Kombinacja jądrowych estymatorów gęstości w klasyfikacji - zastosowanie na sztucznym zbiorze danych

Mateusz Kobos, 07.04.2010
Seminarium Metody Inteligencji Obliczeniowej

Spis treści

- Opis algorytmu i zbioru danych
- Działanie algorytmu w „najlepszym przypadku”
- Działanie algorytmu z właściwym algorytmem uczenia
  - Wersja z leave-one-out cross-validation
  - Wersja z 10-fold cross-validation

Opis algorytmu

- Klasyfikacja punktu ***x***:
  - Estymuj gęstość w ***x*** za pomocą uśrednienia paru (aktualnie dwóch (E=2)) estymatorów jądrowych

$$\hat{p}(\boldsymbol{x}|\omega_i;\boldsymbol{h}) = \frac{1}{E}\sum_{j=1}^{E}\hat{p}(\boldsymbol{x}|\omega_i;h_{i,j})$$

  - Użyj wzoru Bayesa, by uzyskać prawdopodobieństwa klas w punkcie ***x***

$$d_B(\boldsymbol{x},\boldsymbol{h}) = \arg\max_{\omega_i}\hat{P}(\omega_i|\boldsymbol{x};\boldsymbol{h}) = \arg\max_{\omega_i}\frac{\hat{p}(\boldsymbol{x}|\omega_i;\boldsymbol{h})\hat{P}(\omega_i)}{\sum_{i=1}^{c}\hat{p}(\boldsymbol{x}|\omega_i;\boldsymbol{h})\hat{P}(\omega_i)}$$

    - Gdy mianownik =0 (na skutek ograniczonej precyzji maszynowej), zwracamy prawdopodobieństwa z algorytmu 1-NN

Opis algorytmu

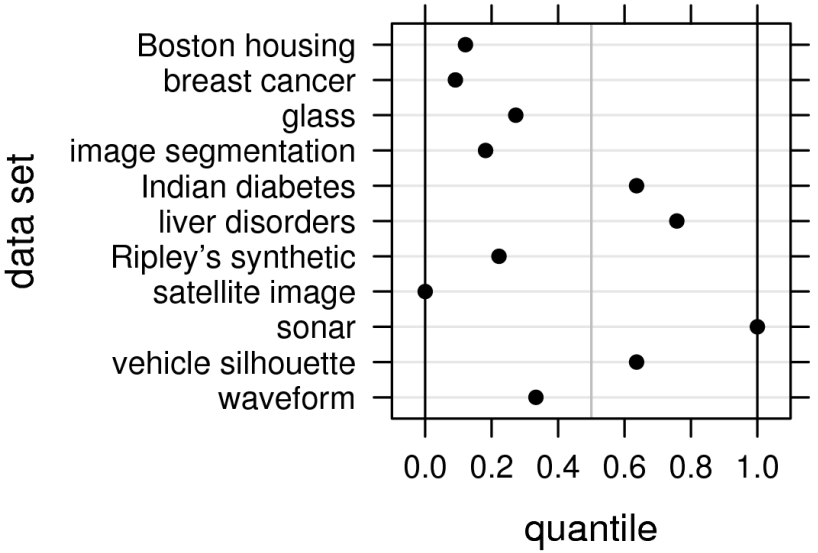
- Uczenie:
  - Minimalizuj błąd średniokwadratowy (MSE) (obliczany za pomocą 10-fold cross-validation) ze względu na parametry wygładzania estymatorów

$$\mathrm{MSE}(\hat{P}(\cdot), \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \sum_{i=1}^{c} (\hat{P}(\omega_i|\mathbf{x}) - \mathbf{t}_i(\mathbf{x}))^2$$

  - Przy minimalizacji używamy pseudo-Newtonowskiego algorytmu L-BFGS-B [Zhu97]
    - Dla 2 estymatorów wybieramy punkt startowy (w przestrzeni poddanej standaryzacji):
      - x0=[1.1, 1]
    - Przedziały poszukiwań współczynników wygładzania [h_min, h_max]:
      - h_max= 99. percentyl odległości między przykładami
      - h_min= 1. percentyl * 1/(promień kuli zawierającej 99% masy rozkładu normalnego)

Wyniki algorytmu na popularnych zbiorach danych

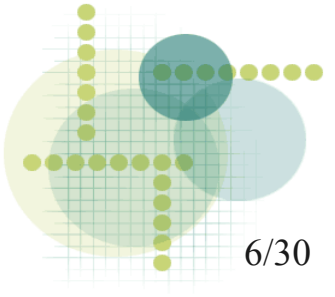
- Dobre wyniki dla wersji E=2 w porównaniu z wynikami literaturowymi:
  - 7/11 wyników należy do 50% najlepszych wśród porównywanej literatury
  - 1 wynik lepszy od najlepszego z porównywanej literatury
  - 1 wynik gorszy od najgorszego z porównywanej literatury



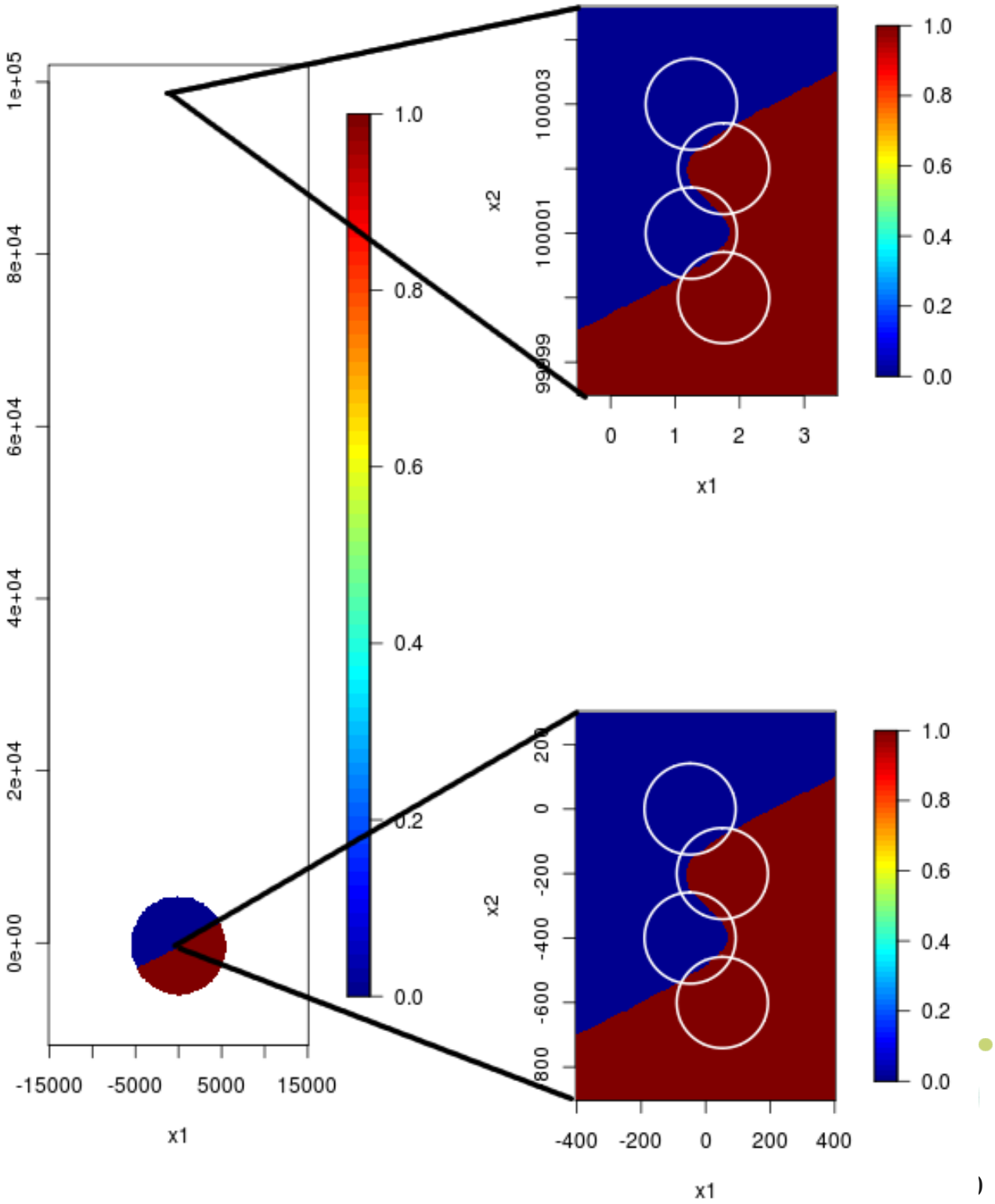
- Na tych zbiorach danych wersja E=2 daje wyniki statystycznie istotnie lepsze niż E=1 (test rangowy Wilcoxsona [Demsar06], p=0.02)

Postulowana własność zbioru danych

- Postulowana własność zbioru danych, na którym wersja z E=2 estymatorami powinna być lepsza od wersji z E=1 estymatorem:
  - Granica decyzyjna powinna przechodzić przez regiony o dużej i małej gęstości
    - Estymator o małym wsp. wygładzania będzie modelował dobrze region o dużej gęstości, a estymator o dużym wsp. wygładzania – o małej gęstości
    - Uśrednianie połączy te 2 modele



Rozkład przykładowego zbioru danych (mieszanina rozkł. Gaussowskich)

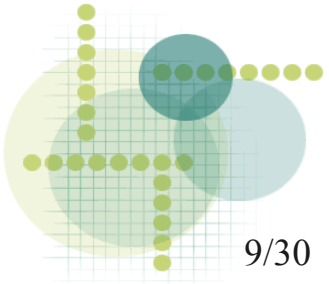


Założenie: wsp. wygładzania równe dla obu klas

- Dla badanego przykładu, optymalne rozwiązanie jest znajdowane w przypadku, gdy wsp. wygładzania dla obu klas są równe
  - Dzięki temu:
    - Zawężamy przestrzeń poszukiwań
    - Funkcja błędu dla przypadku E=2 zależy tylko od 2 wsp. wygładzania – można łatwo narysować
  - Uzasadnienie tej własności:
    - Obie struktury z wykresu są praktycznie odseparowane – można je analizować oddzielnie
    - W obrębie każdej struktury rozkład jednej z klas jest przesuniętą wersją rozkładu drugiej – w tym przypadku stosowanie wspólnego wsp. wygładzania jest uzasadnione [Ghosh04]

Sposób optymalizacji

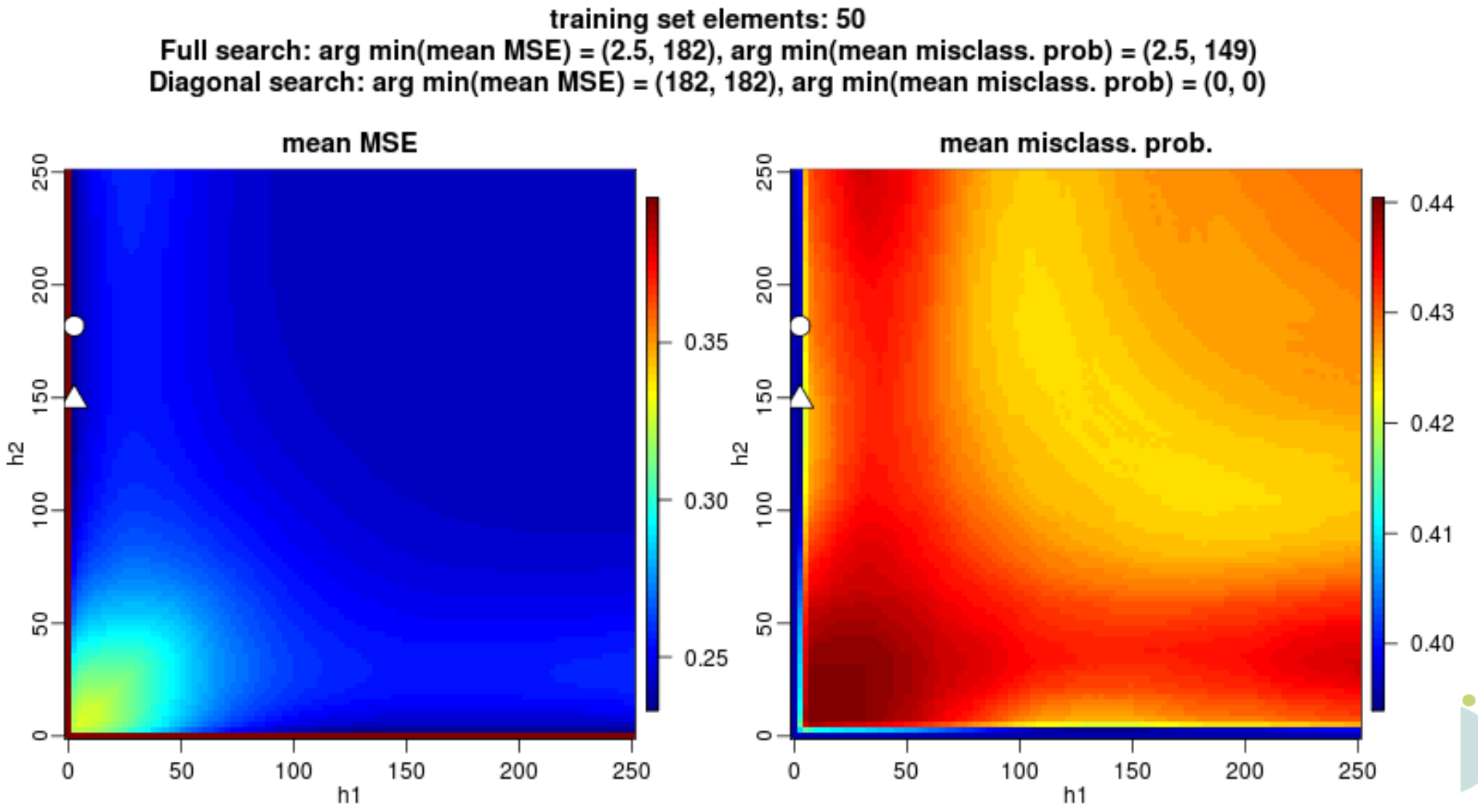
- Chcemy porównać wyniki dla E=1 i E=2 w „najlepszym przypadku” dla każdej z wersji
  - Dobieramy parametry optymalnie przy pomocy zbioru testowego (czyli optymalizujemy na zbiorze testowym)



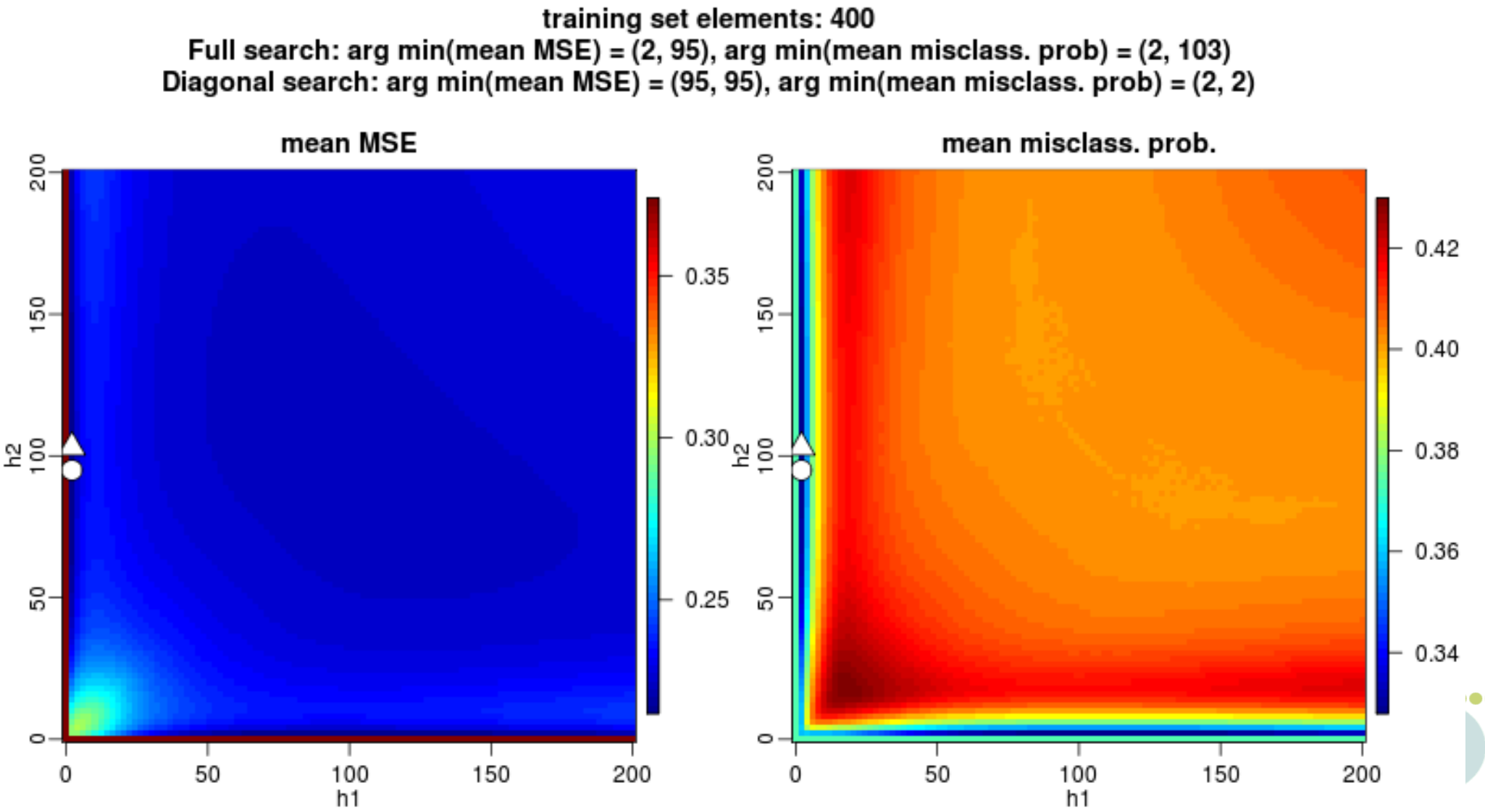
Parametry przykładów

- Wielkość zb. testującego: 2000
- Wielkość zb. uczących: 50-3200
  - Liczba zbiorów uczących, po których uśredniamy: 20
- h_min=0, h_max – dobierane ręcznie dla każdego zb. uczącego (tak, by minimum było w środku przedziału a nie na brzegu)
- Siatka wykresu błędu: 100x100 punktów

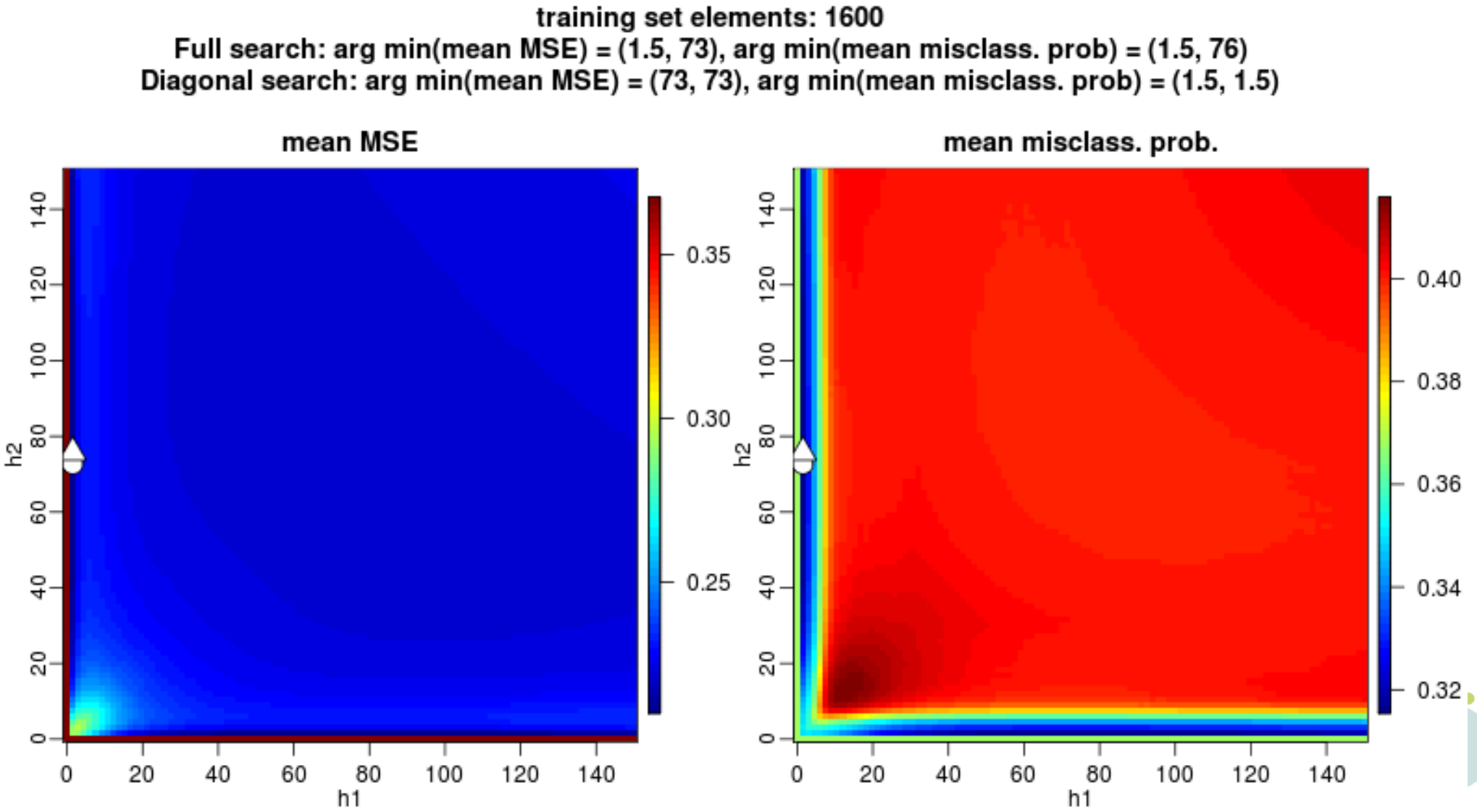
Funkcja błędu



Funkcja błędu

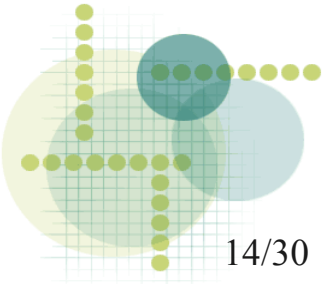


Funkcja błędu

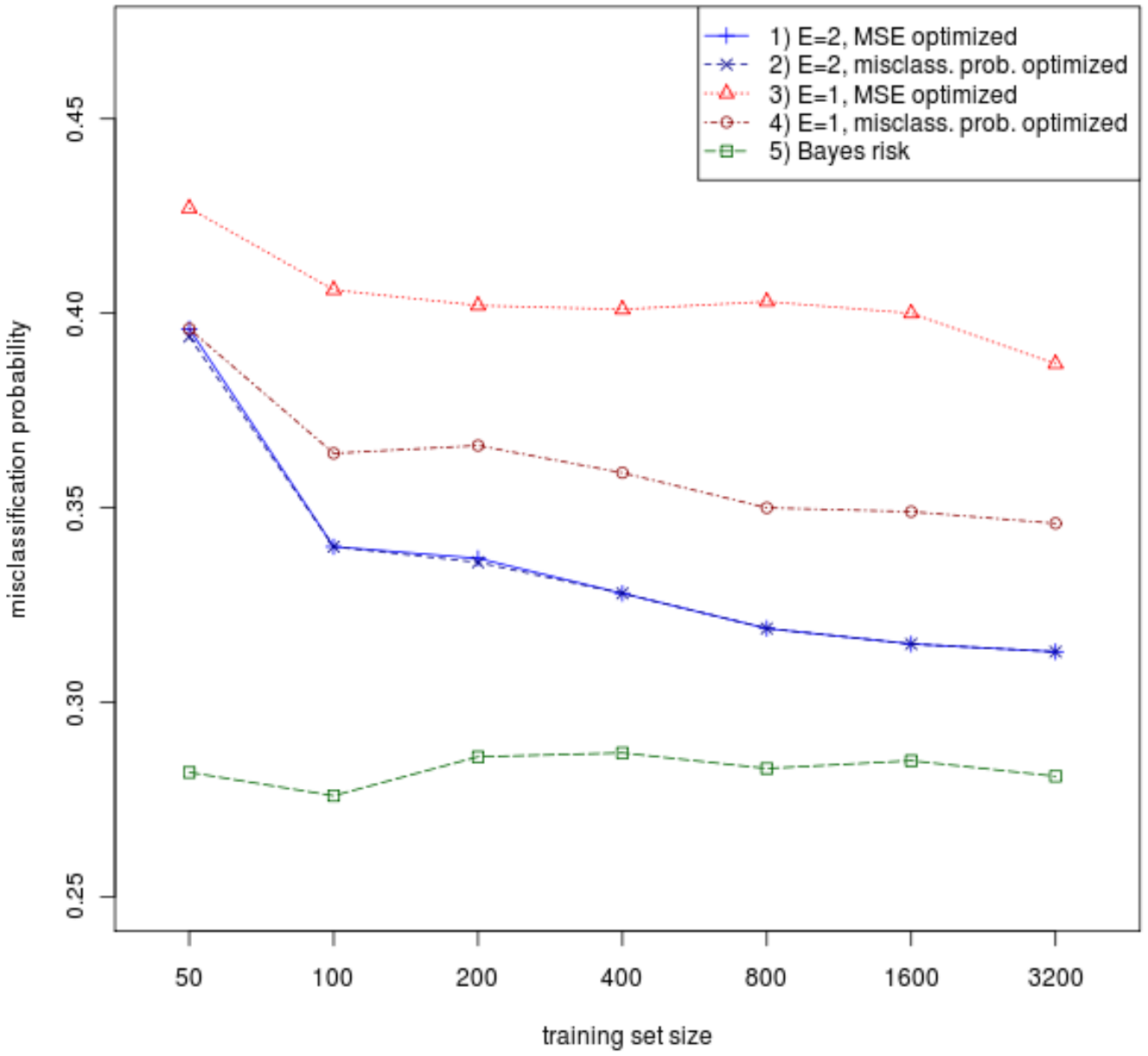


Wnioski z analizy wykresów

- Minimum globalne znajduje się poza przekątną → używanie E=2 estymatorów może dać lepsze wyniki niż E=1 estymatora
- Optymalne rozwiązanie składa się z małego i dużego wsp. wygładzania – potwierdzenie naszej intuicji
- Minimum MSE jest b. blisko minimum błędu klasyfikacji (a podczas właściwego uczenia optymalizujemy właśnie MSE)



Wykres poziomu błędów



Wnioski z analizy wykresu

- Wersja E=2 daje wyniki lepsze od wersji E=1 (poza najmniejszym zb. trenującym)
- Dla wersji E=2, minimalizacja MSE daje w przybliżeniu takie same, dobre wyniki jak bezpośrednia minimalizacja błędu klasyfikacji

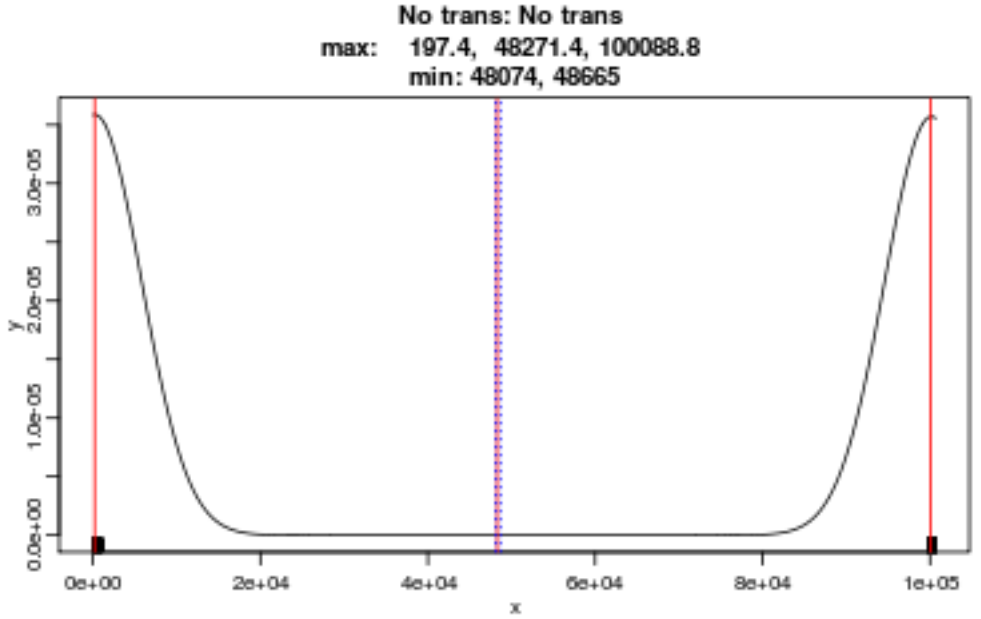
Statystyczna istotność różnic

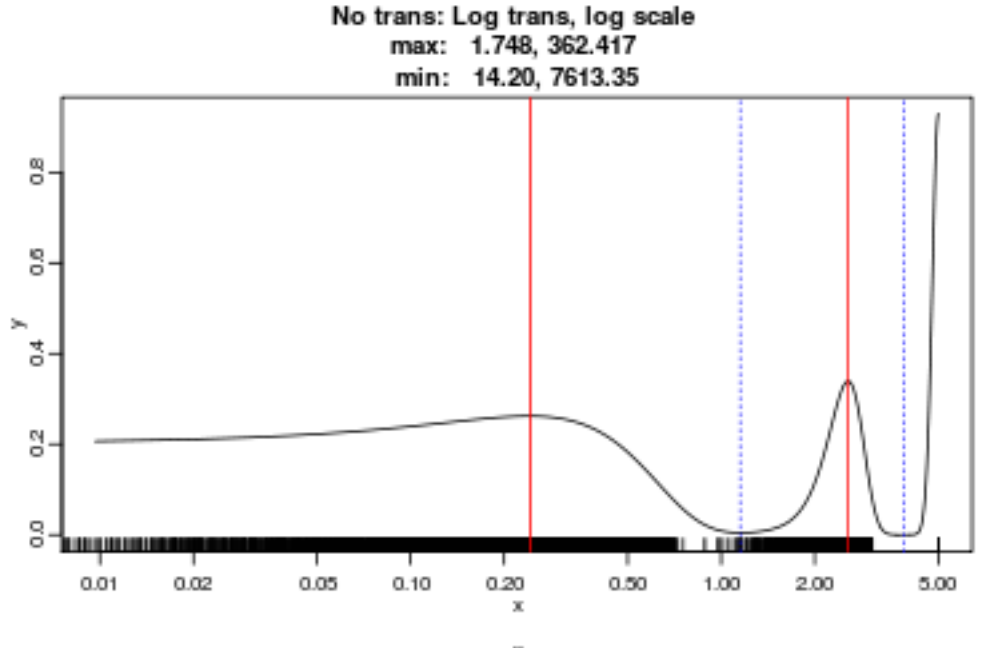
- Gdy porównamy wersję E=2 z E=1 dla minimalizacji różnych funkcji błędu:
  - E=2 (bł. klasyfikacji) vs. E=1 (bł. klasyfikacji)
  - E=2 (MSE) vs. E=1 (bł. klasyfikacji)
  - E=2 (MSE) vs. E=1 (MSE)
- Różnice są statystycznie istotne (paired t-test) dla wszystkich zbiorów uczących oprócz najmniejszego
  - Największe p-value w tych testach to 1.56e-09, więc istotność jest zachowana przy uwzględnieniu, że wykonujemy testy wielokrotne i po wprowadzeniu konserwatywnej poprawki Bonferroni'ego (nowy poziom istotności = poziom istotności/liczba testów = 0.05/20 = 0.0026)

Zastosowanie prawdziwej metody uczenia

- Przy próbie zastosowania opisanej wcześniej metody uczenia pojawia się problem – gradient funkcji błędu w punkcie startowy jest =0
- Pomysł: dobierać punkt startowy bardziej adekwatnie do struktury danych
  - Dokładniej:
    - 1) estymujemy gęstość rozkładu odległości między punktami w zbiorze danych (z logarytmowaniem lub bez, używamy 512 równoodległych punktów),
    - 2) jako punkt startowy wybieramy położenie E maksimów odpowiadającym E najmniejszym położeniom

Rozkład odległości dla wygenerowanych danych

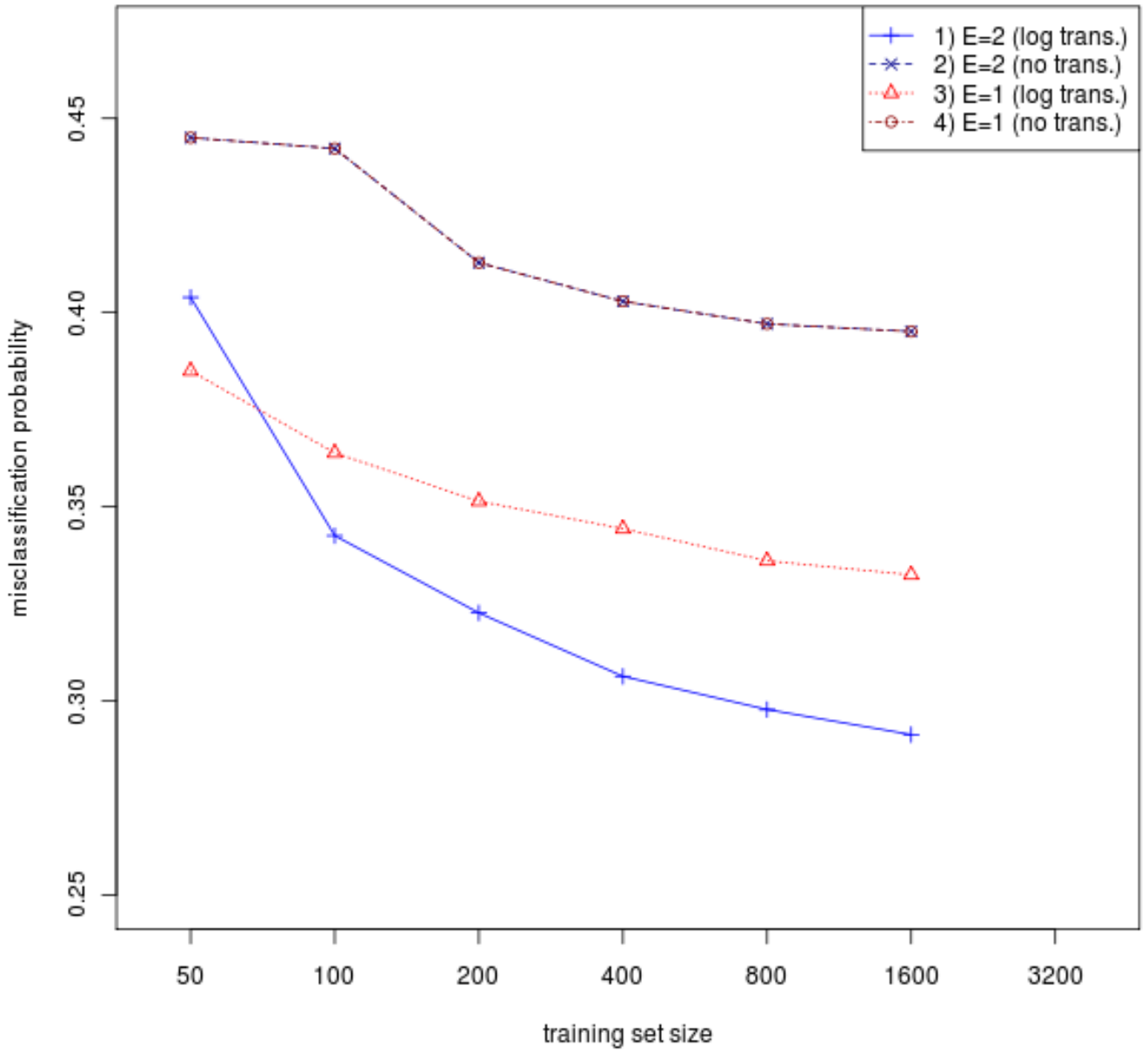




Parametry eksperymentu dla wersji z leave-one-out CV

- Wygenerowane dane:
  - Wielkość zb. testującego: 20000
  - Wielkość zb. uczących: 50-1600
    - Liczba par (zbiór uczący, zbiór testujący), po których uśredniamy dla pojedynczego rozmiaru zb. uczącego: 30
- Wersja algorytmu:
  - Estymacja MSE za pomocą leave-one-out cross-validation
  - Bez transformacji danych
  - h_min=0
  - Dobieranie punktu początkowego na podstawie maksimów w rozkładzie odległości między punktami
    - Z logarytmowaniem
    - Bez logarytmowania
    - Jeśli nie znaleziono liczby maksimów=E, to E jest dopasowywane do liczby znalezionych maksimów

Wykres błędów przy zastosowaniu metody uczenia z LOO CV



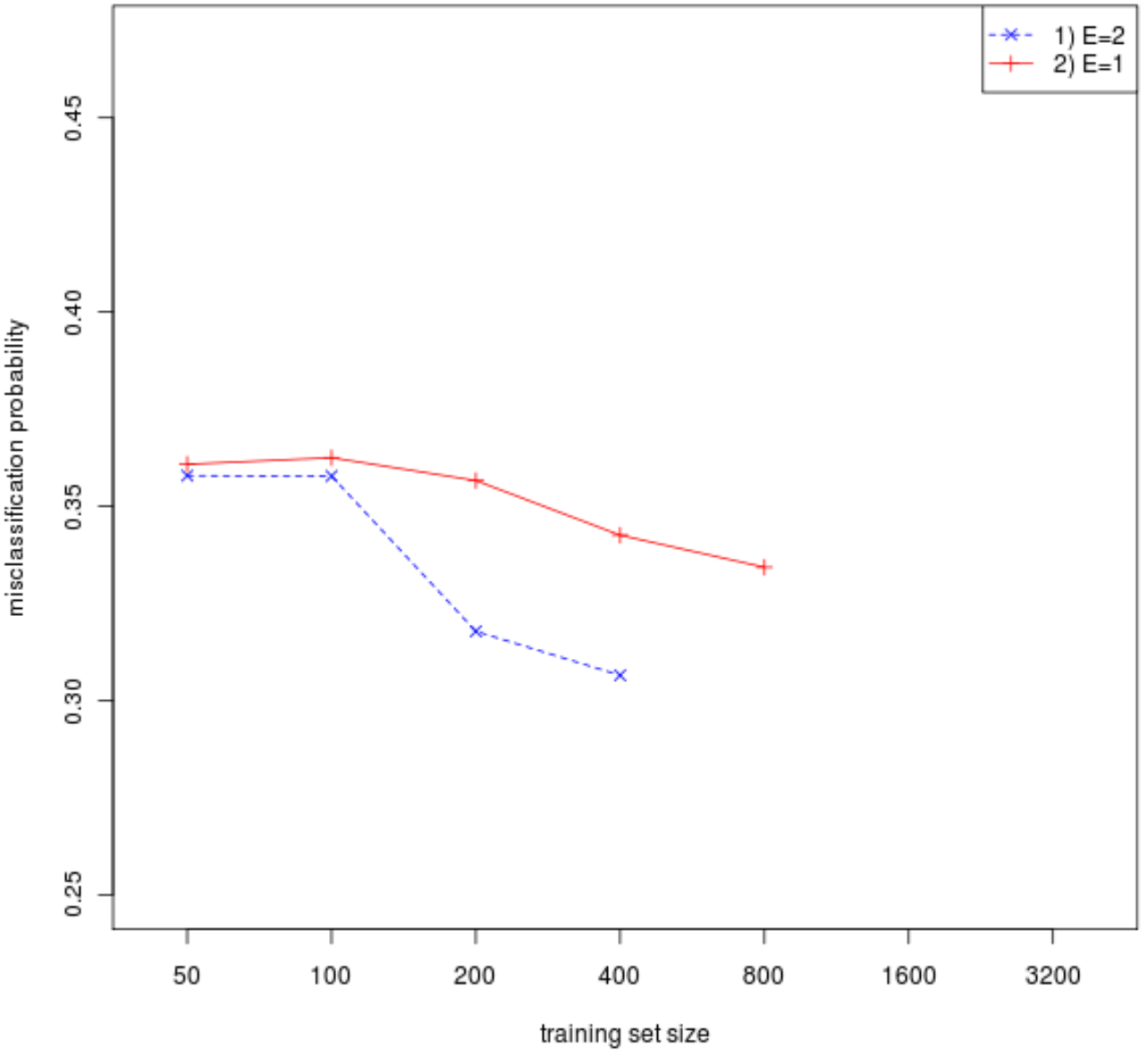
Wnioski z analizy wykresu i statystyczna istotność

- Dla wersji bez logarytmowania: wersja E=2 daje wyniki zbliżone do wersji E=1 (poza najmniejszym zb. trenującym)
  - Różnice nie są statystycznie istotne (paired t-test)
  - Dlaczego? Odp: Bez logarytmowania nie jest wychwytywana różnica między 2 najmniejszymi minimami
- Dla wersji z logarytmowaniem: wersja E=2 daje wyniki lepsze od wersji E=1 (poza najmniejszym zb. trenującym)
  - Różnice są statystycznie istotne (paired t-test) dla wszystkich zbiorów uczących oprócz najmniejszego
    - Największe p-value w tych testach to 5.78e-06, więc istotność jest zachowana przy uwzględnieniu, że wykonujemy testy wielokrotne i po wprowadzeniu konserwatywnej poprawki Bonferroni'ego (nowy poziom istotności = poziom istotności/liczba testów = 0.05/12 = 0.0042)

Parametry eksperymentu dla wersji z 10-fold cross-validation

- Wygenerowane dane:
  - Wielkość zb. testującego: 20000
  - Wielkość zb. uczących: 50-800
    - Liczba par (zbiór uczący, zbiór testujący), po których uśredniamy dla pojedynczego rozmiaru zb. uczącego: 10
      - Dla każdej pary eksperyment powtarzamy 10 razy i uśredniamy
- Wersja algorytmu:
  - Estymacja MSE za pomocą 10-fold cross-validation
  - Bez transformacji danych
  - h_min = 0
  - Dobieranie punktu początkowego na podstawie rozkładu odległości między punktami
    - Z logarytmowaniem

Wykres błędów przy zastosowaniu metody uczenia z 10-fold CV



Wnioski z analizy wykresu i statystyczna istotność

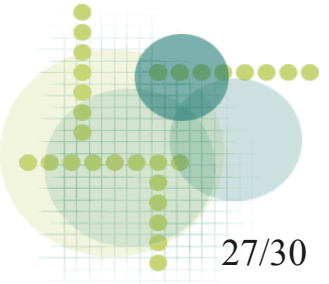
- Wersja E=2 daje wyniki lepsze od wersji E=1 (poza 2 najmniejszymi zb. trenującymi)
- Różnice są statystycznie istotne (paired t-test) dla wszystkich zbiorów uczących oprócz 2 najmniejszych
  - Największe p-value w tych testach to 0.000792, więc istotność jest zachowana przy uwzględnieniu, że wykonujemy testy wielokrotne i po wprowadzeniu konserwatywnej poprawki Bonferroni'ego (nowy poziom istotności = poziom istotności/liczba testów = 0.05/4 = 0.0125)

Pytania i odpowiedzi

- Dlaczego dla najmniejszych zbiorów danych wyniki dla E=2 są słabe?
  - Drugie maksimum (ok. 350) znajduje się zazwyczaj w dużej odległości od optymalnego wsp. wygładzania (ok. 100). W tym miejscu gradient jest zazwyczaj b. mały.
  - Przy małej liczbie punktów, nie ma wystarczającej „siły”, by ściągnąć alg. optymalizacyjny do właściwego optimum.
- Dlaczego wyniki dla E=1 przy optymalizacji MSE są lepsze niż w eksperymentach „najlepszego przypadku”?
  - Punkt startowy optymalizacji dla tego przypadku jest b. mały, co odpowiada minimum błędu klasyfikacji. W trakcie optymalizacji, nie oddalamy się zbyt daleko od tego punktu.

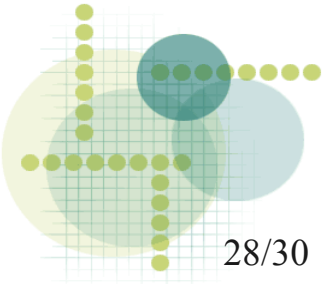
Podsumowanie

- Stosowanie wersji E=2 daje lepsze wyniki niż stosowanie wersji E=1 na sztucznym zbiorze danych o postulowanych właściwościach (poza najmniejszymi zbiorami trenującymi).
  - Przyczyną jest postać funkcji błędu
  - Ta własność zachodzi zarówno dla wersji z zastosowaniem leave-one-out cross-validation jak i 10-fold cross-validation.
- Jeśli wyniki na popularnych zb. danych byłyby dobre dla wersji leave-one-out cross-validation, to należałoby ją stosować
  - Zalety tej wersji:
    - eliminacja losowości z algorytmu
    - prostsza analiza teoretyczna



Główne kierunki badań

- Automatyczne dobieranie liczby estymatorów
- Analiza teoretyczna



Literatura

[Demsar06] Demsar J. „Statistical Comparisons of Classifiers over Multiple Data Sets”, Journal of Machine Learning Research, 2006

[Ghosh04] Ghosh, A.K., Chaudhuri, P.: Optimal smoothing in kernel discriminant analysis. Statistica Sinica 14, 457–483 (2004)

[Zhu97] C. Zhu, R. H. Byrd and J. Nocedal. L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization (1997), ACM Transactions on Mathematical Software, Vol 23, Num. 4, pp. 550 - 560.

Dziękuję za uwagę!