

Zadanie 3 - Inteligentna Analiza Danych

Adam Zambrzycki
Nr indeksu: 216933

Konrad Stępiak
Nr indeksu: 216892

7 czerwca 2019

Kierunek Informatyka
Rok akademicki 2018/19
Semestr 4
Grupa dziekańska 2

Symbol α będzie oznaczał współczynnik nauki, a K liczbę centrów. Współczynnik skalujący w sieci będzie nazywany sigma.

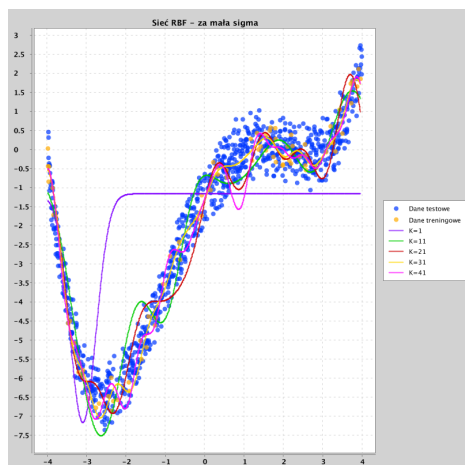
1 Osobna nauka warstw - Aproksymacja

Do nauki wykorzystano następujące parametry: $\alpha = 0.05$, liczba iteracji = 20000. Optymalną sigmę uzyskano ze wzoru:

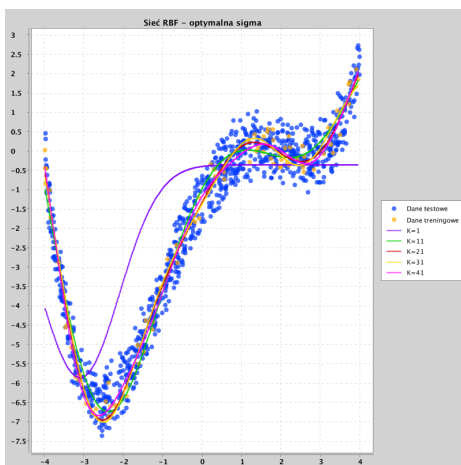
$$\sigma = \frac{d}{\sqrt{2M}} \quad (1)$$

Gdzie d - maksymalna odległość między centrami, a M to liczba centrów.

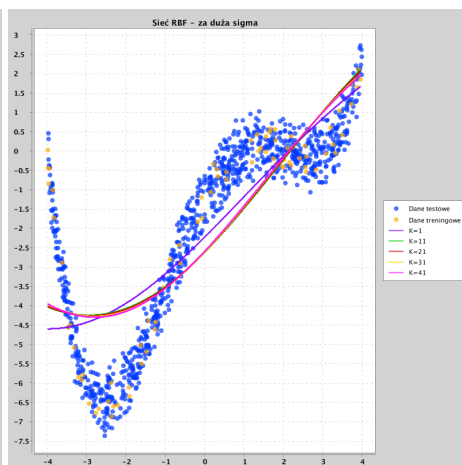
1.1 Podzadanie 1



Rysunek 1: Za mała sigma



Rysunek 2: Optymalna sigma

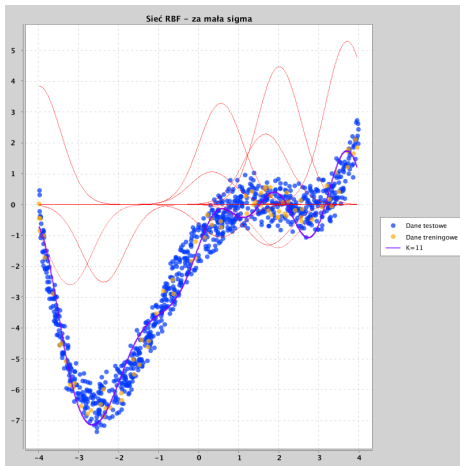


Rysunek 3: Za duża sigma

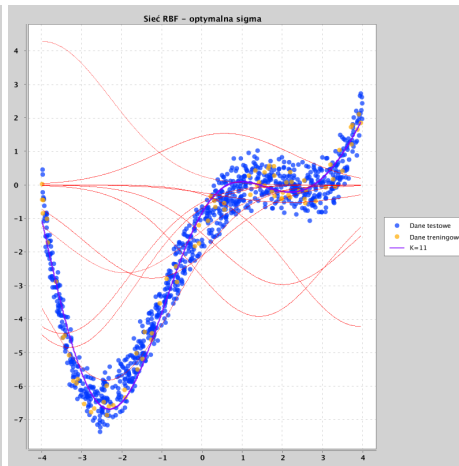
Dla liczby $K \geq 11$, gdy współczynnik skalujący jest zbyt mały sieć przybliża dane treningowe do pewnego stopnia, jednak tworzy pewne zakłócenia, co daje niedokładne przybliżenie. Z drugiej strony, gdy sigma jest zbyt duża sieć praktycznie wcale nie aproksymuje danych jedynie je przecina w pewien sposób. Sieć o $K = 1$ daje takie same wyniki, jak $K = 41$. Gdy sigma jest optymalna, sieć o każdej liczbie neuronów oprócz $K = 1$, aproksymuje tak samo dokładnie.

1.2 Podzadanie 2

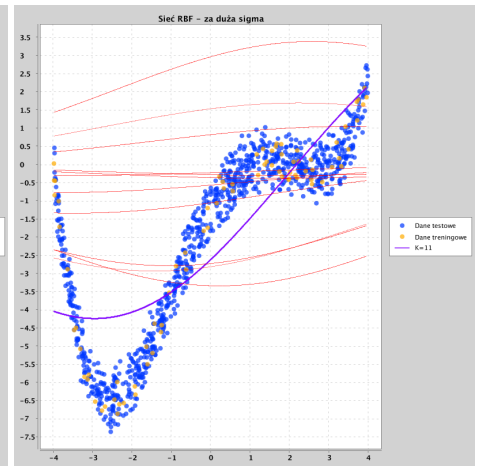
Czerwone linie reprezentują funkcje pojedynczych neuronów. Na podstawie poprzednich wyników wybrano $K = 11$. Na Rysunku 4 widać powód zakłóceń wykrytych na Rysunku 1. Gdy sigma jest za mała funkcje aktywacji neuronów mają zbyt wąski obszar aktywacji. Z drugiej strony na Rysunku 6 widać, że przy dużym



Rysunek 4: Za mała sigma



Rysunek 5: Optymalna sigma



Rysunek 6: Za duża sigma

współczynnikiem skalującym funkcje neuronów są prawie stałe. Nie da się stworzyć z kombinacji funkcji stałych dowolnej funkcji. Jeśli zaś dobierzemy sigmę optymalnie, neurony będą realizowały funkcje nie za wąską i nie za szeroką, co umożliwi bardzo dokładną aproksymację jak na Rysunku 5.

1.3 Podzadanie 3

Symbole ϵ_a , ϵ_b oznaczają błędy średniokwadratowe odpowiednio dla zbioru treningowego i testowego. Symbol σ w Tabeli 1 oznacza odchylenie standardowe. Losowy wybór centrów neuronów wprowadza dodatkowy błąd

Tabela 1: Błąd średniokwadratowy oraz odchylenie dla zbioru treningowego i testowego dla 100 prób nauki

K	$avg(\epsilon_a)$	$\sigma(\epsilon_a)$	$avg(\epsilon_b)$	$\sigma(\epsilon_b)$
1	2.249	0.626	1.955	0.562
6	0.286	0.286	0.252	0.196
11	0.151	0.117	0.168	0.075
16	0.078	0.025	0.115	0.018
21	0.061	0.016	0.107	0.010
26	0.058	0.009	0.108	0.008
31	0.052	0.006	0.102	0.005
36	0.047	0.003	0.098	0.003
41	0.046	0.003	0.097	0.003

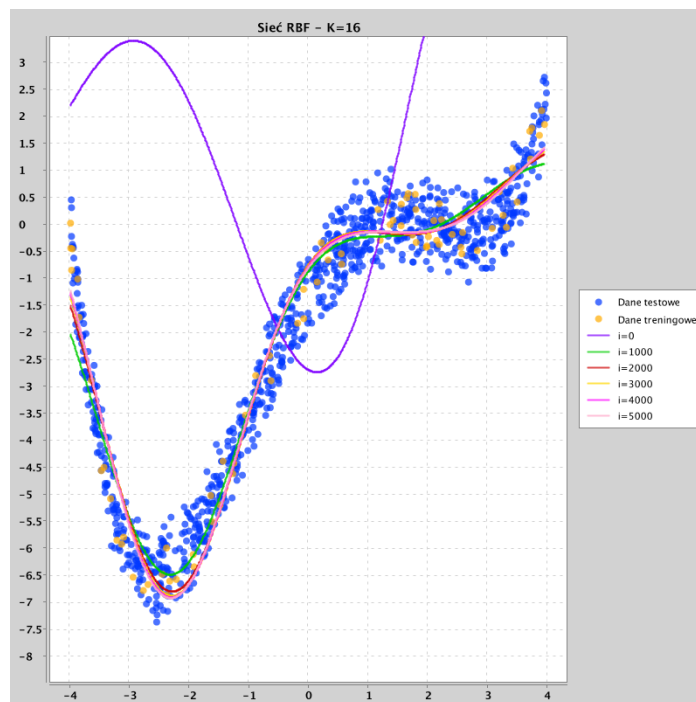
do sieci, gdyż przy małej ich liczbie mogą one zostać rozłożone nierównomiernie po całym zbiorze. Wynika stąd bardzo duże odchylenie standardowe dla $K = 1, 6, 11$. Gdy $K = 11$, błąd jest już relatywnie mały jednak nauka jest niestabilna, przez wspomniane losowanie centrów. Zauważmy, że dla $K = 16$, błąd zmniejsza się już dwukrotnie, a odchylenie standardowe jest bardzo małe. Taka liczba neuronów zapewnia już reprezentatywność danych treningowych. Przy $K = 26, 31, 36, 41$ średni błąd i odchylenie zmienia się nieznaczaco w stosunku do $K = 16$. Za duża liczba neuronów nie zwiększa znacząco możliwości aproksymacyjnych sieci. Błąd i odchylenie na zbiorze testowym zawsze jest większe niż na zbiorze treningowym.

1.4 Podzadanie 4

Na podstawie obserwacji z Tabeli 1, Rysunek 7 został wygenerowany z parametrami $K = 16$, liczba iteracji = 5000. Można na nim wyraźnie zaobserwować, że już po 2000 iteracji funkcja realizowana przez sieć wyglądała tak samo jak dla $i = 5000$. Gdy błąd przestanie się znacząco zmniejszać należy przestać uczyć sieć, gdyż nie da to lepszych efektów.

1.5 Podsumowanie

- Błąd średniokwadratowy z każdą iteracją spada dosyć szybko, jednak w pewnym momencie osiąga próg, po którym dalsze uczenie sieci nie daje znaczących efektów.



Rysunek 7: Zmiana funkcji sieci w różnych momentach

- Potrzebna liczba neuronów w warstwie radialnej jest zależna od charakterystyki danych jakie aproksymujemy. Za mała nie przybliży jej wcale, a za duża daje prawie takie same wyniki jak optymalna, którą należy dobrać eksperymentalnie.
- Sieć można uznać za nauczoną, gdy jej błąd przestanie znacząco maleć. Zależy to od charakterystyki danych. Jest to dość ważny aspekt, gdyż wtedy można trenować sieć np. przy 5000 iteracjach, a nie 20000 iteracji, co zabiera czas i moc obliczeniową.
- Jeśli centra zostaną źle wylosowane to błąd znacząco się zwiększa, dlatego należy wybierać taką ich liczbę aby minimalizować błąd stworzony przez losowanie.
- Współczynnik skalujący ma ogromny wpływ na funkcje realizowaną przez sieć, gdy dobierzemy go źle to przy za małej wartości sieć będzie niedokładna. Przy za dużym współczynniku nie będzie aproksymowała wcale. Należy go dobrać tak, aby był optymalny.

2 Osobna nauka warstw - klasyfikacja

Wykorzystano parametry nauki: $\alpha = 0.05$, optymalna sigma, liczba iteracji = 5000. Centra dobierano za pomocą algorytmu K-Średnich.

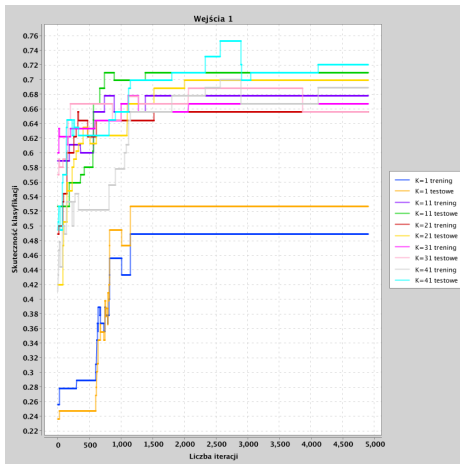
2.1 Podzadanie 1

Dla cech (wejść) 1,2 przedstawionych na Rysunkach 8, 9, skuteczność klasyfikacji stosunkowo niska - 60-70% niezależnie od K . Nie są one reprezentatywne. Z drugiej strony dla cech 3,4 na Rysunkach 10, 11 skuteczność sięga 95% dla $K \geq 11$. Większa liczba neuronów niż 11 nie zwiększa skuteczności. Na podstawie jednej wybranej cechy z czterech można dokonać bardzo dobrej klasyfikacji.

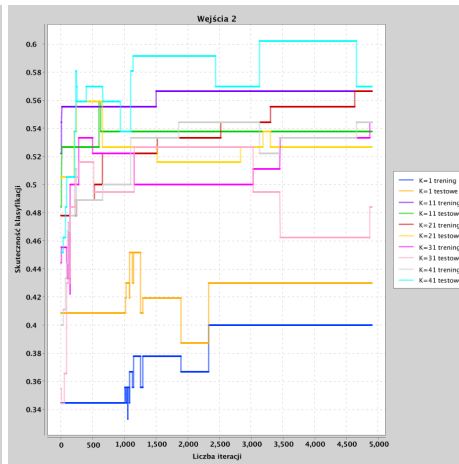
Na Rysunku 12 widać, że po połączeniu 2 słabo reprezentatywnych cech, można zwiększyć zdolności sieci o około 15%. Jednak kombinacja dwóch najbardziej charakterystycznych cech na Rysunku 17 jest skuteczniejsza o tylko 1%.

Kombinacje 3 wejść na rysunkach 18 - 21 oraz 4 wejść na Rysunku 22 również nie polepszają klasyfikacji.

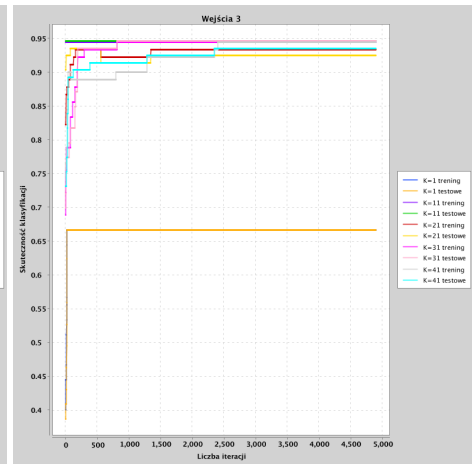
Skuteczność dla każdej z możliwych kombinacji rośnie skokowo. Jeśli cecha nie jest charakterystyczna to czasami nawet spada jak na Rysunku 9. Po około 3000 iteracji przestaje rosnąć i zatrzymuje się w miejscu.



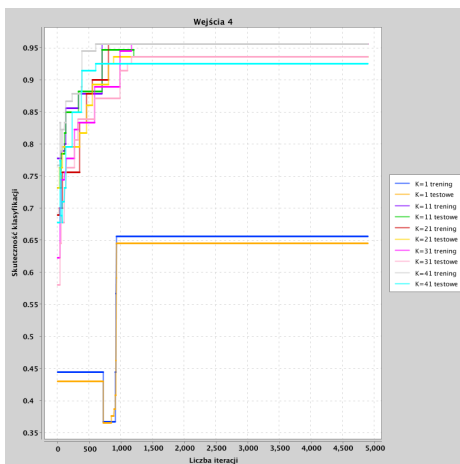
Rysunek 8



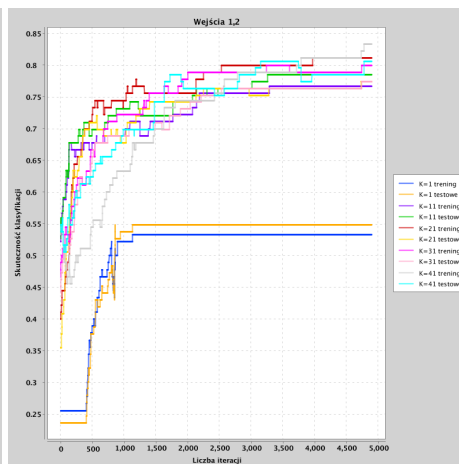
Rysunek 9



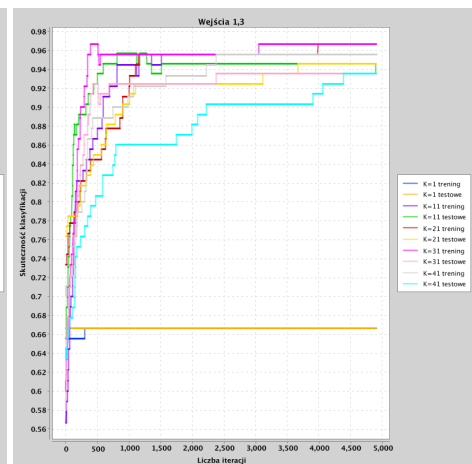
Rysunek 10



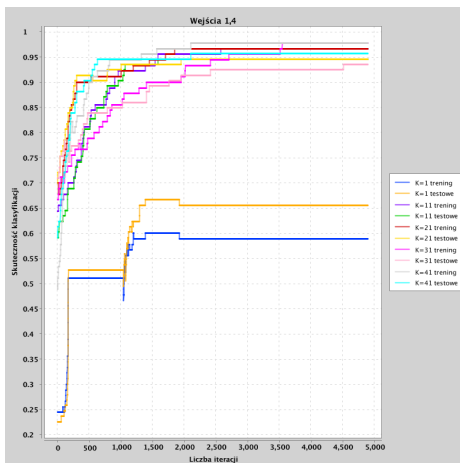
Rysunek 11



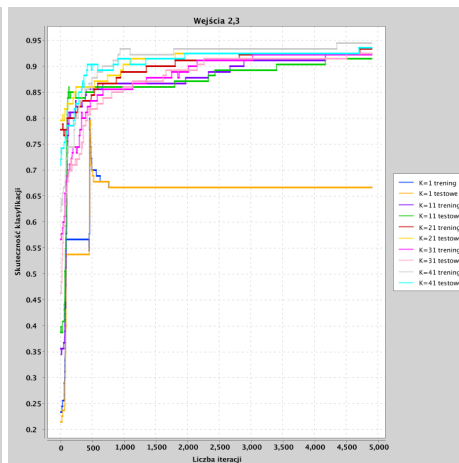
Rysunek 12



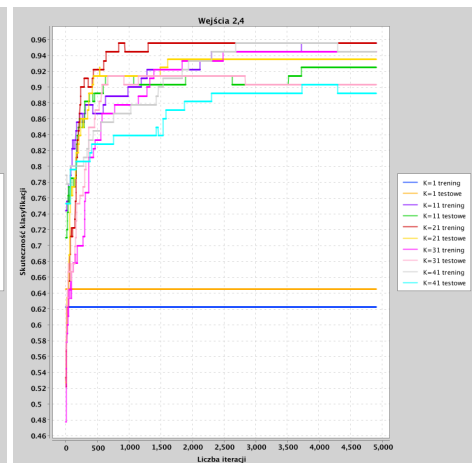
Rysunek 13



Rysunek 14



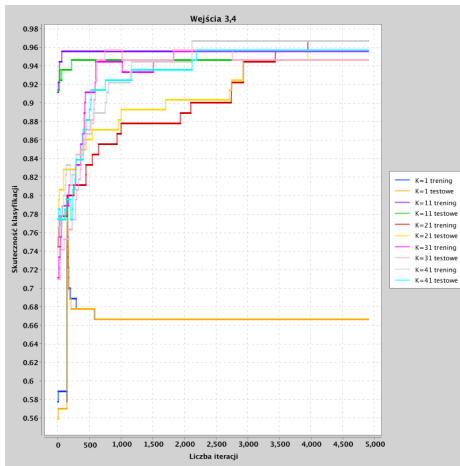
Rysunek 15



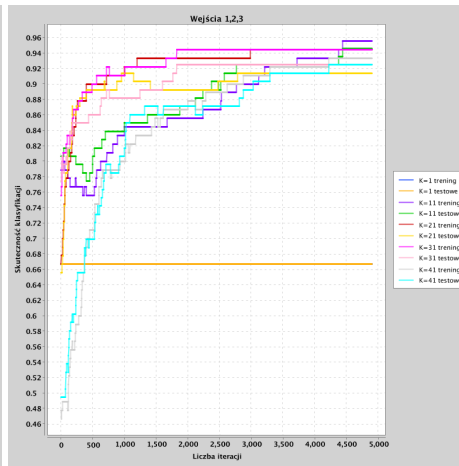
Rysunek 16

2.2 Podzadanie 2

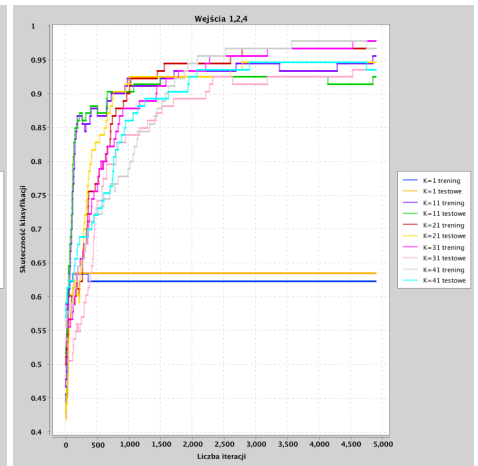
W Tabeli 2 p_a , p_b oznaczają odpowiednio procent klasyfikacji dla zbioru treningowego i testowego. Procent poprawnie sklasyfikowanych obiektów w zbiorze testowym jest zawsze taki sam, albo bardzo zbliżony do zbioru treningowego. Odchylenie standardowe na poziomie 0.01 oznacza, że sieć jest bardzo stabilna i dla każdej próby daje takie samą skuteczność. Dla $K \geq 6$ skuteczność nie zmienia się, tyle neuronów wystarczy aby klasyfikować z dużą dokładnością.



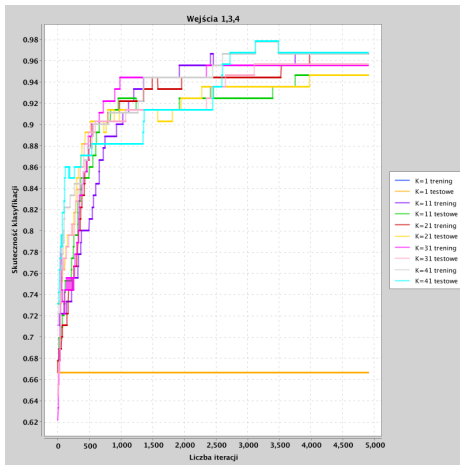
Rysunek 17



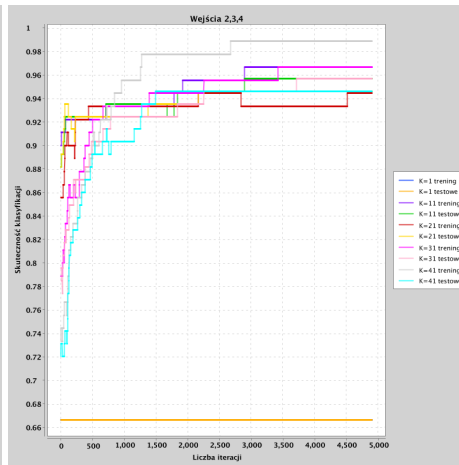
Rysunek 18



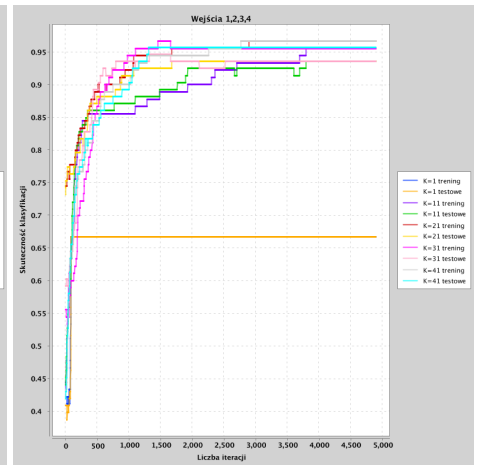
Rysunek 19



Rysunek 20



Rysunek 21



Rysunek 22

Tabela 2: Procent klasyfikacji oraz odchylenie dla zbioru treningowego i testowego dla 100 prób nauki

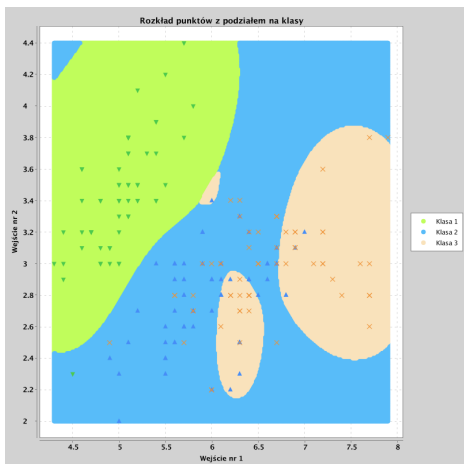
K	$avg(p_a)$	$\sigma(p_a)$	$avg(p_b)$	$\sigma(p_b)$
1	0.67	0.00	0.67	0.00
6	0.94	0.01	0.93	0.01
11	0.94	0.01	0.93	0.01
16	0.94	0.01	0.93	0.01
21	0.94	0.01	0.93	0.01
26	0.94	0.01	0.93	0.01
31	0.94	0.00	0.94	0.01
36	0.94	0.00	0.94	0.01
41	0.94	0.00	0.94	0.01

2.3 Podzadanie 3

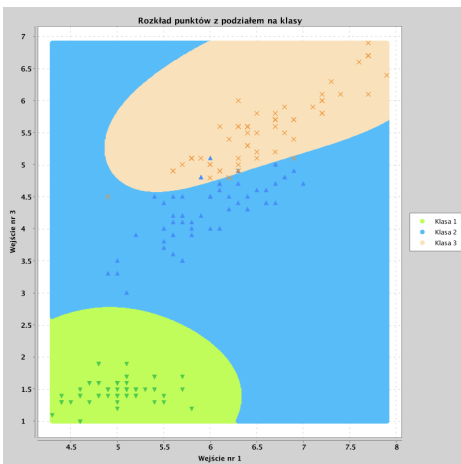
Zaokrąglone obszary wyznaczone na Rysunkach 23 - 28 przypominają kształtem funkcję gaussowską neuronów radialnych. Na granicach obszarów decyzyjnych zachodzą niejednoznaczności w klasyfikacji. Po redukcji, gdy współrzędne w odpowiednich osiach mają taką samą wartość nie ma sposobu na odróżnienie obiektów.

2.4 Podsumowanie

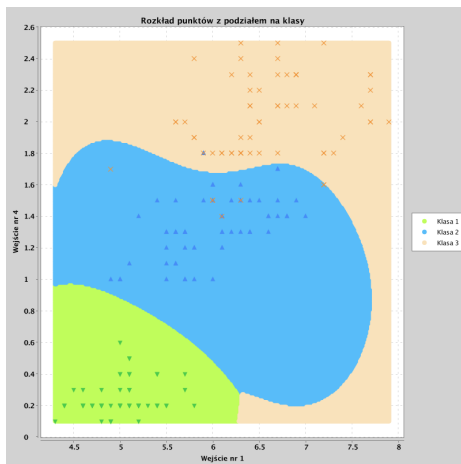
- Skuteczność sieci zmienia się skokowo, jest bardzo zbliżona na zbiorze treningowym i testowym.
- Większa liczba neuronów nie oznacza zwiększenia skuteczności klasyfikacji. Optymalna liczba zależy od charakterystyki zbioru.
- Redukcja wektora wejściowego do różnych kombinacji współrzędnych może wyłonić współrzędne repre-



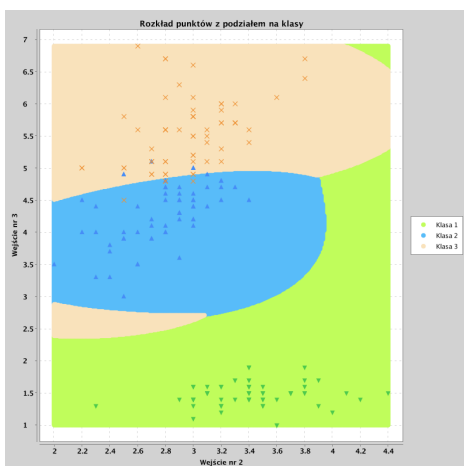
Rysunek 23



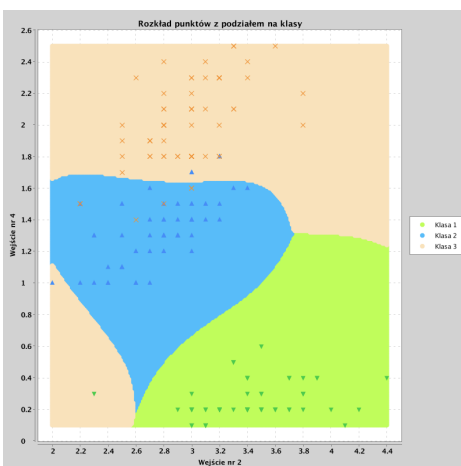
Rysunek 24



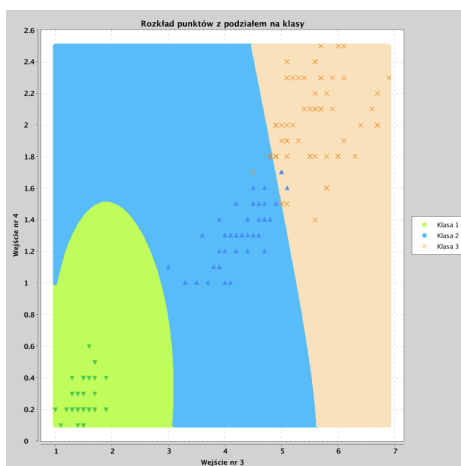
Rysunek 25



Rysunek 26



Rysunek 27



Rysunek 28

zentatywne. Dzięki czemu będzie można rozpoznać obiekty po 1 z cech.

- Gdy znajdziemy jedną cechę charakterystyczną, dodanie większej ich liczby nie zwiększy znacząco skuteczności.
- Sieć po pewnej liczbie iteracji przestaje się uczyć - procent skuteczności przestaje rosnąć.

3 Nauka obu warstw

3.1 Wyprowadzenie wzorów

3.1.1 Pochodna po współczynniku skalującym

$$\begin{aligned}
 \frac{\partial E}{\partial \sigma_k} &= \frac{1}{2N} \sum_{j=1}^N (f(x^j) - y^j)^2 \\
 &= \frac{1}{2N} \sum_{j=1}^N 2(f(x^j) - y^j) f'(x^j) \\
 &= \frac{1}{N} \sum_{j=1}^N f(x^j - y^j) \frac{\partial}{\partial \sigma_k} \sum_{k=0}^K w_k z_k(x^j) \\
 \frac{\partial}{\partial \sigma_k} \sum_{k=0}^K w_k z_k(x^j) &= \sum_{k=0}^K w_k z_k(x^j) = w_k \frac{\partial}{\partial \sigma_k} z_k(x^j)
 \end{aligned}$$

Ponieważ tylko $w_k \frac{\partial}{\partial \sigma_k} z_k(x^j)$ zależy od σ_k , dla $k = 0, \dots, K$

$$\begin{aligned}
\frac{\partial}{\partial \sigma_k} z_k(x^j) &= \frac{\partial}{\partial \sigma_k} \exp(-d^2(x^j, c_k) \frac{1}{2\sigma_k^2}) \\
&= z_k(x^j) \frac{\partial}{\partial \sigma_k} (\frac{1}{2\sigma_k^2}) \\
&= z_k(x^j) \frac{-1}{2} d^2(x^j, c_k) \frac{\partial}{\partial \sigma_k} \sigma_k^{-2} \\
&= z_k(x^j) \frac{-1}{2} d^2(x^j, c_k) - 2\sigma_k^{-3} \\
&= z_k(x^j) d^2(x^j, c_k) \frac{1}{\sigma_k^3}
\end{aligned}$$

Ostatecznie

$$\frac{\partial E}{\partial \sigma_k} = \frac{1}{N} \sum_{j=1}^N f(x^j - y^j) z_k(x^j) w_k d^2(x^j, c_k) \frac{1}{\sigma_k^3} \quad (2)$$

3.1.2 Pochodna po koordynacie centrum neuronu

Analogicznie jak w poprzednim wyprowadzeniu.

$$\begin{aligned}
\frac{\partial}{\partial c_{k,i}} z_k(x^j) &= \frac{\partial}{\partial c_{k,i}} \exp(-d^2(x^j, c_k) \frac{1}{2\sigma_k^2}) \\
&= z_k(x^j) \frac{\partial}{\partial c_{k,i}} - d^2(x^j, c_k) \frac{1}{2\sigma_k^2} \\
&= z_k(x^j) \frac{-1}{2\sigma_k^2} \frac{\partial}{\partial c_{k,i}} d^2(x^j, c_k) \\
\frac{\partial}{\partial c_{k,i}} d^2(x^j, c_k) &= \frac{\partial}{\partial c_{k,i}} \sqrt{\sum_{i=1}^n (x_i - c_{k,i})^2}^2 \\
&= \frac{\partial}{\partial c_{k,i}} \sum_{i=1}^n (x_i - c_{k,i})^2 \\
&= \sum_{i=1}^n 2(x_i - c_{k,i})(-1) \\
&= -2 \sum_{i=1}^n (x_i - c_{k,i}) \\
&= -2(x_i - c_{k,i})
\end{aligned}$$

Ponieważ tylko $x_i - c_{k,i}$ zależy od $c_{k,i}$.

$$\begin{aligned}
\frac{\partial}{\partial c_{k,i}} &= z_k(x^j) \frac{-1}{2\sigma_k^2} (-2)(x_i - c_{k,i}) \\
\frac{\partial}{\partial c_{k,i}} &= z_k(x^j) \frac{1}{\sigma_k^2} (x_i - c_{k,i})
\end{aligned}$$

Ostateczny wzór:

$$\frac{\partial E}{\partial c_{k,i}} = \frac{1}{N} \sum_{j=1}^N f(x^j - y^j) w_k z_k(x^j) \frac{1}{\sigma_k^2} (x_i - c_{k,i}) \quad (3)$$