Sieci Neuronowe w zastosowaniach Biomedycznych
Projekt "Klasyfikacja na podstawie danych pożyczkowych"

Dokumentacja

Igor Markiewicz
Mateusz Majkowski

1 Spis treści

2 Wstępna obróbka danych

W niektórych kolumnach występowały wartości niezidentyfikowane (”NA”). Zostały one zastąpione średnimi arytmetycznymi po odpowiednich kolumnach. Następnie została przeprowadzona normalizacja danych wejściowych :

$$x_{norm} = \frac{x - \overline{x}}{\max(x) - \min(x)}$$

W kolejnym kroku przy użyciu funkcji **trainTestSplit** z pakietu **sklearn**, nastąpił podział na zbiór uczący (70%) i testowy (30%). W ostatnim punkcie wstępnego przetwarzania danych, został zastosowany algorytm **SMOTE** z pakietu **imblearn**, który przy użyciu algorytmu centroidów losuje nowe dane z otoczenia centrum klasy pozytywnej (mniejszościowej) i dąży do zrównoważenia kategorii. W efekcie zamiast wycinać przykłady klasy negatywnej (dostosowując ich liczbę do liczby przykładów klasy mniejszościowej) przeprowadzając podpróbkowanie, następuje nadpróbkowanie powodujące bogaty zestaw przykładów, co przekłada się przeważnie na jakość uczenia sieci.

3 Założenia

- Celem klasyfikacji jest zmienna **SeriousDlqin2yrs** mówiąca o tym czy osoba będzie miała najprawdopodobniej trudności finansowe w ciągu dwóch następnych lat.
- Przykłady do uczenia i testowania są podawane w pakietach (do których następuje każdorazowo losowanie przykładów) o rozmiarze 512, a błąd na wyjściu jest uśredniany
- Sieć posiada dwie warstwy ukryte i warstwę wyjściową (z możliwością rozszerzenia warst ukrytych)
- Dropout - w trakcie uczenia 0.5 a w trakcie testowania 1.0
- Warstwy ukryte posiadają jako funkcję aktywacji elu, natomiast wyjście funkcję sigmoidalną, unipolarną
- Liczba epok jest równa 200
- Jako metodę uczenia wybrano minimalizację funkcji celu (zmodyfikowana norma euklidesowa - zmod. l^2) metodą adaptacyjną ADAM (daje bardzo dobre rezultaty, choć czasem może utknąć w minimach lokalnych)
- Predykcja jest wykonywana na zbiorze testowym o zrównoważonych kategoriach. Jako predykcję przyjmujemy :
  $y \in \{0,1\}$ - prawdziwa klasa
  $yy \in (0,1)$ - wyjście sieci

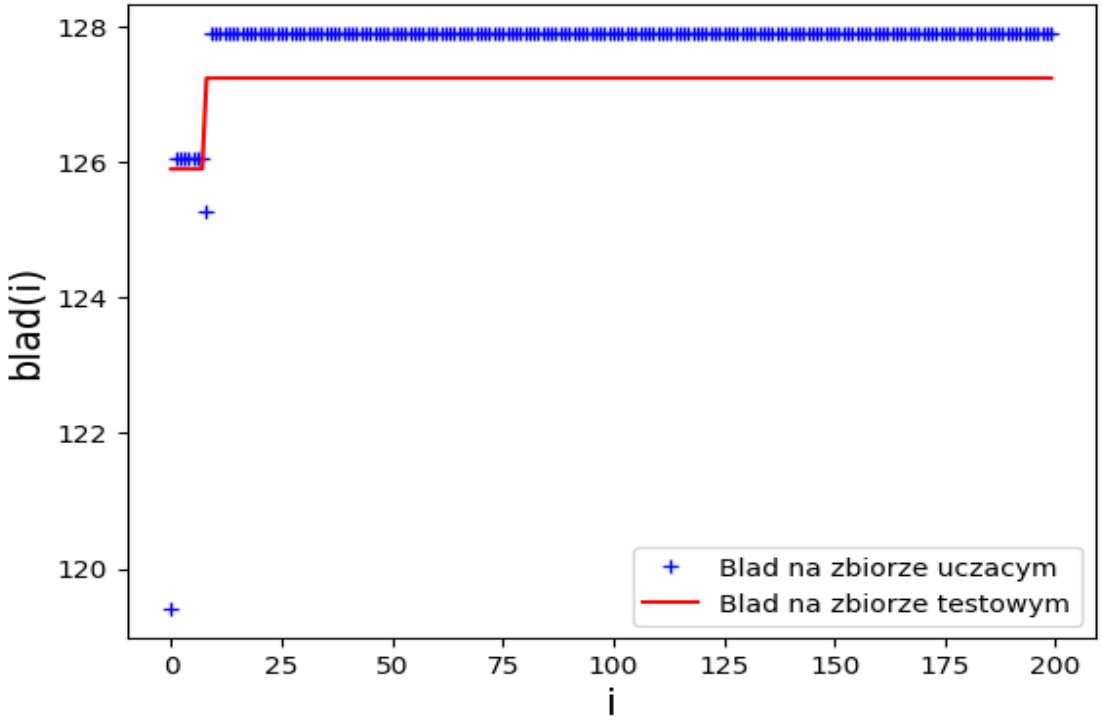
$$prediction(y,yy) = \begin{cases} 1, & \text{gdy } \left(y = 1 \wedge yy \geq \frac{1}{2}\right) \vee \left(y = 0 \wedge yy < \frac{1}{2}\right) \\ 0, & \text{w pozostałych przypadkach} \end{cases}$$

- *i* - numer epoki poczynając od 0

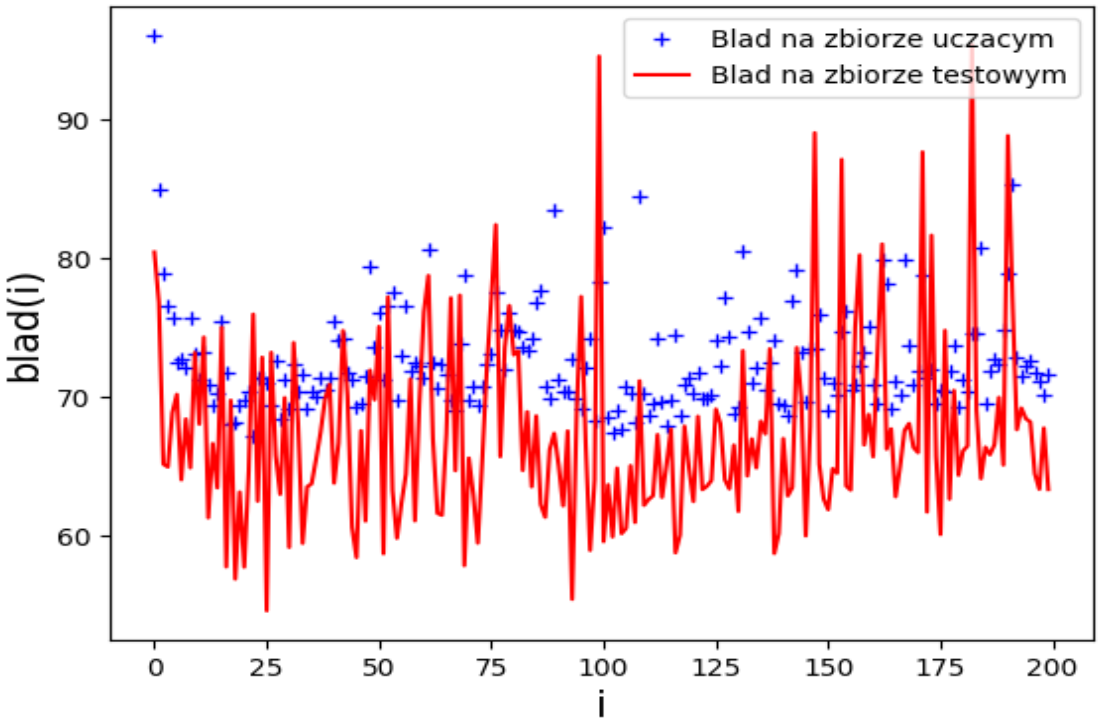
4 Badania i wnioski

4.1 Wpływ współczynnika szybkości uczenia na zachowanie sieci

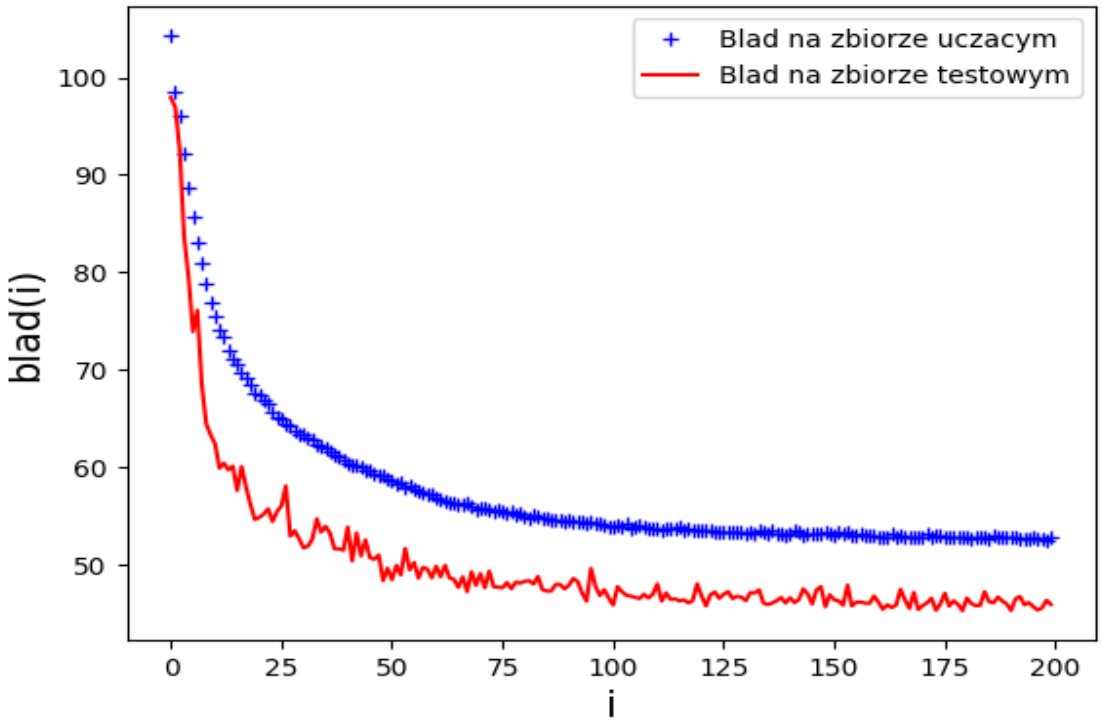
Badania zostały wykonane dla 256 neuronów w obu warstwach ukrytych.



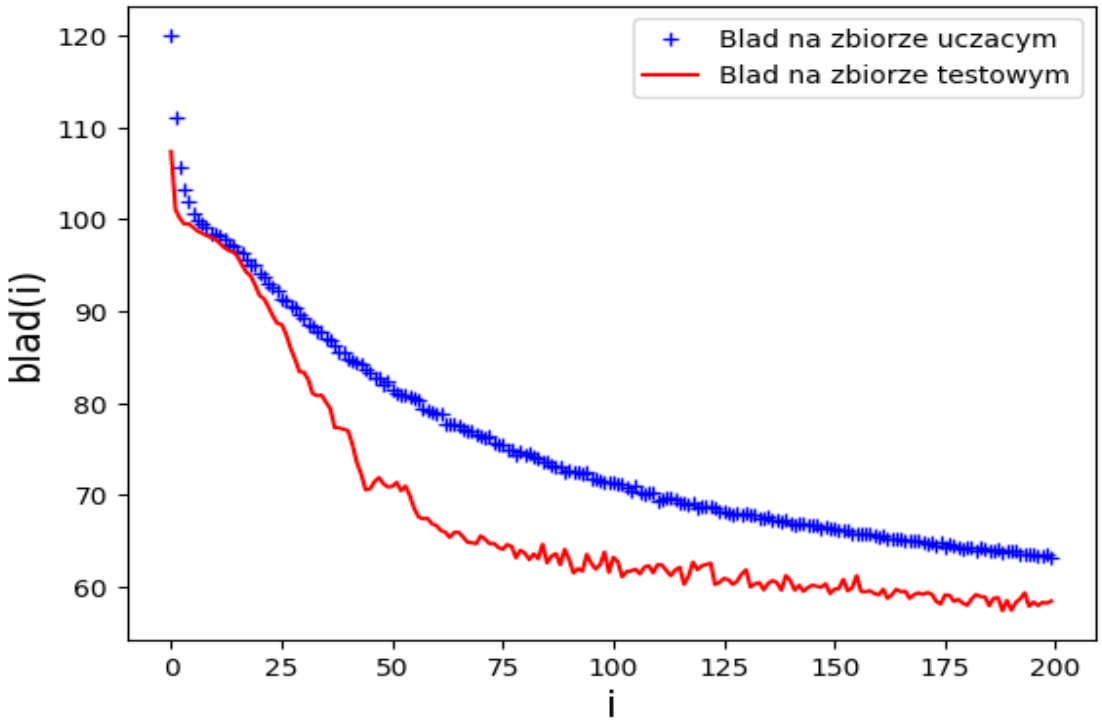
Rysunek 1: Wykres dla learningRate=0.1 Predykcja - 50%



Rysunek 2: Wykres dla learningRate=0.01 Predykcja - 65.12%



Rysunek 3: Wykres dla learningRate=0.001 Predykcja - 81.95%

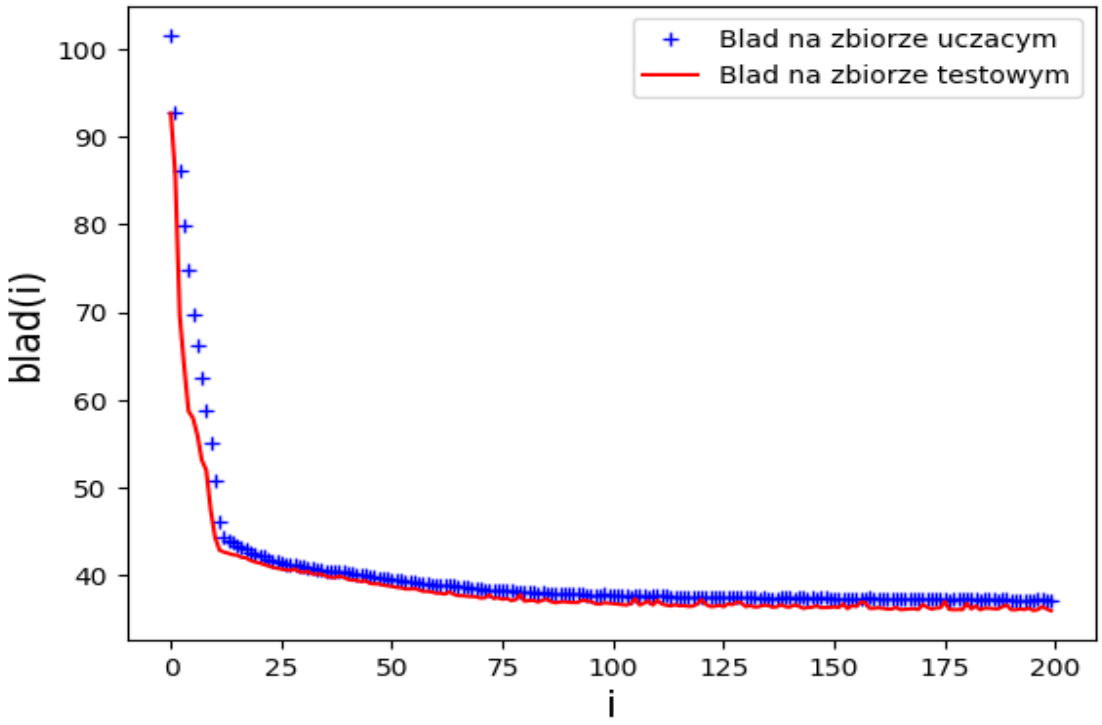


Rysunek 4: Wykres dla learningRate=0.0001 Predykcja - 76.91%

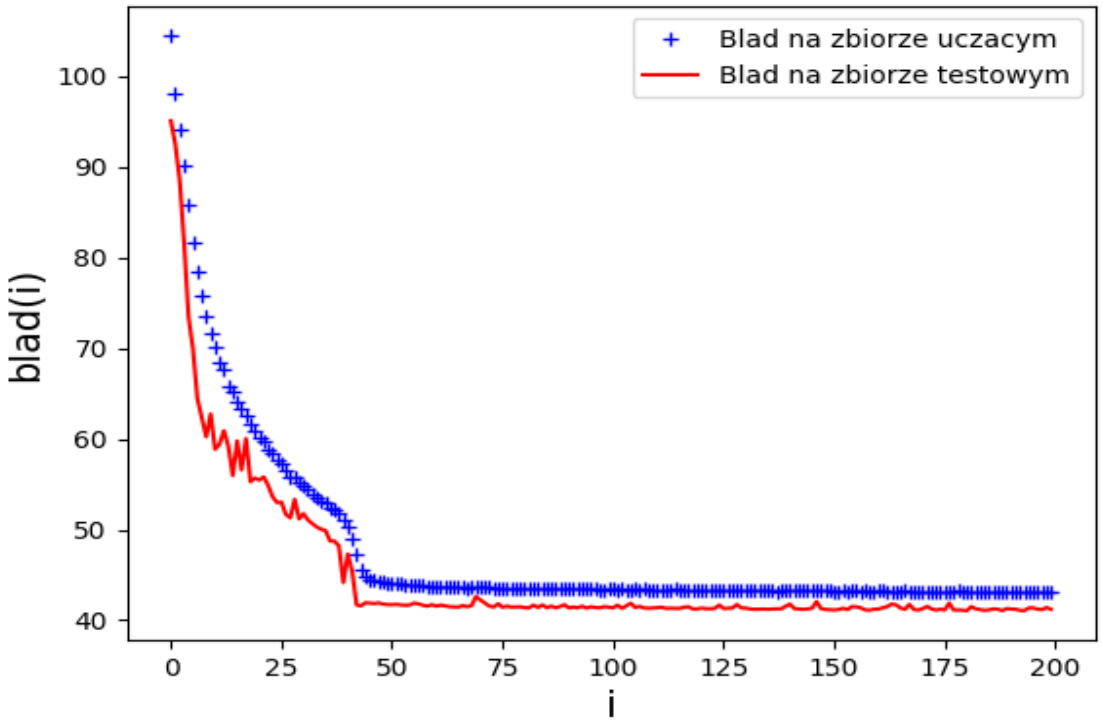
Najlepsze efekty uzyskaliśmy dla learnigRate = 0.001, zarówno jeśli chodzi o spadek błędu jaki i predykcję. Dla learnigRate = 0.0001 błąd na zbiorze testowym zdaje się mieć mniejsze wahania, natomiast zmniejsza się słabiej niż dla learnigRate = 0.001 oraz predykcja jest trochę słabsza. Dla learnigRate = 0.01 predykcja jest całkiem dobra, natomiast błąd zarówno na zbiorze uczącym jak i testowym oscyluje.Z kolei dla learnigRate = 0.1 predykcja jest nie do zaakceptowania, jak również postęp błędów na zbiorze uczącym i testowym.

4.2 Wpływ liczby neuronów na zachowanie sieci

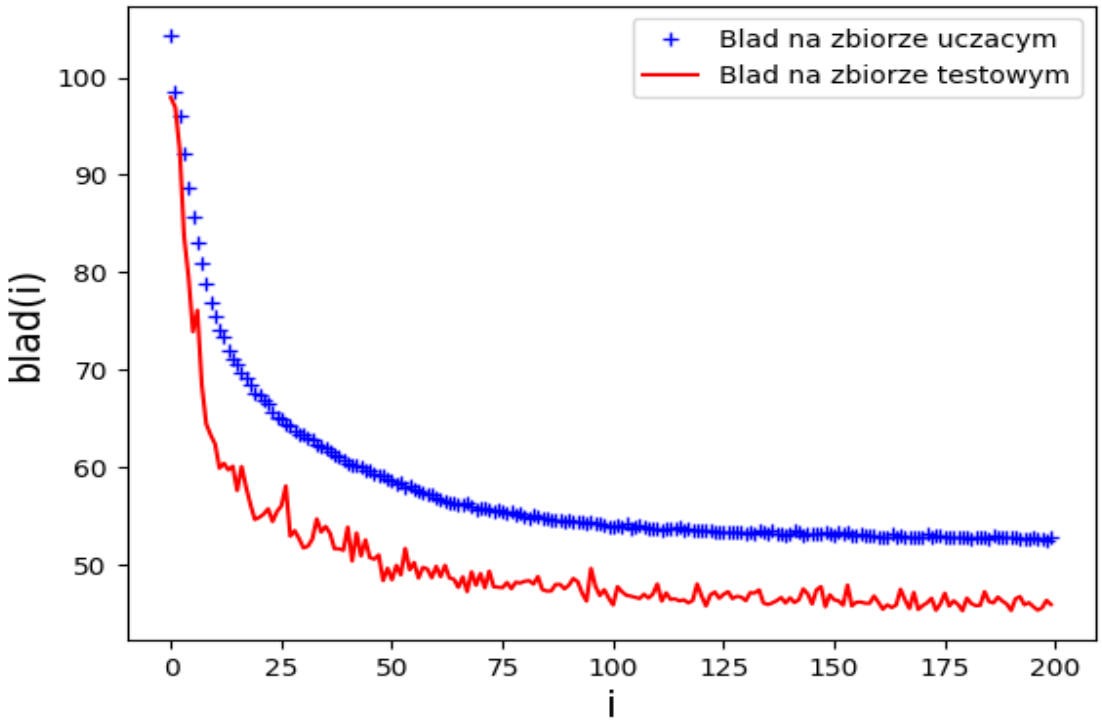
Badania zostały wykonane dla learningRate = 0.001



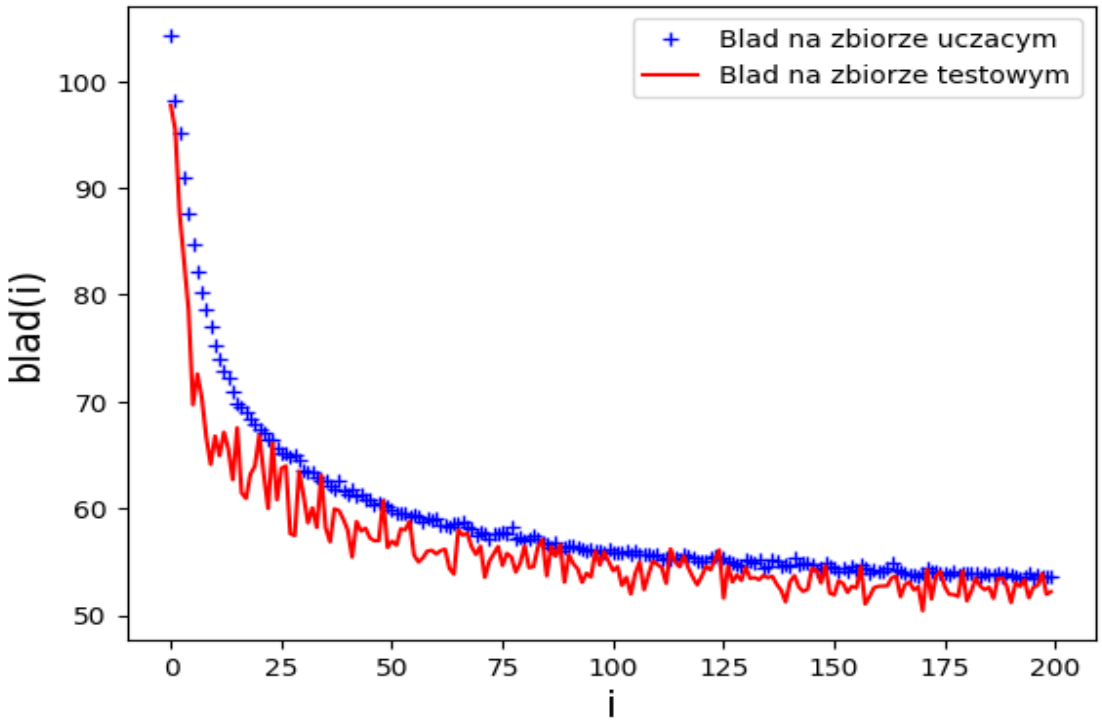
Rysunek 5: Wykres dla 64 neuronów w obu warstwach Predykcja - 80.42%



Rysunek 6: Wykres dla 128 neuronów w obu warstwach Predykcja - 81.05%



Rysunek 7: Wykres dla 256 neuronów w obu warstwach Predykcja - 81.95%



Rysunek 8: Wykres dla 512 neuronów w obu warstwach Predykcja - 79.49%

Największy i najgładszy spadek uzyskaliśmy dla 64 neuronów, zaś najlepszą predykcję dla 256. Dla 512 neuronów błąd na zbiorze testowym wykazuje duże oscylacje w porównaniu z pozostałymi badanymi sytuacjami. Dla wszystkich przypadków jakość predykcji jest jednak bardzo podobna, co może sugerować że możemy z powodzeniem korzystać z sieci o mniejszych rozmiarach. Zarówno podczas badań nad wpływem liczby neuronów jaki i współczynnika szybkości uczenia na jakość sieci, okazuje się że graniczną wartością predykcji jest ok 80%.

5 Bibliografia

[1] Wykłady do przedmiotu Sztuczne Sieci Neuronowe w zastosowaniach Biomedycznych

[2] L.Rutkowski *Metody i techniki sztucznej inteligencji*, PWN, Warszawa 2012

[3] S.Osowski *Sieci neuronowe do przetwarzania informacji*, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 2013

[4] `http://www-users.mat.umk.pl/~rudy/wsn/wyk/wsn-wyklad-05a-propag.pdf`

[5] `http://yann.lecun.com/exdb/publis/pdf/lecun-98b.pdf`

[6] `https://www.jair.org/media/953/live-953-2037-jair.pdf`