Zespoły modeli predykcyjnych
inne podejścia wykład 10

Jerzy Stefanowski

Instytut Informatyki PP

2021

Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI-TECH)
projekt finansowany z środków Programu Operacyjnego Polska Cyfrowa
POPC.03.02.00-00-0001/20

Fundusze Europejskie
Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego

Plan wykładu

- Zróżnicowanie klasyfikatorów składowych
- Generalizacja stosowa (ang. stacking)
- Podejście tzw. mieszanki ekspertów (ang. mixture of experts)
- Podejścia zespołowe do danych silnie wieloklasowych
- Podsumowanie

Motywacje dla Stacking [ang.]

- Alternatywne podejścia do budowania złożonych klasyfikatorów
- Klasyfikatory bazowe – często niejednorodne, uczone różnymi algorytmami
  - Lecz może być użyte do zastąpienia głosowania w zespołach bagging lub boosting
- **Struktura wielopoziomowa** – z różnymi podejściami do rozstrzygania niejednoznaczności wskazań klasyfikatorów bazowych
- Koncepcja tzw. meta-uczenia się (wiedza z odpowiedzi innych klasyfikatorów)
- Możliwe dynamiczne modyfikowanie działania zespołu

Stacking – generalizacja stosowa

Obserwacje – niektóre przykłady mają wysokie prawdopodobieństwo złej klasyfikacji, a inne są częściej dobrze klasyfikowane.

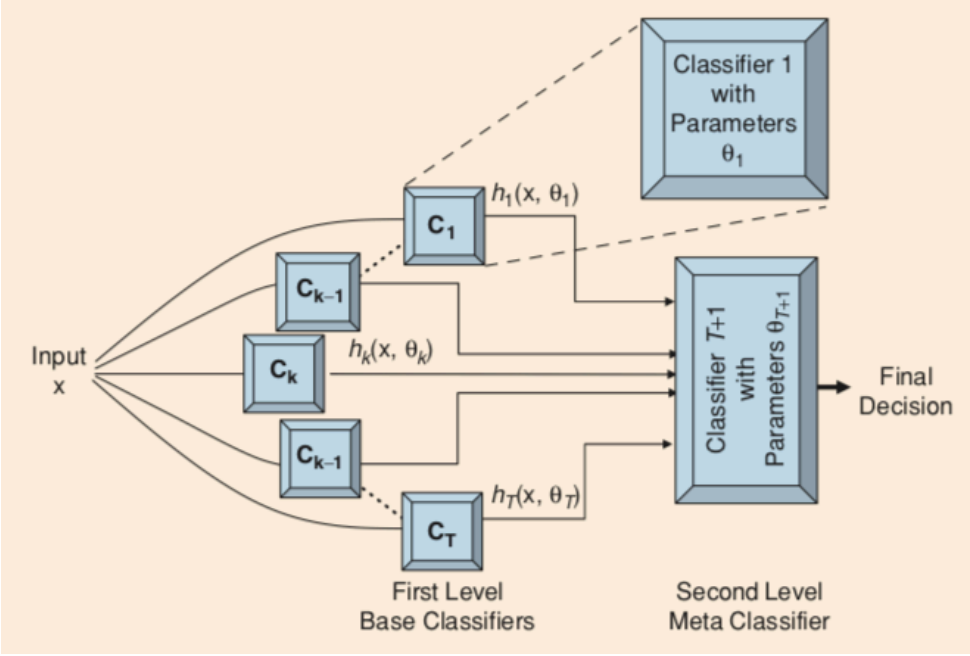
Pytanie – czy można nauczyć się ogólniejszych **meta zasad**, jak dokonać korekty klasyfikacji pewnych modeli (lepiej niż głosowanie większościowe)

Wolpert – **generalizacja stosowa**:

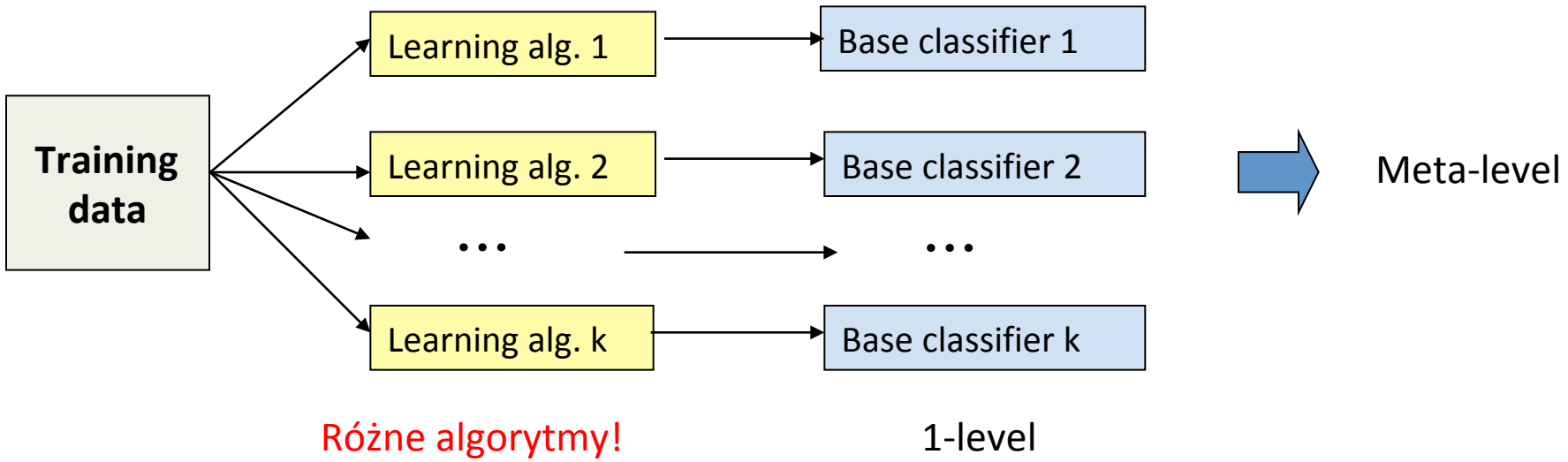
- Wyjścia klasyfikatorów są wejściami dla kolejnych algorytmów (meta-uczących się), w celu nauczenia zasad korekty wcześniejszych predykcji.
- Możliwe jest łączenie wielu warstw algorytmów / klasyfikatorów

Stacked generalization [Wolpert 1992]

- Wykorzystaj ideę meta-uczenia
  - Predykcje tzw. base learners/model (*level-0 models*) przekazane na wejścia kolejnych tzw. meta learner (*level-1 model)*
- Metody uczenia bazowych modeli (poziom 0) są często różnymi algorytmami (podejścia niejednorodnych modeli)
- Różne rozwiązania idei meta-poziomu (poziom 1), np. tzw. combiner albo arbiter



Meta-uczenie -> tzw. combiner

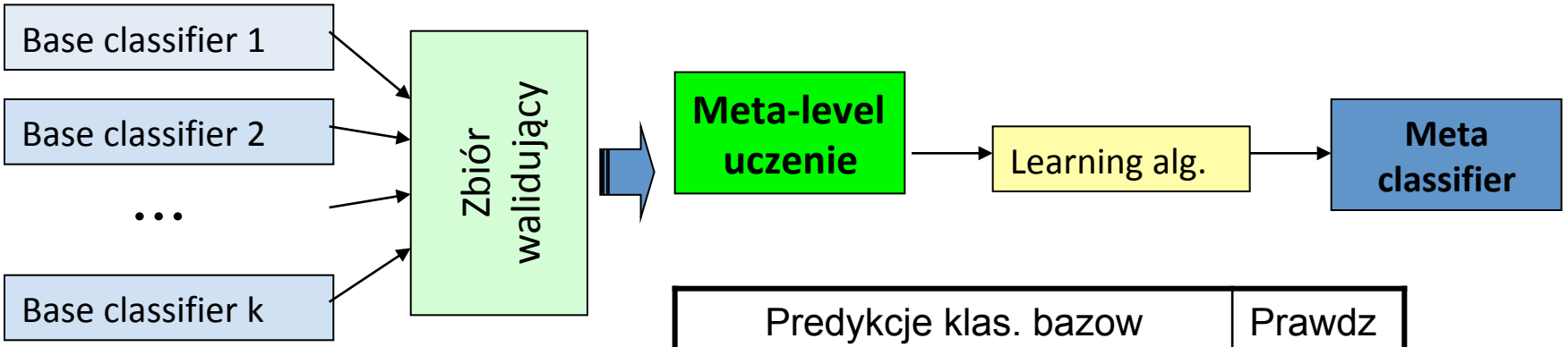


Chan & Stolfo : *Meta-learning* [meta-uczenie]

- Dwa poziomy:
  - 1-level – base classifiers
  - 2-level – meta-classifier
- Różne algorytmy użyte do uczenia klasyfikatorów bazowych (zróżnicowanie klasyfikatorów)

Więcej patrz: Experiments on Multistrategy Learning by Meta-Learning. P.Chan, S.Stolfo 1993

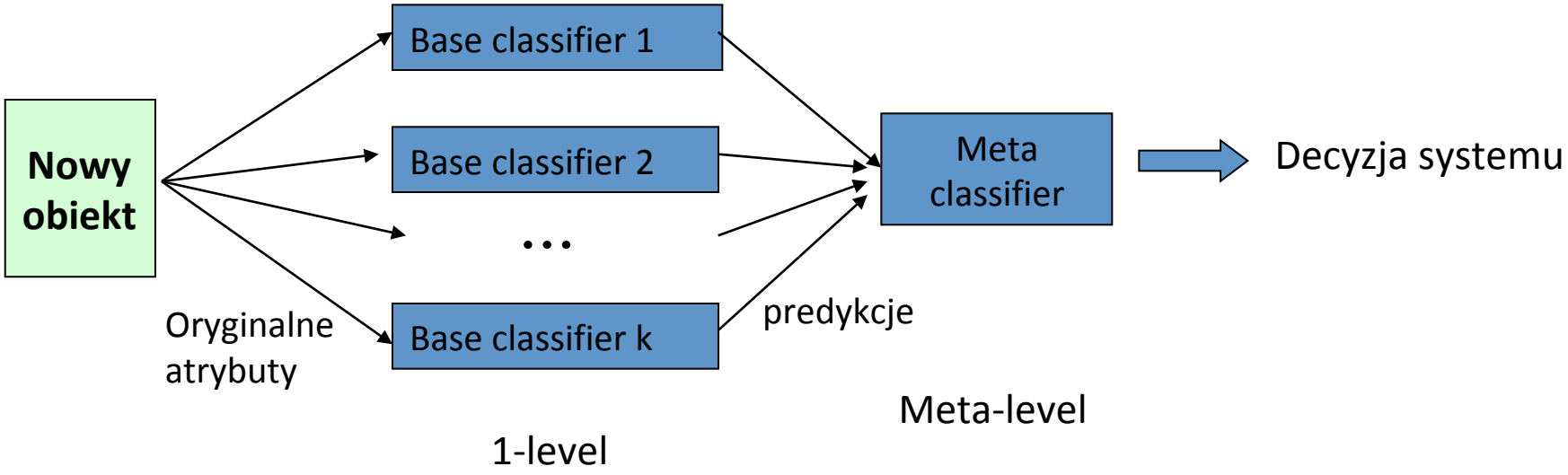
Uczenie meta klasyfikatora



Predykcje klas. bazow				Prawdz etykieta
Cl.1	Cl.2	…	Cl.K	
A	A	…	B	**A**
A	B	…	C	**B**

- Predykcje klasyfikatorów bazowych na zbiorze walidującym (ew. wewnętrzna ocena krzyżowa) wraz z właściwą etykietą – zbiór uczący meta-level (może być rozszerzony przez atrybuty z org. danych)
- Niezależny algorytm uczący meta-klasyfikator
- Celem jest wyuczenie korekty predykcji w b. złożony sposób niż głosowanie większościowe

Predykcja systemy złożonego combiner



Klasyfikacja nowego przykładu

Chan & Stolfo [95/97] : eksperymenty w architekturze ({CART,ID3,K-NN} →NBayes) / [dane biomedyczne] trafność lepsza niż pojedyncze klasyfikatory i ich złożenia poprzez głosowanie większościowe

Comparison of classification accuracy (%)

Data set	K-NN	C4.5	MODLEM	**Combiner**
acl	84.29	85.00	85.00	84.29
bupa	63.19	62.32	68.10	69.12
cleveland	52.10	53.14	54.46	55.66
glass	68.80	65.42	69.63	71.50
hsv	56.56	51.64	55.74	59.02
imidasolium	58.21	58.21	60.70	66.67
…	…	…	…	…
yeast	57.80	52.10	54.30	58.36

More in Nowaczyk, Stefanowski: On Using Rule Induction in Multiple Classifiers with a Combiner Aggregation Strategy. ISDA 2005.

Meta-uczenia : Naive Bayes / sprawdzano też inne

Ogólny przyrost trafności + lepiej niż pojedyncze klasyfikatory składowe

Inne eksperymenty [materiały njt.edu]

Breast Cancer Dataset

Method	Error (%)	Precision (%)	Recall (%)
NB	13.7	86	90
Linear SVM	12.5	92	92
Logistic Regression	10.7	93	95
Random Forest	5.35	95	98
Boosted Trees	3.57	95	95
Stacked Ensemble Classifier	**1.78**	**98**	**98**

Table: Comparison of performance of different classifiers using the Breast Cancer d

Przyrost miar oceny na 6 różnorodnych zbiorach danych, także w porównaniu do standardowych klasyfikatorów

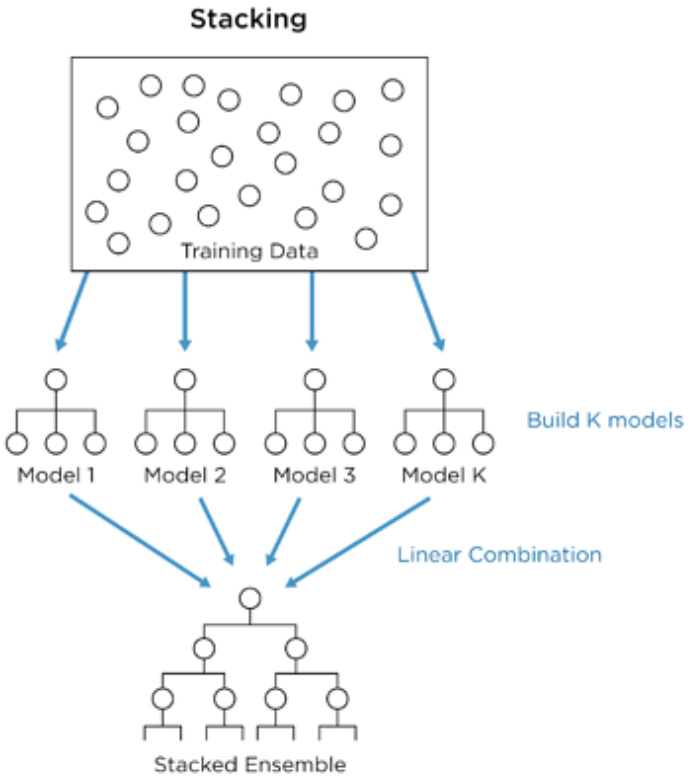
Więcej na linku podanym w ekursy

Stacking - rozszerzenia

- Wyjścia klasyfikatorów są rozkładami prawdopodobieństw [Witten 1999] / także meta-uczenie z odmianą regresyjną tzw. model tree [Dzeroski, Zenko 2004] − złożenie 3 klasyfikatorów, ocena eksperymentalna na 30 zbiorach danych z UCI
- StackingC – przeznaczony dla problemów wieloklasowych. Klasyfikatory bazowe ukierunkowane na rozpoznanie jednej z klas
- SCANN – wykorzystanie przekształcenie wyjść poprzez analizę korespondencji i poszukiwanie bliskości w nowej przestrzeni − odmiana kNN jako finalna predykcja [Merz 1999]

Generalizacja stosowa - więcej

- Literatura naukowa – eksperymentalne doświadczenia z rozszerzaniem klasyfikatorów bagging zamiast niejednorodnych klasyfikatorów badawczych
- Wyższe poziomy mogą być także zespołem klasyfikatorów



Podejście arbitrażu

Inny sposób rozstrzygania niespójnych decyzji łączonych klasyfikatorów − tzw. arbiter trees - także wprowadzony przez Chan i Stolfo [1999]

- Zbiór uczący podzielony na *k* rozłącznych części
- Dla każdej pary klasyfikatorów − uczymy specjalny klasyfikator arbiter do rozstrzygania niezgodności pomiędzy ich decyzjami (specjalne reguła arbitrażu − głosowanie trzech dla przykładu)
- Nowy arbiter (wyższego meta-poziomy) jest uczony z wyjść arbitrów niższego poziomu aż do finalnej decyzji arbitrażowej
- Konstruuje się strukturę drzewa podjęcia decyzji końcowej

Opis podejścia z algorytmem − książka L.Rokach: Pattern classification using ensemble methods (2009)

Mixture of experts

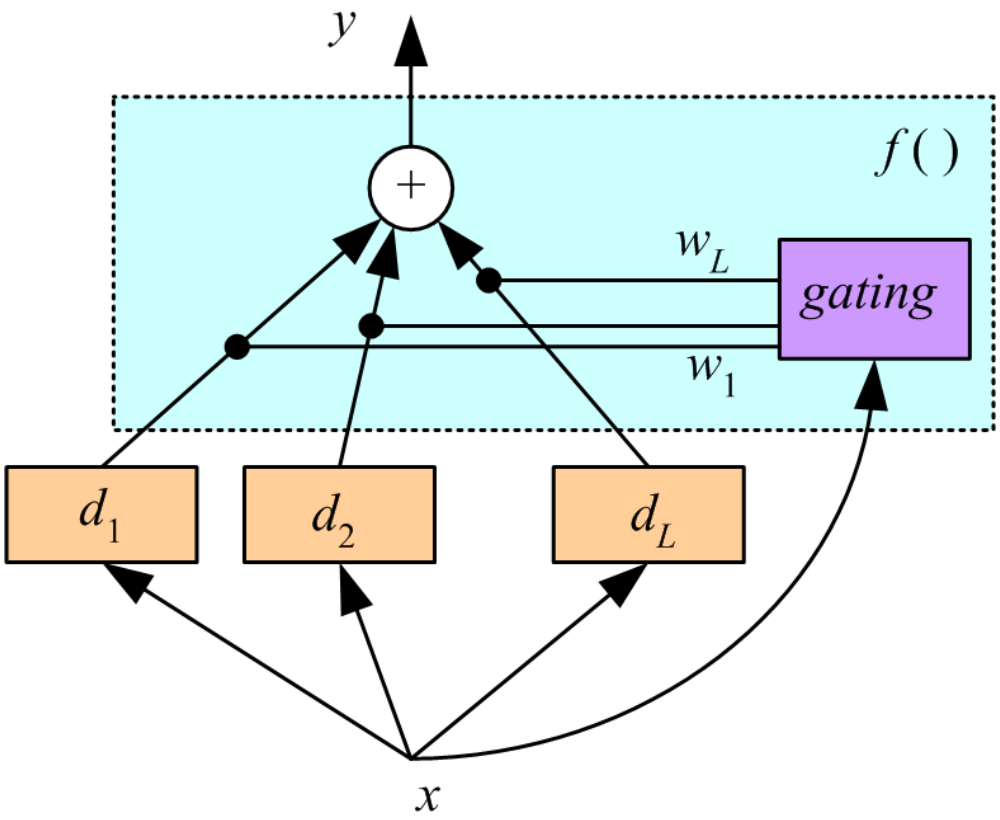
- Motywacja - różne obszary przestrzeni cech pokrywane są przez różne modele / będącymi „ekspertami” od tych pod-problemów
- Ich predykcje (dla nowego / klasyfikowanego przykładu **x**) są “miękko” składane lub specjalny składnik (tzw. ang. gating network) wybiera najbardziej kompetentnych ekspertów dla danego obszaru i przykładu x
- Proces uczenia − obejmuje zarówno modele bazowe jak i tzw. gating network
- Różne rozwiązania:
  - np. probabilistyczne modele generatywne
  - Sieci neuronowe -> soft max element
- Sieć neuronowa RBF może być prostą realizacją idei „mieszanki ekspertów”

Mieszanka ekspertów z dynamicznym wagowaniem ich predykcji

Tzw. gating network

Przykład – przypisanie wag do predykcji klasyfikatorów / dostosowanych do klasyfikowanego przykładu x

$$y = \sum_{j=1}^{L} w_j(x) d_j(x)$$

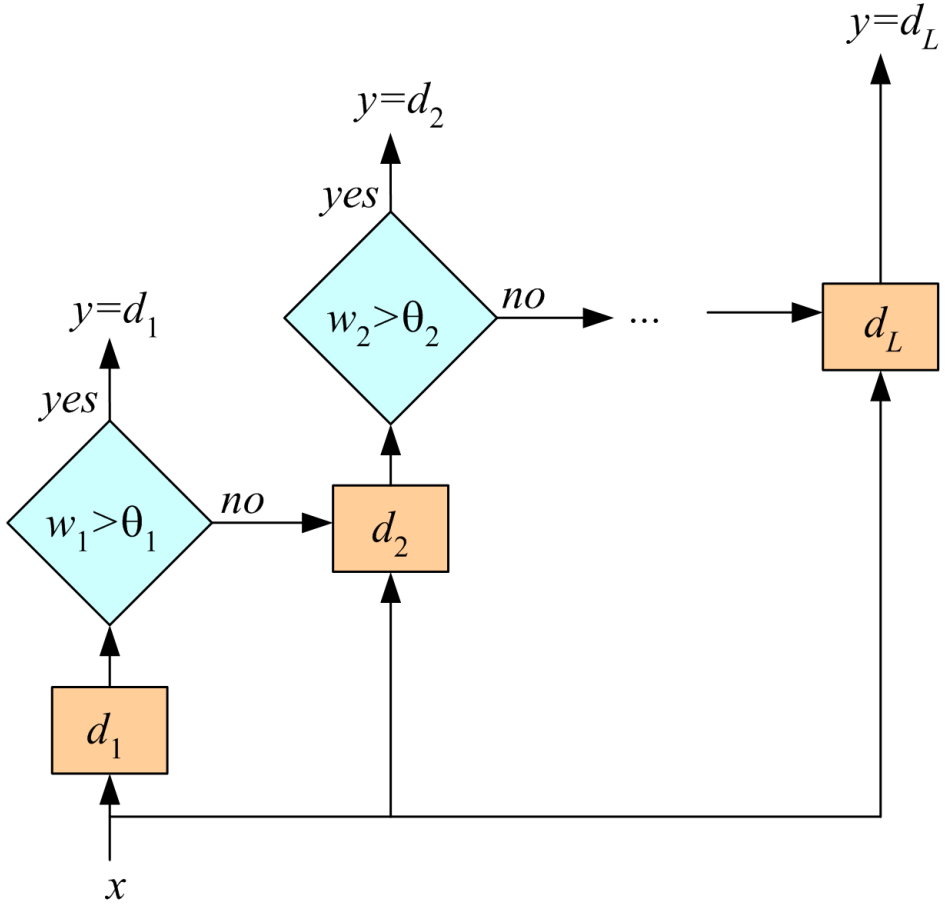


Rozwiązanie kaskadowe

Cascade models:

Stwórz kaskadę modeli predykcyjnych (zróżnicowanych) w celu poradzenia sobie ze złożonymi zadaniami / specjalizują się w trudnościach

Wykorzystaj kolejny model i jego predykcję d_j , jeśli poprzednicy są niepewni (ang. not sufficiently confident)



Kaskadowa sekwencja

- Bazowe klasyfikatory są dodawane sekwencyjnie
  - jeśli pewność predykcji pierwszego klasyfikatora jest duża, to jego decyzja jest uznawana za ostateczną
  - W innym przypadku decyzja jest przekazywana do kolejnego klasyfikatora, itd.
- Model kaskadowy znajduje zastosowanie w systemach czasu rzeczywistego, jeśli decyzja powinna być szybka, podjęta maksymalnie przez kilka klasyfikatorów

Miary pewności predykcji zespołu

W przypadku prostych złożeń (bagging, stacking, …) ocenia się tzw. margines decyzji zespołu

- Różnica głosów za zwycięska i drugą klasą
  - Np. dla problemu 3 klasowego bagging z 17 drzewami ma rozkład predykcji: C1 − 10 głosów, C2 − 5 i C3 − 2 głosów -> margines 10-5=5
- Analogiczna różnica dla miar zagregowanych po prawdopodobieństwach lub innych „scores”, także ważonych

Poczekaj do wykładu nt. aktywnego uczenia z techniką “Query by Committee”, gdzie wykorzystuje się takie marginesy

Miary zróżnicowania

- Wiele propozycji – patrz książka L.Kuncheva Combining Pattern Classifiers
- Rozważmy parę klasyfikatorów C_i oraz C_j (tzw. pairwise measures) + decyzje binarne (poprawny lub błędny)

	Cj jest poprawny	Cj jest błędny
Ci jest poprawny	a	b
Ci jest błędny	c	d

- **Q Statistics** $Q_{ij}=(ad-bc)/(ad+bc)$
- Dodatnie wartości Q: jeśli przykłady są poprawnie klasyfikowane przez oba klasyfikatory, ujemne dla odwrotnych klasyfikacji. Maksymalne zróżnicowanie predykcji -> Q=0

Miary zróżnicowania klasyfikatorów

- Miary ang. disagreement i double fault

  $D_{ij} = b + c \qquad DF_{ij} = d$

- oraz inne Kappa, korelacja odpowiedzi,…
  - Przegląd w książce Ludmila Kuncheva
- Dla T klasyfikatorów − mamy $T(T\text{-}1)/2$ miar zróżnicowania par
- Najczęściej uśrednia się je do jednej globalnej wartości

Niesparowane miary zróżnicowania

Niech dla i-tego przykładu, e_i jest liczbą klasyfikatorów z *T*, które błędnie klasyfikują ten przykład, wtedy

Entropia (0 − klasyfikatory podejmują te same decyzje, 1 są maksymalnie zróżnicowane):

$$E = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T-(T/2)}\min\{e_i, T-e_i\}$$

Kohavi-Wolpert variance

$$KW = \frac{1}{nT^2}\sum_{i=1}^{n} e_i \cdot (T-e_i)$$

Dobre i złe zróżnicowanie klasyfikatorów składowych

- **Postulat zróżnicowania** (ang. diversity) łączonych klasyfikatorów ->
  nie mogą być zbyt podobne do siebie (mieć podobnych predykcji) + dostatecznej ich jakości (dla obserwacji z rozkładu przestrzeni cech, rozumianych że wystarczająco dużo klasyfikatorów powinno podjąć prawidłową predykcje)
  - Zbyt wiele miar zróżnicowania i brak uniwersalnej miary
- Tzw. złe zróżnicowanie (ang. bad diversity) − pewna różnorodność klasyfikatorów może wpływać negatywnie na ich połączenie
- Więcej informacji w pozycjach:
  - “Good” and “ bad” diversity in majority vote ensembles. G Brown, L Kuncheva MCS 2010

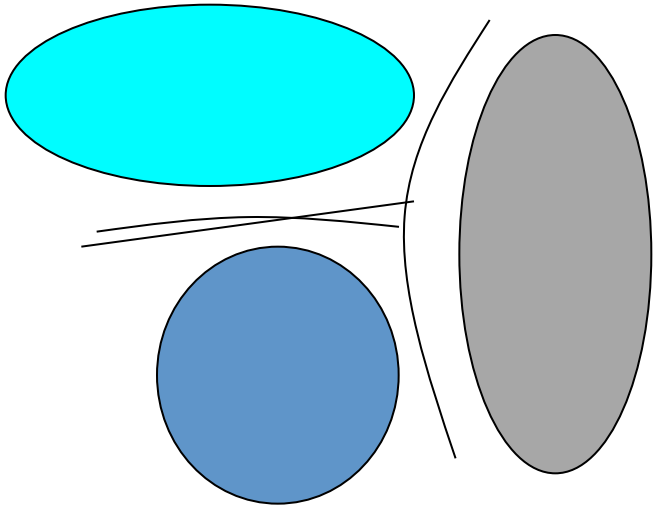
Wykorzystanie miar zróżnicowania

- Analiza budowy i działania nauczonych już zespołów
- Wykorzystanie do redukcji zbyt licznych zespołów (ang. ensemble pruning), tj. wyboru „najbardziej wartościowych” klasyfikatorów składowych i poprawy predykcji takiego złożenia kombinowanego
- Przegląd podejść [M.Woźniak]:
  1. Rank-based pruning – ustaleniu ranking klasyfikatorów z wykorzystaniem parowych miar zróżnicowania i wyboru najlepiej ocenianych
  2. Optimization based pruning – sformułowanie problemu optymalizacji (łączone kryteria) i zastosowania podejść ewolucyjnych
  3. Clustering based pruning – grupowanie podobnych klasyfikatorów i wybór ich reprezentantów

Więcej: M.Woźniak: Zespoły klasyfikatorów − aktualne kierunki badań (2015).

Specjalne zespoły dla problemów wieloklasowych

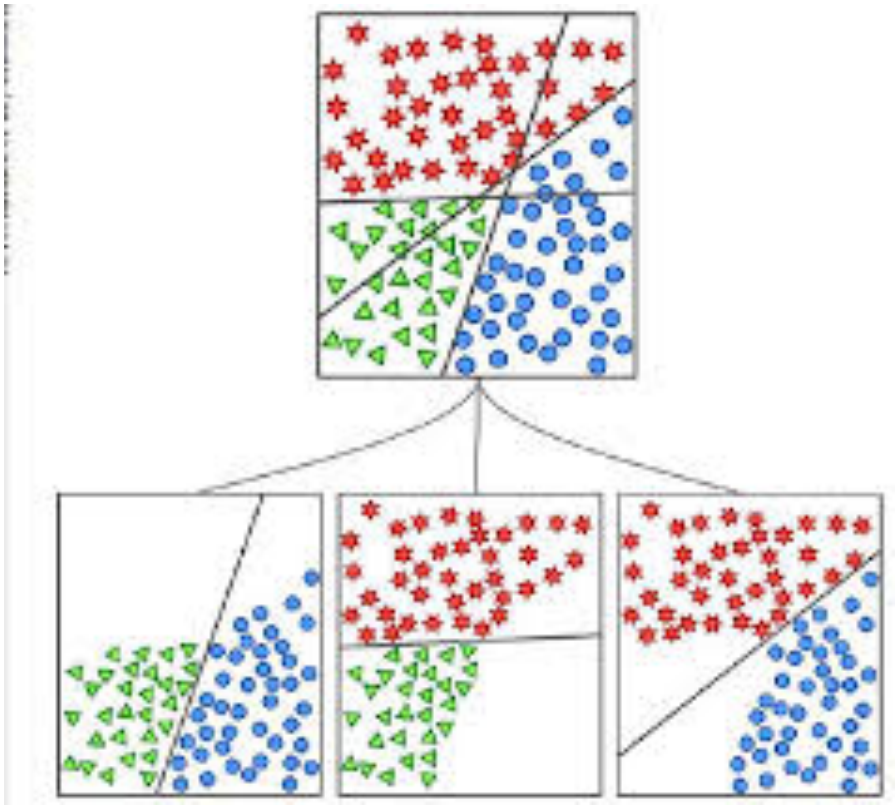
- Zadanie klasyfikacji do wielu klas (*n* >>2 kategorii)
- Część rzeczywistych danych – klasy trudne do nauczenia (pojęcia o nieliniowych granicach i trudnych rozkładach w przestrzeni atrybutów)
- Często dekompozycja na podproblemy – łatwiej się nauczyć
- Przykład problemy trzy klasowego (tzw. pairwise decision boundaries between each pairs of classes are simpler)



Dekompozycja binarna

Podejście z j. ang. one-against-one / pairwise coupling

Rozważ wszystkie kombinacje dwóch klas i wyucz specjalizowane klasyfikatory binarne

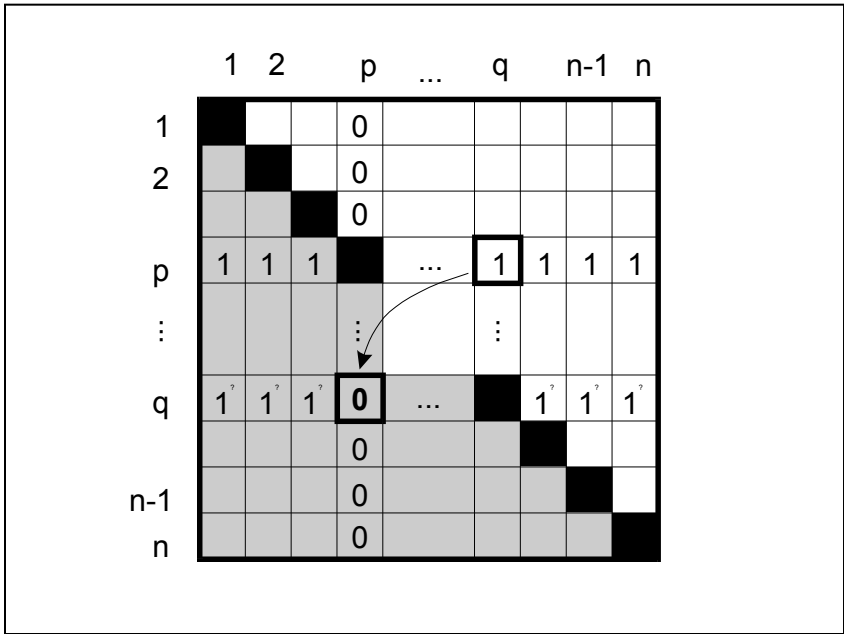


Pairwise coupling - n2-classifier

Zespół złożony z $(n^2-n)/2$ *binarnych klasyfikatorów* (wszystkie połącznia par z n klas)

- Każda para klas (i,j), gdzie $i,j \in [1..n]$, $i \neq j$, rozróżniana przez niezależny klasyfikator C_{ij}
- Uczenie C_{ij} – tylko przykłady z klas i,j;
- Wszystkie klasyfikatory uczone tym samym algorytmem
- klasyfikator C_{ij} wskazuje dwie decyzje (1 or 0), klasyfikatory C_{ij} and C_{ji} równoważne

$$C_{ji}(\mathbf{x}) = 1 - C_{ij}(\mathbf{x})$$



Własny rysunek z pracy hab. 2001

Zasady klasyfikacji w pairwise coupling

- Nowy przykład **x**, przekazany na wszystkie klasyfikatory $C_{ij}(\mathbf{x})$ w strukturze n^2 (one against one)= konieczna reguła agregacji dla wypracowania decyzji i rozstrzygania konfliktów
- Najczęstsza reguła – wybierz klasę, która wygrała w największej licznie porównać parami klas /odmiana majority voting/.
- Możliwe rozszerzenia:
  - Oszacuj wiarygodność klasyfikatora binarnego P_{ij} (np. w trakcie uczenia)
  - Reguła - a weighted majority rule:
    - Wybierz klasę „*i*” która maksymalizuje $\sum_{j=1,i\neq j}^{n} P_{ij} \cdot C_{ij}(x)$
- Wprowadź odpowiedź „nie wiem” lub zasadę dynamicznego ważenia głosów

Przykład oceny klasyfikatora n^2 z drzewa c4.5

Data set	Classification accuracy *DT* (%)	Classification accuracy n^2 (%)	Improvement n^2 vs. *DT* (%)
Automobile	85.5 ± 1.9	87.0 ± 1.9	**1.5***
Cooc	54.0 ± 2.0	59.0 ± 1.7	**5.0**
Ecoli	79.7 ± 0.8	81.0 ± 1.7	**1.3**
Glass	70.7 ± 2.1	74.0 ± 1.1	**3.3**
Hist	71.3 ± 2.3	73.0 ± 1.8	**1.7**
Meta-data	47.2 ± 1.4	49.8 ± 1.4	**2.6**
Primary Tumor	40.2 ± 1.5	45.1 ± 1.2	**4.9**
Soybean-large	91.9 ± 0.7	92.4 ± 0.5	**0.5***
Vowel	81.1 ± 1.1	83.7 ± 0.5	**2.6**
Yeast	49.1 ± 2.1	52.8 ± 1.8	**3.7**

Własna praca wspólnie z J.Jelonkiem

Inne zagadnienia

Inne warianty podstawowych algorytmów

- Modyfikacje bagging (Pasting small votes)
- Pośrednie modele (Arc-c4), inne boosting, gradient boosting
- Rotation Forest

ECOC – popularna wersja wieloklasowa

Agregacja odpowiedzi liczbowych (ang. numeric predictions)

- Zespoły specjalizowane dla trudnych danych
- …

Inne ciekawe wykorzystanie zespołów

Uogólnienia dla trudnych danych, np.

- Niezbalansowanie klas (zwłaszcza uogólnienia bagging)

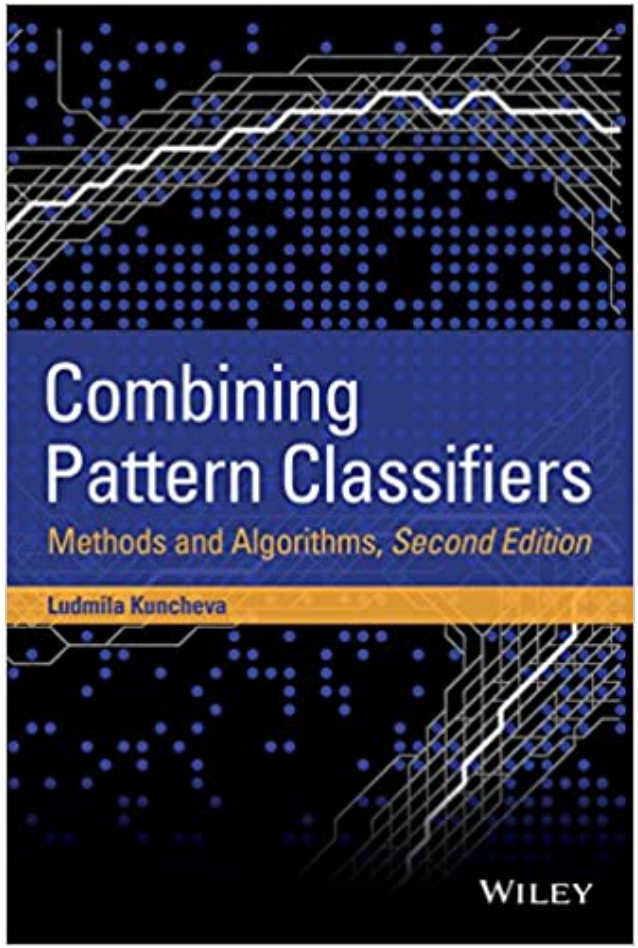
Zespoły klasyfikatorów dla **zmiennych strumieni danych** (wiele rozwiązań, nie tylko uogólnienia jak online bagging)

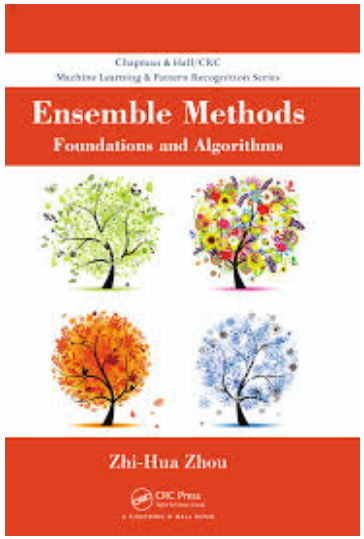
Klasyfikacja wielo-etykietowa (multi-labeled) − tzw. łańcuchy klasyfikatorów

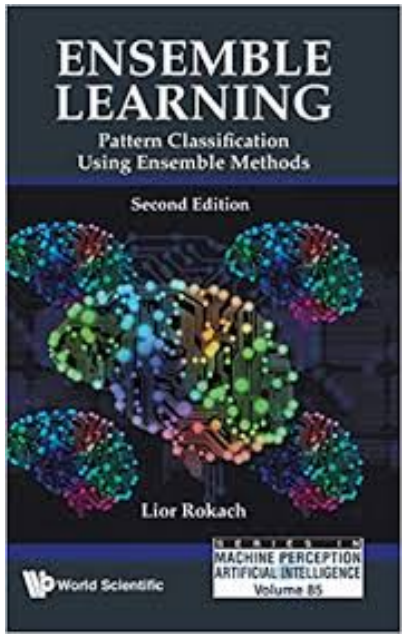
Multi-view learning (lub odmiany self-learning) dla trybu uczenia częściowo-etykietowanego

Querry by committee w aktywnym uczeniu się

Więcej odpowiedzi, radzę książki







Odnośniki do literatury

- Intensywny rozwój od lat 90 poprzedniego wieku
- Wiele różnych propozycji
- Przykładowe pozycje:
  - L.Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, 2004 (large review + list of bibliography).
  - T.Dietterich, Ensemble methods in machine learning, 2000.
  - Using Correspondence Analysis to Combine Classifiers. C.Merz Machine Learning J. (1997)
  - Is Combining Classifiers with Stacking Better than Selecting the Best One? S.Dzeroski, B.Zenko, Machine Learning J. (2004)
  - G.Valentini, F.Masulli, Ensemble of learning machines, 2001 [obszerna lista referencyjna]
  - R.Polikar,Ensemble based systems in decision making, IEEE Circuits and Systems Magazine, vol. 6, no. 3, pp. 21–45, 2006.
  - W Polsce − przykładowo prace M.Woźniak i współpracownicy

Pytanie i komentarze?

Dalszy kontakt:

jerzy.stefanowski@cs.put.poznan.pl

http://www.cs.put.poznan.pl/jstefanowski/

Fundusze Europejskie Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska Europejski Fundusz Rozwoju Regionalnego