Warszawa, 16.04.2021

Politechnika Warszawska

Wydział Elektroniki i Technik Informacyjnych

Sieci neuronowe w zastosowaniach biomedycznych (SNB)

Projekt: (35) Diagnostyka raka piersi w badaniach mammograficznych za pomocą sieci MLP

ETAP I

Prowadzący: dr inż. Paweł Mazurek

Wykonawca:

Zespół **4**

Jagoda Adamczyk, Aleksandra Krakowiak

Spis treści

1 Wstęp

Celem wykonywanego projektu jest stworzenie sieci neuronowej MLP. Zadaniem stworzonej sieci będzie zaklasyfikowanie kobiet do jednej z dwóch grup – łagodnej zmiany mammograficznej lub złośliwej. Powyższy podział umożliwiłby zmniejszenie liczby niepotrzebnych biopsji piersi. Klasyfikacja będzie odbywać się na podstawie zebranych informacji podczas wykonywania badań mammograficznych. Do stworzenia modelu klasyfikacji wykorzystywana będzie baza – *Mammographic Mass Data Set*[1]. Baza została opublikowana 29 października 2007 roku. Zawiera ona dane od 961 pacjentów (kobiet), gdzie 516 danych opisują łagodne zmiany mammograficzne, a 445 – złośliwe zmiany. Zestaw danych wejściowych ma charakter wielowymiarowy.

Pierwszy etap projektu będzie się skupiał na analizie danych oraz zaplanowaniu modelu i działania sieci, której stworzenie będzie realizowane w kolejnym etapie projektu.

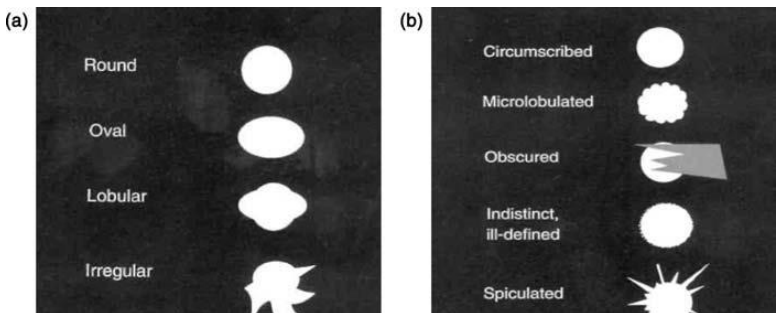
Projekt będzie wykonywany przy wykorzystaniu oprogramowania MATLAB R2021a przy wykorzystaniu Matlab Driver. Matlab Driver umożliwia współdzielenie pliku, dzięki czemu możliwa jest praca na jednym pliku.

2 Analiza danych

2.1 Lista nazw cech zawartych w wektorze wejściowym

Wektor wejściowy zawiera 5 cech (opis kolejnych kolumn), które mają charakter liczb całkowitych. Poniżej przedstawiono cechy wektora wejściowego:

1. **BI – RADS assessment** – skala opisująca ustalenia i wyniki badania mammograficznego;
   Skala:
   1 – negatywny, brak nieprawidłowości
   2 – łagodne nienowotworowe stwierdzenie; wynik negatywny, ale opisanie objawów
   3 – prawdopodobne wykrycie; zalecana obserwacja
   4 – podejrzana nieprawidłowość; zalecenie wykonania biopsji
   5 – silne nieprawidłowości; zalecenie wykonania biopsji[2]
2. **Age** – wiek badanej kobiety wyrażonej w latach
3. **Shape** – opisuje kształt wykrytej zmiany podczas badania wyrażony w sposób nominalny;
   Kształt:
   okrągły - 1,
   owalny – 2,
   płacikowaty – 3,
   nieregularny – 4



Rysunek 1. (a) kształt masy, (b) krawędzie masy[3]

4. **Margin** – opisuje krawędzie wykrytej zmiany podczas badania wyrażony w sposób nominalny;
   Krawędzie:
   poprawne - 1,
   mikrolobulowane – 2,
   zasłonięte – 3,
   źle zdefiniowane – 4,
   spiczaste - 5
5. **Density** – opisuje gęstość wykrytej zmiany podczas badania wyrażony w sposób nominalny;
   Wartość gęstości:
   duża - 1,
   zgodna z normami ISO - 2,
   mała – 3,
   zawierająca tłuszcz - 4

Każdy wektor wejściowy opisany jest etykietą, wskazującą na łagodne lub złośliwe zmiany w piersi. Etykieta o wartości 0 opisuje łagodne zmiany (*benign*), etykieta o wartości 1 – złośliwe zmiany (*malignant*).

2.2 Zakres zmienności cech zawartych w wektorze wejściowym

2.2.1 Wartości maksymalne i minimalne, średnia, mediana oraz odchylenie standardowe

Ze względu na pewną charakterystykę bazy danych wartości maksymalne, minimalne, średnia, mediana i odchylenie standardowe zostały wyznaczone tylko dla cechy *Age* (Tab. 1) z podziałem na trzy grupy: cała baza danych, grupa o łagodnej zmianie i grupa o złośliwej zmianie. W pozostałych przypadkach wyliczono medianę danej cechy. Wyliczone wartości mediany zostały przedstawione w tabeli (Tab. 2) z podziałem na wyżej wymienione grupy.

Tab 1. Zestawienie wyliczonych wartości dla cechy AGE.

	WIEK (*Age*)		
	OGÓŁ	GRUPA O ŁAGODNEJ ZMIANIA	GRUPA O ZŁOŚLIWEJ ZMIANIE
WARTOŚĆ MAX	96	86	96
ŚREDNIA	55	49	62
WARTOŚĆ MIN	18	18	28
MEDIANA	57	50	63
ODCHY. STAND.	14	13	12

Tab. 2. Zestawienie wartości mediany dla poszczególnych cech.

	MEDIANA		
	ogół	grupa o łagodnej zmianie	grupa o złośliwej zmianie
BI-RADS assessment	4	4	5
SHAPE	3	2	4
MARGIN	3	1	4
DENSITY	3	3	3

2.2.2 Histogramy

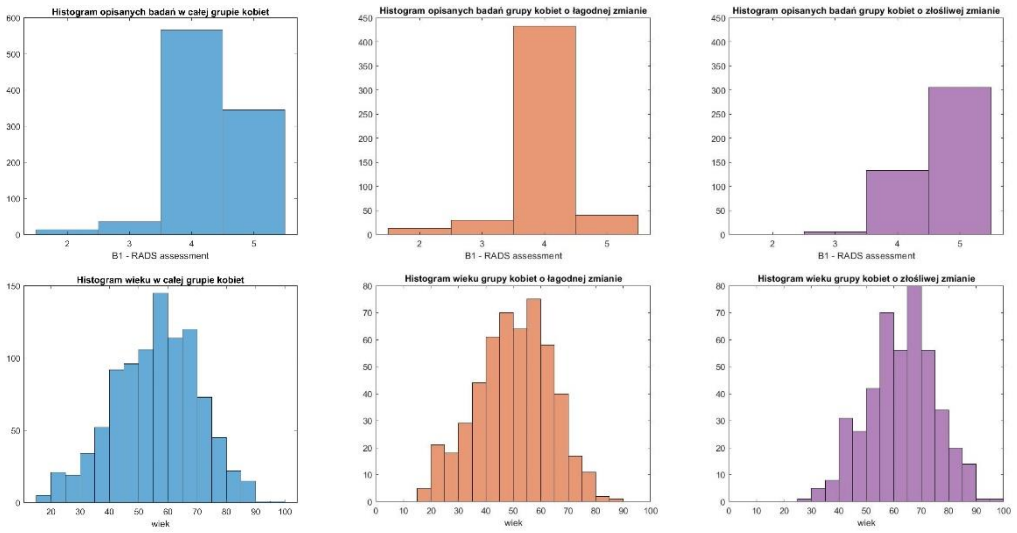
Na podstawie danych z wykorzystywanej bazy zostały stworzone histogramy. Histogramy opisują daną cechę z podziałem na trzy grupy: ogół (kolor niebieski), grupa o łagodnej zmianie (kolor pomarańczowy), grupa o złośliwej zmianie (kolor fioletowy). Histogramy podzielone zostały na dwa zbiory, pierwszy ze zbiorów przedstawia histogramy utworzone ze zbioru danych, w którym wartości niekompletne zostały uzupełnione:

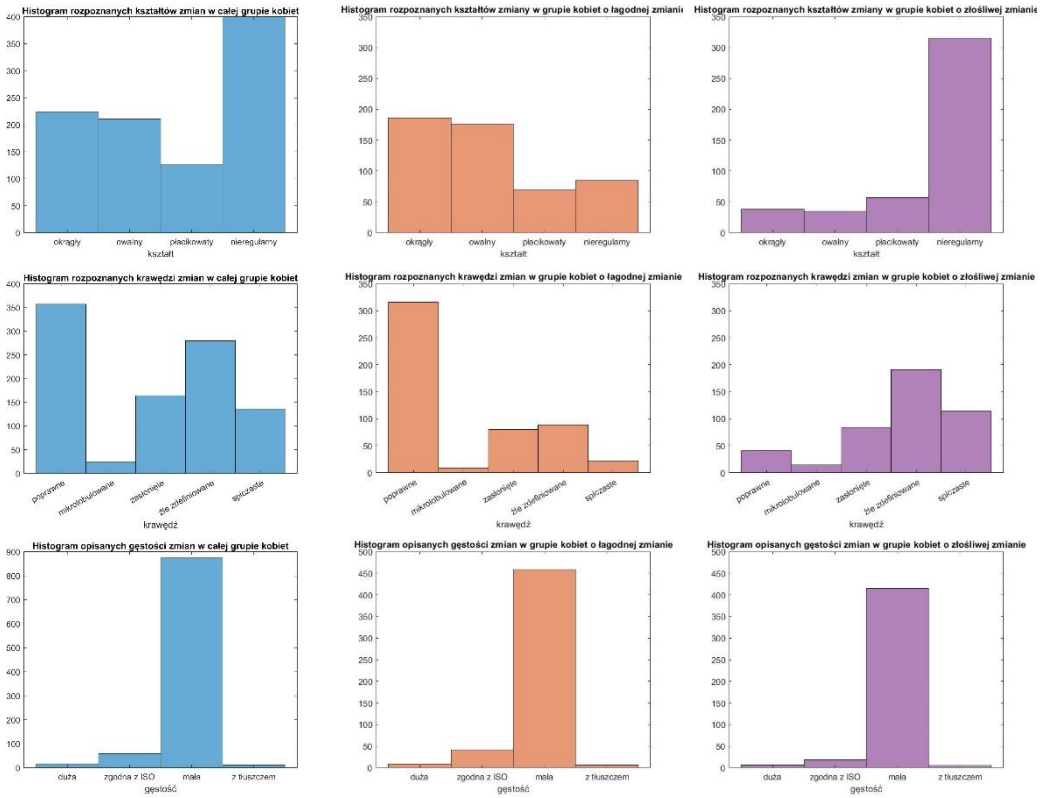
1. Medianą wyliczoną z całej grupy - w przypadku *BI-RADS*, *Shape*, *Margin* i *Density*,
2. Średnią wyliczoną z całej grupy – w przypadku wieku.

Drugi zbiór histogramów reprezentuje zbiór danych, w którym wartości niekompletne uzupełnione zostały:

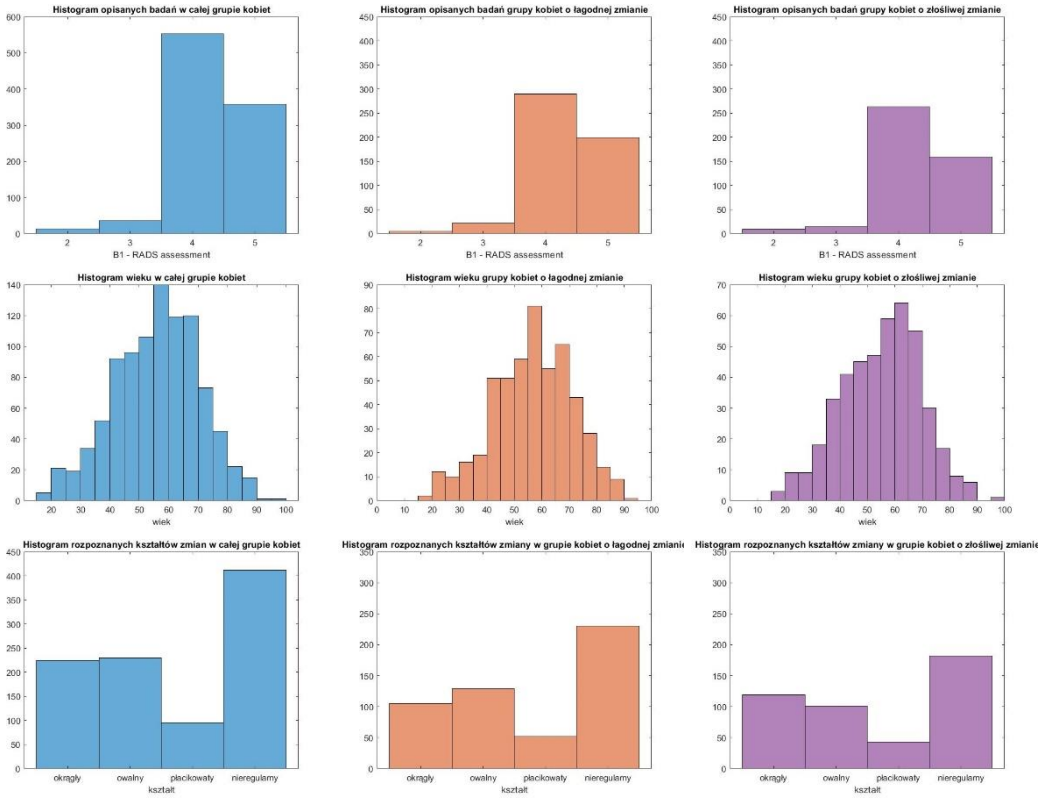
1. Medianą wyliczoną z klasy, którą reprezentuje dany rekord (łagodna zmiana/złośliwa zmiana) - w przypadku *BI-RADS*, *Shape*, *Margin* i *Density*,
2. Średnią wyliczoną z klasy, którą reprezentuje dany rekord (łagodna zmiana/złośliwa zmiana) - w przypadku wieku.

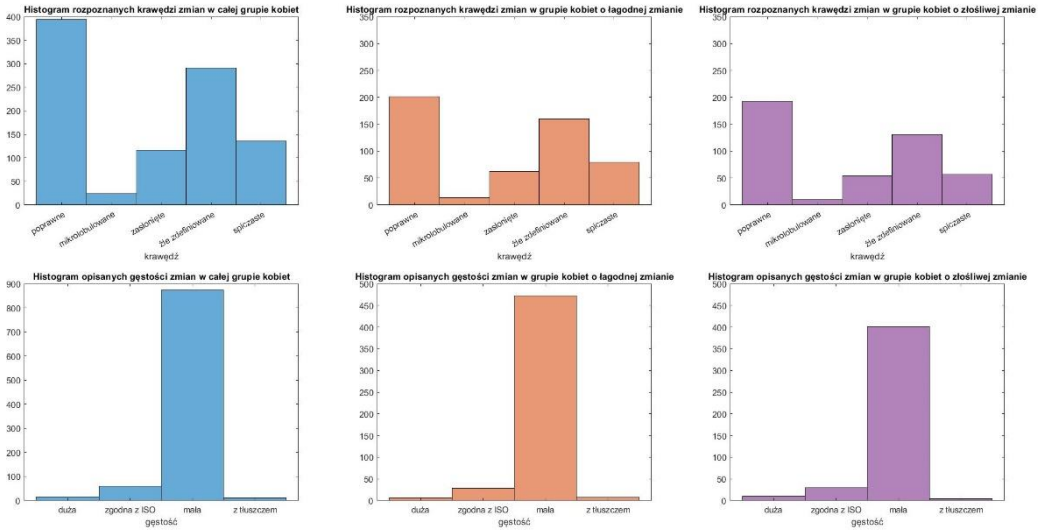
Zbiór 1.




Histogram rozpoznanych kształtów zmian w całej grupie kobiet
okrągły
owalny
płaskowaty
nieregularny
kształt
Histogram rozpoznanych kształtów zmiany w grupie kobiet o łagodnej zmianie
okrągły
owalny
płaskowaty
nieregularny
kształt
Histogram rozpoznanych kształtów zmiany w grupie kobiet o złośliwej zmianie
okrągły
owalny
płaskowaty
nieregularny
kształt
Histogram rozpoznanych krawędzi zmian w całej grupie kobiet
poprawne
mikrolobulowane
zaokrąglone
źle zdefiniowane
spiczaste
krawędź
Histogram rozpoznanych krawędzi zmian w grupie kobiet o łagodnej zmianie
poprawne
mikrolobulowane
zaokrąglone
źle zdefiniowane
spiczaste
krawędź
Histogram rozpoznanych krawędzi zmian w grupie kobiet o złośliwej zmianie
poprawne
mikrolobulowane
zaokrąglone
źle zdefiniowane
spiczaste
krawędź
Histogram opisanych gęstości zmian w całej grupie kobiet
duża
zgodna z ISO
mała
z tłuszczem
gęstość
Histogram opisanych gęstości zmian w grupie kobiet o łagodnej zmianie
duża
zgodna z ISO
mała
z tłuszczem
gęstość
Histogram opisanych gęstości zmian w grupie kobiet o złośliwej zmianie
duża
zgodna z ISO
mała
z tłuszczem
gęstość

Zbiór 2.


Histogram opisanych badań w całej grupie kobiet
B1 - RADS assessment
Histogram opisanych badań grupy kobiet o łagodnej zmianie
B1 - RADS assessment
Histogram opisanych badań grupy kobiet o złośliwej zmianie
B1 - RADS assessment
Histogram wieku w całej grupie kobiet
wiek
Histogram wieku grupy kobiet o łagodnej zmianie
wiek
Histogram wieku grupy kobiet o złośliwej zmianie
wiek
Histogram rozpoznanych kształtów zmian w całej grupie kobiet
okrągły
owalny
płaskowaty
nieregularny
kształt
Histogram rozpoznanych kształtów zmiany w grupie kobiet o łagodnej zmianie
okrągły
owalny
płaskowaty
nieregularny
kształt
Histogram rozpoznanych kształtów zmiany w grupie kobiet o złośliwej zmianie
okrągły
owalny
płaskowaty
nieregularny
kształt



2.3 Kodowanie danych nienumerycznych

W wykorzystywanej bazie twórcy z góry nałożyli pewną implementację danych nienumerycznych. Dzięki czemu nie musimy wykonywać tych operacji ponownie. Pracujemy na bazie, która zawiera liczby całkowite (stworzenie pewnych skal).

2.4 Metoda wstępnego przetwarzania danych

2.4.1 Standaryzacja

Ze względu na rożne zakresy wartości poszczególnych cech dokonana będzie standaryzacja danych. Standaryzację przeprowadzimy zgodnie z poniższymi wzorami[4].

$$\overline{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_j^i \qquad (2.1)$$

$$\sigma_i(x) = \sqrt{\frac{\sum_{j=1}^{n}(x_j^i - \overline{x}_i)^2}{n-1}} \qquad (2.2)$$

$$z_i = \frac{x_i - \overline{x}_i}{\sigma_i(x)} \qquad (2.3)$$

gdzie:

$x_i - i - ta\ cecha\ wektora\ wejściowego$
$\overline{x}_i - wartość\ średnia\ i - tej\ cechy$
$x_j^i - wartość\ i - tej\ cechy\ j - tego\ elemnetu\ wektora\ wejściowego$
$\sigma_i(x) - odchylenie\ standardowe\ i - tej\ cechy$
$z_i - przeskalowana\ wartość\ cechy\ x_i$

Dzięki takiej transformacie otrzymamy cechy o wartości średniej równej 0 oraz odchyleniu standardowym równym 1. Należy również podkreślić, że w czasie procesu uczenia bierzemy pod uwagę wszystkie dostępne dane.

2.4.2 Dane niekompletne

Baza *Mammographic Mass Data Set* posiada dane niekompletne. Usunięcie tych wierszy spowodowałoby utratę dużej ilości danych. Aby uniknąć odrzucenia niekompletnych rekordów, rekordy te powinny zostać uzupełnione. W wyniku analizy histogramów stworzonych po próbie uzupełnienia danych niekompletnych dwoma sposobami, ostatecznie dane niekompletne zostaną zamienione na wartości mediany/średniej poszczególnej cechy w danej klasie (zmiana łagodna/zmiana złośliwa). Procedura uzupełniania polega na podzieleniu danej kolumny na dwa osobne zbiory ze względu na zmianę łagodną lub złośliwą. Następnie w każdym z tych zbiorów wyznaczana jest wartość mediany/średniej z pominięciem wartości NaN. Następnie w każdym zbiorze wyszukiwano pozycji niekompletnej i przypisywano wartość mediany/średniej w danej klasie. Znalazło się również kilka pozycji, które odbiegały od przyjętych z góry wartości (np. wartość 55 w pierwszej kolumnie, gdzie pierwsza kolumna ma zakres 1-5). Takie rekordy zostawały również zmienione na wartości mediany/średniej badanej cechy w klasie.

Ze względu na z góry założoną skalę przy cechach *BI-RADS assessment, Shape, Margin, Density* wartości niekompletne zostały uzupełnione wartościami mediany. Natomiast nieokreślone dane w kolumnie *Age* zostały przypisane wartości średniej tej kolumny w danej klasie.

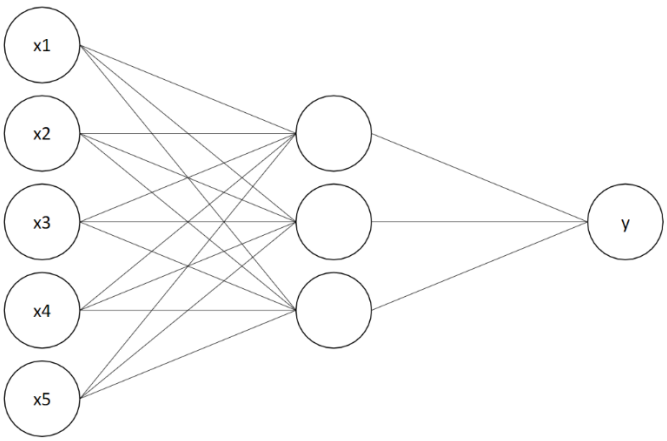
2.5 Podział danych na dane treningowe i testowe

Dane treningowe to zbiór losowo wybranych danych z dostępnej bazy. Wartość tego zbioru to 15% z udostępnionej bazy. Pozostałe 85% to zbiór danych testowych.

3 Koncepcja realizacji sieci neuronowej

3.1 Struktura sieci

Zaprojektowana sieć będzie siecią wielowarstwową perceptronową jednokierunkową.



Rysunek 2. Schemat planowanej sieci

Liczba wejść modelu odpowiada ilości danych wejściowych - 5 cech, na podstawie których odbywa się klasyfikacja. Na wyjściu będzie jeden neuron, gdyż klasyfikacja odbywa

się poprzez przypisanie danego pacjenta o kategorii zmiana łagodna (0) lub zmiana złośliwa (1). Ilość neuronów w warstwie ukrytej została obliczona ze wzorów:

$$n_{ukr} = \sqrt{n_{wej} * n_{wyj}} = \sqrt{5 \cdot 1} \approx 2{,}24 \qquad (3.1)$$

$$n_{ukr} = \frac{n_{wej}}{2} + n_{wyj} = \frac{5}{2} + 1 = 3{,}5 \qquad (3.2)$$

gdzie:

$n_{wej} - liczba\ neuronów\ wejściowych$
$n_{wyj} - liczba\ neuronów\ wyjściowych$
$n_{ukr} - liczba\ neuronów\ w\ warstwie\ ukrytej$

Następnie dokonane zostało uśrednienie wartości dwóch otrzymanych wyników i ostatecznie ilość neuronów w warstwie ukrytej zostało oszacowane na liczbę 3. Dobranie odpowiedniej ilości neuronów w warstwie ukrytej jest znaczącym elementem prawidłowego działania sieci oraz uniknięcia niepożądanych efektów jakim może być na przykład przetrenowanie sieci.

Przedstawiona powyżej koncepcja sieci jest modelem wstępnym, w czasie trwania projektu może ulec zmianie.

3.2 Funkcja aktywacji

3.2.1 Funkcja aktywacji warstwy ukrytej

Do aktywacji warstwy ukrytej użyjemy funkcji ReLu, która jest określona wzorem:

$$f(x) = \begin{cases} x\ dla\ x > 0 \\ 0\ dla\ \ x \leq 0 \end{cases} \qquad (3.3)$$

Funkcja ta jest najczęściej stosowana do aktywacji warstw ukrytych. Dzięki niej gradient nie znika (ograniczenie problemu znikającego gradientu) oraz jest efektywna obliczeniowo. Problemem jednak jest, że funkcja nie jest różniczkowalna dla x = 0.[5]

3.2.2 Funkcja aktywacji warstwy wyjściowej

Ze względu na charakter wartości wyjściowych (0 lub 1) funkcją aktywacji warstwy wyjściowej będzie funkcja sigmoidalna unipolarna:

$$y_m = \frac{1}{1+exp(-\beta * v_m)} \qquad (3.4)$$

gdzie:

$\beta \in (0,1]$

3.3 Algorytm uczenia

Algorytmem wykorzystywanym do uczenia sieci będzie algorytm wstecznej propagacji błędów. Metoda ta umożliwia uczenie sieci wielowarstwowych poprzez propagację różnicy pomiędzy pożądanym a otrzymanym sygnałem na wyjściu sieci.

Algorytm:

1. Przyjęcie małych, losowych wartości wag $w_{l,n}$ oraz $\omega_{m,l}$.
2. Podanie na wejście sieci losowy wektor uczący $x^k = \left[x_1^k \;\; x_2^k \;\; \ldots \;\; x_N^k\right]^T$.
3. Obliczenie pobudzenia neuronów warstwy ukrytej, przy pomocy wzoru:

$$v^k = \sum_{n=1}^{N} w_{l,n} \cdot x_n^k \tag{3.5}$$

4. Obliczenie stanu wyjść neuronów warstwy ukrytej:

$$\xi_l^k = f\left(v_l^k\right) \tag{3.6}$$

5. Obliczenie pobudzenia neuronów warstwy wyjściowej, przy pomocy wzoru:

$$v_m^k = \sum_{l=1}^{L} \omega_{m,l} * \xi_l^k \tag{3.7}$$

6. Obliczenie stanu wyjść neuronów warstwy wyjściowej:

$$y_m^k = f\left(v_m^k\right) \tag{3.8}$$

7. Obliczenie sygnału błędy dla warstwy wyjściowej:

$$\delta_m^{(o)k} = \left(d_m^k - y_m^k\right) * f'(v_m^k) \tag{3.9}$$

8. Obliczenie sygnału błędy dla warstwy ukrytej:

$$\delta_l^{(h)k} = f'\left(v_l^k\right) * \sum_{m=1}^{M} \left(\omega_{m,l} * \delta_m^{(0)k}\right) \tag{3.10}$$

9. Modyfikacja wagi warstwy wyjściowej:

$$\omega_{m,l}(t+1) = \omega_{m,l}(t) + \eta * \delta_m^{(o)k} * \xi_l^k \tag{3.11}$$

10. Modyfikacja wagi warstwy ukrytej:

$$w_{l,n}(t+1) = w_{l,n}(t) + \eta * \delta_l^{(h)k} * x_n^k \tag{3.12}$$

W przypadku poziomu błędu, który nie jest zadowalający, następuję powtórzenie całego cyklu. Funkcja błędu dla k-tego wzorca opisana jest wzorem:

$$E^k = \frac{1}{2} \cdot \sum_{m=1}^{M} \left(d_m^k - y_m^k\right)^2 \qquad (3.13)$$

Dzięki tej funkcji istnieje możliwość obserwacji jak zmienia się całkowita wartość błędu danych wejściowych. Jeśli błąd danych testowych będzie wzrastał powinnyśmy rozważyć zmniejszenie liczby węzłów warstwy ukrytej.

Należy również brać pod uwagę zatrzymanie się sieci w minimum lokalnym funkcji minimalizującej błąd. W takim przypadku należy wielokrotnie powtarzać proces uczenia, przyjmować mniejsze wartości początkowych wag oraz zmieniać wartości współczynnika uczenia η.

4 Bibliografia

1. UCI Machine Learning Repository: Mammographic Mass Data Set
2. https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms/understanding-your-mammogram-report.html
3. https://www.researchgate.net/figure/Different-breast-abnormalities-a-mass-shapes-and-b-mass-margins_fig4_221728770
4. https://www.naukowiec.org/wzory/statystyka/test-z_125.html
5. https://www.is.umk.pl/~grochu/wiki/lib/exe/fetch.php?media=zajecia:nn_2018_1:nn-wyklad.pdf
6. "Sieci neuronowe do przetwarzania informacji" Stanisław Osowski
7. https://www.cri.agh.edu.pl/uczelnia/tad/inteligencja_obliczeniowa/07%20-%20Wyb%C3%B3r%20struktury%20sieci.pdf?fbclid=IwAR0NoYUNOhX__WISRWocfx0iS7BJ8nCdGlJjjm3P8qr9POW7C91pldTB8Rs
8. prezentacje do wykładu *SNB - Sieci neuronowe w zastosowaniach biomedycznych*

5 Oświadczenie

Oświadczam, że niniejsza praca stanowiąca podstawę do uznania osiągnięcia efektów uczenia się z przedmiotu *Sieci neuronowe w zastosowaniach biomedycznych (SNB)* została wykonana przeze mnie samodzielnie.

Adamczyk Jagoda, 289214
Krakowiak Aleksandra, 290292