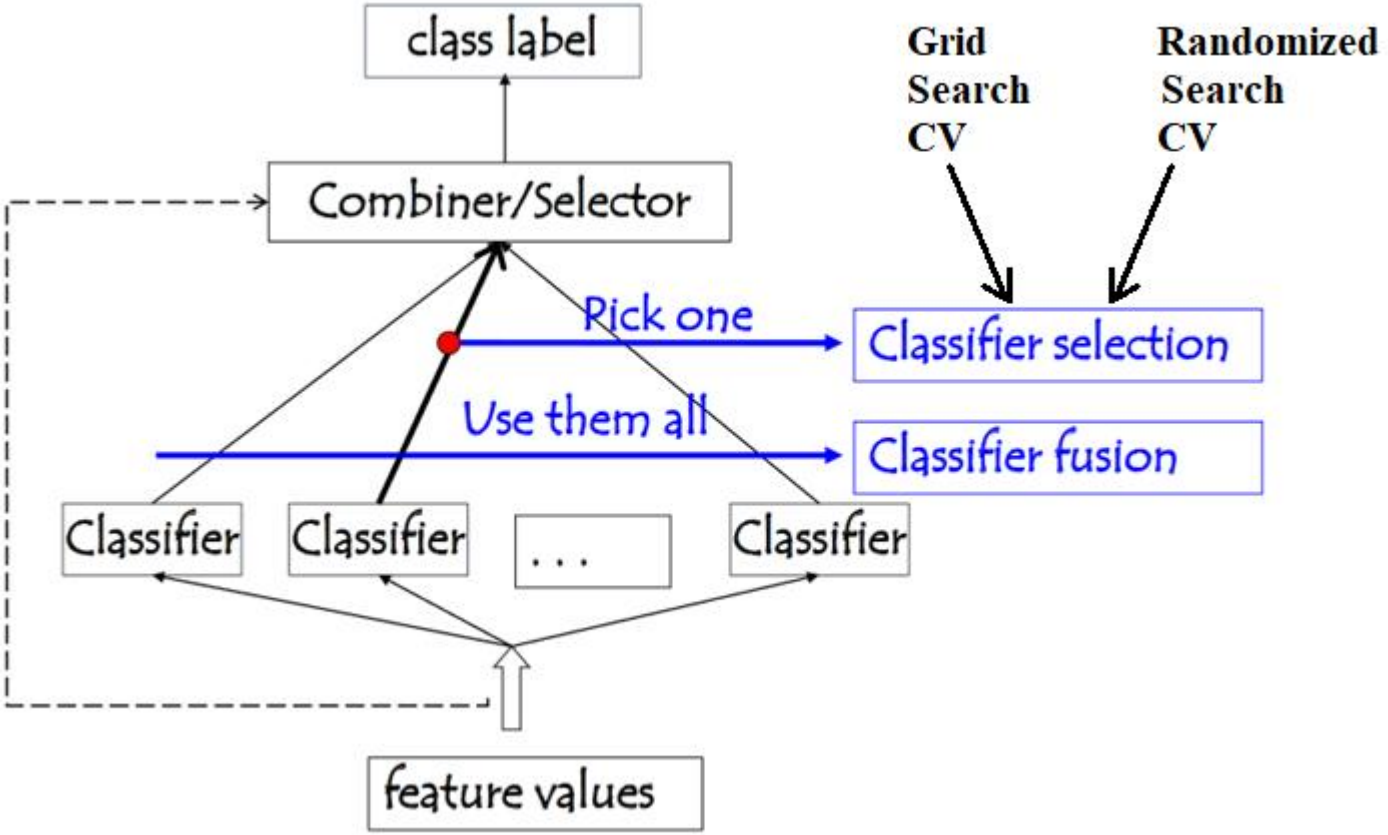
Zespoły modeli – przegląd literatury

Stanisław Kaźmierczak

Agenda

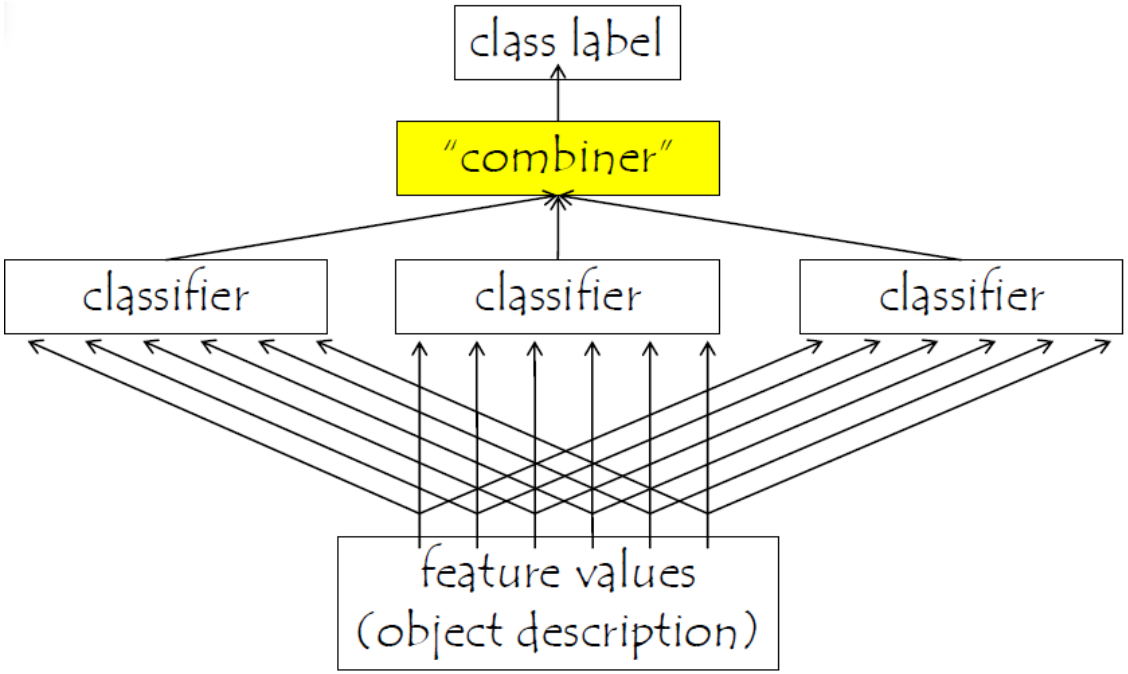
- Motywacja
- Obszary tworzenia zespołów
- Faza wyboru instancji
- Faza selekcji cech
- Faza tworzenia modeli
- Faza łączenia modeli
- Różnorodność

Łączenie vs selekcja (1)



[1]

Łączenie vs selekcja (2)



[1]

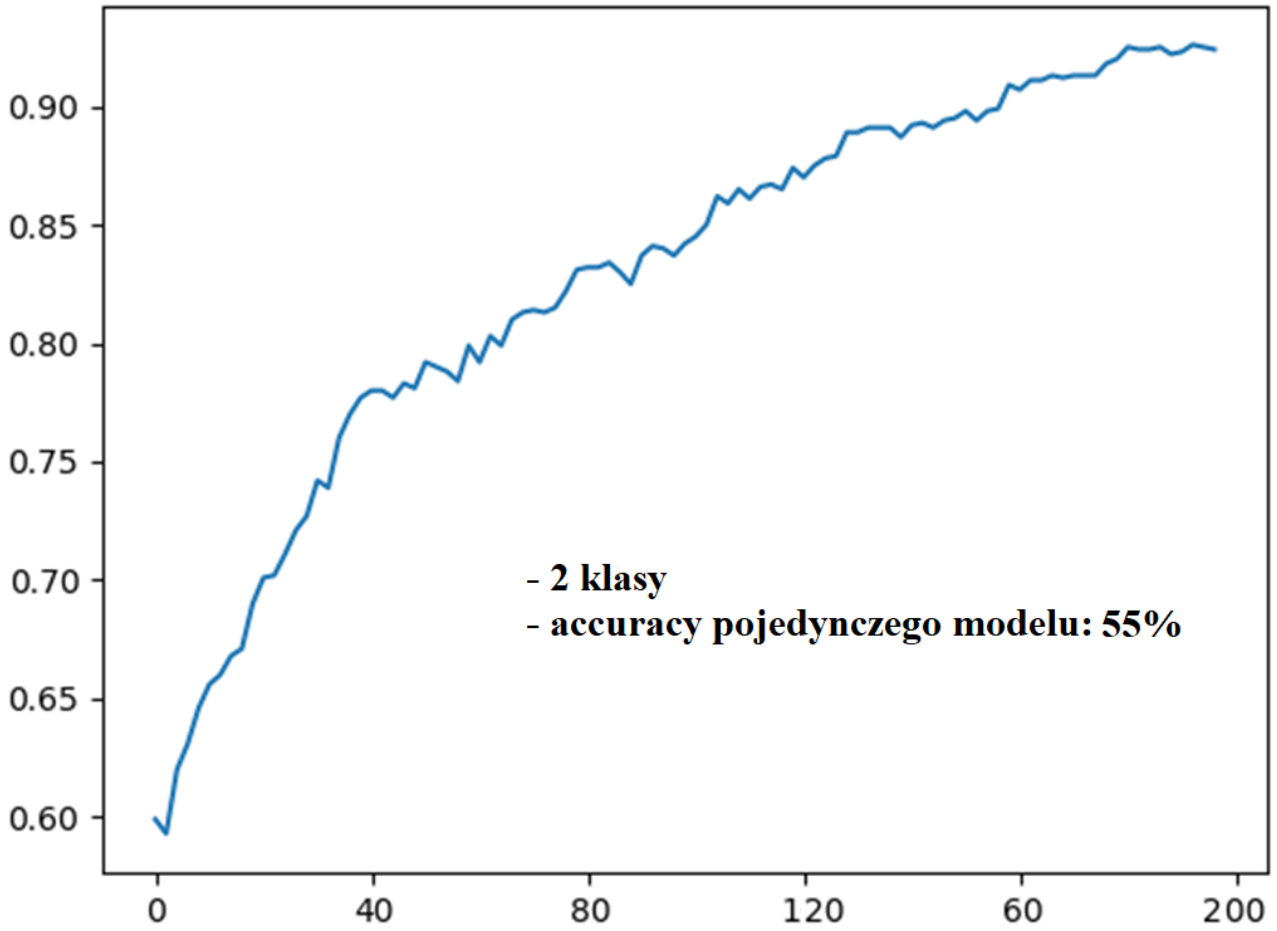
- Łączenie modeli w zespół prawie zawsze daje lepsze wyniki niż najlepszy z pojedynczych modeli
  - Analogia do zespołu ekspertów, konsylium lekarskiego

Intuicja (1)

Poprawna klasa	Model 1	Model 2	Model 3	Zespół
A	A	B	A	A
A	A	A	B	A
B	A	B	B	B
A	B	B	B	B
B	B	B	A	B
B	A	B	B	B
B	B	A	B	B
A	B	A	A	A
B	B	B	A	B
A	A	A	A	A
Accuracy	60%	70%	60%	90%

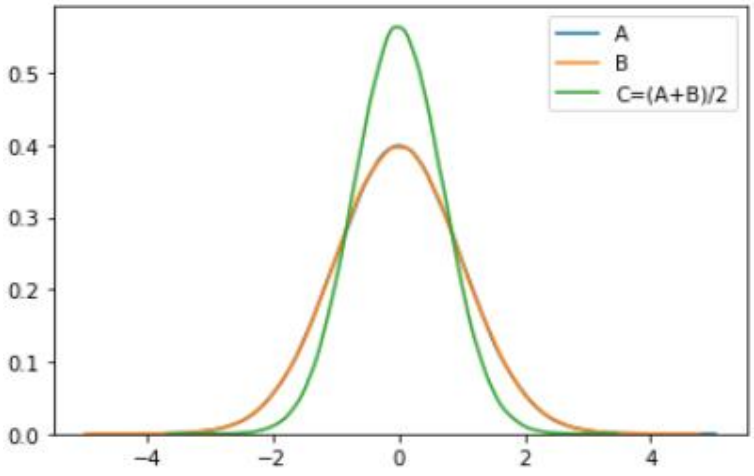
Intuicja (2)

- *Weak learner* – model dający nieco lepszy rezultat niż losowa predykcja



Wzór Bienaymé

- Wartości zmiennych losowych mogą symulować wielkość błędów, które popełniają modele
- Przy założeniu, że wartość oczekiwana zmiennej losowej wynosi 0, wartość błędu średniokwadratowego (MSE) jest równa wariancji błędu (wariancji zmiennej losowej)
- Zmniejszenie wariancji jest tożsame ze spadkiem MSE
- Dla zbioru nieskorelowanych zmiennych losowych X_i o tym samym rozkładzie zachodzi:



$$\operatorname{Var}\left(\overline{X}\right) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}\left(X_i\right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Modele w pełni skorelowane

Poprawna klasa	Model 1	Model 2	Model 3	Zespół
A	A	A	A	A
A	A	A	A	A
B	A	A	A	A
A	B	B	B	B
B	A	A	A	A
B	B	B	B	B
B	B	B	B	B
A	B	B	B	B
B	B	B	B	B
A	A	A	A	A
Accuracy	60%	60%	60%	60%

Modele częściowo skorelowane

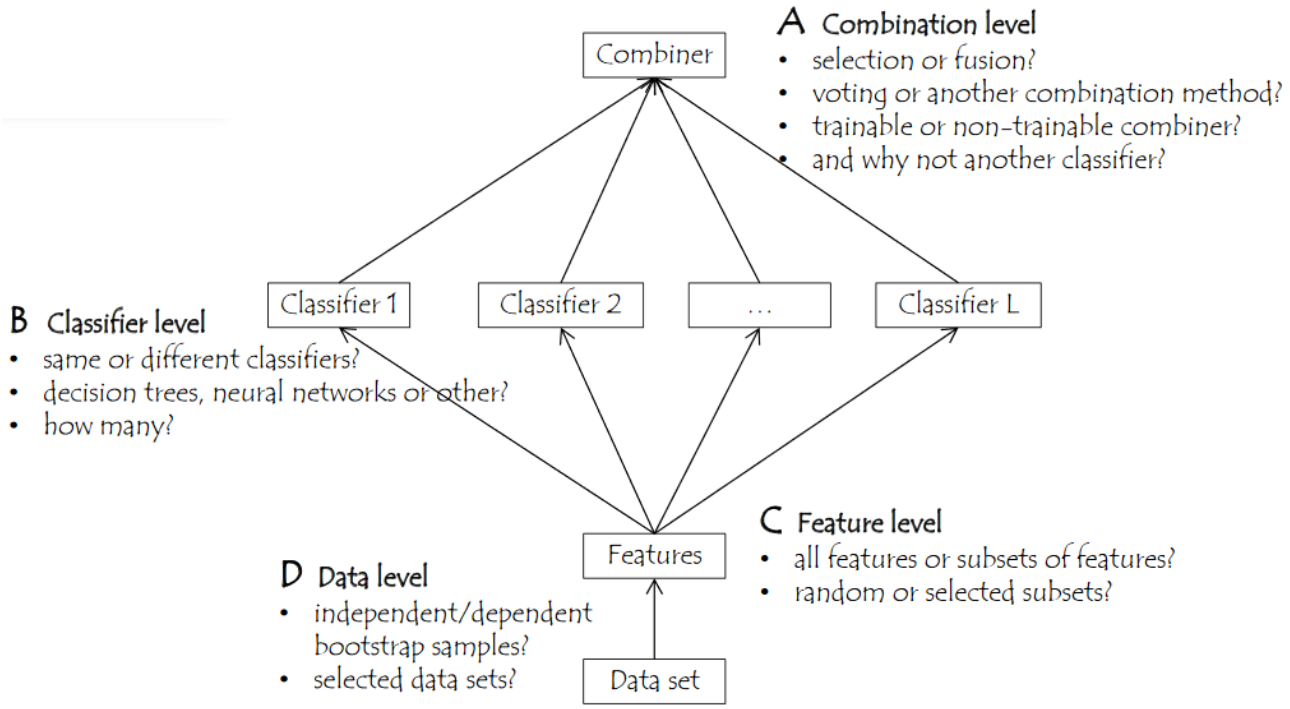
- Rzeczywiste modele są z reguły dość mocno (ale nie w pełni) skorelowane
  - Trudne instancje są trudne dla (prawie) wszystkich modeli
  - Podobnie wygląda sytuacja dla prostych instancji

Poprawna klasa	Model 1	Model 2	Model 3	Zespół
A	A	A	A	A
A	A	A	A	A
B	A	A	A	A
A	B	B	B	B
B	B	B	A	B
B	A	B	B	B
B	B	A	B	B
A	B	B	B	B
B	B	B	B	B
A	A	A	A	A
Accuracy	60%	60%	60%	70%

Jakość zespołu

- O sile zespołu modeli decydują dwa główne czynniki
  - Zróżnicowanie modeli składowych
  - Jakość modeli składowych
- *Tradeoff* pomiędzy zróżnicowaniem i jakością pojedynczych modeli
- W przypadku *weak learners* zespoły osiągają większe zróżnicowanie modeli składowych, ale zespół musi być liczniejszy

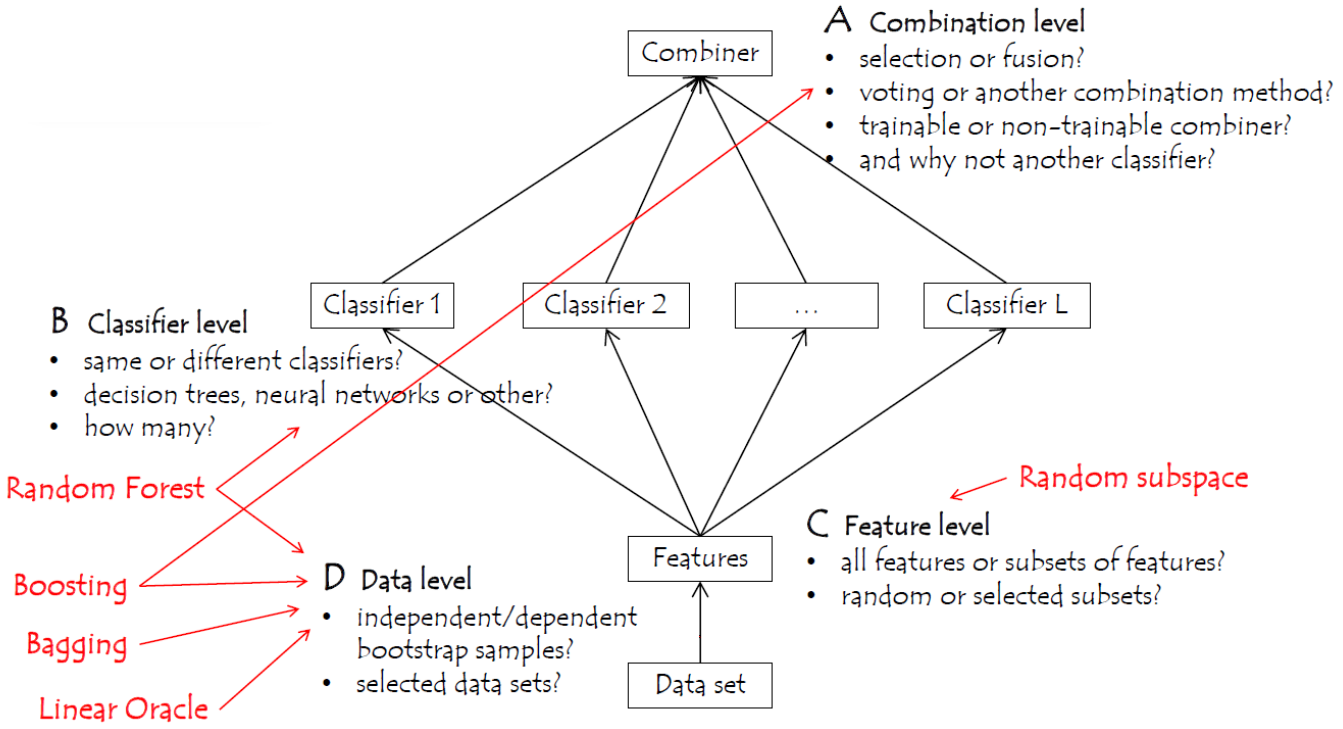
Budowanie modelu (1)



[1]

- Większość technik z różnych obszarów można stosować niezależnie od siebie

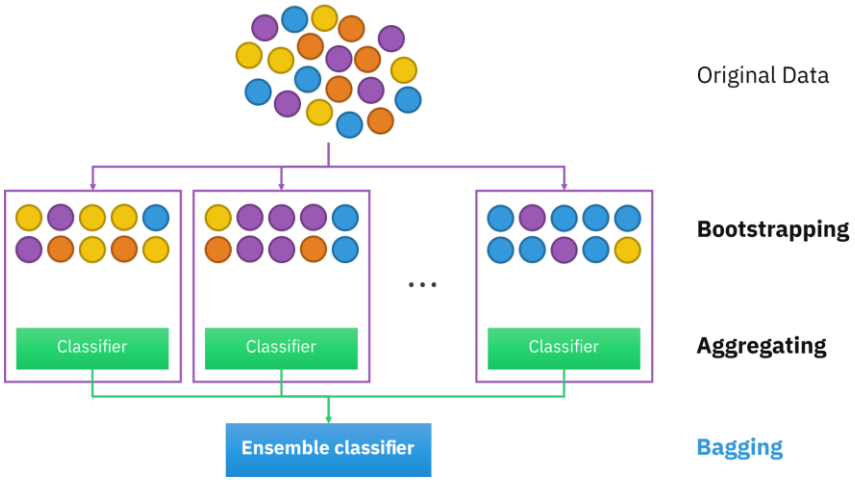
Budowanie modelu (2)



[1]

Bagging

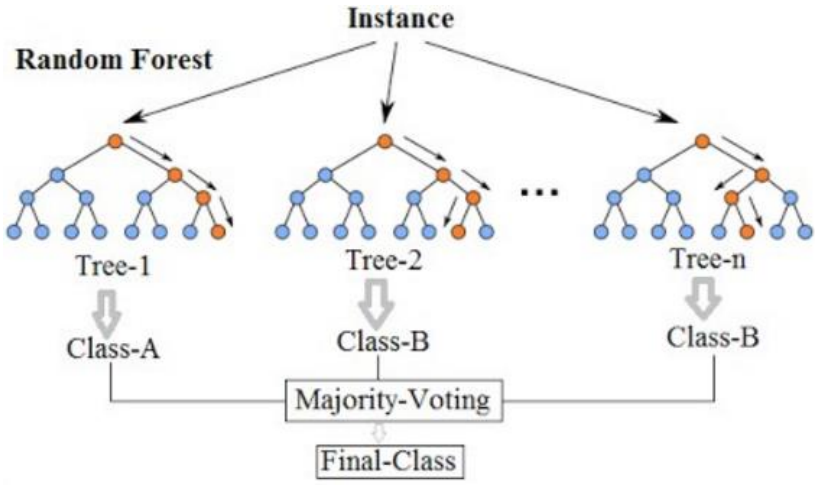
- *Bagging* = ***b**ootstrap **agg**regat**ing***
- Budowa zespołu bazuje na utworzeniu grupy modeli na różnych zestawach danych wygenerowanych z oryginalnego zbioru uczącego
- Każdy z zestawów zawiera tę samą liczbę rekordów i powstaje w wyniku losowania ze zwracaniem z oryginalnego zbioru uczącego
- Pojedynczy zestaw danych zawiera zdublowane rekordy, jak również nie zawiera pewnych rekordów z oryginalnego zbioru



[2]

Lasy losowe (1)

- Zespół modeli budowany wg następujących założeń:
  - Modelami składowymi są drzewa decyzyjne
  - Dane treningowe służące do skonstruowania kolejnych drzew losowane są ze zwracaniem ze zbioru uczącego (*bootstrapping*)
  - Drzewa są jedynie konstruowane, nie jest stosowane ich przycinanie
- Las losowy jest więc formą baggingu, w którym pojedyncze modele są lekko zmodyfikowanymi drzewami decyzyjnymi
- Las losowy może być wykorzystany do oceny znaczenia poszczególnych cech



[3]

Lasy losowe (2)

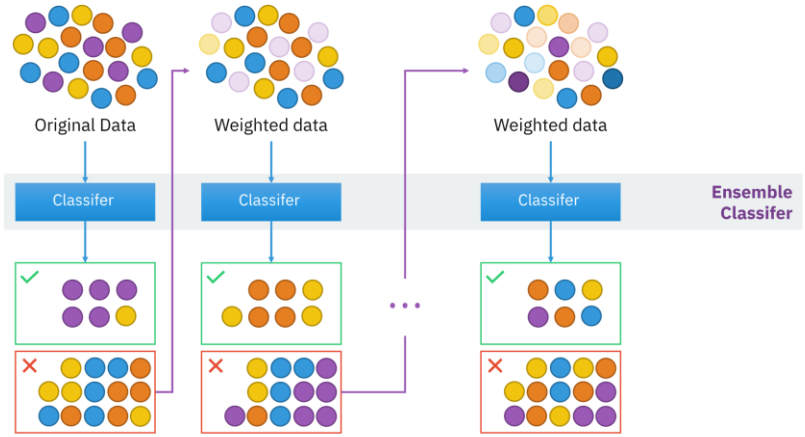
- Klasyczny algorytm konstrukcji lasu losowego wygląda następująco:

Input: D - zbiór uczący, p - liczba zmiennych niezależnych, T - liczba tworzonych drzew (wielkość zespołu), $d < p$ - wymiar podprzestrzeni
Result: L - las losowy
begin
  for $t = 1$ *to* T **do**
    utwórz zestaw danych (ang. bootstrap sample) D_t poprzez wylosowanie $card(D)$ rekordów ze zbioru D ze zwracaniem;
    zbuduj drzewo M_t na podstawie D_t, nie stosując przycinania;
    W każdym węźle wylosuj d spośród p zmiennych i rozważ podziały bazujące na tych d zmiennych
  end
  return $\{M_t : t = 1, \ldots, T\}$
end

[4, 5]

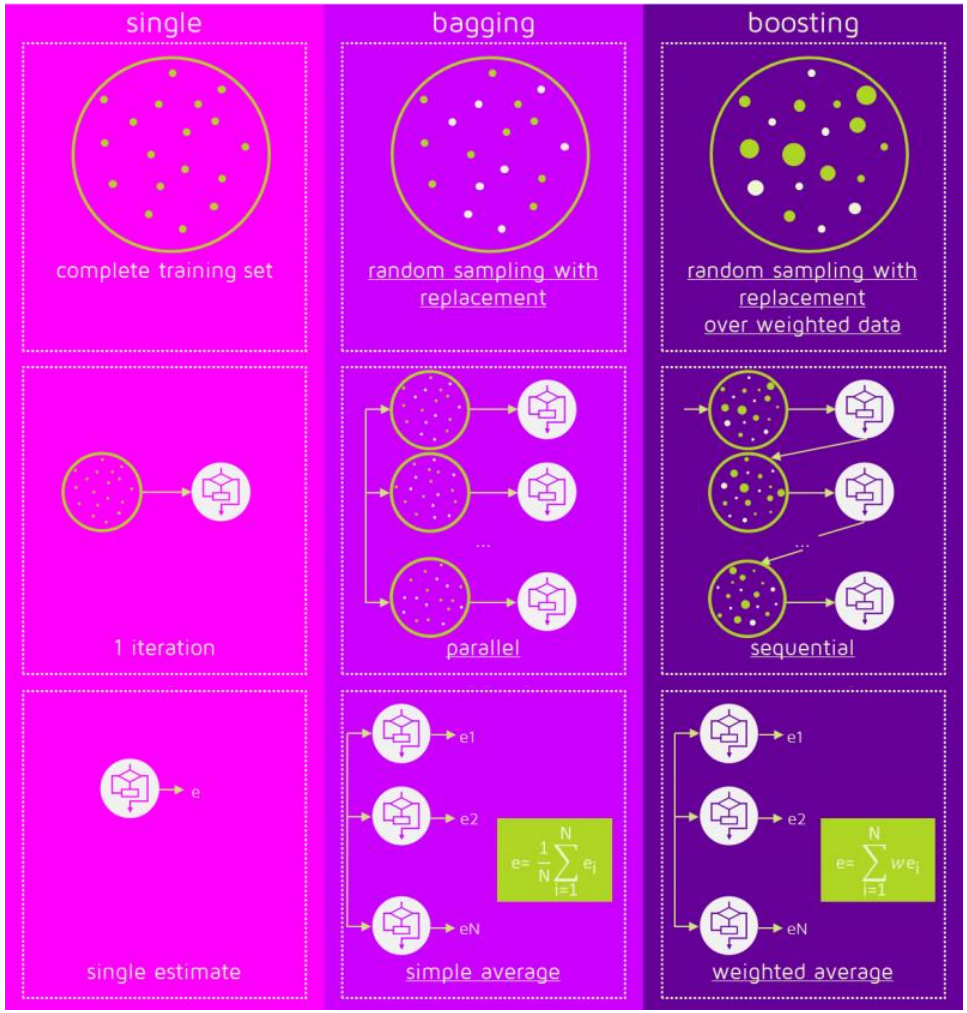
Boosting

- Zbiory treningowe generowane są poprzez losowanie ze zwracaniem
- Nowy model budowany jest tak, aby lepiej poradzić sobie z instancjami, w których poprzedni model zawiódł
- Ostateczna predykcja podejmowana jest poprzez uśrednienie predykcji (lub głosowanie większościowe) bazowych klasyfikatorów
  - Średnia ważona na podstawie jakości predykcji poszczególnych modeli



[6]

Pojedynczy model/bagging/boosting



[7]

Selekcja/ekstrakcja cech

- Redukując liczbę wejściowych cech, adresujemy problem wysokowymiarowych i rzadkich danych
  - *Random subspace*
    - podzbiór cech jest losowo wybierany i na ich podstawie konstruowane są klasyfikatory bazowe
    - predykcja z reguły przy użyciu ważonego głosowania, gdzie wagą pojedynczego modelu jest jego *accuracy*
  - *Input decimation*
    - cechy wybierane są w sposób nielosowy
    - algorytm redukuje korelację między błędami klasyfikatorów bazowych poprzez trenowanie ich na różnych podzbiorach cech wejściowych
    - rozszerzenia metody polegające na jednoczesnym sprawdzaniu *accuracy* i różnorodności, jak również zastosowaniu algorytmów ewolucyjnych

Poziom modeli

- Modele niegeneratywne
  - Opierają się na zbiorze bazowych, dobrze nauczonych, klasyfikatorów/regresorów
  - Nie służą do tworzenia nowych bazowych modeli, ale starają się w odpowiedni sposób połączyć/wyselekcjonować już istniejące
- Modele generatywne
  - Tworzą nowe bazowe modele
  - Na nowopowstający model mają wpływ wcześniej utworzone modele
  - Nowopowstający model tworzony jest zgodnie z ogólną regułą maksymalizacji jakości własnej predykcji oraz różnorodności zespołu

Łączenie klasyfikatorów (1)

- *Combiner* dostaje na wejściu wyjście z bazowych (pojedynczych) modeli, np. wybraną etykietę (klasę), ranking etykiet lub wektor prawdopodobieństw dla każdej klasy
- Zwraca ostateczną etykietę danej instancji
- Trenowalne i nietrenowalne *combinery*
- Metody
  - *Majority voting*/ważone *Majority voting*
  - Bazujące na regułach Bayesa
  - Wykorzystujące różne operatory algebraiczne (maksimum, minimum, mediana, średnia, itp.)

Łączenie klasyfikatorów (2)

- Metody (cd.)
  - Bazujące na teorii zbiorów rozmytych
  - *Meta-learning*

    - trenowalne *combinery*

    - output bazowych modeli staje się cechami *combinera*
- Jako wejście *combinera*, oprócz wyjść z pojedynczych modeli, mogą podane być również cechy wejściowe pojedynczych modeli

Łączenie klasyfikatorów (3)

- „The combiner matters (a lot)” [8]
- „The trained combiner works better” [8]
- Train the combiner if you have *enough* data [9]

Selekcja modeli (1)

- Selektor wybiera podzbiór zbioru bazowych modeli
- Metody
  - *Test and select*
    - podejście zachłanne; nowy klasyfikator jest dodawany do zespołu tylko wtedy, gdy zmniejsza błąd średniokwadratowy
    - dowolna metoda doboru kolejnego klasyfikatora do zespołu, analizowane były m.in. algorytmy ewolucyjne
  - *Cascading classifiers*
    - bazowe klasyfikatory są dodawane sekwencyjnie
    - jeśli pewność predykcji pierwszego klasyfikatora jest duża, jest uznawana za ostateczną
    - Wpp. decyzja jest przekazywana do kolejnego klasyfikatora, itd.
    - model kaskadowy znajduje szczególne zastosowanie w systemach czasu rzeczywistego, ponieważ decyzja powinna być szybka, podjęta maksymalnie przez kilka klasyfikatorów

Selekcja modeli (2)

- Metody cd.
  - *Dynamic Classifier Selection*
    - kompetencje każdego klasyfikatora bazowego są oceniane na bieżąco
    - wybierane są te, które najbardziej nadają się do danego wejścia
    - funkcja oceniająca klasyfikatory priorytetyzuje te, które dobrze radziły sobie z przypadkami, które znajdują się blisko w przestrzeni cech z testowanym egzemplarzem
  - *Metody klasteringowe*
    - podejście dynamiczne może być w niektórych przypadkach zbyt kosztowne obliczeniowo
    - klastry tworzą modele bazowe, które dawały podobne rezultaty
    - z każdego klastra wybierany jest klasyfikator(y), co zapewnia odpowiednie zróżnicowanie zespołu

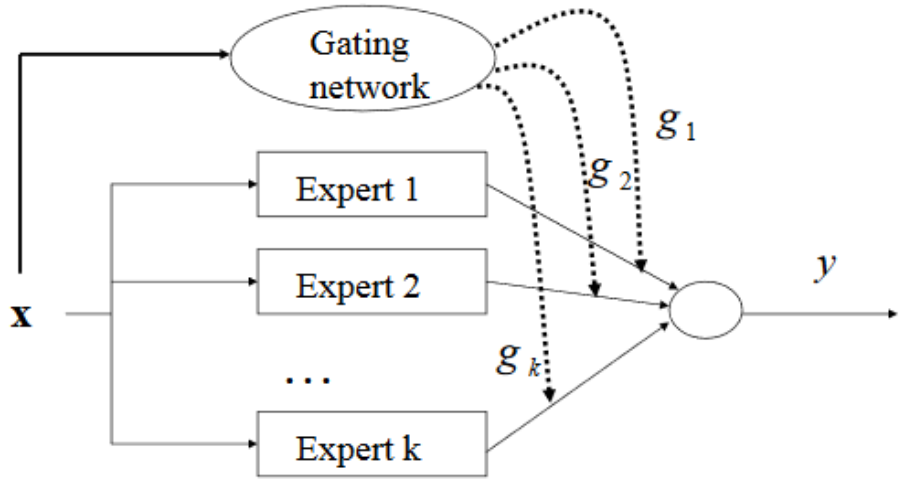
Selekcja modeli (3)

- Metody cd.
  - *Orientation ordering*
    - algorytmy zachłanne, które preferują dołożyć do zespołu te modele, które skorygują dotychczasowe błędne predykcje zespołu
  - Procedury statystyczne
    - wybierają z bazy modele z wysokim *accuracy*, a ostateczna predykcja dokonywana jest przy użyciu *majority voting*
  - Bazujące na metodach selekcji cech
    - Np. *Forward selection/backward elimination* dodaje/usuwa klasyfikatory tak, aby minimalizować określoną funkcję celu

Mixture of experts

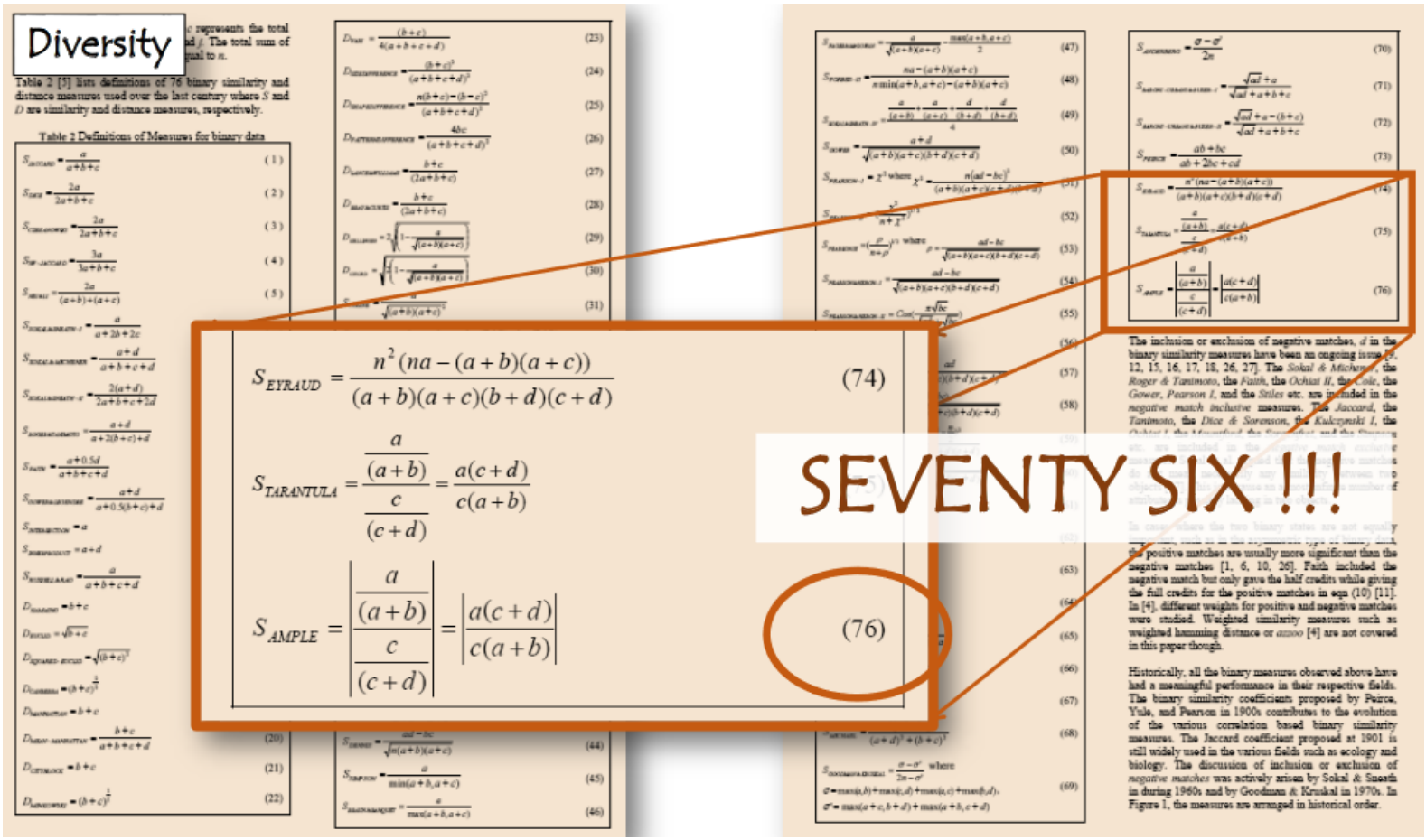
- Przestrzeni cech dzielona jest na jednolite regiony
- Zespół bazowych klasyfikatorów jest nadzorowany przez tzw. *gating network*
- *Gating network* decyduje, który klasyfikator (lub w jakim stopniu) ma być użyty do którego regionu
- W procesie nauki uczone są zarówno modele bazowe jak i *gating network*

$g_1, g_2, \ldots g_k$ - gating functions



[4]

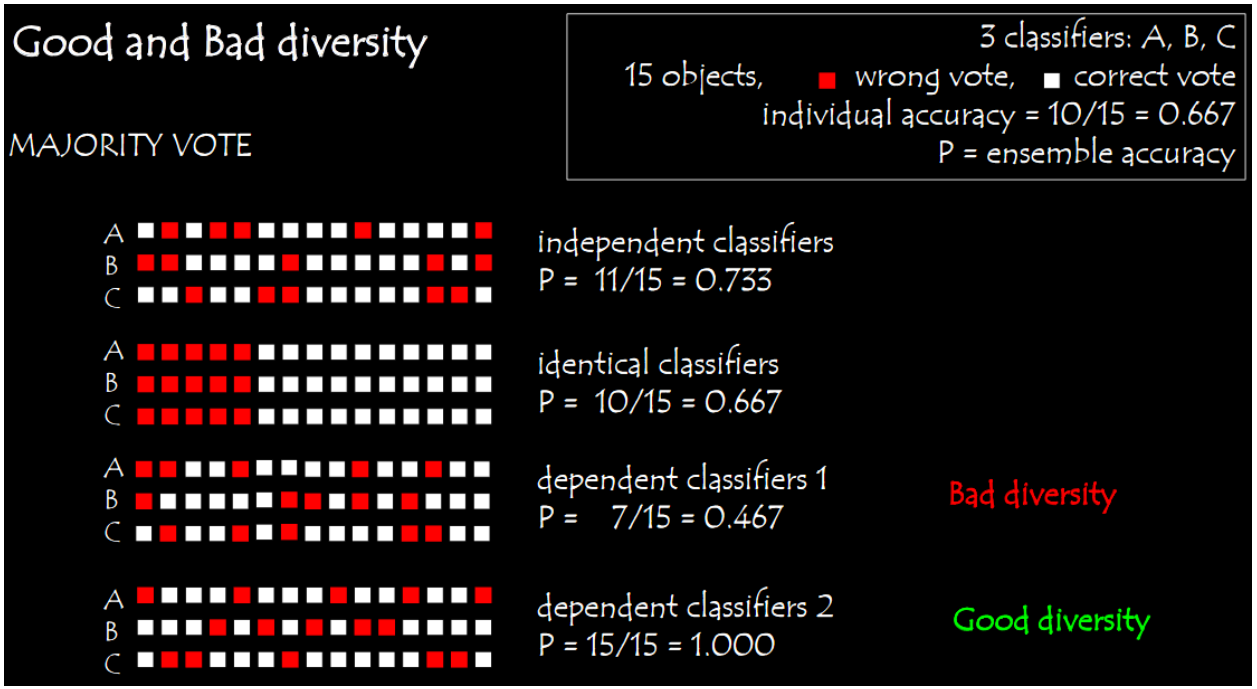
Różnorodność (1)



[1, 9]

Różnorodność (2)

- Nie ma monotonicznej zależności pomiędzy różnorodnością modeli w zespole, a jego *accuracy*
  - A może istnieje taka miara?



[1]

„Last step: ensemble it”

Best Kagglers

* Wystarczy losowość na poziomie modelu oraz majority/soft-voting

Źródła (1)

1. http://www.icpram.org/Documents/Previous_Invited_Speakers/2016/ICPRAM2016_Kuncheva.pdf
2. https://en.wikipedia.org/wiki/Bootstrap_aggregating
3. https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d
4. Learning, M. (2012). The Art and Science of Algorithms that Make Sense of Data.
5. M. Grzenda, slajdy z wykładów przedmiotu *Metody Data Science*
6. https://en.wikipedia.org/wiki/Boosting_(machine_learning)
7. https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/

Źródła (2)

8. Duin, R. P. (2002). The combining classifier: to train or not to train?. In Object recognition supported by user interaction for service robots (Vol. 2, pp. 765-770). IEEE.
9. Choi, S. S., Cha, S. H., & Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, *8*(1), 43-48.
10. https://people.cs.pitt.edu/~milos/courses/cs2750-Spring04/lectures/class22.pdf