Jacek Bartman

Wpływ opisu danych na efektywność uczenia oraz pracy sztucznej sieci neuronowej na przykładzie identyfikacji białek

Edukacja - Technika - Informatyka 4/2, 358-365

2013

Jacek BARTMAN
Uniwersytet Rzeszowski, Instytut Techniki, Polska

Wpływ opisu danych na efektywność uczenia oraz pracy sztucznej sieci neuronowej na przykładzie identyfikacji białek

Wstęp

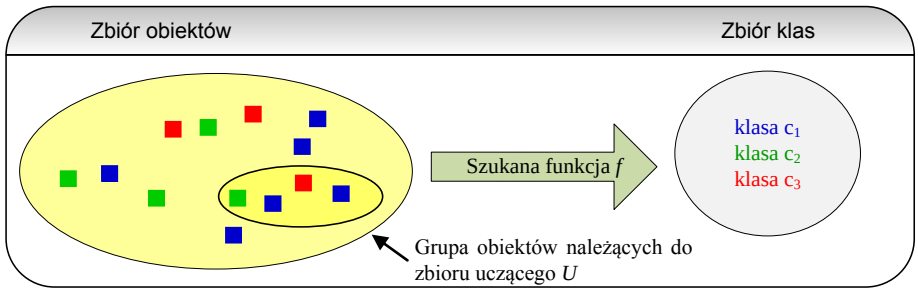
Sztuczne sieci neuronowe stanowią jedną z metod implementacji sztucznej inteligencji. Inspiracją do ich budowy były układy nerwowe istot żywych. Dlatego też sztuczne sieci neuronowe, podobnie jak ich biologiczne pierwowzory, zbudowane są z neuronów, a nieodzownym etapem ich funkcjonowania jest uczenie, które stanowi podstawową metodę zdobywania przez sieć wiedzy [Tadeusiewicz 1993: 10]. Uczenie sieci polega na modyfikacji wag neuronów, tak aby wyposażyć je w niezbędną wiedzę, gdyż to właśnie w wagach zgromadzona jest cała wiedza sieci [Hebb 1949: 2]. Sam proces uczenia sieci neuronowej może mieć charakter uczenia nadzorowanego bądź uczenia nienadzorowanego [Tadeusiewicz 1993: 10] i w pewnym sensie można go traktować jako odpowiednik programowania znanego z informatyki klasycznej. Przebieg procesu uczenia sztucznej sieci neuronowej ma charakter w dużej mierze stochastyczny, a więc w pewnym zakresie nieprzewidywalny. Projektując sztuczną sieć neuronową, dążymy do tego, aby zakres nieprzewidywalności procesu uczenia był jak najmniejszy – niestety, literatura nie podaje gotowych metod, co robić, aby sieć na pewno się nauczyła i aby jej działanie było maksymalnie efektywne. Autorzy opracowań poruszających zagadnienie uczenia sztucznych sieci neuronowych podkreślają jednak, iż zdecydowanie większą efektywność i przewidywalność uczenia sieci uzyskuje się stosując metody uczenia nadzorowanego.

Uczenie nadzorowane

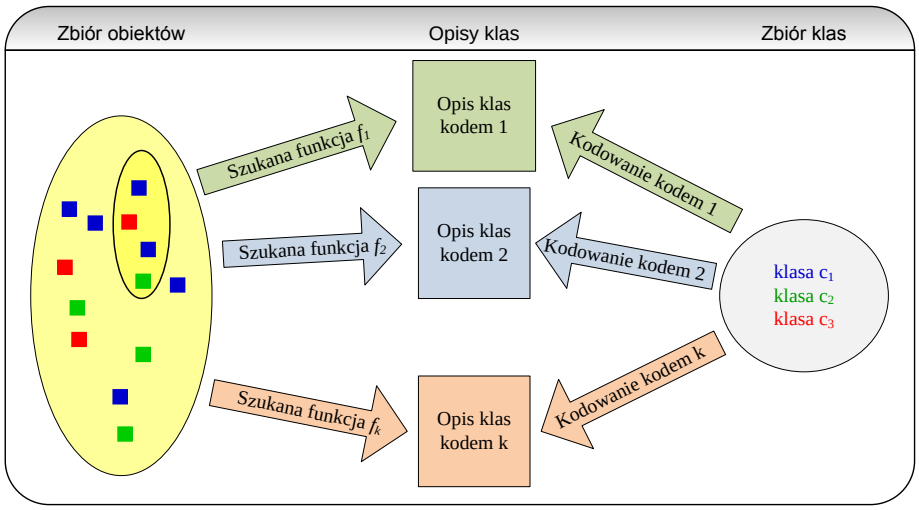
Uczenie nadzorowane polega na prezentacji sieci neuronowej opisu obiektu o_i wraz z etykietą t_j klasy c_j, do której jest on zaliczony [Stąpor 2005: 5]. Zadaniem systemu w naszym przypadku sieci neuronowej jest znalezienie funkcji f przyporządkowującej każdemu obiektowi odpowiednią klasę (rys. 1). Zbiór uczący w przypadku uczenia nadzorowanego można formalnie opisać zależnością:

$$U = \left\{\left(\boldsymbol{x}_i = \ opis(o_i),\ \boldsymbol{t}_j = f(\boldsymbol{x}_i)\right)\right\}_{i=1}^{N} \qquad (1)$$

gdzie: U – zbiór uczący,
N – liczba obiektów w zbiorze uczącym,
o_i – i-ty obiekt należący do zbioru uczącego (opis obiektu),
$\boldsymbol{x}_i$ – wektor cech i-tego obiektu,
$\boldsymbol{t}_j$ – etykieta j-tej klasy (klasy do której należy i-ty obiekt),
f – funkcja kwalifikująca.



Rys. 1. Schemat zależności pomiędzy zbiorem przykładów, zbiorem klas oraz szukaną funkcją segregującą (kwalifikującą)



Rys. 2. Schemat zależności pomiędzy zbiorem przykładów, zbiorem klas oraz szukaną funkcją segregującą (kwalifikującą) uwzględniający różne sposoby kodowania klas

Zauważmy, iż zgodnie z zależnością (1) funkcja f przyporządkowuje obiekt do etykiety klasy, a nie do klasy. Etykieta wprawdzie jednoznacznie określa klasę, jednak łatwo sobie wyobrazić, iż można stosować różne formy opisu etykiet klas –

różne metody ich kodowania. W konsekwencji prowadzi to do sytuacji, że w zależności od przyjętego sposobu kodowania szukana jest inna funkcja kwalifikująca *f* (rys. 2). Tym samym oczywiste wydaje się, iż dobór sposobu opisu klas (sposobu ich zakodowania) może mieć istotny wpływ na to, czy istnieje satysfakcjonująca nas funkcja kwalifikująca *f* oraz jak łatwo można ją odnaleźć.

Dla nietrywialnych przypadków najczęściej nie jest możliwe znalezienie funkcji segregującej *f*, która prawidłowo klasyfikowałaby wszystkie obiekty. Dlatego też w procesie uczenia dopuszcza się pewną liczbę błędnych klasyfikacji – poszukując przybliżenia funkcją φ szukanej funkcji *f*. Bardzo ważne jest dobranie właściwego kryterium oceny przybliżenia – tak aby możliwe było jego znalezienie, a jednocześnie aby było ono jak najlepsze [Stąpor 2005: 9].

Zakres i metodologia badań

Celem pracy było zbadanie, jaki wpływ na skuteczność uczenia nadzorowanego sieci oraz na efektywność jej pracy może mieć sposób, w jaki zakodowano dane wykorzystywane do jej uczenia. Testy przeprowadzono poddając identyfikacji 8 rodzajów białek: albuminę ludzką, kazeinę, hemoglobinę α, hemoglobinę β, inwolukrynę, klaudynę-4, albuminę bydlęcą oraz akwaporynę-1. Identyfikację przeprowadzono na podstawie dwóch cech białek: liczby aminokwasów oraz masy cząsteczkowej. Zarówno liczba aminokwasów, jak i masa cząsteczkowa stanowiące opis białek (patrz zależność 1) są wartościami liczbowymi, można więc wprost podać je na wejście sieci. Nieco inaczej jest z opisem klas, do których one należą. Klasy te podane są w postaci nazw białek, nie jest więc możliwe wykorzystanie ich w takiej formie do uczenia sieci – konieczne jest zakodowanie opisu klas w postaci liczbowej. Metod kodowania jest bardzo wiele, w pracy zdecydowano się wykorzystać trzy (tabela 1):

- kodowanie liniowe – białka zostały ponumerowane liczbami całkowitymi od 0 do 7,
- kodowanie binarne – białka podobnie jak w poprzedniej metodzie ponumerowano od 0 do 7, ale liczby zapisano w kodzie binarnym,
- kodowanie 1 z N – białka opisano liczbami złożonymi z 8 cyfr (bo tyle białek należało rozróżnić), ale w całej liczbie tylko jedna cyfra miała wartość 1, a pozostałe 0.

Zestaw 1639 białek użytych do uczenia oraz symulacji pracy sieci zaczerpnięto z pracy poświęconej identyfikacji białek [Madej 2013: 6], w której wykorzystano materiały ze specjalistycznych baz danych, artykułów oraz literatury [www.mybiosource.com; www.uniprot.org; www.ncbi.nlm.nih.gov/pubmed; www.drugbank.ca; www.ionsource.com].

Tabela 1

Kody opisujące rodzaje białek

Nazwa białka	Kod liniowy	Kod binarny	Kod 1 z N
Albumina ludzka	0	000	10000000
Kazeina	2	001	01000000
Hemoglobina α	2	010	00100000
Hemoglobina β	3	011	00010000
Inwolukryna	4	100	00001000
Klaudyna-4	5	101	00000100
Albumina bydlęca	6	110	00000010
Akwaporyna-1	7	111	00000001

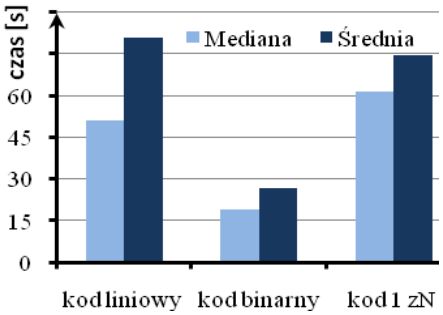
Wyniki badań

Uczenie sieci oraz symulację jej pracy przeprowadzono wykorzystując pakiet MatLab wyposażony w toolbox Neural Networks, pracujący pod kontrolą systemu operacyjnego Windows 8 na komputerze z procesorem Intel i5 3317 oraz 8GB pamięci RAM.

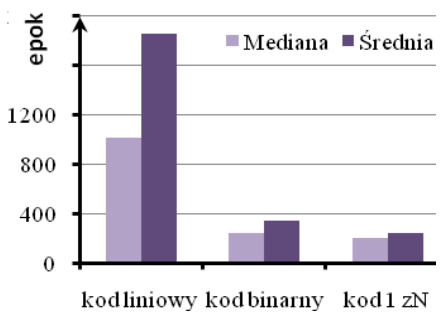
Tabela 2

Zestawienie wyników badań dla wszystkich analizowanych kodów

		Kod liniowy	Kod binarny	Kod 1 z N
Liczba epok uczenia	Min.	356	138	76
	Max	8899	6376	1501
	Średnia	1858	346	245
	Mediana	1017	244	203
Czas uczenia [s]	Min.	18,19	10,87	22,80
	Max	447,40	487,04	475,91
	Średni	80,72	26,64	74,66
	Mediana	51,04	18,88	61,29
Efektywność pracy [%]	Min.	100	99,39	96,94
	Max	100	100	100
	Średnia	100	99,74	99,43
Niepowodzenie uczenia [%]		3	0	0



Rys. 3. Porównanie liczby epok uczenia identyfikacji białek w zależności od sposobu zakodowania danych



Rys. 4. Porównanie czasów uczenia identyfikacji białek w zależności od sposobu zakodowania danych

Badano wpływ sposobu zakodowania danych uczących (zakodowania nazw białek) sieci na:

– czas uczenia mierzony w sekundach,
– liczbę epok uczenia,
– efektywność pracy nauczonej sieci.

Jak wspomniano wcześniej, uczenie sztucznej sieci neuronowej nie jest procesem powtarzalnym – wpływ na jego przebieg mają wartości początkowe wag neuronów, które są losowane oraz zależny od nich kształt funkcji celu. W celu uzyskania rzetelnych wyników dla każdego z wymienionych wyżej sposobów kodowania 100-krotnie powtórzono uczenie sieci oraz zbadano efektywność jej pracy. W tabeli 2 zebrano wartości minimalne, maksymalne, średnie oraz mediany każdego z badanych parametrów dla każdego ze sposobów kodowania.

Najczęściej mówiąc o szybkości uczenia sieci neuronowej, podaje się tylko liczbę epok. Liczba epok mówi nam jedynie, ile iteracji uczących zostało wykonanych, aby sieć się nauczyła. Nie uwzględnia ona złożoności obliczeniowej algorytmu uczącego oraz wynikającej z różnicy w kodowaniu wyjść sieci zmiany rozmiaru jej parametrów. W omawianym przypadku złożoność algorytmu nie ma znaczenia, gdyż każdorazowo do uczenia sieci wykorzystywano metodę Levenberga-Marquardta. Jednak zmiana sposobu opisu nazw białek skutkuje zmianą rozmiaru macierzy wag warstwy wyjściowej sieci, gdyż w przypadku kodowania liniowego wykorzystywany jest tylko 1 neuron wyjściowy, kodowania binarnego 3 neurony, a kodowania 1 z N aż 8 neuronów. Powyższe przesłanki spowodowały, iż w niniejszej pracy ocena szybkości uczenia sieci opiera się na analizie liczby epok uczących oraz czasu ucznia.

Oceniając uczenie sieci na podstawie liczby epok uczenia (tabela 2, rys. 3), można stwierdzić, iż najlepiej wypada kodowanie 1 z N (do nauczenia sieci potrzeba średnio 245 epok), niewiele odbiega kodowanie binarne (336 epok), zdecydowanie najwięcej iteracji uczących wymaga kodowanie liniowe (1858 epok).

Nieco inaczej wyglądają czasy uczenia (tabela 2, rys. 4), najszybciej uczy się sieć wykorzystująca kod binarny (średni czas uczenia to 26,64 s), pozostałe dwie metody wypadają znacznie gorzej: uczenie wykorzystujące kodowanie liniowe potrzebuje średnio 80,72 s, a kodowanie 1 z N 74,66 s. Wyniki te pokazują jednoznacznie, iż wzrost rozmiaru macierzy wag warstwy wyjściowej negatywnie wpłynął na czas uczenia sieci.

Rys. 5. Porównanie efektywności pracy sieci w zależności od sposobu zakodowania danych

Efektywność pracy nauczonej sieci była bardzo dobra. W przypadku kodowania liniowego uzyskano 100% efektywność identyfikacji, zaś w przypadku kodowania binarnego wynosiła ona 99,74%, a kodowania 1 z N 99,53% (tabela 2, rys. 5). Warto przypomnieć, iż wykonano po 100 testów identyfikacji 1639 białek.

Podsumowanie

Przeprowadzone badania miały na celu zbadanie, czy i jaki wpływ na skuteczność uczenia sieci oraz na efektywność jej pracy może mieć sposób, w jaki zakodowano dane wykorzystywane do jej uczenia. Rezultaty pozwalają stwierdzić, iż w badanym przypadku:

- sposób zakodowania danych uczących wpływa na łatwość uczenia się sieci, nie stwierdzono jednak, aby wpływał na zdolność sieci do nauczenia się,
- przyporządkowanie jednemu neuronowi wyjściowemu wielu klas powoduje, iż uczenie sieci jest bardzo powolne i wymaga wykonania dużej liczby iteracji (porównaj: kodowanie liniowe),
- przyporządkowanie każdemu neuronowi warstwy wyjściowej tylko jednej klasy powoduje duży wzrost liczby wag warstwy wyjściowej, co w konsekwencji wydłuża bezwzględny czas uczenia (patrz: kodowanie 1 z N),

Przeprowadzone badania dotyczyły tylko jednego zadania identyfikacji, dlatego też – mimo iż są one zgodne z naszą intuicją – w celu wyciągnięcia ogólnych wniosków należałoby je rozszerzyć.

Literatura

Bielecki A. (2003), *Mathematical model of architecture and learning processes of artificial neural networks*, „TASK Quarterly 7", no. 1.

Hebb D. (1949), *The Iragnization of Behavior*, New York.

Korbicz J., Obuchowicz A., Uciński D. (1994), *Sztuczne sieci neuronowe. Podstawy i zastosowania*, Warszawa.

Kwater T., Bartman J., Atamanyuk I., Sidenko E. (2011), *Diagnosis of apples by automatic classification of objects*, Computing in Science and Technology, Monographs in applied informatics.

Madej K. (2013), *Projekt neuropodobnego systemu identyfikacji białek*. Praca inżynierska, Rzeszów.

McClelland T.L., Rumelhart D.E. and the PDP Research Group (1986), *Paralell Distributed Processing*, MIT Press, Cambrideg, Mass.

Osowski St. (1996), *Sieci neuronowe w ujęciu algorytmicznym*, Warszawa.

Rosenblatt F. (1958), *The perceptron: A probalistic model for information storage and organization in the brain*, „Psychology Review", no. 65.

Stąpor K. (2005), *Automatyczna klasyfikacja obiektów*, Wydawnictwo Exit.

Tadeusiewicz R. (1993), *Sieci neuronowe*, Warszawa.

Żurada J., Barski M., Jędruch W. (1996), *Sztuczne sieci neuronowe. Podstawa teorii i zastosowania*, Warszawa.

www.mybiosource.com

www.uniprot.org

www.ncbi.nlm.nih.gov/pubmed

www.drugbank.ca

www.ionsource.com

Streszczenie

Uczenie jednokierunkowych wielowarstwowych sztucznych sieci neuronowych jest zagadnieniem szeroko omawianym w literaturze. Autorzy większości opracowań skupiają się na metodach uczenia, zdecydowanie mniej prac poświęconych jest wpływowi preprocesingu danych na uczenie i efektywność pracy sieci.

Skoro uczenie sztucznych sieci neuronowych jest szukaniem funkcji odwzorowującej zbiór danych wejściowych w zbiór oczekiwanych odpowiedzi, to czego możemy oczekiwać, jeżeli zmienimy opis danych uczących? Zmienia się funkcja odwzorowująca, a więc szukamy innej funkcji, zatem jest możliwe, iż sposób kodowania danych wpływa na efektywność uczenia i pracy sieci. Niniejsza praca dotyka przedstawione zagadnienie badając wpływ sposobu zakodowa-

nia opisu białek na efektywność uczenia oraz pracy sieci neuronowej identyfikującej rodzaj białka.

Słowa kluczowe: sztuczna sieć neuronowa, uczenie.

Influence of data description on efficiency of learning and job artificial neural network on example of identification of proteins

Abstract

Learning feedforward multilayer neural networks is an issue widely discussed in the literature. The authors of the most works focus on methods of learning, much less work is devoted to the influence of data preprocessing on learning and the efficiency of the network.

If learning of artificial neural networks is finding the mapping function set of input data into a set of expected responses, what you can expect if you change the description of the data learners? Changes of mapping functions, and so we are looking for another function, so it is possible that the encoding of data affects the efficiency of learning and job of the network. This paper touches the issue presented by examining the impact of coding method information about the proteins on the effectiveness of learning and the work of the neural network identifies the type of protein.

Key words: artificial neural network, learning.