Kombinacja jądrowych estymatorów gęstości w klasyfikacji - własności teoretyczne wraz z testami na sztucznych i referencyjnych zbiorach danych

Mateusz Kobos

Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska

28.02.2011

promotor: prof. Jacek Mańdziuk

Seminarium Metody Inteligencji Obliczeniowej

Spis treści

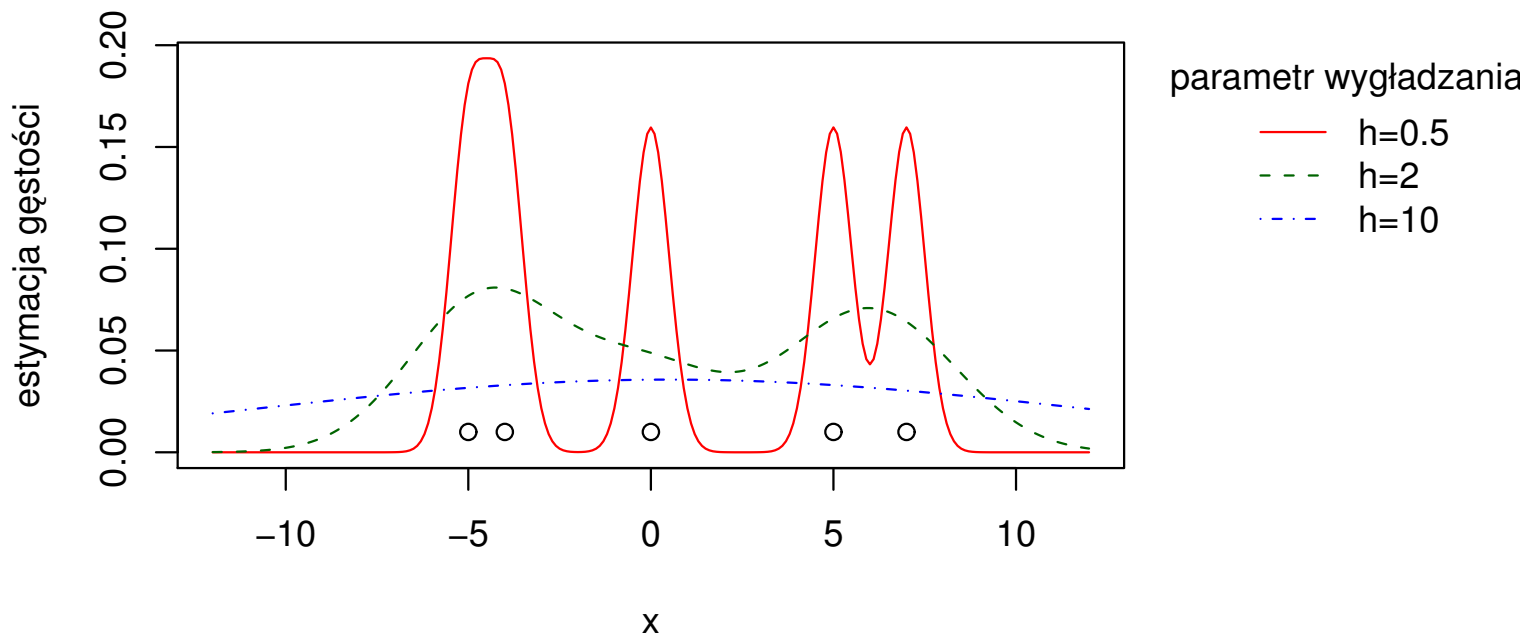
Spis treści

Wprowadzenie

- Klasyfikacja – zadanie automatycznego przypisana etykiety klasy obiektowi na podstawie zbioru uczącego zawierającego obiekty o znanych etykietach klas.
- Podejścia do wydobywania informacji nt wewnętrznej struktury danych:
  - Ekstrakcja i selekcja cech
  - Kwantyzacja wektorowa
  - Dyskretyzacja
  - Przekształcenia liniowe i nieliniowe przestrzeni cech
  - Inne
    - podejście z wykorzystaniem wielu „rozdzielczości” / „skal” spojrzenia na dane

Estymator jądrowy - kernel density estimator (KDE)

- Służy do oszacowania gęstości rozkładu zmiennej losowej na podstawie otrzymanych realizacji.
- Pojedynczy estymator: $\hat{f}(\mathbf{x}; h) = \frac{1}{D} \sum_{\mathbf{x}' \in D} \frac{1}{h^d} \phi(\frac{\mathbf{x}-\mathbf{x}'}{h})$
- Gdzie
  - $\mathbf{x} \in \mathbb{R}^d$
  - $\phi(\cdot)$ – funkcja gęstości np. $\phi(\mathbf{x}) = (2\pi)^{-d/2} exp(-\frac{1}{2}\mathbf{x}^T\mathbf{x})$
  - h – współczynnik wygładzania/szerokość jądra



Idea

- Proponujemy nowe podejście do interpretacji i realizacji wykorzystania pomysłu podejścia „wielorozdzielczościowego” / „wieloskalowego”. W praktyce:
  - Wykorzystujemy średnią jądrowych estymatorów gęstości (kernel density estimators, KDEs).
  - Wartość współczynnika wygładzania każdego z estymatorów odpowiada pojedynczej „rozdzielczości”.
  - Używamy wzoru Bayesa by otrzymać prawdopodobieństwa przynależności do klas.
  - Parametry dobierane tak, by optymalizować jakość klasyfikacji.

Ogólne rozważania

- Czy to podejście do klasyfikacji ma sens? Przecież:
  - Estymacja gęstości jako problem regresji jest trudniejsza niż klasyfikacja[1].
  - Optymalne parametry modelu dla problemu estymacji gęstości nie muszą być optymalne dla klasyfikacji[2].
- Tak, bo optymalizujemy jakość klasyfikacji, ignorujemy jakość estymacji gęstości.
- Dodatkowo, dla skrajnych wartości wsp. wygładzania klasyfikator oparty na KDE zachowuje się jak inne skuteczne algorytmy[3]:
  - mały współczynnik wygładzania: 1-NN
  - duży współczynnik wygładzania: average linkage

---

[1] L. Devroye, L. Gyorfi, G. Lugosi: **A probabilistic theory of pattern recognition**, Springer-Verlag (1996), sect.6.7

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

[3] D. W. Scott: **Multivariate Density Estimation: Theory, Practice, and Visualization**, Wiley, New York (1992), p.251

Spis treści

Literatura

Najbardziej zbliżone algorytmy z literatury można podzielić na 2 grupy:

- zastosowanie kombinacji estymatorów gęstości do optymalizacji jakości estymacji gęstości, np.
  - M. Di Marzio, C. C. Taylor, Boosting kernel density estimates: A bias reduction technique?, **Biometrika** 91 (1) (2004) 226–233.
  - P. Smyth, D. Wolpert, Linearly combining density estimators via stacking, **Machine Learning**, 36 (1999) 59–83.
  - D. J. Marchette, C. E. Priebe, G. W. Rogers, J. L. Solka, Filtered kernel density estimation, **Computational Statistics** 11 (2) (1996) 95–112.
- zastosowanie kombinacji klasyfikatorów opartych na estymacji gęstości do optymalizacji jakości klasyfikacji, np.
  - M. Di Marzio, C. C. Taylor, On boosting kernel density methods for multivariate data: density estimation and classification, **Statistical Methods and Applications** 14 (2005) 163–178.
  - M. Di Marzio, C. C. Taylor, Kernel density classification and boosting: an l2 analysis, **Statistics and Computing** 15 (2005) 113–123.

Literatura

Proponowany algorytm znajduje się „pomiędzy” tymi grupami:

- zastosowanie kombinacji estymatorów gęstości ale z parametrami dobieranymi, by bezpośrednio optymalizować jakość klasyfikacji. Jedyne inne algorytmy, o których nam wiadomo że też należą do tej kategorii to
  -  A. K. Ghosh, P. Chaudhuri, D. Sengupta, Classification using kernel density estimates: Multiscale analysis and visualization, **Technometrics** 48 (1) (2006) 120–132.
  -  A. K. Ghosh, P. Chaudhuri, C. A. Murthy: Multiscale Classification Using Nearest Neighbor Density Estimates, **IEEE Transactions on Systems, Man, and Cybernetics** 36 (5) (2006) 1139-1148.
  - Nasze podejście jest inne i dużo prostsze.

Literatura

Literatura związana z uczeniem maszynowym, w której wykorzystuje się pomysł wielorozdzielczościowego/wieloskalowego podejścia do danych:

- Wykorzystanie przekształcenia falkowego (ang. wavelet transform), np.
  -  G. Sheikholeslami, S. Chatterjee, A. Zhang, WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases, In **Proc. 1998 Int. Conf. Very Large Data Bases**, (1998) 428–439 New York, USA, 1998
- Analiza obrazów i sygnałów, np.
  -  T. Lindeberg: **Scale-space theory in computer vision**, Springer, (1994)
- Inne (podział przestrzeni cech w sposób hierarchiczny, „pyramid match kernel" w SVM w reprezentacji obrazów typu „bag-of-features", ...)

W tej literaturze nie znaleźliśmy rozwiązania podobnego do proponowanego, tj. stosowanie kilku „rozdzielczości", które są automatycznie dostosowywane do danych.

Spis treści

Kontekst literaturowy

- W literaturze nie ma powszechnie przyjętych ogólnych modeli matematycznych, które wyjaśniają, czemu łączenie klasyfikatorów daje dobre wyniki[1].
- W dalszej części zajmiemy się modyfikacją modelu algorytmu uśredniającego pr. a posteriori *n* klasyfikatorów zaproponowanego w
  - K. Tumer, J. Ghosh: Analysis of decision boundaries in linearly combined neural classifiers, **Pattern Recognition** 29 (1996) 341–348.
  - K. Tumer, J. Ghosh: Error correlation and error reduction in ensemble classifiers, **Connection Science** 8 (1996) 385–404.
- i dalej rozwijanego przez autorów w
  - K. Tumer, J. Ghosh: Linear and order statistics combiners for pattern classification, in: A. Sharkey (Ed.), **Combining Artificial Neural Nets**, Springer-Verlag (1999), pp. 127–155.
  - K. Tumer, J. Ghosh, Robust combining of disparate classifiers through order statistics, **Pattern Analysis and Applications** 5 (2002) 189–200.

[1] Kuncheva, **Combining Pattern Classifiers**, Wiley (2004), p.267.

- Cel: obliczyć $\frac{\bar{E}_{\text{add}}}{E_{\text{add}}}$, gdzie:
  - $\bar{E}_{\text{add}}$ – błąd „dodany” klasyfikatora opartego na średniej estymatorów gęstości
  - E_{add} – błąd „dodany” klasyfikatora opartego na pojedynczym estymatorze gęstości
- Rozważmy jednowymiarowy problem klasyfikacji.
  - Chcemy uzyskać regułę decyzyjną $d(x) : \mathbb{R} \to \{1, 2, \ldots, c\}$, gdzie:
    - x – obserwacja/punkt; $1, 2, \ldots, c$ – etykiety klas.
  - Używamy zbioru trenującego t składającego się z n obserwacji (dopasowujemy parametry reguły decyzyjnej do obserwacji z t).
  - t jest realizacją pewnej zmiennej losowej T.
- By otrzymać regułę decyzyjną, można wykorzystać klasyfikator Bayesa:

$$d_B(x) = \arg\max_k p_k(x), \quad \text{gdzie} \quad p_k(x) = \frac{f_k(x) P_k}{f(x)} \ ,$$

gdzie:

- $p_k(x)$ – pr. a posteriori przynależności do klasy k,
- $f_k(x)$ – gęstość klasy k,
- P_k – pr. a priori klasy k,
- $f(x) = \sum_l f_l(x) P_l$ – całkowita gęstość w punkcie x.

- W praktyce nie znamy prawdziwych gęstości klas – musimy użyć estymatorów

$$\hat{d}_B(x) = \arg\max_k \hat{p}_k(x) \quad \text{gdzie} \quad \hat{p}_k(x) = \frac{\hat{f}_k(x) P_k}{\hat{f}(x)} \ ,$$

gdzie:

  - $\hat{p}_k(x)$ – estymator pr. a posteriori przynależności do klasy k,
  - $\hat{f}_k(x)$ – estymator gęstości klasy k,
  - P_k – pr. a priori klasy k,
  - $\hat{f}(x) = \sum_l \hat{f}_l(x) P_l$ – estymator całkowitej gęstości w punkcie x.
- Stosujemy inny model błędu niż w modelu Tumer&Ghosh:
  - Tutaj: $\hat{f}_k(x) = f_k(x) + \varepsilon_k(x)$, gdzie
    - $\varepsilon_k(x)$ – błąd estymacji
  - Tumer&Ghosh: $\hat{p}_k(x) = p_k(x) + \tilde{\varepsilon}_k(x)$

- Rozważmy przedział $[x_1, x_2]$, w którym
  - istnieje dokładnie jedna optymalna granica decyzyjna ozn. x^*.
  - Granica ta oddziela klasy 1 i 2.
  - W takim punkcie $f_1(x^*)P_1 = f_2(x^*)P_2$.
    - (równość „ważonych gęstości")
- Dodatkowe założenia (analogiczne są przyjmowane w modelu Tumer&Ghosh).
  - Każda realizacja T prowadzi do utworzenia dokładnie jednej estymowanej granicy decyzyjnej ozn. x_b leżącej w przedziale $[x_1, x_2]$. Granica ta oddziela klasy 1 i 2.
    - (zachodzi $\hat{f}_1(x_b)P_1 = \hat{f}_2(x_b)P_2$).
  - Punkt x_b powstaje z przesunięcia punktu x^*.

- Zauważmy, że
  - błąd klasyfikacji klasyfikatora = nieredukowalny błąd Bayesa + błąd dodany klasyfikatora
- Błąd dodany ma tutaj postać

$$A(b) = \int_{x*}^{x^*+b} (p_2(x) - p_1(x)) f(x)\, dx \ ,$$

gdzie

  - $x_b = x^* + b$, tj. b – różnica między punktem optymalnej i estymowanej granicy decyzyjnej.
- Po podstawieniu wzoru Bayesa otrzymujemy

$$\begin{aligned} A(b) &= \int_{x^*}^{x^*+b} \frac{f_2(x)P_2 - f_1(x)P_1}{f(x)} f(x)\, dx \\ &= \int_{x^*}^{x^*+b} (f_2(x)P_2 - f_1(x)P_1)\, dx \ . \end{aligned}$$

- Do tej pory rozważaliśmy sytuację dla pojedynczej realizacji t zmiennej losowej T. Jednak w praktyce jesteśmy zainteresowani wartością oczekiwaną błędu dodanego (liczoną po wszystkich realizacjach T) postaci

$$E_{\text{add}} \stackrel{df}{=} E(A(B)) = \int_{-\infty}^{\infty} A(b) f_B(b)\, db \ ,$$

gdzie

  - b – realizacja odpowiedniej zmiennej losowej B,
  - f_B – gęstość B.

- Tutaj punkt estymowanej granicy decyzyjnej jest zdefiniowany jako $X_b = x^* + B$.

- Rozważania podobne do tych w pracy Tumer&Ghosh prowadzą do wzoru na oczekiwany błąd dodany
  - przy założeniu, że błędy $\varepsilon_1(X_b)$ i $\varepsilon_2(X_b)$ nie są skorelowane
  - i przy przybliżeniu wartości gęstości w pobliżu x^* funkcją liniową

$$E_{\text{add}} = \frac{1}{2s}((P_1\beta_1 - P_2\beta_2)^2 + P_1^2\sigma_1^2 + P_2^2\sigma_2^2) \ ,$$

gdzie

  - $\beta_k = E(\varepsilon_k(X_b))$ jest obciążeniem estymatora tj. wartością oczekiwaną błędu estymatora dla klasy k po wszystkich estymowanych punktach granicy decyzyjnej,
  - $\sigma_k^2 = \mathrm{Var}(\varepsilon_k(X_b))$ jest wariancją błędu estymatora dla klasy k po wszystkich estymowanych punktach granicy decyzyjnej.

- Rozważmy model z uśrednianiem:

$$\bar{f}_k(x) = \frac{1}{N_k} \sum_{i=1}^{N_k} \hat{f}_{k,i}(x) \ ,$$

gdzie

  - N_k – liczba estymatorów dla klasy k,
  - $\hat{f}_{k,i}(x) = f_k(x) + \varepsilon_{k,i}(x)$ – estymacja gęstości klasy k dokonywana przez i-ty estymator przypisany tej klasie,
  - $\varepsilon_{k,i}(x)$ – błąd tego estymatora.
  - $\bar{\varepsilon}_k(x) = 1/N_k \sum_{i=1}^{N_k} \varepsilon_{k,i}(x)$ – błąd modelu z uśrednianiem
- Powtarzając poprzednie rozumowanie otrzymujemy
  - przy założeniu, że błędy $\bar{\varepsilon}_1(\bar{X}_b)$ i $\bar{\varepsilon}_2(\bar{X}_b)$ nie są skorelowane

$$\bar{E}_{\text{add}} = \frac{1}{2s}((P_1\bar{\beta}_1 - P_2\bar{\beta}_2)^2 + P_1^2\bar{\sigma}_1^2 + P_2^2\bar{\sigma}_2^2) \ ,$$

gdzie

  - $\bar{\beta}_k = E(\bar{\varepsilon}_k(\bar{X}_b)) = \frac{1}{N_k} \sum_{i=1}^{N_k} E(\varepsilon_{k,i}(\bar{X}_b))$
  - $\bar{\sigma}_k^2 = \frac{1}{N_k^2} \sum_{i=1}^{N_k} \sigma_{k,i}^2 + \frac{1}{N_k^2} \sum_{i=1}^{N_k} \sum_{j \neq i} \rho_{k,i,j}\sigma_{k,i}\sigma_{k,j}$, gdzie
    - $\rho_{k,i,j}$ – korelacja między $\varepsilon_{k,i}(\bar{X}_b)$ i $\varepsilon_{k,j}(\bar{X}_b)$,
    - $\sigma_{k,i}^2 = \mathrm{Var}(\varepsilon_{k,i}(\bar{X}_b))$.

- Mając wzory na błędy oczekiwane dla obu modeli, możemy przeanalizować iloraz

$$\frac{\bar{E}_{\text{add}}}{E_{\text{add}}} = \frac{(P_1\bar{\beta}_1 - P_2\bar{\beta}_2)^2 + P_1^2\bar{\sigma}_1^2 + P_2^2\bar{\sigma}_2^2}{(P_1\beta_1 - P_2\beta_2)^2 + P_1^2\sigma_1^2 + P_2^2\sigma_2^2},$$

- Rozważmy sytuację, w której
  - wyrażenia odpowiadające obciążeniu w obu modelach są równe zero, tj. $P_1\beta_1 - P_2\beta_2 = 0$ oraz $P_1\bar{\beta}_1 - P_2\bar{\beta}_2 = 0$,
  - liczba estymatorów na klasę jest równa tj. $N_1 = N_2 = N$,
- Dla uproszczenia załóżmy, że
  - możemy dobrać estymatory w modelu z uśrednianiem tak, by ich średnia wariancja była równa wariancji pojedynczego estymatora, tj. $1/N \sum_{i=1}^{N} \sigma_{k,i}^2 = \sigma_k^2$.
- Przy dodatkowym założeniu, że korelacja między błędami różnych estymatorów w modelu z uśrednianiem jest zerowa, tj. dla $k \in \{1, 2\}$ oraz $i, j \in \{1, \ldots, N\}$ mamy $\rho_{k,i,j} = 0$
  - (co może być w przybliżeniu spełnione w przypadku estymatorów zwracających bardzo różne estymacje)

- otrzymujemy analogiczny wynik do tego w pracach Tumer&Ghosh

$$\frac{\bar{E}_{\text{add}}}{E_{\text{add}}} = \frac{1}{N} \ .$$

- W praktyce otrzymanie tak dobrego wyniku byłoby prawie niemożliwe,
  - np. zapewnienie zerowej korelacji między błędami byłoby trudne nawet dla dwóch estymatorów.
- Uzyskana intuicja co do pożądanych własności klasyfikatora opartego na średniej estymatorów gęstości:
  - wyrażenia związane z obciążeniem estymatora $P_k\bar{\beta}_k$ w obu klasach powinny przyjmować zbliżone wartości,
  - korelacja między błędami różnych estymatorów powinna być jak najmniejsza.

Spis treści

- Każdy algorytm klasyfikacyjny ma 2 tryby działania:
  - klasyfikacja/odtwarzanie/zastosowanie modelu
  - nauka/trenowanie/dobór parametrów modelu

- By klasyfikować punkt $\mathbf{x} \in \mathbb{R}^d$ wykorzystujemy wzór Bayesa

$$\bar{d}_B(\mathbf{x}) = \arg\max_k \bar{p}_k(\mathbf{x}) \quad \text{gdzie} \quad \bar{p}_k(\mathbf{x}) = \frac{\bar{f}_k(\mathbf{x}) P_k}{\bar{f}(\mathbf{x})} \ ,$$

- gdzie zamiast pojedynczej estymacji gęstości (tak jak w standardowym podejściu) używamy średniej

$$\bar{f}_k(\mathbf{x}) = \frac{1}{N_k} \sum_{i=1}^{N_k} \hat{f}_{k,i}(\mathbf{x}) \ .$$

- Jako pojedynczego estymatora gęstości dla danej klasy używamy najpopularniejszego estymatora nieparametrycznego, mianowicie KDE postaci

$$\hat{f}_{k,i}(\mathbf{x}; h_{k,i}) = \frac{1}{|\mathcal{D}_k|} \sum_{\mathbf{x}' \in \mathcal{D}_k} \frac{1}{(h_{k,i})^d} \phi\Big(\frac{\mathbf{x} - \mathbf{x}'}{h_{k,i}}\Big) \ ,$$

gdzie

- $h_{k,i}$ – współczynnik wygładzania i-tego estymatora jądrowego związanego z klasą k,
- $\mathcal{D}_k$ – zbiór obserwacji, które należą do klasy k,
- $\phi : \mathbb{R}^d \to [0, \infty)$ – funkcja gęstości zwaną „funkcją jądrową”.

- funkcja jądrowa $\phi(\mathbf{x})$ – gęstość rozkładu normalnego
- Uproszczenia:
  - ten sam współczynnik wygładzania dla każdego wymiaru
  - w funkcji jądrowej przyjmujemy $\mathbf{\Sigma} = \mathbf{I}$, więc $\phi(\mathbf{x}) = (2\pi)^{-d/2} \exp(-1/2\mathbf{x}^T\mathbf{x})$
    - podejście częściowo uzasadnione, bo dokonujemy wstępnej transformacji przestrzeni cech (używamy standaryzacji)

Nauka – ogólnie

Celem etapu nauki jest znalezienie optymalnej kombinacji estymatorów jądrowych. Etap ten można podzielić na następujące kroki.

1. Losowo permutuj instancje w zbiorze uczącym.
   - Tego kroku wymaga metoda 10-krotnej walidacji krzyżowej, którą używamy w dalszej części.
2. Transformuj dane.
   - Domyślnie jako transformacji używa się standaryzacji.
3. Dobierz punkt startowy algorytmu optymalizacyjnego i ograniczenia optymalizacyjne.
4. Iteracyjnie minimalizuj funkcję błędu.

Jako wynik etapu nauki, otrzymujemy zestaw optymalnych (lokalnie) parametrów wygładzania.

Minimalizacja funkcji błędu

- Minimalizujemy błąd średniokwadratowy (MSE, Mean Squared Error), który jest szacowany przy pomocy metody 10-krotnej walidacji krzyżowej. Funkcja ta dla pojedynczego zbioru walidacyjnego jest zdefiniowana jako

$$\mathrm{MSE}(\mathcal{D}^v, \mathbf{h}) = \frac{1}{|\mathcal{D}^v|} \sum_{\mathbf{x} \in \mathcal{D}^v} \sum_{k=1}^{c} (\hat{p}_k(\mathbf{x}; \mathbf{h}) - \mathbf{t}_k(\mathbf{x}))^2 \ ,$$

  gdzie
  - $\mathcal{D}^v$ – walidacyjny zbiór danych,
  - $\mathbf{h}$ – wektor składający się z wartości wszystkich współczynników wygładzania,
  - $\hat{p}_k(\mathbf{x}; \mathbf{h})$ – estymacja pr. a posteriori klasy k,
  - $\mathbf{t}_k(\mathbf{x})$ – wektor, którego k-ty element, gdzie k odpowiada prawdziwej klasie k, jest równy 1 a wszystkie inne elementy są równe 0.
- Używamy funkcji MSE zamiast bezpośrednio optymalizować funkcję błędu klasyfikacji ze względu na różniczkowalność MSE (wymagane przez stosowany algorytm optymalizacyjny). To samo podejście jest stosowane np. w algorytmie MLP.

Minimalizacja funkcji błędu

- Używany algorytm optymalizacyjny: L-BFGS-B[1]
  - pseudo-Newtonowski, wykorzystuje wartość funkcji i jej gradient
  - uwzględnia proste ograniczenia typu $x > a$ nałożone na zmienne
  - jak w każdym algorytmie iteracyjnym, należy ustalić punkt startowy
- Dobieramy w pewien sposób przedział przeszukiwanych sensownych wartości współczynników wygładzania.

[1] R. H. Byrd, P. Lu, J. Nocedal, C. Zhu, A limited memory algorithm for bound constrained optimization, **SIAM Journal on Scientific and Statistical Computing** 16 (1995) 1190–1208.

Punkt startowy algorytmu optymalizacyjnego

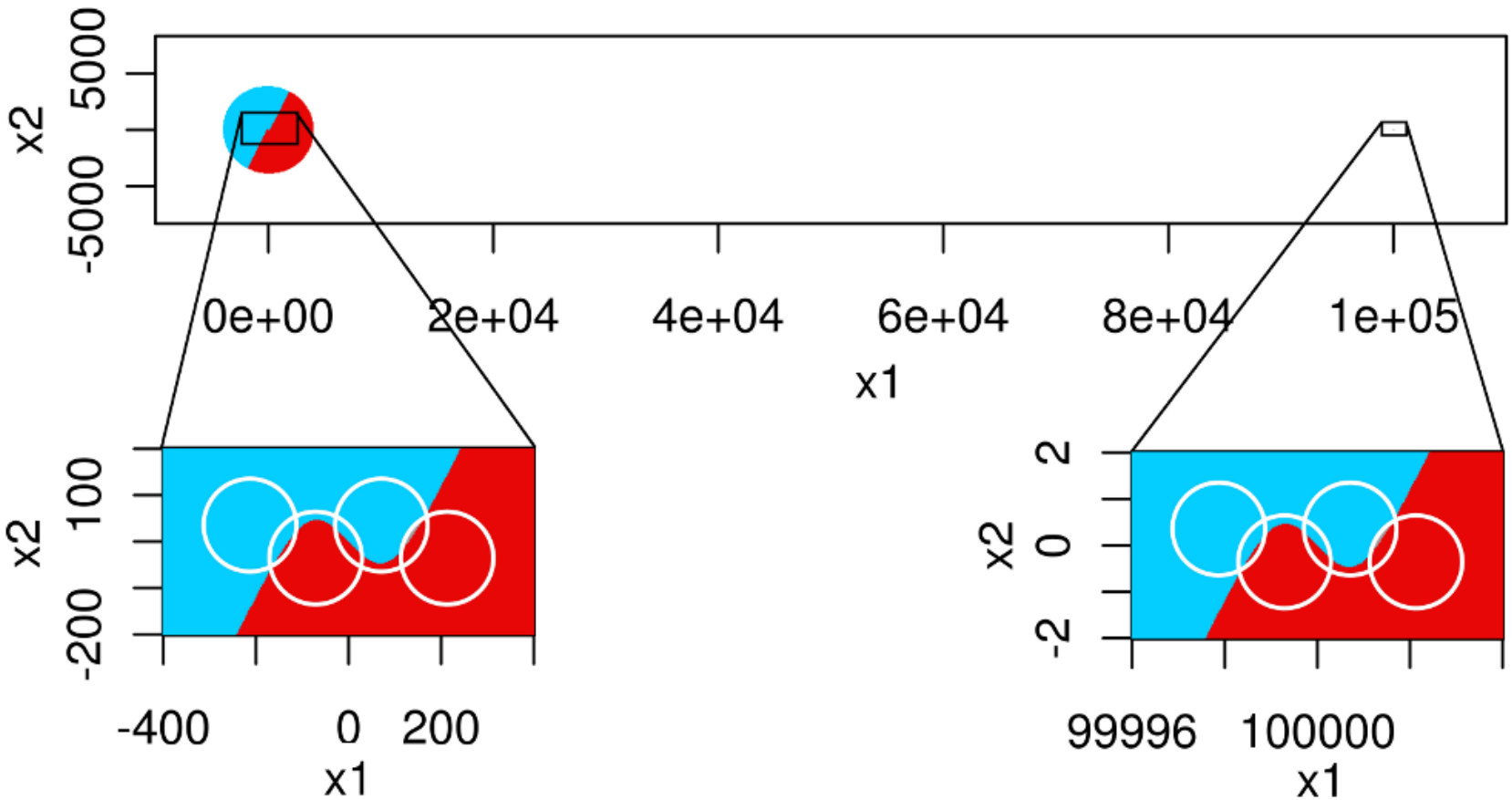
Stosujemy 2 heurystyczne metody doboru punktu startowego:

- „oparta na odchyleniu standardowym”
  - W przypadku dwóch estymatorów na klasę, dla każdej klasy wybieramy punkt startowy leżący w pobliżu punktu (σ, σ), gdzie σ jest odchyleniem standardowym uśrednionym po wszystkich wymiarach dla punktów danej klasy.
- „oparta na rozkładzie odległości”
  - analizujemy gęstość rozkładu zlogarytmowanych odległości między punktami należącymi do danej klasy w zbiorze danych.

Spis treści

- Intuicja: by jak najlepiej wykorzystać podejście wielorozdzielczościowe, badany problem klasyfikacyjny musi mieć naturę wielorozdzielczościową, tj. dane muszą być dobrze aproksymowane przy użyciu różnych rozdzielczości.
- By potwierdzić tą intuicję, zaproponowaliśmy zbiory danych posiadające wspomnianą własność i wykonaliśmy na nich eksperymenty.

- Rozważmy dwuwymiarowy „bazowy” problem klasyfikacji binarnej zdefiniowany poprzez określenie funkcji gęstości dla każdej z klas.
  - 2 skupienia generowane przez mieszaninę rozkładów gaussowskich (GMM):
    - skupienie o małej gęstości ($\sigma_1 = 100$)
    - skupienie o dużej gęstości ($\sigma_2 = 1$)



Podczas projektowania tego problemu chcieliśmy osiągnąć cele:

- nietrywialna granica decyzyjna (tutaj: kształt „fali"),
- występowanie obszarów, gdzie KDE o b. różnych współczynnikach wygładzania dobrze modelują granicę decyzyjną,
- optymalne wyniki są uzyskiwane w przypadku równych współczynników wygładzania dla obu klas.
  - Można to uzyskać poprzez:
    - odseparowanie skupień i zapewnienie, że
    - w każdym ze skupień rozkład jednej klasy jest przesuniętą wersją rozkładu drugiej klasy (por.[1]).
  - Dzięki temu:
    - zawężenie przestrzeni poszukiwań optymalnych współczynników,
    - w przypadku stosowania dwóch wsp. wygładzania na klasę: błąd klasyfikacji zależy od 2 parametrów – łatwo go wizualizować.

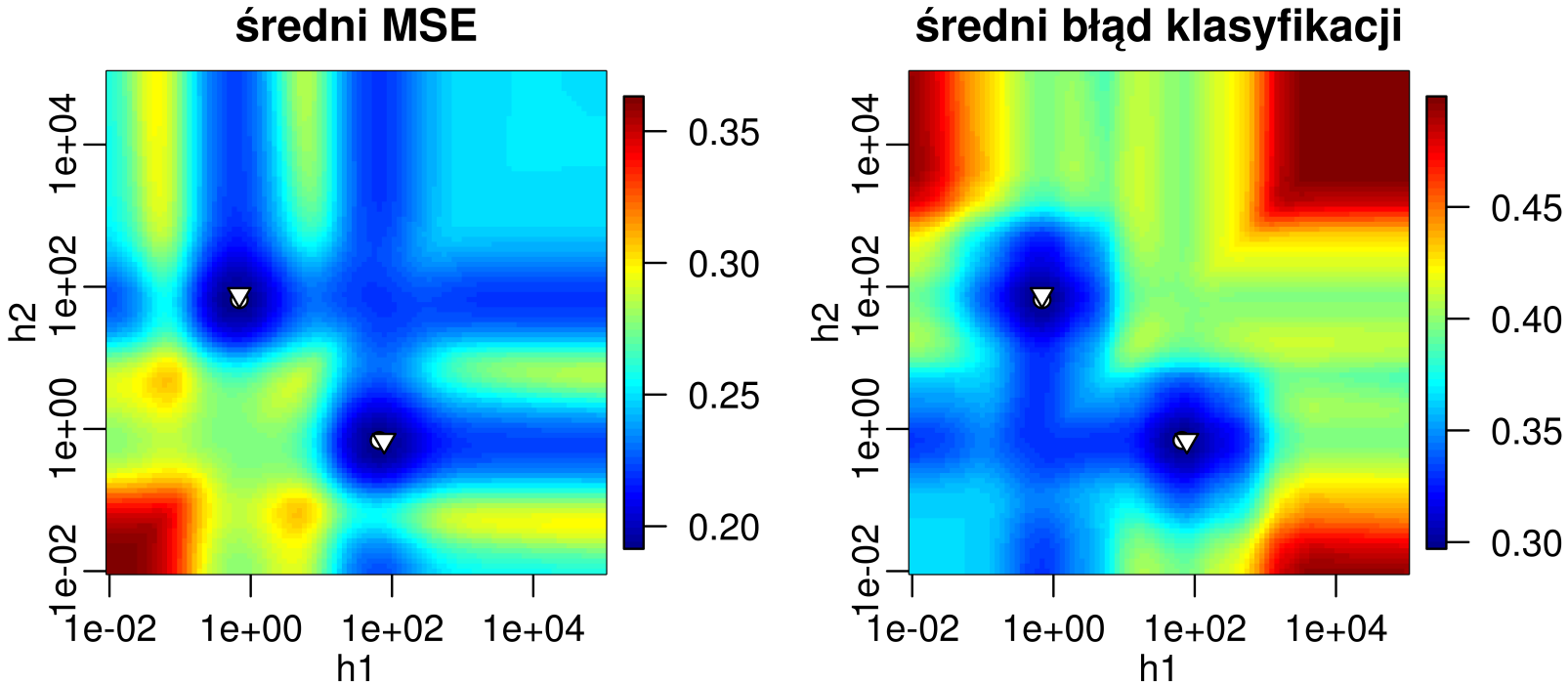
---

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

Przeprowadzaliśmy 2 rodzaje eksperymentów:

- „optymistyczny”
  - współczynniki wygładzania dobierane optymalnie na zbiorze testowym
    - za pomocą „brutalnej” metody grid search
  - wykonywany dla problemów, gdzie rozwiązanie optymalne składało się z równych współczynników wygładzania dla obu klas.
  - Dla każdego badanego parametru eksperymentu, eksperyment był wielokrotnie powtórzony, a wynik uśredniony
    - (dla każdego parametru eksperymentu wygenerowaliśmy 20 zbiorów trenujących (każdy z nich o domyślnym rozmiarze 400 punktów), stosowano 1 zbiór testowy o rozmiarze 2 000).
- „realistyczny”
  - współczynniki wygładzania dobierane na zbiorze treningowym
    - za pomocą algorytmu optymalizacyjnego L-BFGS-B
  - Dla każdego parametru eksperymentu, eksperyment był wielokrotnie powtórzony, a wynik uśredniony
    - (dla każdego parametru eksperymentu wygenerowaliśmy 20 par (zbiór trenujący, zbiór testowy), dla każdej z par eksperyment był powtarzany 10 razy, zbiór trenujący zawierał domyślnie 400, punktów zbiór testowy zawierał 20 000 punktów).

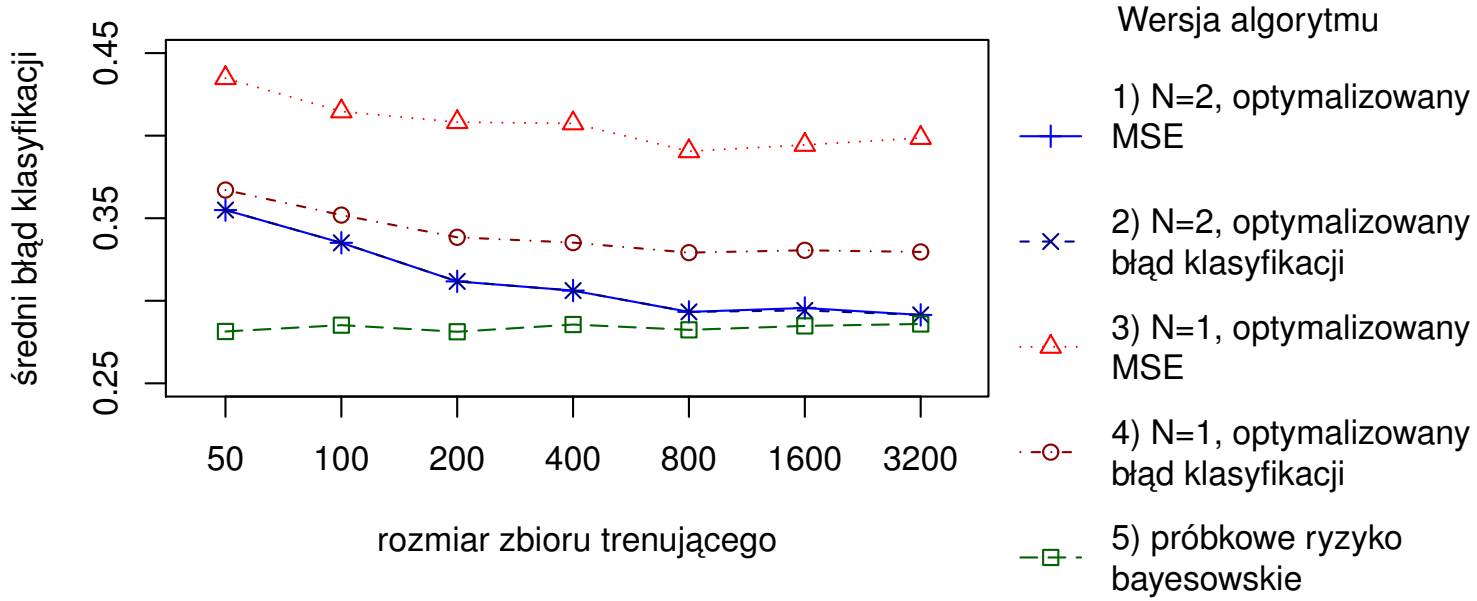
Przykładowa funkcja błędu dla bazowego problemu



Zauważmy:

- podstawowy algorytm ($N = 1$) może osiągnąć jedynie te wartości funkcji błędu, które leżą na przekątnej
- położenie optimów jest zgodne z naszymi oczekiwaniami
- ciekawostka: minima MSE znajdują się prawie w tych samych punktach co minima błędu klasyfikacji

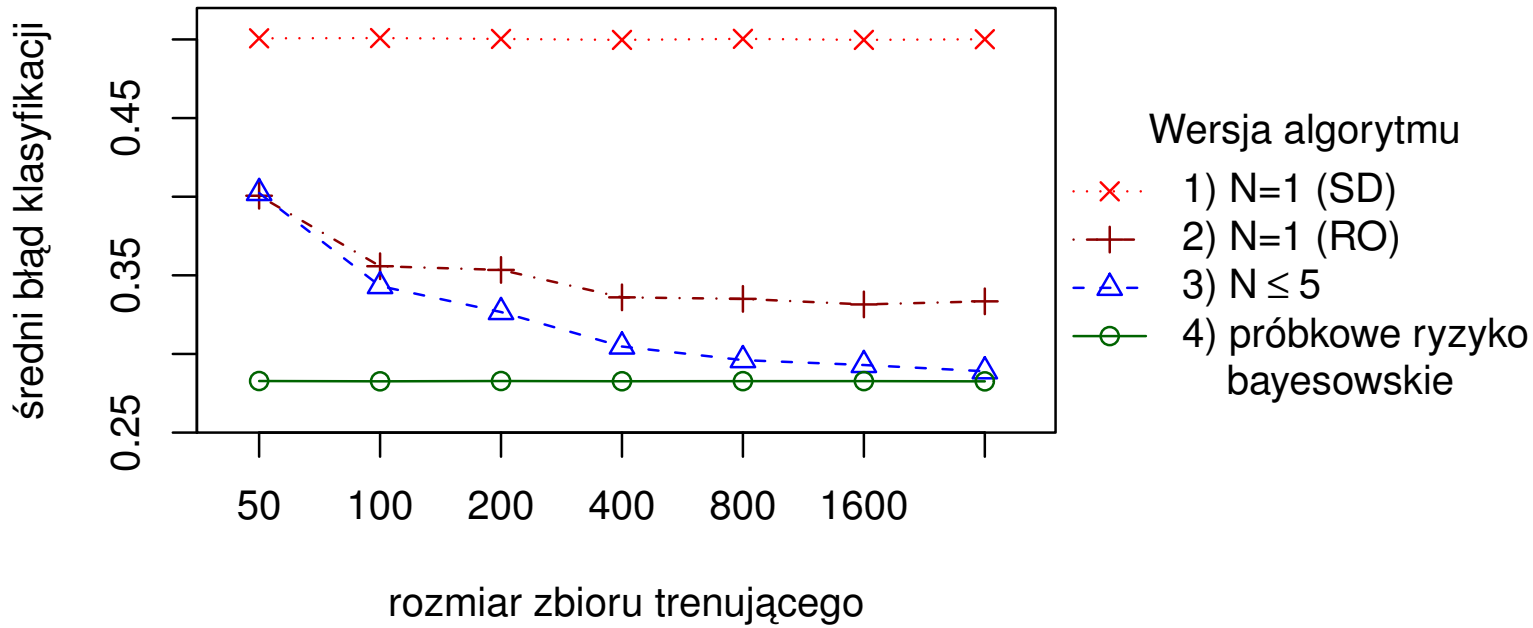
Eksperyment „optymistyczny” dla bazowego problemu



Analiza wyników

- $N = 2$ uzyskuje statystycznie istotnie lepsze wyniki niż $N = 1$ (test t dla prób zależnych, $p \leqslant 0.0004$ dla liczby punktów większej niż 100)
- $N = 2$ zbiega szybciej do optymalnej wartości błędu niż $N = 1$ (prawdopodobnie dlatego, że stosujemy odpowiedni model do problemu).
- dla $N = 2$ nie ma znaczącej różnicy między optymalizacją MSE i błędu klasyfikacji (test t dla prób zależnych, $p > 0.05$)

Eksperyment „realistyczny” dla bazowego problemu



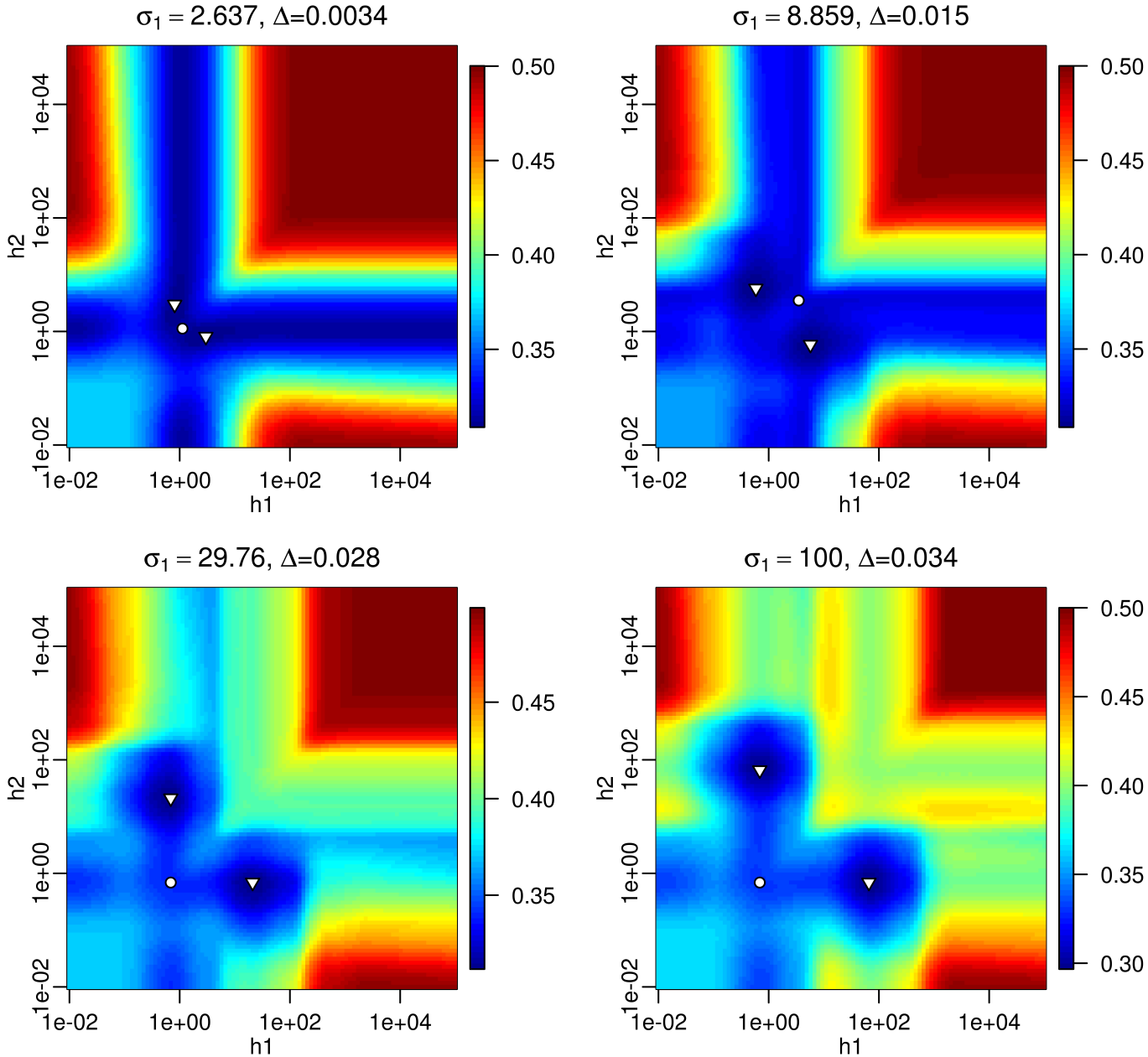
Analiza wyników

- $N \leqslant 5$ uzyskuje statystycznie istotnie lepsze wyniki niż $N = 1$ (RO) (test t dla prób zależnych, $p > 0.0037$ dla liczby punktów większej niż 50)
- $N \leqslant 5$ zbiega szybciej do optymalnej wartości błędu niż $N = 1$ (RO)

Eksperyment dla różnych parametrów skali

- Potwierdziliśmy, że proponowany algorytm dobrze działa na zbiorze danych z dwoma skupieniami o bardzo różnych gęstościach.
  - Pytanie: jak zmienia się efektywność algorytmu, gdy zmienia się stosunek gęstości w obu skupieniach?
  - Odpowiedź: będziemy zmieniali wartość parametru skali większego skupienia σ_1 (parametr skali mniejszego skupienia będzie stały $\sigma_2 = 1$).

Eksperyment „optymistyczny” dla różnych param. skali



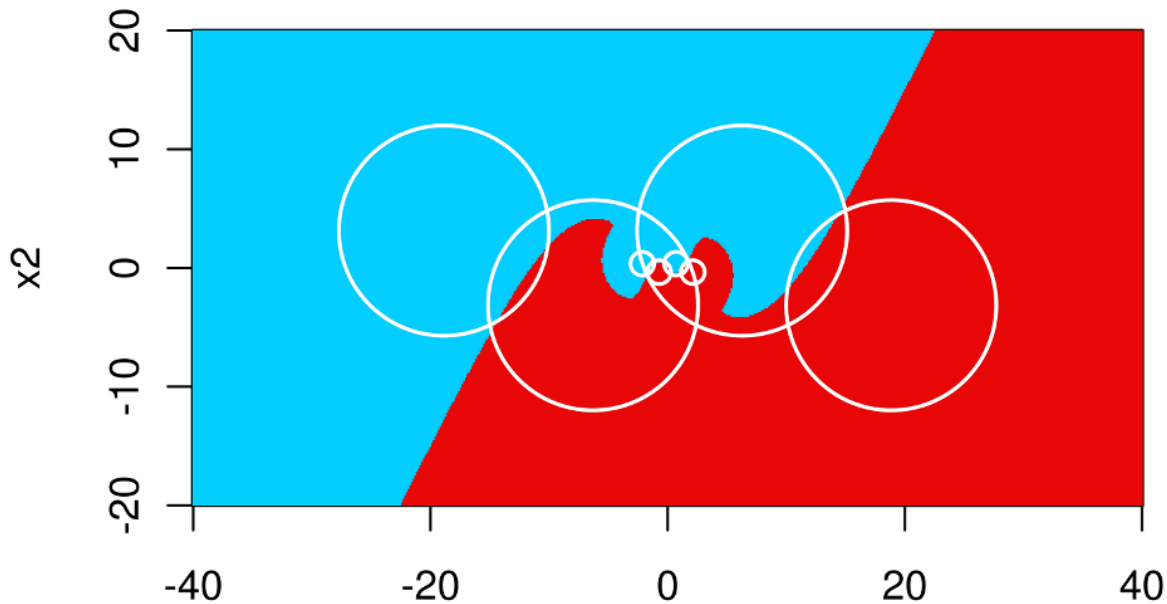
Eksperyment dla różnych parametrów skali

Analiza poziomu błędów (analogicznie do poprzedniej) przy zmieniającym się parametrze skali pozwala wyciągnąć wnioski:

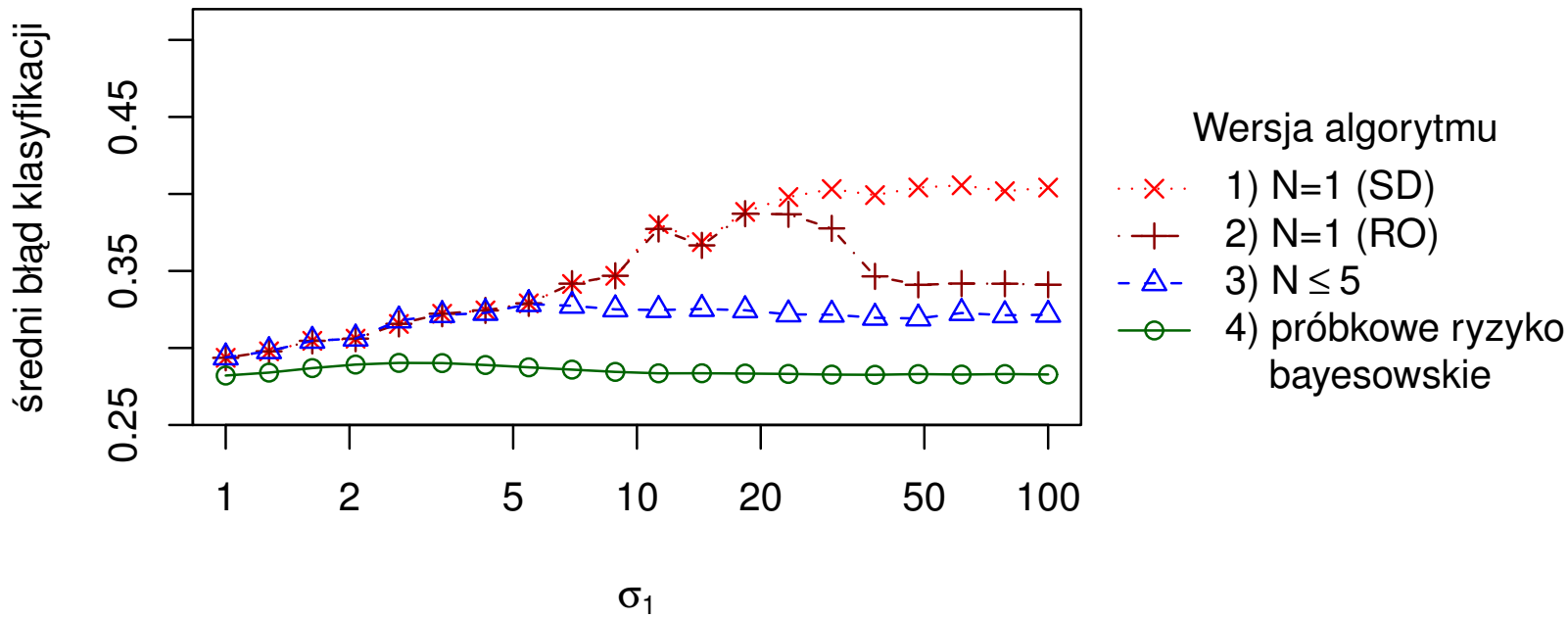
- Im, większe różnica między parametrami skali, tym większa przewaga proponowanego algorytmu nad wersją bazową
- Dla eksperymentu „realistycznego" wersja $N \leqslant 5$ jest statystycznie istotnie lepsza od wersji $N = 1$ (RO) dla $\sigma_1 \geqslant 11.29$ (test t dla prób zależnych, $p < 7.97 \cdot 10^{-7}$).

Eksperyment dla wycentrowanych skupień z różnymi param. skali

- Zajmowaliśmy się problemami, w których skupienia o różnych gęstościach były odseparowane.
  - Pytanie: czy otrzymane wyniki będą podobne jeśli separacja nie będzie tak silna?
  - Odpowiedź: będziemy badali problem, gdzie środki skupień znajdują się w tym samym punkcie. Będziemy zmieniali parametr skali większego skupienia σ_1 (parametr skali mniejszego skupienia będzie stały $\sigma_2 = 1$).



Eksperyment „realistyczny” dla wycentrowanych skupień z różnymi param. skali



- Wyniki podobne do poprzednich, ale przewaga wersji $N \leqslant 5$ nad wersją podstawową $N = 1$ (RO) nie jest tak duża jak poprzednio.
- Wersja $N \leqslant 5$ jest statystycznie istotnie lepsza od wersji $N = 1$ (RO) dla $\sigma_1 \geqslant 8.86$ (test t dla prób zależnych, $p < 8.93 \cdot 10^{-4}$).

Spis treści

- Algorytm był testowany na 19 popularnych zbiorach danych.
  - Dane te nie były dobierane pod względem posiadania jakichkolwiek specjalnych własności, które sprzyjałyby proponowanemu algorytmowi.
- Przetestowaliśmy dwa rodzaje algorytmu
  - $N = 1$ – wersja bazowa z 1 estymatorem jądrowym na klasę
  - $N = 2$ – wersja z 2 estymatorami jądrowymi na klasę
  - transformacja danych: standaryzacja
  - obliczanie punktu startowego: metoda „oparta na odchyleniu standardowym”
- Eksperymenty „realistyczne”
- Cele eksperymentów:
  - Czy używanie zaledwie dwóch różnych estymatorów jądrowych da wyniki lepsze niż w algorytmie bazowym?
  - Czy proponowany algorytm jest konkurencyjny względem innych metod opisanych w literaturze?

Używane zbiory danych

nazwa	klasy	atrybuty	instancje	zb. treningowy	zb. testowy
blood transfusion	2	5	748		
Boston housing	3	13	506		
breast cancer	2	9	683		
ecoli	8	7	336		
glass	6	5	214		
heart	2	44	267	80	187
image segmentation	7	19	2310		
Indian diabetes	2	7	532		
ionosphere	2	34	351		
iris	3	4	150		
liver disorders	2	6	345		
Ripley's synthetic	2	2	1250	250	1000
satellite image	6	36	6435	4435	2000
sonar	2	20	208	104	104
vehicle silhouette	4	18	846		
vowel Deterding	11	10	990	528	462
waveform	3	21	3600	600	3000
wine	3	13	178		
yeast	10	8	1484		

Porównanie z wersją bazową

- Porównaliśmy wersje algorytmu: $N = 2$ z $N = 1$.
- Dla każdej pary: algorytm i zbiór danych, wykonano 10-krotnie powtórzony eksperyment 10-krotnej walidacji krzyżowej by uzyskać wynik średni.
- Proponowany algorytm $N = 2$ osiągnął statystycznie istotnie lepsze wyniki niż wersja podstawowa $N = 1$ (rangowy test Wilcoxsona[1], $p \approx .02$, przedział ufności: $[.0007, .0067]$).

[1] J. Demsar: Statistical comparisons of classifiers over multiple data sets, **Journal of Machine Learning Research** 7 (2006) 1–30.

Porównanie z wersją bazową

Tabela: Porównanie błędów klasyfikacji (pr. błędnej klasyfikacji)

zb. danych	$N = 1$	$N = 2$
blood transfusion	.2231	.2202
Boston housing	.2340	.2323
breast cancer	.0328	.0315
ecoli	.1357	.1324
glass	.2592	.2906
heart	.2087	.2094
image segmentation	.0385	.0357
Indian diabetes	.2458	.2433
ionosphere	.0729	.0500
iris	.0547	.0473
liver disorders	.4060	.3684
Ripley's synthetic	.0959	.1001
satellite image	.0875	.0876
sonar	.1467	.1386
vehicle silhouette	.2926	.2906
vowel Deterding	.0144	.0146
waveform	.1607	.1586
wine	.0365	.0252
yeast	.3969	.3963

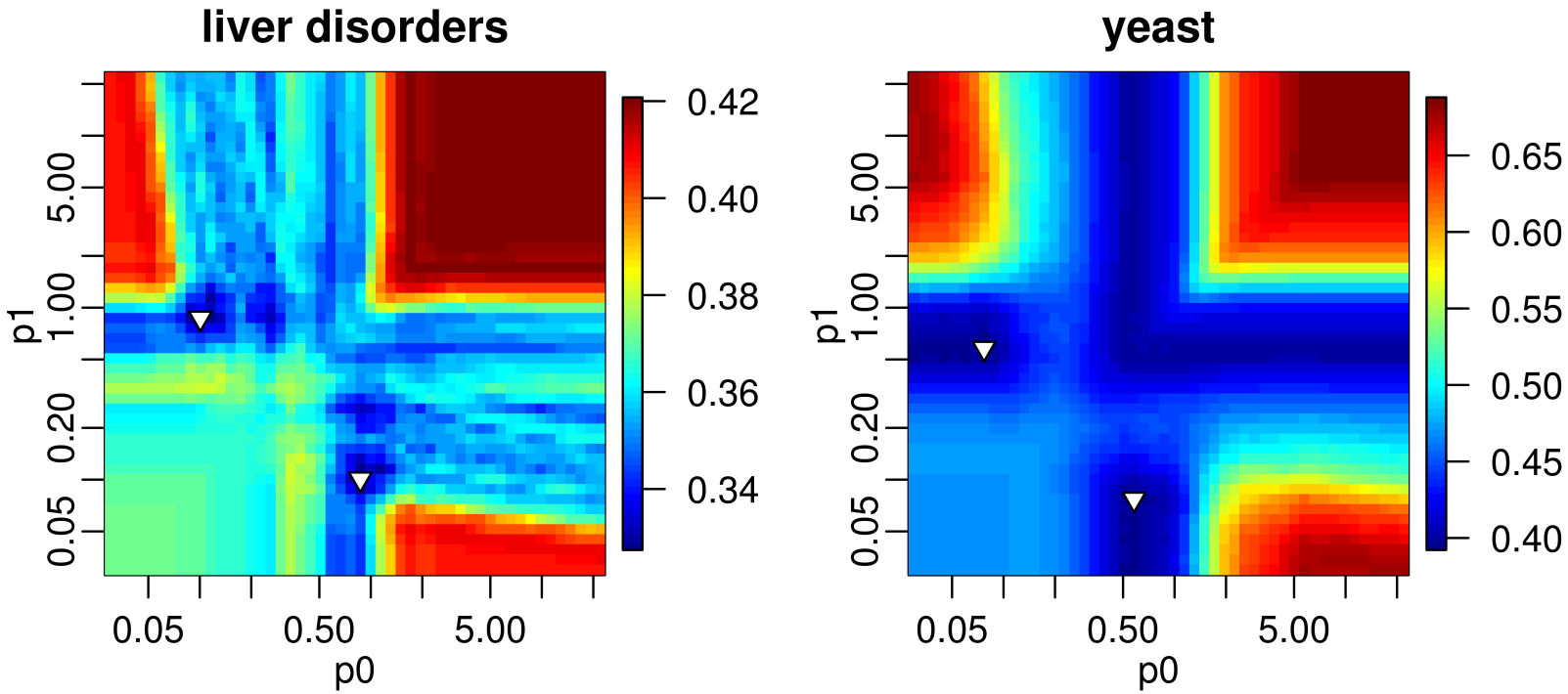
Porównanie z wersją bazową – dyskusja

Różnica między błędami klasyfikacji jest względnie duża dla części zbiorów danych (*liver disorders*, *ionosphere*, *wine*, *glass*), a dla innych praktycznie znika (*satellite image*, *vowel Deterding*, *yeast*, *heart*).

- Prawdopodobna przyczyna: różna struktura geometryczna zbiorów danych.
  - Część zbiorów danych może mieć własność wielorozdzielczości, która sprzyja algorytmowi wielorozdzielczościowemu, inne dane mogą nie mieć tej własności.
  - Nawet jeśli dane mają tą własność, różnica między gęstościami skupień występujących w danych może nie być wystarczająco duża.

Przykład postaci funkcji błędu klas. dla dwóch różnych zbiorów

- Uśredniony błąd klasyfikacji na zbiorze testowym (5 razy powtórzony eksperymenty 10-krotnej walidacji krzyżowej).
- Wersja $N = 2$, gdzie współczynnik wygładzania jest taki sam dla wszystkich klas.



Komentarz:

- *liver disorders* – widoczne minima globalne poza przekątną
- *yast* – minima na przekątnej i poza nią mają podobną wartość

Porównanie z wersją bazową – dyskusja

Dla części zbiorów danych wyniki wersji $N = 2$ są gorsze niż wyniki $N = 1$. A przecież przestrzeń rozwiązań $N = 1$ jest zawarta w przestrzeni rozwiązań $N = 2$. Więc czemu tak się dzieje?

- Prawdopodobne przyczyny:
  - Przeuczenie – dane nie mają charakterystyki, do której dopasowujemy parametry modelu.
    - (mimo stosowania walidacji krzyżowej)
  - Stosujemy parę przybliżeń:
    - Optymalizacja lokalna zamiast globalnej
    - Optymalizacja MSE zamiast błędu klasyfikacji
    - Optymalizacja na zb. testowym zamiast na treningowym

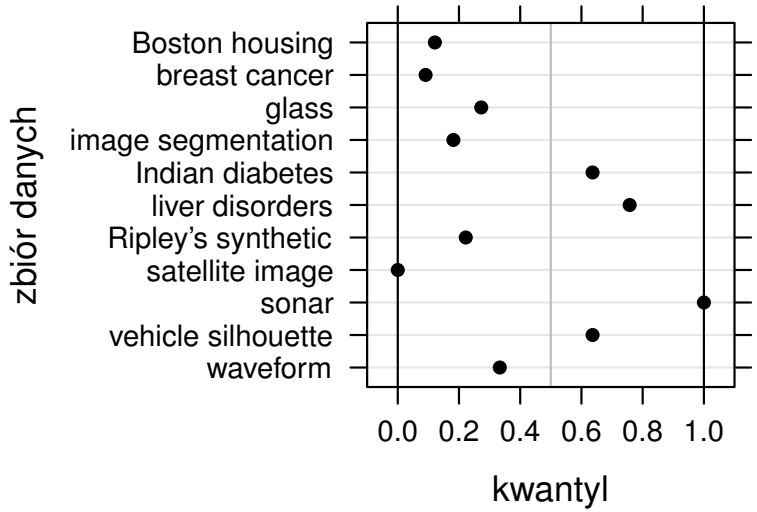
Porównanie z wynikami literaturowymi

Algorytm $N = 2$ porównaliśmy z wynikami umieszczonymi w:

-  T.-S. Lim, W.-Y. Loh, Y.-S. Shih, A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, **Machine Learning** 40 (2000) 203–228.
  - Zawiera wyniki 33 algorytmów klasyfikacyjnych (wśród nich: CART, LDA, QDA, 1-NN z odległością Mahalanobisa, sieć neuronowa LVQ, radialna sieć neuronowa (RBF neural network)) przetestowanych na różnych zbiorach danych.
-  A. K. Ghosh, P. Chaudhuri, D. Sengupta, Classification using kernel density estimates: Multiscale analysis and visualization, **Technometrics** 48 (1) (2006) 120–132.
  - Zawiera wyniki nowego algorytmu opartego na estymatorach jądrowych wraz z kilkoma wynikami literaturowymi.

Postępowaliśmy zgodnie z metodologią opisaną w artykułach, poza tym, że każdy eksperyment był powtarzany 10 razy, by otrzymać lepszą estymację błędu.

Porównanie z wynikami literaturowymi



Rysunek: Porównanie wyników algorytmu w wersji $N = 2$ z wynikami literaturowymi. Każdy punkt odpowiada pozycji kwantylowej eksperymentu w porównaniu z wynikami literaturowymi.

Komentarz:

- Algorytm konkurencyjny w porównaniu z innymi popularnymi klasyfikatorami.
- 7/11 wyników otrzymanych przez nasz model znajduje się w przedziale 50% najlepszych wyników.

Spis treści

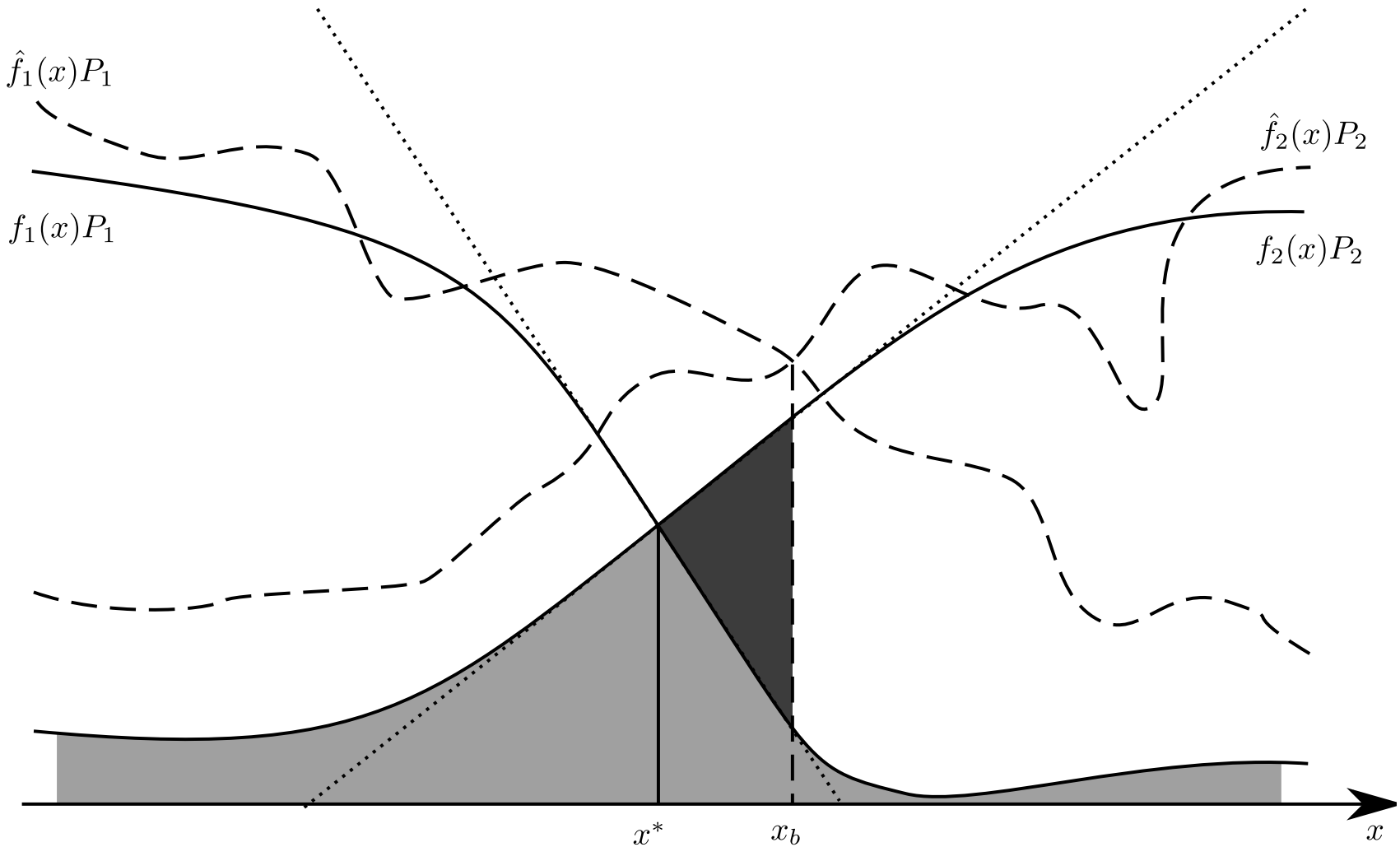
Podsumowanie

- Przedstawiliśmy nowe podejście do klasyfikacji opartej na „wielorozdzielczościowym” spojrzeniu na dane.
- Pokazaliśmy formalnie, że, pod pewnymi warunkami, ogólna klasa tego typu metod może charakteryzować się niskimi błędami klasyfikacji.
- Algorytm dobrze działa na sztucznych zbiorach danych o specyficznej własności „wielorozdzielczości” – dane składają się ze skupień o różnej gęstości.
  - Im większa różnica między gęstościami, tym lepsze wyniki.
- W eksperymentach na zbiorach referencyjnych (bez doboru pod względem specyficznych własności) przewaga podejścia wielorozdzielczościowego nad wersją podstawową jest mała. Mimo to, algorytm dobrze radzi sobie w porównaniu z innymi klasyfikatorami.

Publikacje

-  M. Kobos, J. Mańdziuk: Classification Based on Multiple-Resolution Data View, **Lecture Notes in Computer Science** 6354 (20th International Conference on Artificial Neural Networks), Springer, 124–129, 2010
-  M. Kobos, J. Mańdziuk: Classification Based on Combination of Kernel Density Estimators, **Lecture Notes in Computer Science** 5769 (19th International Conference on Artificial Neural Networks), Springer, 125–134, 2009
-  M. Kobos: Combination of Independent Kernel Density Estimators in Classification, **International Multiconference on Computer Science and Information Technology**, 4th International Symposium Advances in Artificial Intelligence and Applications, Mrągowo, Polska, 2009, 57–63
-  M. Kobos: Classification based on combination of two kernel density estimators, **4th International PhD Students and Young Scientists Conference: Young scientists towards the challenges of modern technology**, Warszawa, Polska, 2009, 327–333

Rysunek: błąd dodany



Rysunek: obszar jasnoszary – błąd Bayesa (ryzyko bayesowskie), obszar ciemnoszary – błąd dodany, linia ciągła – ważona gęstość, linia przerywana – estymowana ważona gęstość, linia kropkowana – liniowa aproksymacja ważonej gęstości w pobliżu x^*