Grupowanie i klasyfikacja dużych zbiorów danych z wykorzystaniem wzorców częstych

prof. dr hab. inż. Tadeusz Morzy
dr inż. Maciej Piernik

Instytut Informatyki
Politechnika Poznańska

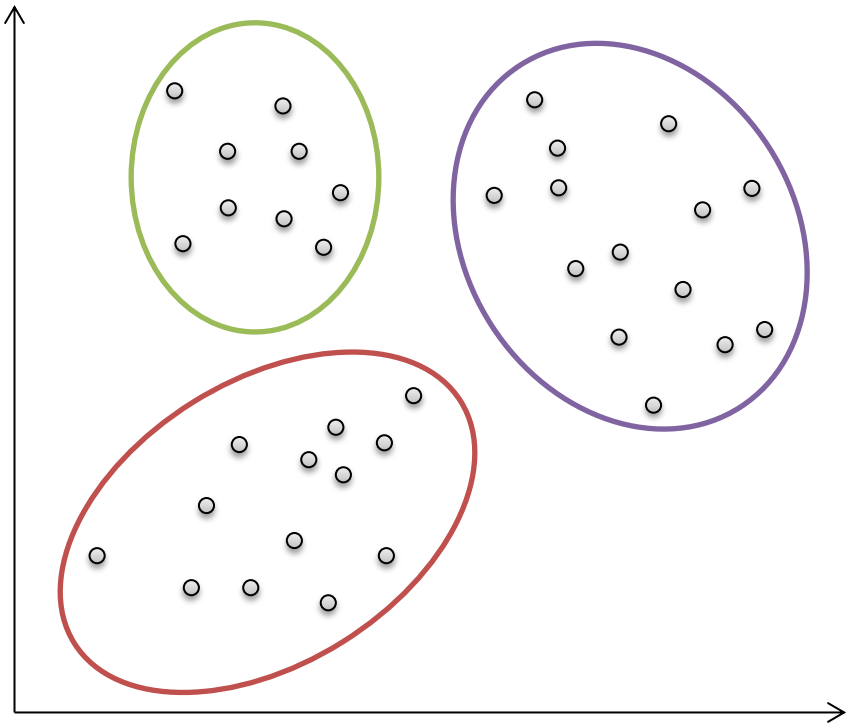
Seminarium naukowe
Big Data: przetwarzanie i eksploracja
Poznań, 22.04.2016

Plan prezentacji

- Grupowanie przez wzorce
- Klasyfikacja przez wzorce
- Wzorce w danych strumieniowych

Grupowanie

Automatyczny podział zbioru obiektów na podzbiory zgodnie z ich podobieństwem.

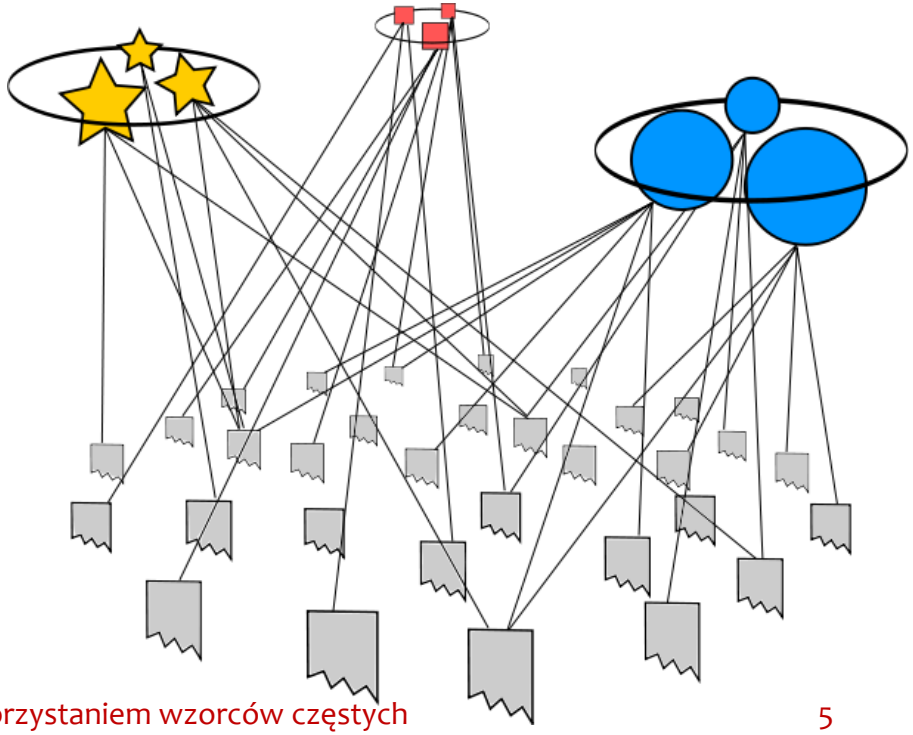


Problemy badawcze

- Brak formalnej metodyki grupowania wykorzystującej informację globalną
- Brak gwarancji interpretowalności wyników
- Parametryzacja

Proponowane rozwiązanie

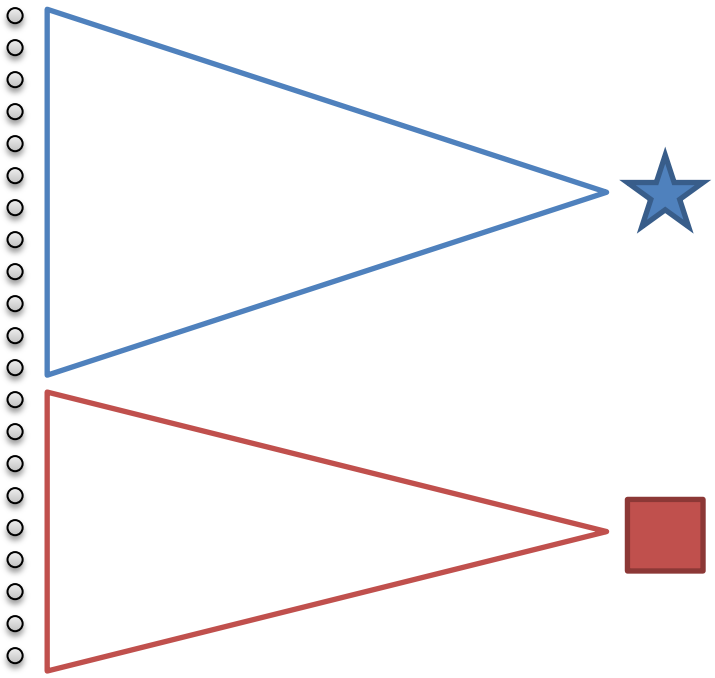
- Framework XPattern
  1. Transformacja danych
  2. Odkrywanie wzorców
  3. Grupowanie wzorców
  4. Przypisanie obiektów



Porównanie różnych definicji wzorców

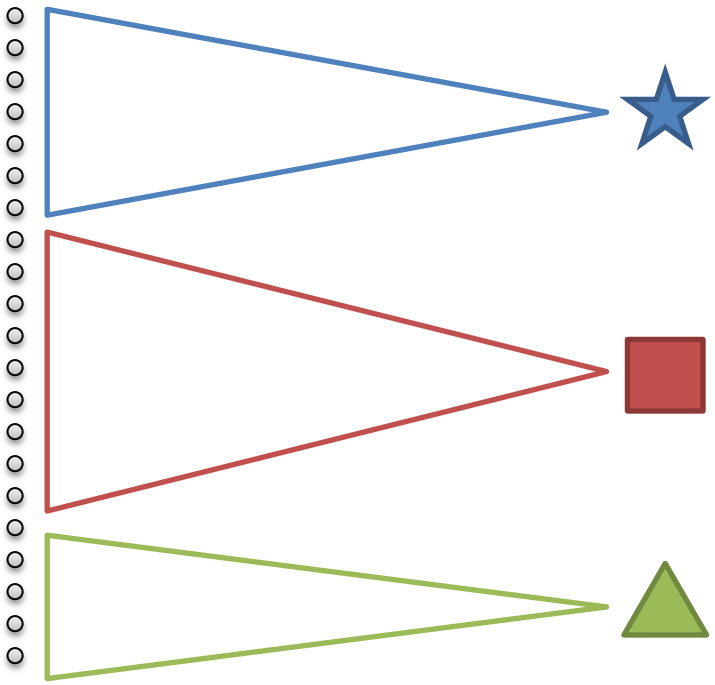
Zbiór danych	sig	het	hom	db0	db1	db2	db3
Wzorzec				Precision			
Poddrzewo	**1.00**	**1.00**	0.90	-	-	-	-
Podścieżka	**1.00**	**1.00**	**0.92**	0.66	**0.71**	**0.66**	**0.64**
Etykieta	0.51	**1.00**	0.35	**0.73**	0.69	0.45	0.44
Metadane	0.98	0.45	0.36	0.22	0.18	0.15	0.17
Wzorzec				Czas [s]			
Poddrzewo	0.07	1.40	3.24	-	-	-	-
Podścieżka	1.57	0.09	0.17	65.25	137.76	311.64	403.38
Etykieta	0.08	1.06	0.05	21.44	25.62	42.66	44.07
Metadane	**0.02**	**0.03**	**0.01**	**0.18**	**0.24**	**0.40**	**0.38**

Automatyczna detekcja liczby skupień



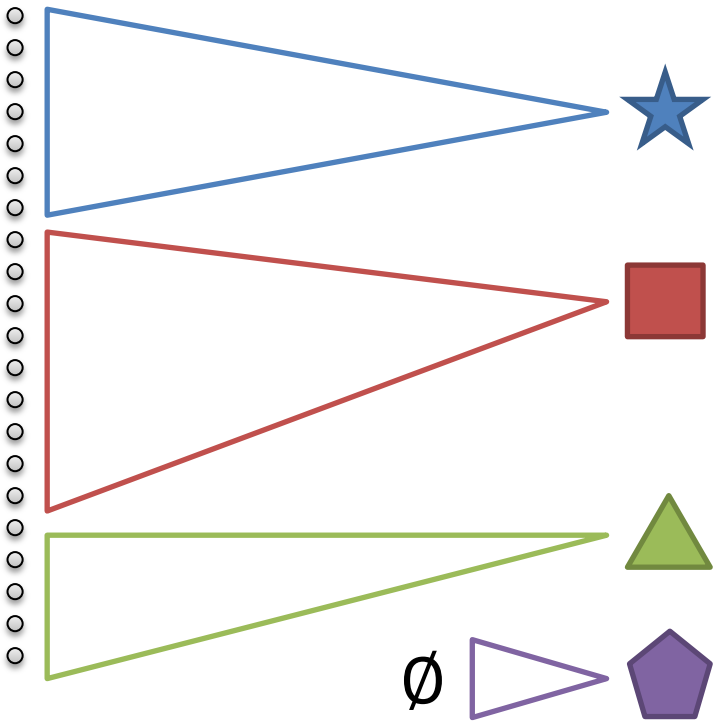
k = 2

Automatyczna detekcja liczby skupień



k = 3

Automatyczna detekcja liczby skupień



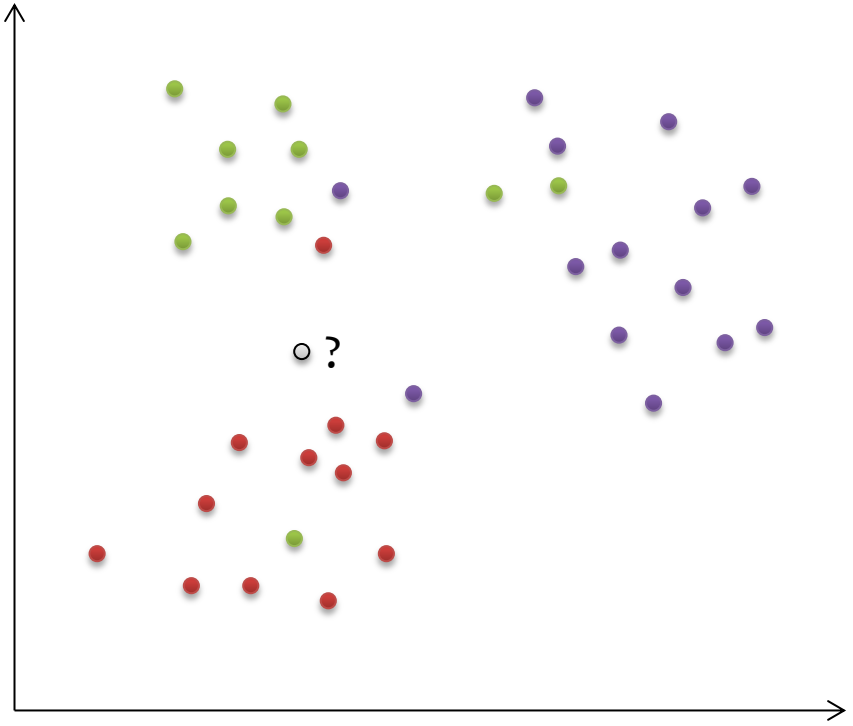
k = 4

Automatyczna detekcja liczby skupień – ocena

Zbiór danych	sig	het	hom	db0	db1	db2	db3
Algorytm				Liczba skupień			
PathXP	2	10	3	11	11	11	11
*PathXP**	2	11	9	22	23	14	18
				Czas [s]			
PathXP	2	<1	<1	65	138	312	403
*PathXP**	1	19	1	1538	3176	3647	5999

Klasyfikacja

Predykcja klas obiektów za pomocą klasyfikatora skonstruowanego w procesie uczenia na wcześniej sklasyfikowanych przykładach.



Dodatkowy problem badawczy

Brak dopasowania nowych obiektów do odkrytych wzorców

Proponowane rozwiązanie

- Algorytm K-Nearest Patterns
  - Trening
    - Odkrycie wzorców częstych w każdej klasie
  - Klasyfikacja
    - Przypisanie obiektu do klasy na podstawie głosowania większościowego k najbliższych wzorców

Wyniki

Algorytm	*Klasyczne reguły*	*KNP*
Zbiór danych	Accuracy [%]	
DS1	64.54 [63%]	**73.25**
DS2	79.77 [58%]	**81.64**
DS3	56.77 [63%]	**66.35**
DS4	60.32 [66%]	**64.04**
CS1-2	**80.37** [50%]	80.33
CS2-3	79.67 [61%]	**79.72**
CS3-1	79.16 [63%]	**79.22**
CS12-3	79.33 [58%]	**79.40**

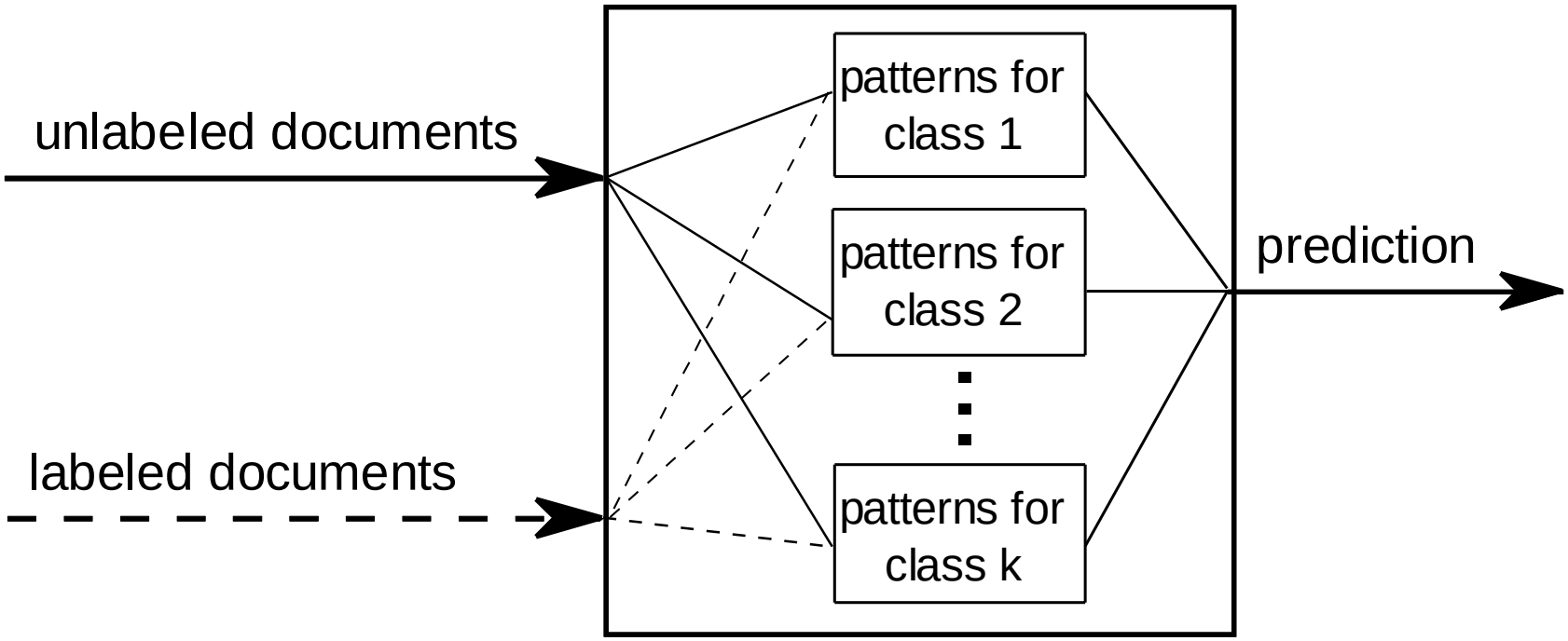
[] – wartości w nawiasach przedstawiają procent dokumentów sklasyfikowanych przy pomocy reguły domyślnej

Strumienie danych

- Ciągły napływ obiektów
- Ograniczony czas
- Ograniczona pamięć
- Reakcja na zmiany

Proponowane rozwiązanie

Przyrostowe, blokowe odkrywanie wzorców częstych i równoległa klasyfikacja nowych przykładów



Wyniki

- Wysoka średnia jakość klasyfikacji
- Model zajmuje mało miejsca
- Zdolny do reakcji na różne typy zmian
- Dłuższy czas przetwarzania

Podsumowanie

- Łatwe do interpretacji wyniki
- Elastyczny klasyfikator
- Klasyfikator strumieniowy reagujący na zmiany

Publikacje

- M. Piernik, D. Brzezinski, T. Morzy, **Clustering XML Documents by Patterns**, *Knowledge and Information Systems*, vol. 46, no. 1, pp. 185-212, 2016.
- D Brzezinski, M Piernik, **Structural XML Classification in Concept Drifting Data Streams**, *New Generation Computing*, vol. 33, no. 4, pp. 345-366, 2015.
- M. Piernik, D. Brzezinski, T. Morzy, A. Lesniewska, **XML Clustering: A Review of Structural Approaches**, *Knowledge Engineering Review*, vol. 30, no. 3, pp. 297-323, 2015.
- D. Brzeziński, M. Piernik, **Adaptive XML Stream Classification Using Partial Tree-Edit Distance**, *Proceedings of 21st International Symposium on Methodologies for Intelligent Systems*, ISMIS 2014, Roskilde, Denmark, June 25-27, 2014.
- D. Brzezinski, A. Lesniewska, T. Morzy, M. Piernik, **XCleaner: A New Method for Clustering XML Documents by Structure**, *Control and Cybernetics*, vol. 40(3), pp. 877-89, 2011.
- D. Brzezinski, A. Lesniewska, T. Morzy, M. Piernik, **Clustering XML Documents by Patterns**, *Proceedings of III Polish National Conference on Data Processing Technologies*, KKNTPD 2010, Poznan, Poland, 2010.

Dziękuję za uwagę