17.567	23.842	...	0.657	12.46	2
GSM4	2.112	8.432	54.849	...	87.656	45.32	1
...	...	...	...	...	...	...	...
GSM14	8.453	10.087	8.678	...	2.986	9.656	3

Podobieństwo

Uczenie się

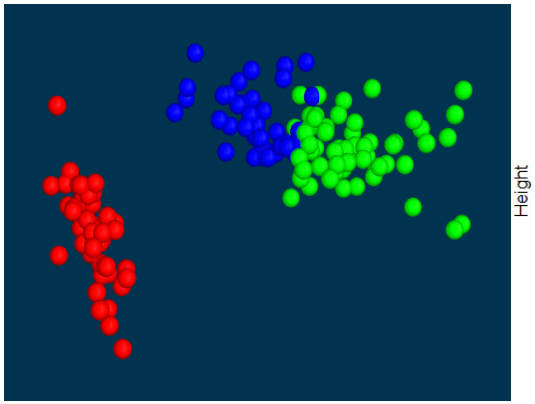
Rozwiązywanie problemów

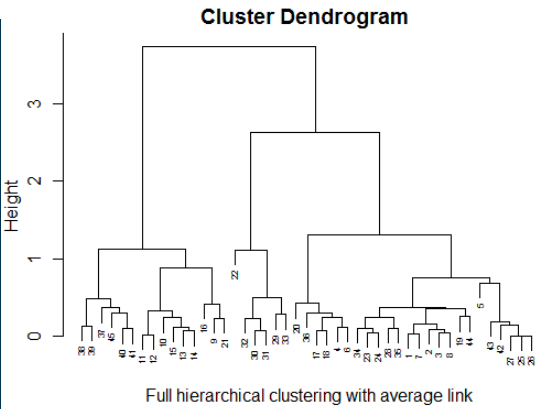


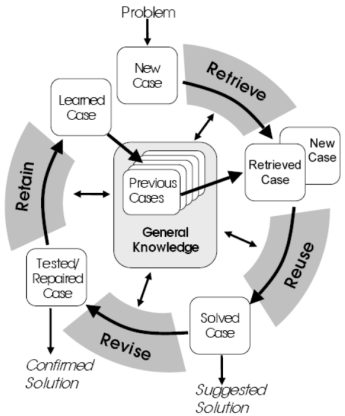
Zastosowania modeli podobieństwa

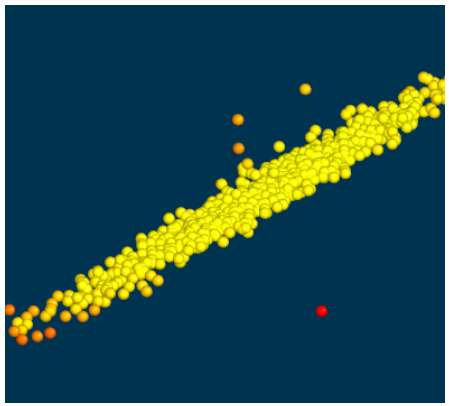
Przykłady:

- klasyfikacja i regresja,
- segmentacja danych,
- planowanie, rozwiązywanie problemów,
- wykrywanie nietypowych obiektów,
- wizualizacja i streszczanie danych.









Podstawowa zasada:

Podobne obiekty powinny być traktowane podobnie (np. należeć do tej samej klasy decyzyjnej, czy grupy).

Czym tak naprawdę jest podobieństwo?

Trudności ze ścisłą definicją podobieństwa:

- relacja, czy funkcja?
- obiektywne, czy subiektywne?
- bezkontekstowe, czy kontekstowe?
- globalne, czy lokalne?

Czynniki, które wpływają na kontekst to:

- cel lub zadanie, któremu służy ewaluacja podobieństwa,
- wiedza o innych znanych obiektach.

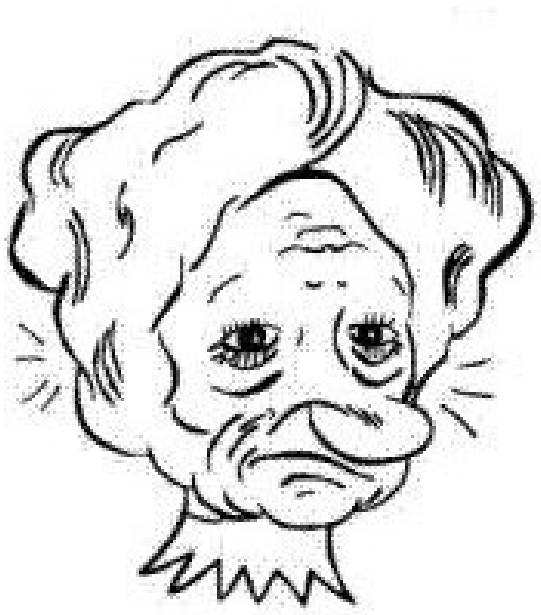
Czym tak naprawdę jest podobieństwo?

Trudności ze ścisłą definicją podobieństwa:

- relacja, czy funkcja?
- obiektywne, czy subiektywne?
- bezkontekstowe, czy kontekstowe?
- globalne, czy lokalne?

Czynniki, które wpływają na kontekst to:

- cel lub zadanie, któremu służy ewaluacja podobieństwa,
- wiedza o innych znanych obiektach.



Idea Tversky-ego

Model kontrastu cech:

- obiekty postrzegane są jako zbiory cech jakościowych,
- cechy są zazwyczaj na wyższym poziomie abstrakcji niż dane "sensoryczne", np.
  dwa samochody są podobne ponieważ są małe i szybkie,
- ważne są zarówno wspólne jak i wyróżniające cechy obiektów,
- $S(a,b) = \theta f(A \cap B) - \alpha f(A \setminus B) - \beta f(B \setminus A)$, gdzie $\theta, \alpha, \beta \geqslant 0$

Idea Tversky-ego

Model kontrastu cech:

- obiekty postrzegane są jako zbiory cech jakościowych,
- cechy są zazwyczaj na wyższym poziomie abstrakcji niż dane "sensoryczne", np.
  dwa samochody są podobne ponieważ są małe i szybkie,
- ważne są zarówno wspólne jak i wyróżniające cechy obiektów,
- $S(a,b) = \theta f(A \cap B) - \alpha f(A \setminus B) - \beta f(B \setminus A)$, gdzie $\theta, \alpha, \beta \geqslant 0$

Model Tversky-ego trudno jest zaaplikować do rzeczywistych danych:

- jak definiować wysokopoziomowe cechy?
- jak wybrać te istotne w danym kontekście?

Propozycja: można wykorzystać teorię zbiorów przybliżonych!

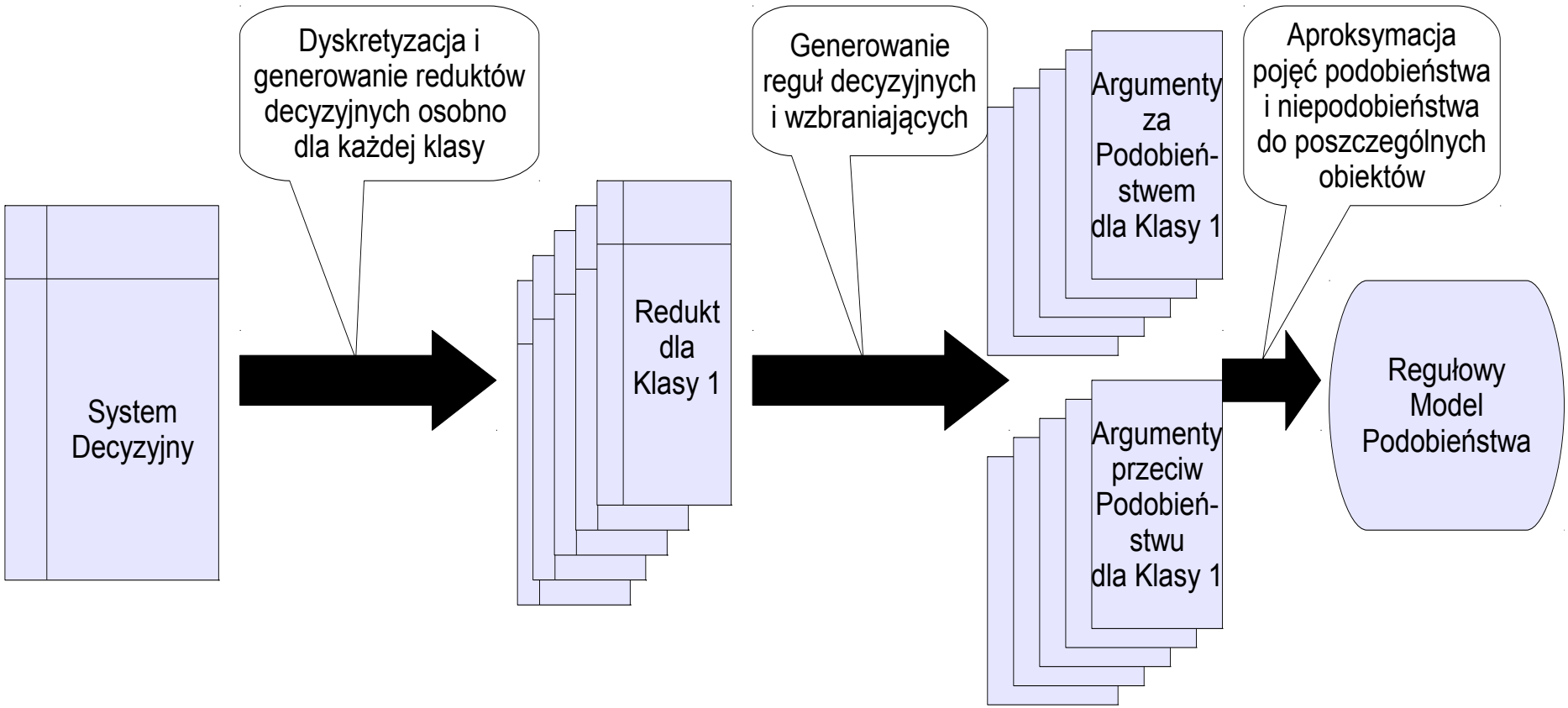
Założenia proponowanego modelu podobieństwa:

Uczenie się podobieństwa w języku zbiorów przybliżonych:

wybór istotnych aspektów podobieństwa	$\leftrightarrow$	wybór przestrzeni aproksymacji
wysokopoziomowe cechy	$\leftrightarrow$	lewe strony reguł
agregacja argumentów za i przeciw podobieństwu	$\leftrightarrow$	aproksymacja pojęć bycia podobnym i niepodobnym do obiektu
funkcja podobieństwa	$\leftrightarrow$	funkcja przynależności do aproksymacji pojęcia

- Wysokopoziomowe cechy można traktować jak argumenty za lub przeciw podobieństwu obiektów!
- Aproksymacja podobieństwa do obiektu to zbiór obiektów, do który pasują argumenty za podobieństwem a nie pasują argumety przeciwko.

Konstrukcja proponowanego modelu podobieństwa



Opis formalny modelu

Aproksymacja podobieństwa i niepodobieństwa:

$F^+_{(i)}$ oraz $F^-_{(i)}$ – zbiory cech dla i-tej klasy decyzyjnej, wyznaczone przez reguły decyzyjne i wzbraniające;

$$F^+_{(i)} = \{f : \Big(f \rightarrow (d = i)\Big) \in RuleSet(DR_i)\};$$

$$F^-_{(i)} = \{f : \Big(f \rightarrow (d \neq i)\Big) \in RuleSet(DR_i)\};$$

$$SIM_{(i)}(u) = \bigcup_{f \in F^+_{(i)} \wedge f(u)=1} [u]_f \quad \Big| \quad DIS^0_{(i)}(u) = \bigcup_{f \in F^-_{(i)} \wedge f(u)=0} U \backslash [u]_f \quad \Big| \quad DIS^1_{(i)}(u) = \bigcup_{f \in F^-_{(i)} \wedge f(u)=1} [u]_f$$

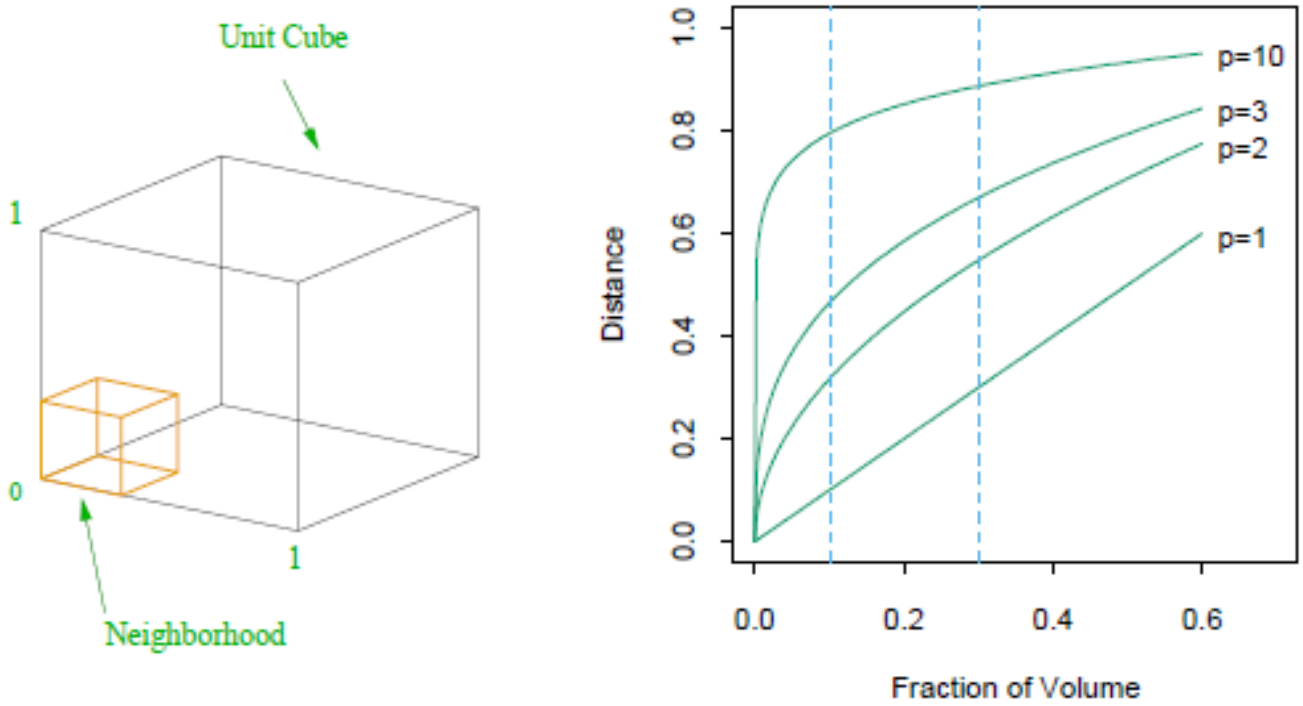
Przynależność do $SIM_{d(u_1)}(u_1)$:

$$\mu(u_2, SIM_i(u_1)) = \frac{|SIM_i(u_1) \cap SIM_i(u_2)|}{|SIM_i(u_1)|}$$

Przynależność do $DIS^0_{d(u_1)}(u_1)$:

$$\psi(u_2, DIS^0_i(u_1)) = \frac{\left|DIS^0_i(u_1) \cap DIS^1_i(u_2)\right|}{\left|DIS^0_i(u_1)\right|}$$

Dlaczego dane wielowymiarowe?



Rysunek: Ilustracja "przekleństwa wielu wymiarów" (z książki *Elements of Statistical Learning: Data Mining, Inference and Prediction*).

- typowe metody nie radzą sobie z problemem *niewielu obiektów o dużej liczbie cech*,
- duża złożoność obliczeniowa algorytmów uczenia się podobieństwa z danych wielowymiarowych.

Rozszerzenia modelu dla danych wielowymiarowych

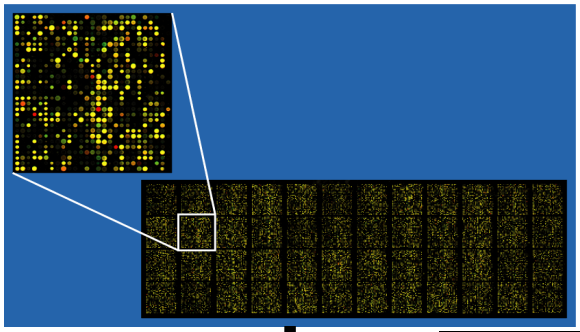
Główna idea:

W przypadku danych wielowymiarowych konieczne jest rozpatrywanie wielu lokalnych modeli podobieństwa, które można interpretować jako autonomicznych agentów z własnymi preferencjami i doświadczeniem.

Dwa typy wielowymiarowych danych

Dane mikromacierzowe:	Dane tekstowe:
uczenie z nadzorem	uczenie bez nadzoru
redukty dynamiczne	biredukty informacyjne
reguły decyzyjne i wzbraniające	pojęcia z ontologii dziedzinowej

Opis eksperymentów na danych mikromacierzowych



Microarray data:
few-objects-many-attributes problem

≈40k genes (attributes)

sampleID	AFFX-3_at	3322_i_at	4969_s_at	...	22095_s_at	22379_at	Diagnosis
GSM1.CEL	4.010	12.434	32.443	...	1.665	12.434	3
GSM2.CEL	5.314	43.765	5.763	...	3.567	7.645	2
GSM3.CEL	3.275	17.567	23.842	...	0.657	12.446	2
GSM4.CEL	2.112	8.432	54.849	...	87.656	45.324	1
...	...	...	...	...	...	...	...
GSM149.CEL	8.453	10.087	8.678	...	2.986	9.656	3

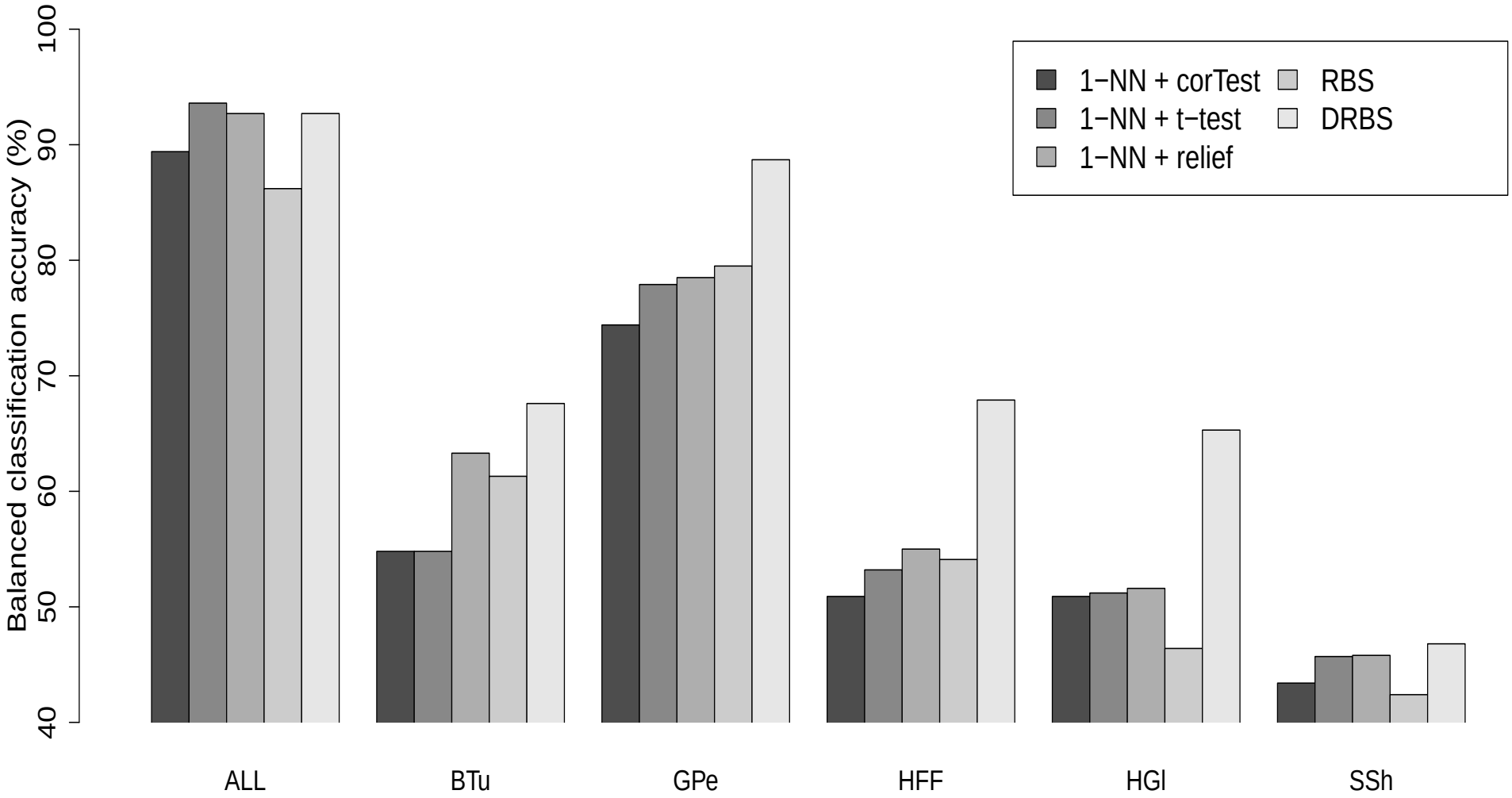
Opis danych

- 11 zbiorów mikromacierzy,
- liczba obiektów: 124 – 284,
- liczba atrybutów: $22k$ – $61k$,
- zbiory pochodzą z repozytorium ArrayExpress.

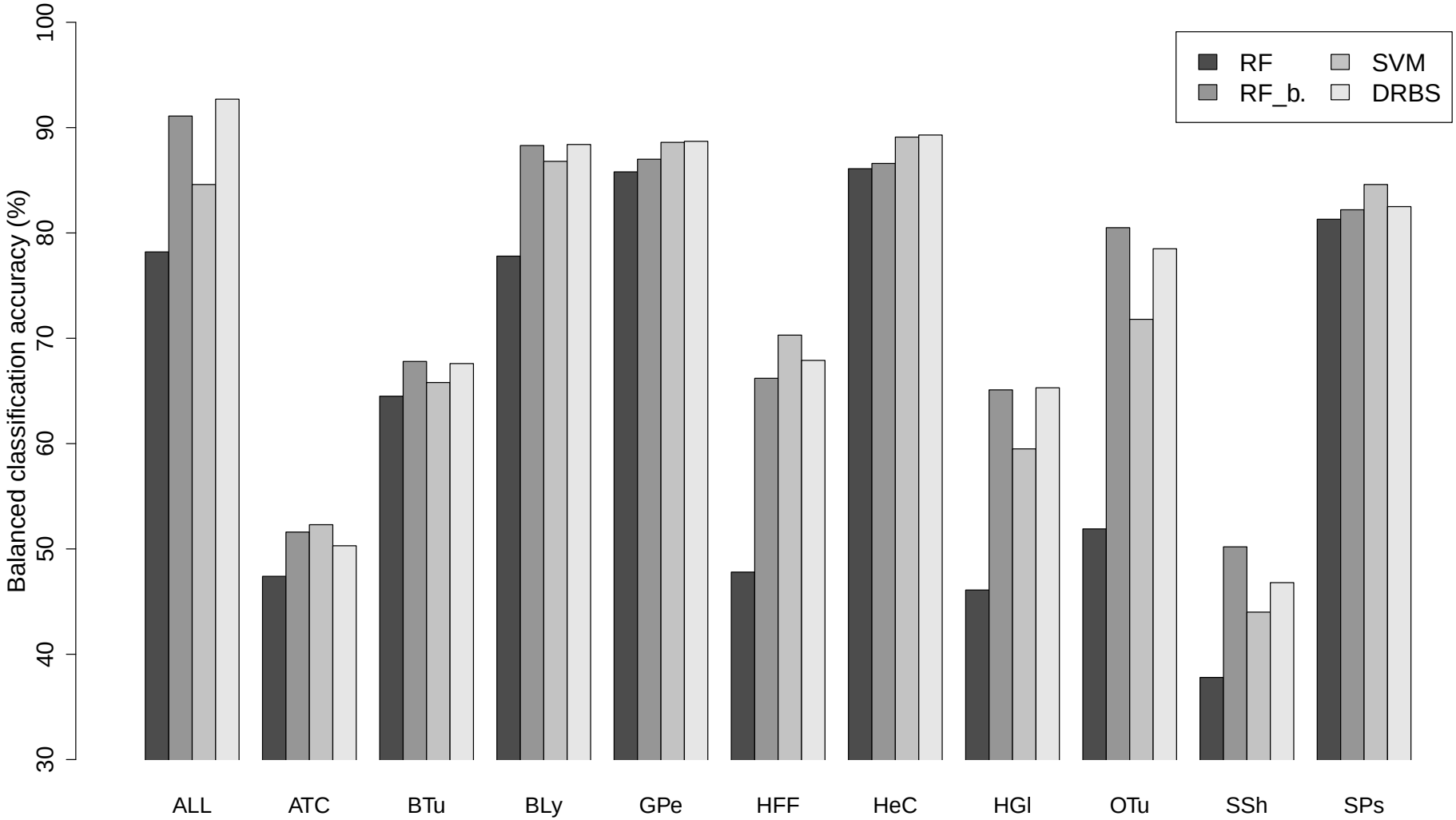
Opis eksperymentu

- wielokrotnie powtarzana weryfikacja krzyżowa,
- miary jakości: ACC i BAC,
- porównywane klasyfikatory: *k-NN**, *RF*, *SVM*.

Wyniki porównania z wybranymi modelami podobieństwa



Wyniki porównania z wybranymi metodami klasyfikacji



Opis eksperymentów na danych tekstowych

Opis danych

- zbior 1000 artykułów naukowych z repozytorium PubMed Central,
- ontologia dziedzinowa MeSH ($\approx 26k$ pojęć),
- metoda automatycznego etykietowania: ESA,
- zbiory etykiet nadanych przez ekspertów.

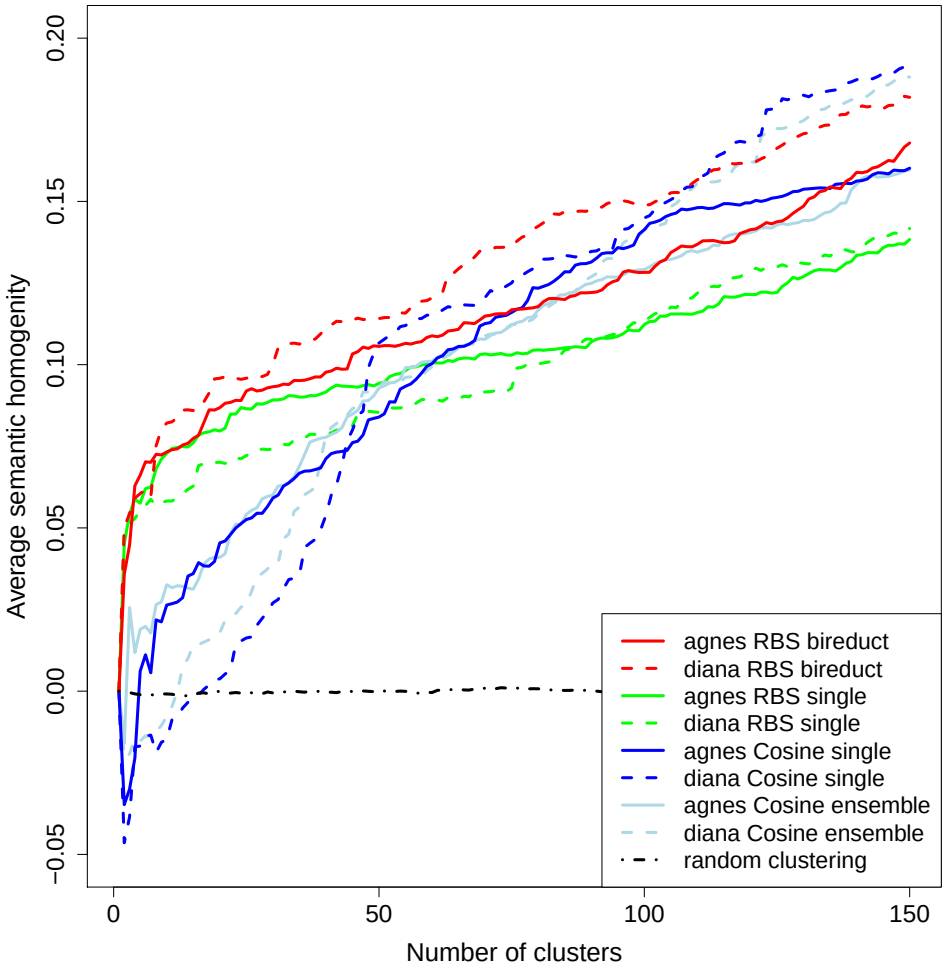
Opis eksperymentu

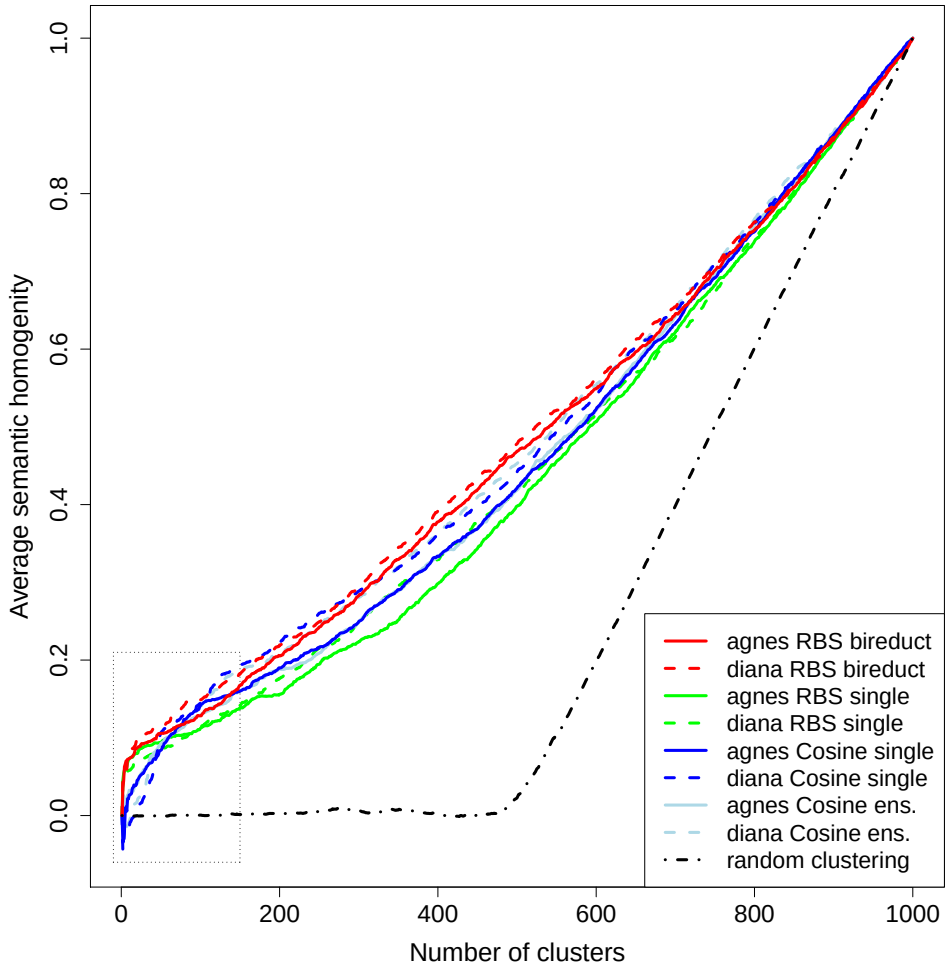
- grupowanie hierarchiczne artykułów,
- stosowane algorytmy: *agnes* i *diana*,
- porównywane modele: dwa oparte o miarę kosinusową,
- zewnętrzna miara oceny jakości grupowania.

Ewaluacja wyników

Miara zgodności etykiet nadanych przez ekspertów wewnątrz grup.

Wyniki ewaluacji modelu





Podsumowanie

Co się udało?

- dokonano interpretacji problemu uczenia się podobieństwa z punktu widzenia teorii zbiorów przybliżonych,
- zaproponowano intuicyjny i elastyczny model uczenia się podobieństwa z danych,
- opracowano efektywne algorytmy działające dla wielowymiarowych zbiorów danych,
- przeprowadzono dokładną ewaluację zaproponowanego podejścia.

Kierunki na przyszłość:

- lepsze wykorzystanie wiedzy dziedzinowej,
- optymalizacja wydajności obliczeniowej dla dużych zbiorów danych,
- stworzenie wysokopoziomowego środowiska do eksperymentów.

Dziękuję za uwagę!