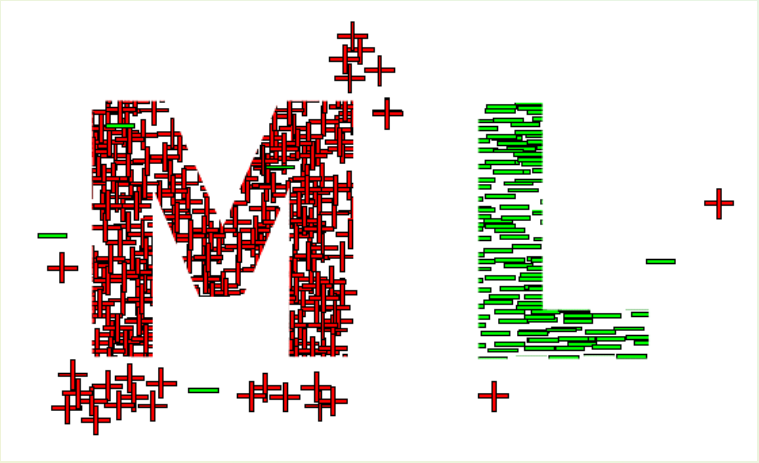
Badania w zakresie systemów uczących się w Zakładzie ISWD



Politechnika Poznańska
Instytut Informatyki

Seminarium ML - Poznań, 3 04 2013

Informacje ogólne

- Politechnika Poznańska
- Wydział Informatyki, Instytut Informatyki
  - Zakład ISWD (jeden z 4 zakładów Inst.)
  - Ponadto grupy:
    - Semantic Data and Web Mining (A.Ławrynowicz)
    - Zespół baz danych (T.Morzy)

Modele regułowe i wspomaganie decyzji

R.Słowiński, J.Błaszczyński, K.Dembczyński, M.Kadziński, W.Kotłowski, R.Susmaga, I.Szczęch, M.Szeląg, Sz.Wilk, P.Zielniewicz

- Główne obszary badań
  - Modelowanie różnych typów niedoskonałości informacji → „rough set & fuzzy set theories”
  - Dominacyjna teoria zbiorów przybliżonych (DRSA)
  - Uczenie się preferencji z danych porządkowych
  - Klasyfikacja porządkowa, wieloetykietowa i ranking wieloatrybutowy
  - Indukcja reguł decyzyjnych dla klasyfikacji i rankingu oraz ocena jakości reguł
  - Wspomaganie decyzji w warunkach ryzyka i niepewności za pomocą modeli decyzyjnych odkrytych z danych
  - Aksjomatyka modeli decyzyjnych
  - …

Indukcja reguł i inne kierunki

Jerzy Stefanowski i inni:

Reguły

- Algorytmy indukcji reguł (MODLEM, EXPLORE, …)
- Strategie klasyfikacyjne dla zbiorów reguł (nearest rules)
- Uwzględnianie nieznanych wartości atrybutów o różnej semantyce (JS i Alexis Tsoukias 2001-2004)
- Wykorzystanie eksperckich uzasadnień (argument based rule learning – ABMODLEM z K.Napierałą)
- Porządkowe reguły asocjacyjne (JS z S.Greco, R.Słowiński)
- Ocena ważności warunków w regułach (JS z S.Greco, R.Słowiński)

Grupowanie danych tekstowych (JS z Dawid Weiss)

– Description Comes First -> algorytm LINGO (także z S.Osiński)

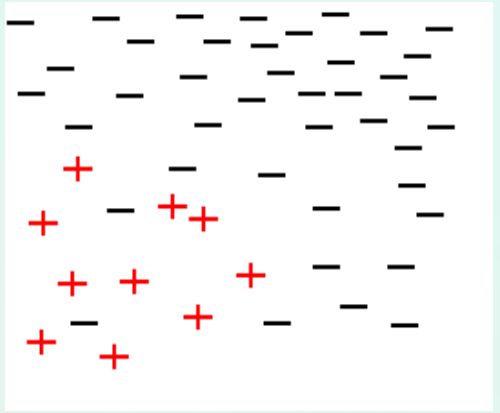
- System Carrot (grupowanie wyników wyszukiwań w WWW)

Złożone klasyfikatory

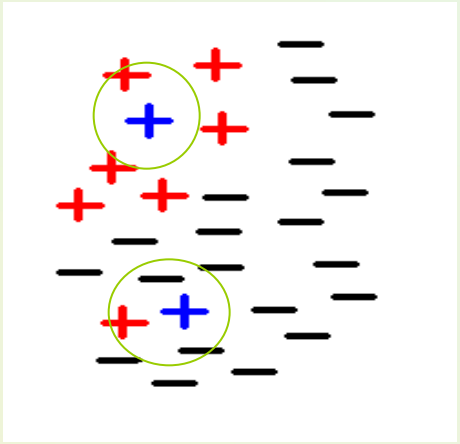
- Klasyfikator n2 dla problemów wieloklasowych (Jelonek, Stefanowski, ECML98)
- Rozszerzenia doboru podzbiorów cech BagFS (JS 2007)
- Rodziny reguł: reguła jako bazowy klasyfikator w boostingu, MLRules, ENDER (Dembczyński, Kotłowski, Słowiński 2008, DAMI 2010), LPRules (Kotłowski, Słowiński, ICML 2009)
- Uogólnienia baggingu ze modyfikacją metod agregacji (J.Stefanowski 2006)
- Abstaining classifiers (Błaszczyński et al. ISMIS 2009)
- Variable consistency sampling dla baggingu (Błaszczyński, Stefanowski ECML 2009)

Uczenie się z niezrównoważonych danych

Jerzy Stefanowski, Krystyna Napierała, Szymon Wilk



- Badania nad naturą problemu
  - Identyfikacja typów rozkładów danych (2010-12)
- Metody przetwarzania wstępnego
  - Metoda SPIDER (ECML 2007)
  - Uogólnienia metod SMOTE → LN-SMOTE (2011) oraz SMOTE-IPF (2012)



Dataset	Base	Oversampling	Filtr Japkowicz	NCR	SPIDER
subclus-0	0.9540	0.9500	0.9500	0.9460	**0.9640**
subclus-30	0.4500	0.6840	0.6720	0.7160	**0.7720**
subclus-50	0.1740	0.6160	0.6000	0.7020	**0.7700**
subclus-70	0.0000	0.6380	0.7000	0.5700	**0.8300**
clover-0	0.4280	0.8340	**0.8700**	0.4300	0.4860
clover-30	0.1260	0.7180	0.7060	0.5820	**0.7260**
clover-50	0.0540	0.6560	0.6960	0.4460	**0.7700**
clover-70	0.0080	0.6340	0.6320	0.5460	**0.8140**
paw-0	0.5200	**0.9140**	0.9000	0.4900	0.5960
paw-30	0.2640	0.7920	0.7960	0.8540	**0.8680**
paw-50	0.1840	0.7480	0.7200	0.8040	**0.8320**
paw-70	0.0060	0.7120	0.6800	0.7460	**0.8780**

Klasyfikatory z niezrównoważonych danych

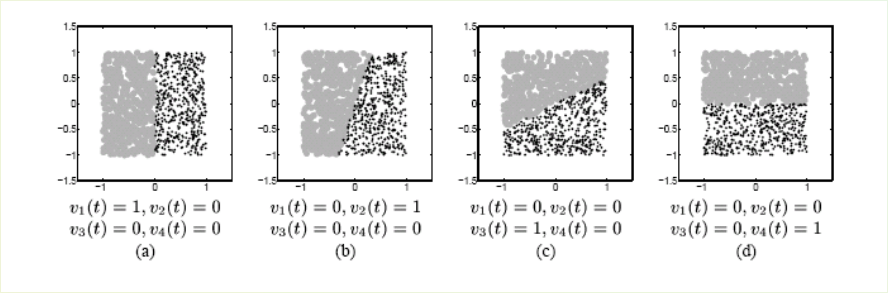
- Poprawa klasyfikatorów regułowych
  - Modyfikacje indukcji reguł i strategii klasyfikacyjnej (Grzymała, Stefanowski, Wilk 2004)
  - Specjalizowany algorytm BRACID (Napierała, Stefanowski 2011)
  - Wykorzystanie argumentacji trudnych przykładów ABMODLEM (Napierała 2010)
- Uogólnienia klasyfikatorów złożonych
  - IIvotes – modyfikacja losowania ważnościowego ze SPIDERem (Błaszczyński, Deckert, Stefanowski, Wilk 2010)
  - Local vs All over bagging (Błaszczyński, Stefanowski 2013)

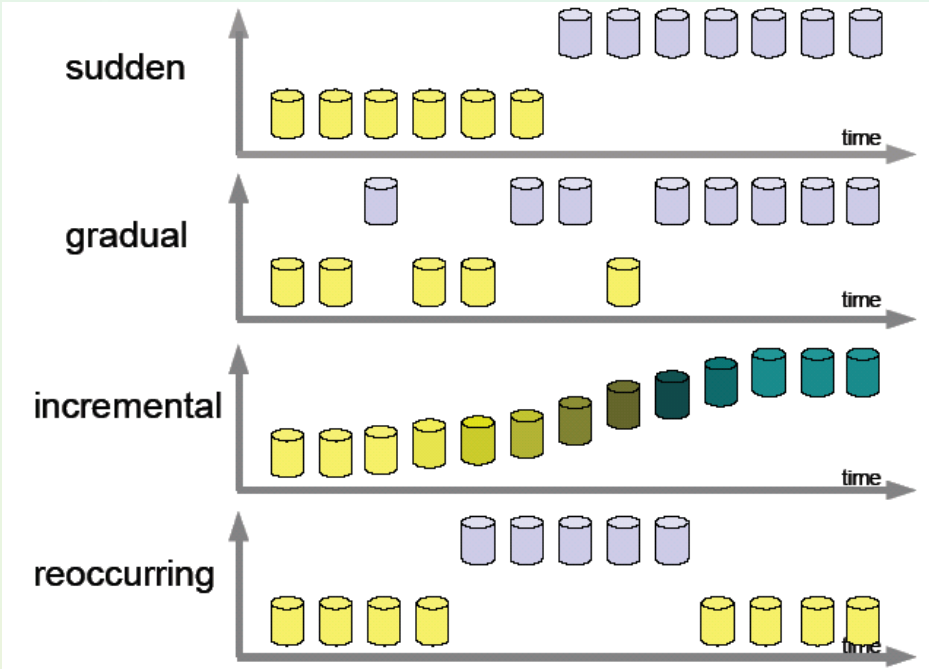
Uczenie się ze zmiennych strumieni danych

Jerzy Stefanowski, Dariusz Brzeziński, Magdalena Deckert

- Aspekt czasu → przetwarzanie przyrostowe
- Wymagania wydajnościowe
  - potencjalnie duże rozmiary danych, szybkość przetwarzania, jednokrotny dostęp
- Zmienność danych → ang. „concept drift”

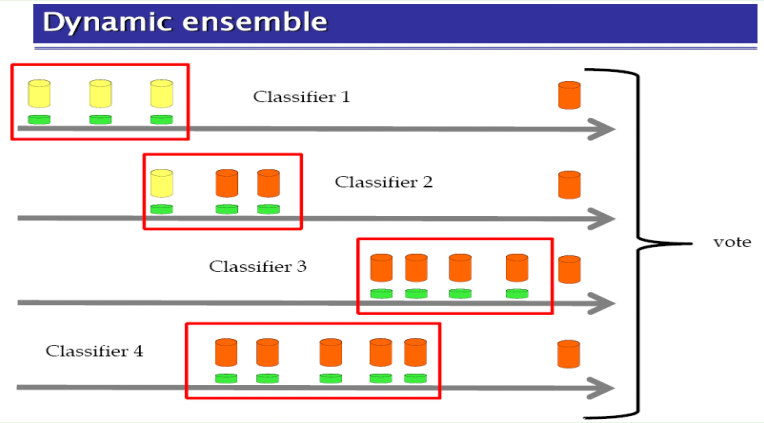
- „Concept drift means that the concept about which data is obtained may shift from time to time, each time after some minimum permanence”





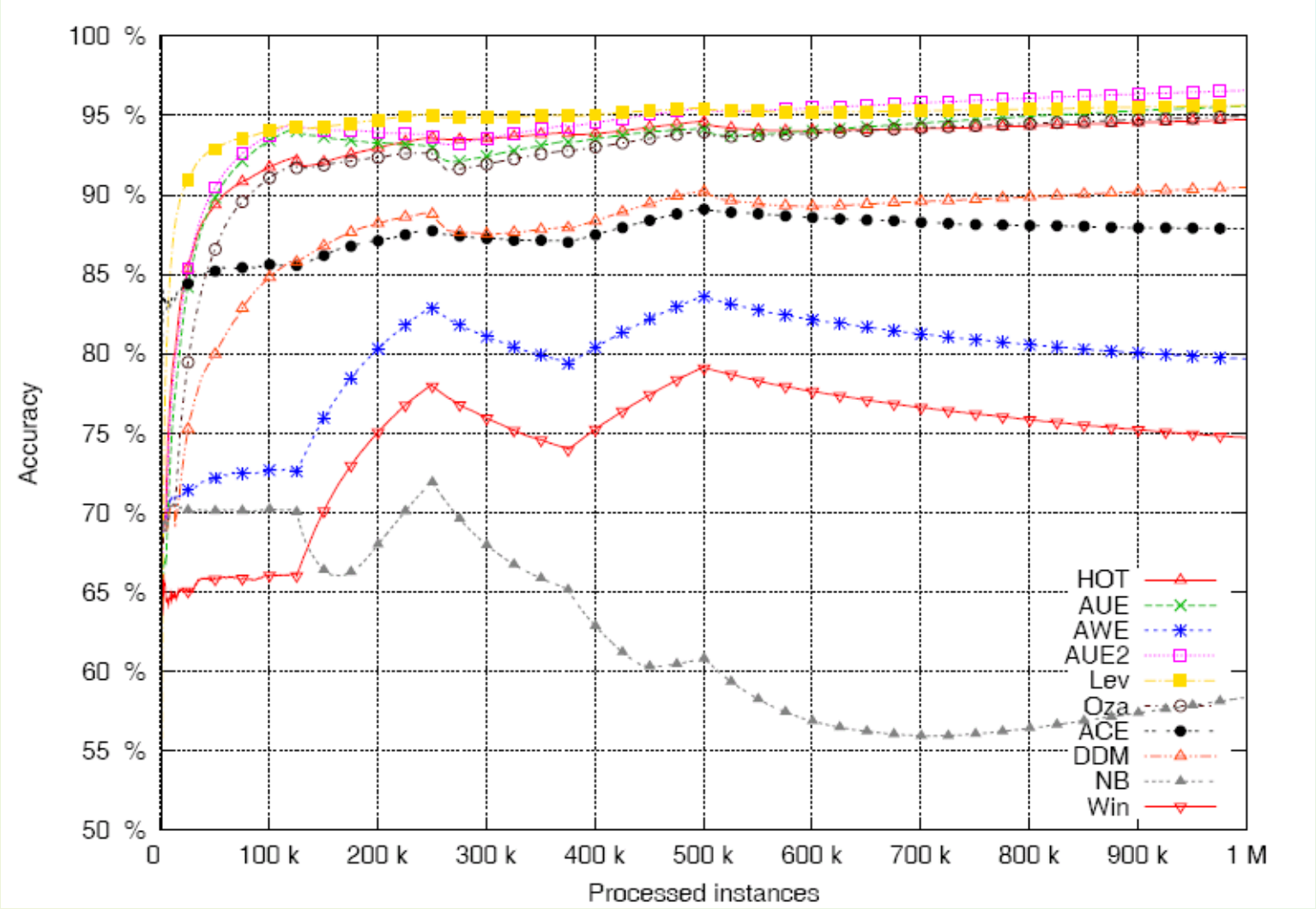
Uczenie się ze zmiennych strumieni danych

- Metody detekcji zmian i ich wykorzystanie
  - Dane częściowo etykietowane → obserwacja zmian modelu (M.Kmieciak, J.Stefanowski 2010)
  - Wykorzystanie detektorów w modyfikacji struktury blokowego klasyfikatora BDWM (M.Deckert 2011)
- Specjalizowane klasyfikatory złożone
  - Accuracy Updated Ensemble (D.Brzeziński 2011-2012)
  - Wersja w pełni przyrostowa OAUE (D.Brzeziński 2013)



Przykład porównania klasyfikatorów

Trafność dla danych RBF-GR



Teoria uczenia się

- Teoria przyrostowego uczenia się (*online learning*):
  - Analiza algorytmów przyrostowego uczenia się rodzin wykładniczych (Grünwald, Kotłowski, COLT 2010; Kotłowski, Grünwald COLT 2011)
  - Algorytmy przyrostowego uczenia się macierzy (Kotłowski, Warmuth COLT 2011; Koolean, Kotłowski, Warmuth NIPS 2011)
  - ,,Kernelizacja'' algorytmów (Warmuth, Kotłowski, Zhou ALT 2012)
- Analiza spójności rangowych funkcji strat:
  - Dembczyński et al., ICML 2009; Dembczyński et al., NIPS 2011; Dembczyński et al., ECML 2011; Kotłowski, Dembczyński, Hüllermeier, ICML 2011; Dembczyński, Kotłowski, Hüllermeier, ICML 2012; Dembczyński et al., Machine Learning 2012
- Statystyczna teoria uczenia się:
  - Klasyfikacja porządkowa z ograniczeniami monotonicznymi (Kotłowski, Słowiński ICML 2009; Kotłowski, Słowiński TKDE 2013)
  - Uczenie się dyskryminacji macierzy gęstości stanów kwantowych (Guta, Kotłowski NJP 2010)

Uczenie się preferencji

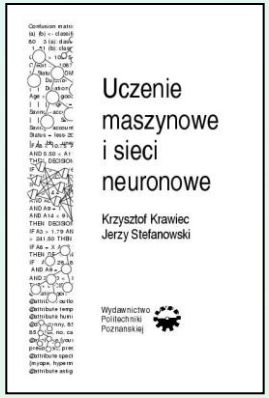
- Algorytmy rodzin reguł decyzyjnych dla problemów uczenia się preferencji:
  - ORDER (Dembczyński, Kotłowski, Słowiński, 2009),
  - RankRules (Dembczyński, Kotłowski, Szeląg, Słowiński 2010),
  - MORE (Dembczyński, Kotłowski, Słowiński 2009, 2010),
  - LPRules (Kotłowski, Słowiński, ICML 2009)
- Choqistyczna regresja
  - Wykorzystanie całki Choquet do budowy modeli monotonicznych (Fallah, Cheng, Dembczyński, Hüllermeier, ECML 2011, Machine Learning 2012)
- Analiza spójności rangowych funkcji strat
  - Problem klasyfikacji binarnej (Kotłowski, Dembczyński, Hüllermeier, ICML 2011)
  - Problem rankingu wieloetykietowego (Dembczyński, Kotłowski, Hüllermeier, ICML 2012)

Klasyfikacja Wieloetykietowa

- Probabilistyczne łańcuchy klasyfikatorów:
  - Wprowadzenie PCC (Dembczyński, Cheng, Hüllermeier, ICML 2009),
  - Analiza teoretyczna łańcuchów klasyfikatorów (Dembczyński, Waegeman, Hüllermeier, ECAI 2012 BPA),
- Analiza teoretyczna
  - Postać klasyfikatora bayesowskiego dla różnych funkcji strat (Dembczyński, Cheng, Hüllermeier, ICML 2009, Dembczyński, Cheng, Waegeman, Hüllermeier, NIPS 2011)
  - Analiza „żalu” pomiędzy funkcjami strat (Dembczyński, Cheng, Waegeman, Hüllermeier, ECML 2011, Dembczyński, Cheng, Waegeman, Hüllermeier, Machine Learning 2012)
- Algorytmy redukcji
  - Problem rankingu wieloetykietowego (Dembczyński, Kotłowski, Hüllermeier, ICML 2012)
  - Empiryczna analiza (Dembczyński, Cheng, Waegeman, Hüllermeier, Machine Learning 2012)

Inne aktywności dydakt. popularyzatorskie

- Książki i skrypty
  - Krawiec K., Stefanowski J., Uczenie maszynowe i sieci neuronowe, Wyd. Politechniki Poznańskiej, 2004.



- Wykłady i warsztaty
  - Wykłady zaproszone (R.Słowiński, K.Krawiec, M.Komosiński, J.Stefanowski, K. Dembczyński …)
  - Warsztaty:
    - Combined learning models (RSCTC 2010)
    - Mining complex and stream data (ADBIS 2012)
    - Class imbalances: Past, Present and Future (ICMLA 2012)
  - Co-Chairs konferencji (JRST 2007, RSCTC 2012, EuroGP2013,…)
  - Semantic data mining tutorial (ECML&PKDD'2011)
  - Warsztaty IRMLES 2009-2011 (International Workshop on Inductive Reasoning and Machine Learning on the Semantic Web)
  - Preference Learning Stream na EURO 2012 i EURO 2013 (Dembczyński, Waegeman, Słowiński)
  - ICML Tutorial 2013 Multi-target Prediction (Waegeman, Dembczyński, Hüllermeier)
- Członkowie komitetów programowych: ICML, NIPS, AAAI, IJCAI, UAI, COLT, ECML/PKDD