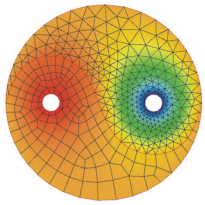
Publishing House
AKAPIT

COMPUTER METHODS IN MATERIALS SCIENCE

Informatyka w Technologii Materiałów

Vol. 6, 2006, No. 2



OPTYMALIZACJA MODELI NEURONOWYCH NA PRZYKŁADZIE OCENY AKTYWNOŚCI BIOLOGICZNEJ ZWIĄZKÓW CHEMICZNYCH

MACIEJ SZALENIEC[1], RYSZARD TADEUSIEWICZ[2], ANDRZEJ SKOCZOWSKI[3]

[1] Instytut Katalizy i Fizykochemii Powierzchni PAN, 30-239 Kraków, ul. Niezapominajek 8
[2] Akademia Górniczo-Hutnicza im. Stanisława Staszica, 30-059 Kraków, al. Mickiewicza 30
[3] Instytut Fizjologii Roślin im. Franciszka Górskiego PAN, 30-239 Kraków, ul. Niezapominajek 21

OPTIMIZATION OF NEURAL MODELS USING EVALUATION OF BIOLOGICAL ACTIVITY OF CHEMICAL SUBSTANCES AS AN EXAMPLE

Abstract

Neural networks (NNs) are tools that are very frequently successfully applied in the modeling of various phenomena and processes. This is due to combination of characteristic for NNs wide approximation capabilities (manifesting especially in nonlinear modeling tasks) with their flexibility and high performance in fitting the model to the real data during the learning process. Taken together these features make NNs one of the best modeling tools available. However, it is a common practice to achieve success with neural network technique in a modeling of particular system while confining the research only to neural model selection, optimization of parameters and validation of the NN performance goodness. Frequently, neural models predictions are analyzed and compared with other modeling techniques or other neural systems. In this paper we provide a complementary approach to the above-mentioned scheme. We took one non-trivial modeling task as an example (i.e. prediction of biological activity of chemical compounds based on their structure and properties) and studied various types of neural networks in order to determine the optimal type of NN, which deals with modeling problem in the most efficient way. We analyzed both linear and non-linear neural networks of MLP and GRNN type. In non-linear MLP systems the linear or non-linear output layers were tested. Moreover a hybrid neural system was developed that joins results of architecture optimization of MLP and GRNN. The paper addresses also the issue of input parameters selection, optimal number of hidden neurons and data representation, especially in terms of an output results. A dozen or so thousands of neural models were developed, providing a rich dataset for assessment of neural networks usefulness. It seems that such a comparative study can be of a high value for other researchers using neural systems in modeling studies. It should allow to chose a type and size of NN used based less on arbitrary and more on rational basis. Our results provide also better understanding into the character and cause-result relationship of processes that take place in neural networks.

Słowa kluczowe: sieci neuronowe, sieci klasyfikacyjne, sieci regresyjne, sieci realizujące uogólnioną regresję, przewidywanie aktywności biologicznej, GRNN, MLP

1. WPROWADZENIE

Praca ta ma na celu sprawdzenie przydatności różnych typów i różnych struktur sieci neuronowych oraz różnych technik ich uczenia w kontekście zadania przewidywania właściwości określonych związków chemicznych jeszcze przed przebadaniem empirycznym tych właściwości metodami laboratoryjnymi. Jest oczywiste, że wobec ogromnej liczby i różnorodności możliwych do syntetyzowania związków, takie przewidywanie ich właściwości będące przedmiotem zainteresowania na drodze modelowania komputerowego jest bardzo atrakcyjną alternatywą dla kosztownych badań eksperymental-

ISSN 1641-8581

nych. Metodyka opisana w tym artykule może być użyteczna do przewidywania **różnych** właściwości **różnych** grup związków, ale konkretne badania przedstawione niżej dotyczyć będą przewidywania aktywności chemicznej związków określanych jako alkiloaromatyczne i alkiloheterocykliczne w enzymatycznej reakcji katalizowanej przez dehydrogenazę etylobenzenową. Z punktu widzenia chemicznego wyniki uzyskane w tych badaniach przedstawiono i omówiono w pracach (Szaleniec i in., 2006a, 2006b), natomiast ten artykuł ma za zadanie przedstawienie tego problemu (i uzyskanych rozwiązań) z punktu widzenia techniki sieci neuronowych i optymalizacji obliczeń neuronowych.

Przydatność i wielostronną użyteczność sieci neuronowych (Weiss i in., 2006) wykazano już na setkach problemów dotyczących różnych, często bardzo odległych od siebie dziedzin. Trudno w tym zakresie dokonać jakiegokolwiek bardziej systematycznego przeglądu w sytuacji, kiedy dosłownie każdego dnia ukazują się nowe prace donoszące o sukcesach (rzadziej publikowane są prace o niepowodzeniach, chociaż i one także się zdarzają!) niezliczonych badań, prowadzonych na całym świecie w tej dziedzinie. Próba poszukiwania w Internecie artykułów związanych z hasłem *Neural Networks Applications* przynosi około 20 milionów odnośników do różnych informacji, poczynając od krótkich wzmianek sygnalnych o prowadzonych pracach i osiąganych wynikach, poprzez liczne dobrze opracowane artykuły w czasopismach i internetowe vortale tematyczne poświęcone sieciom neuronowym oraz ich zastosowaniom, aż do obszernych książkowych i sieciowych opracowań monograficznych, dających syntetyczny pogląd na wiele aspektów rozważanego tu zagadnienia.

Nawet ograniczając się tylko do prac najnowszych, to znaczy pochodzących z okresu ostatniego roku (poprzedzającego napisanie tej pracy), i do zagadnień mających pośredni lub bezpośredni związek z chemią, można przywołać szereg przykładowych publikacji, w których pokazano wielokrotnie i na bazie wielu zadań, że modele budowane w oparciu o technikę sieci neuronowych potrafią dobrze opisywać różne skomplikowane zjawiska, dostarczając możliwości ich lepszej kontroli, prognozy i sterowania – niż techniki konkurencyjne.

Najbliższa rozważanemu tu zagadnieniu wydaje się praca (Mei i in., 2005), w której opisano użycie sieci neuronowej do rozwiązania zagadnienia typu QSAR (*Quantitative Structure-Activity Relationships*) w odniesieniu do wybranej grupy 48 bipeptydów ocenianych z punktu widzenia ich aktywności biologicznej jako inhibitorów konwertazy angiotensyny. Nieco podobna jest praca Plewczynskiego i Kocha (Plewczynski i Koch 2006), w której prognozowano za pomocą sieci neuronowych aktywność biologiczną określonych ligandów w odniesieniu do pięciu rodzajów obiektów biologicznych. W pracy tej udało się pokazać, że spośród licznych testowanych metod automatycznej klasyfikacji (algorytm *k*-średnich, metoda SVM, techniki oparte na uczących się drzewach decyzyjnych oraz na algorytmach genetycznych itp.) – to właśnie sieci neuronowe okazały się najskuteczniejszym narzędziem dla uzyskania wiarygodnej prognozy poszukiwanej aktywności. Z kolei w pracy (Hong i Chunsheng, 2006) zaprezentowano model („metaforę", jak to nazwano) przebiegu reakcji chemicznej, nawiązujący od strony implementacyjnej do techniki agentowej, ale w istocie silnie korzystający z możliwości modelowania złożonych zjawisk i procesów, jakie stwarzają właśnie sieci neuronowe.

Pewne podobieństwo do zagadnienia rozważanego w przedstawianej tu pracy ma także praca (Fogelman i in. 2006), w której pokazano, jak bardzo sieci neuronowe przewyższają inne metody określania (prognozowania) zapotrzebowania na określone substancje (w rozważanym przypadku – tlen) w określonym typie reaktora chemicznego (w rozważanym przypadku - oczyszczalni ścieków) oraz praca (Pianese i in., 2006), w której podobne zagadnienie rozważano w kontekście proporcji paliwa do powietrza w silnikach spalinowych z zapłonem iskrowym. Na marginesie można dodać, że zagadnienia optymalizacji konstrukcji oraz sterowania różnych typów aparatury chemicznej z wykorzystaniem modeli tworzonych z użyciem sieci neuronowych są obecnie bardzo często podejmowanym tematem różnych publikacji (dla przykładu można wskazać dosyć typową dla tej sfery zastosowań pracę (Yixing i in., 2006)). Wspominamy o tym fakcie, bo rysuje się możliwość korzystania z jednorodnych metod opartych na sieciach neuronowych w pełnym cyklu zadań: poczynając od wyboru struktury pożądanego (w kontekście jakiegoś zastosowania) związku chemicznego, poprzez określenie szczegółów reakcji prowadzących do jego syntezy, aż do projektowania aparatury technologicznej i sterowania procesem wytwarzania.

Jak z tego wynika, podjęty tu temat jest popularny i został już stosunkowo dobrze przebadany. Niemniej większość badaczy zmierzła do osiągnięcia celu praktycznego i używane modele neuronowe traktowała czysto narzędziowo: wybierano arbitral-

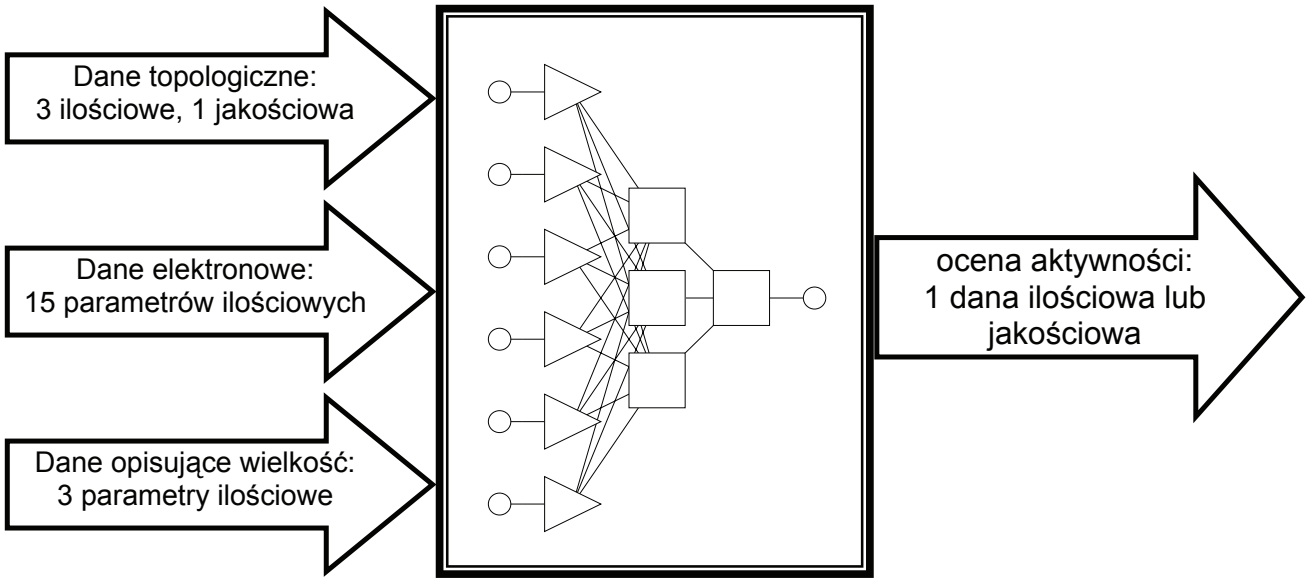
nie jakąś sieć, uzyskiwano przy jej pomocy jakieś wyniki i przedstawiano te wyniki, pomijając lub bardzo ograniczając dyskusję tego, jakie sieci były stosowane, w jaki sposób były wybierana i co by można było w tym samym problemie osiągnąć stosując jeszcze inne sieci (albo inne metody nieneuronowe, na przykład regresyjne).

W tej sytuacji każdy kolejny badacz, podejmujący podobny problem, staje przed poważną rozterką metodologicznej natury: jaką wybrać sieć, w jaki sposób ją uczyć i jak reprezentować dane, żeby uzyskać możliwie najlepsze wyniki. W tej pracy przedstawione zostaną wyniki badań, w których do tego samego (trudnego) problemu przewidywania aktywności chemicznej sporej grupy związków chemicznych – stosowano różne sieci, uzyskując przy ich pomocy różne wyniki. Na tej podstawie zaproponowane zostaną wnioski, które sieci i które metody uczenia okazały się lepsze, a które gorsze w rozważanym zagadnieniu. Wniosków tych nie można mechanicznie uogólniać, bo z każdym właściwie problemem zastosowania sieci neuronowych wiąże się jego unikatowa specyfika, ale autorzy tej pracy wyrażają nadzieję, że ich badania, prowadzone bardzo szerokim frontem i dokładnie dokumentowane, okażą się użyteczne dla innych osób chcących zastosować sieci neuronowe i rozważających, od jakiego modelu rozpocząć swoje poszukiwania.

2. PROBLEM, NA BAZIE KTÓREGO ZEBRANO DOŚWIADCZENIA

Pełny opis badanego zagadnienia od strony sformułowania rozwiązywanego problemu chemicznego oraz od strony dyskusji uzyskanych szczegółowych wyników znaleźć można w przywołanych już publikacjach (Szaleniec i in., 2006a i b). Z punktu widzenia tej pracy ważne wydają się jedynie następujące okoliczności:

- Rozwiązywane zadanie charakteryzuje się tym, że wnioskowanie opiera się na dużej liczbie przesłanek (w maksymalnym zestawie jest ich 22 – patrz rysunek 1). Wśród tych przesłanek występują dane dotyczące topologii rozważanej cząsteczki chemicznej (3 dane ilościowe i jedna dana opisowa, określająca lokalizację podstawnika w stosunku do aktywnego centrum cząsteczki w trzech kategoriach: *para, meta, orto*), dane dotyczące jej parametrów elektronowych (15 danych ilościowych wynikających z obliczeń kwantowo-chemicznych) oraz parametry opisujące jej wielkość (3 wartości ilościowe, w tym dwie pochodzące z obliczeń kwantowo-chemicznych). Od strony wejść do sieci jej złożoność może więc maksymalnie osiągać poziom **24 neuronów wejściowych**, co wynika z faktu, że przy kodowaniu danej jakościowej metodą 1-z-N wejście określające lokalizację podstawnika w pierścieniu aromatycznym wymaga wykorzystania trzech neuronów (dla trzech możliwych kategorii) w wejściowej warstwie.
- Ze względu na różne oceny przydatności (ważności) różnych danych wejściowych przy prowadzeniu niektórych eksperymentów liczba sygnałów wejściowych była ograniczana, stąd wiele opisywanych dalej sieci ma mniejszą liczbę wejść, niż by to wynikało z bilansu danych pokazanych na rysunku 1.
- Redukcja liczby danych wejściowych bywała nie tylko punktem wyjścia do budowy sieci, ale



Rysunek 1. *Ogólny schemat rozważanego w pracy modelu neuronowego.*
Figure 1. *General scheme of neural model.*

także często ważnym wynikiem jej działania. Względna ocena ważności poszczególnych wejść była bowiem możliwa *ex post*, na podstawie analizy parametrów wytrenowanej sieci albo na podstawie analizy wrażliwości neuronowego modelu na zmiany jego poszczególnych wejść. Porównanie działania sieci przy różnej liczbie danych wejściowych jest jednym z wyników badawczych, dyskutowanych w tej pracy.

- Model zawsze miał tylko jeden sygnał wyjściowy, jednak zależnie od sposobu sformułowania problemu wyjście to mogło mieć interpretację zmiennej typu ilościowego lub typu jakościowego. Z faktu, że rozważana sieć ma zawsze tylko jedno wyjście, wynikała między innymi ta okoliczność, że przy wyborze architektury i sposobu działania sieci można było korzystać także z tych struktur, które są przeznaczone wyłącznie do modeli o jednym wyjściu (na przykład GRNN). Natomiast dzięki temu, że ocena aktywności mogła być rozważana raz jako dana ilościowa (gdy brane były pod uwagę wyznaczone doświadczalnie wartości miary poziomu aktywności), a innym razem aktywność była kwantyfikowana w kilku umownych kategoriach np. jako ujemna (inhibitory), mała, średnia i duża, możliwe było porównywanie sprawności działania rozważanych sieci działających w jednym ze znanych reżimów – albo regresyjnym, albo klasyfikacyjnym.

Badaniom poddane zostały różne aspekty działania sieci, które będą kolejno dyskutowane w następnych rozdziałach tej pracy. Prezentując te wyniki autorzy wyrażają przypuszczenie, że wciąż jeszcze mogą być wartościowe i ciekawe wyniki badań, pokazujących jak radzą sobie sieci neuronowe z pewnymi szczególnymi zadaniami, dotyczącymi modelowania złożonych, nieliniowych związków przyczynowo-skutkowych, ważnych z punktu widzenia określonych obszarów zastosowań. Wyniki takie wzbogacają nie tylko konkretną dziedzinę, dla której dedykowane są tworzone neuronowe modele, ale także tworzą empiryczne zręby wiedzy na temat nerurocomputingu jako takiego, ważne i potrzebne dlatego, że ogólna metodologia tworzenia i eksploatacji sieci neuronowych nie jest wciąż jeszcze definitywnie opracowana.

Dla pełnego scharakteryzowania problemu, który był rozwiązywany, podamy jeszcze wykaz związków chemicznych, które tworzyły zbiór uczący (wraz z ich względnymi aktywnościami – tablica 1).

***Tablica 1**. Zawartość zbioru uczącego. Aktywność (stała szybkości reakcji k_{cat}) podana w %, przyjmując aktywność etylobenzenu za 100%.*

***Table 1**. Dataset used in the study. Biological activity (kinetic constant, kcat) was provided in relative % scale, taking the activity with ethylbenzene as 100%.*

Nazwa związku	Względna aktywność rk_{cat}
1,2-dietylobenzen	0
1,4-dietylobenzen	35
1-etylonaphtalene	0
2-etyloanilina	94.53
2-etylofuran	133.92
2-etylonaftalen	9.3
2-etylofenol	56.14
2-etylopirydyna	0
2-etylopirol	234.63
2-etylotiofen	242.92
2-etylotoluen	3.8
2-metylofuran	0
2-metylopirol	0
2-metylotiofen	0
3-etylofenol	24.31
3-etylopirydyna	16.94
3-etylotoluen	10
4-etylofenol	259.01
4-etylopirydyna	0
4-etylotoluen	28
4-fluoretylobenzen	15
etylobenzen	100
n-propylbenzen	14
toluen	0
4-etyloanilina	134
4-propylofenol	180

W zbiorze tym podczas badań wydzielano (losowo) część danych używaną wprost do uczenia sieci, część używaną jako zbiór walidacyjny dla ustalenie momentu przerwania uczenia w celu uniknięcia efektu zaniku zdolności do generalizacji oraz część testową, na której badano końcowy wynik w celu sprawdzenia, czy przypadkowa koincydencja danych uczących i walidacyjnych nie doprowadziła do zafałszowania wyników uczenia i funkcjonowania sieci. Proporcje, w jakich dzielono posiadany zasób danych na wymienione podzbiory przyjmowana była w sposób następujący: 14:2:2 (U:W:T) dla sieci liniowych przewidujących wyłącznie aktywność substratów (18 związków) oraz 18:4:4 (U:W:T) w przypadku sieci MLP przewidujących aktywność zarówno substratów jak i inhibitorów (26 związków).

W badaniach, w których używane były sieci klasyfikacyjne powyższe dane skonwertowano w taki sposób, że utworzono cztery klasy aktywności, oznaczone odpowiednio kodami:

-1 - inhibitory
0 – aktywność do 50%,
1 – 51-150 %,
2 – powyżej 150 %

liczność klas była następująca:

klasa „-1" – 8 związków
klasa „0" – 9 związków
klasa „1" – 5 związków
klasa „2" – 4 związków

Wszystkie sieci były tworzone, uczone i badane za pomocą programu *Statistica® Neural Networks 7.1* [www.statsoft.pl], przy czym dla określonego typu sieci (na przykład MLP) korzystano z możliwości, jakie w tym programie daje opcja *Automatyczny projektant* i typowo tworzono oraz badano co najmniej 1000 modeli sieci wybranego typu. W wyniku optymalizacji architektury oraz wektora wejściowego przez *Automatycznego projektanta* otrzymywano 50 najlepszych sieci, z których do dalszej analizy wyselekcjonowano wyłącznie tę sieć, która w testach uzyskała na posiadanych danych najlepsze wyniki.

3. WYBÓR STRUKTURY I TYPU SIECI

3.1. Sieć liniowa

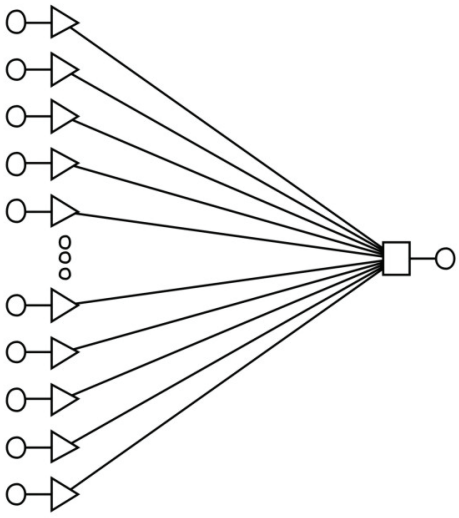
Rozważany w pracy problem był raczej dość skomplikowany, więc z góry zakładano, że konieczne będzie odwołanie się do sieci wielowarstwowych, a więc nieliniowych. Dla uzyskania pełnej orientacji na temat tego, jak zachowywać się będą w tym zadaniu różnego typu sieci, a także w celu umożliwienia porównywania wyników dostarczanych przez sieć neuronową z rozwiązaniami, jaki dostarczają metody QSAR szeroko używane w chemii, zbudowano także i przebadano odpowiednią dla rozważanego problemu sieć liniową (Weiss i in. 2006), przedstawioną na rysunku 2. Warto może w tym miejscu odnotować fakt, że po okresie totalnego odrzucania w dziedzinie sieci neuronowych możliwości korzystania z sieci liniowych (na skutek ich oczywistych ograniczeń), obserwuje się szereg prac (na przykład Hartono i in. 2005), w których postuluje się powrót do sieci liniowych – chociaż często z różnymi modyfikacjami.

W podpisie rysunku 2 (i w wielu innych miejscach tej pracy) przyjęto ogólnie akceptowany sposób zapisywania struktury sieci. Sposób ten polega na podaniu kolejno: liczby sygnałów wejściowych, po dwukropku liczby neuronów wejściowych (liczby te mogą się różnić w przypadku korzystania z danych jakościowych, reprezentowanych metodą 1-z-N), następnie po znakach pauzy kolejne liczby podają liczebności neuronów kolejnych warstw aż do warstwy wyjściowej (można z tego także odczytać liczbę warstw składających się na sieć). Na końcu po podaniu liczby neuronów warstwy wyjściowej po dwukropku podawana jest liczba danych wyjściowych. Brzmi to skomplikowanie i wygląda w pierwszej chwili trochę zagadkowo, ale jest bardzo wygodne i pozwala opisać dowolną sieć. Na przykład zapis (dotyczący jednej z później omawianych tu sieci, mający postać:

15:15-10-4:1

należy interpretować następująco: sieć wykorzystuje 15 sygnałów wejściowych i używa do tego 15 neuronów, co oznacza, że dane wejściowe wszystkie są typu ilościowego. Następnie ma jedną warstwę ukrytą, w której jest 10 neuronów. I na koniec ma warstwę wyjściową złożoną z 4 neuronów, które produkują jednak tylko jeden sygnał wyjściowy – co oznacza, że tym sygnałem wyjściowym jest zmienna jakościowa o 4 możliwych kategoriach.

Wracając od tego ogólnego stwierdzenia dotyczącego notacji do rozważanej tu sieci liniowej można obejrzeć jej strukturę na rysunku 2 i odczytać liczby zaangażowanych neuronów w podpisie rysunku.

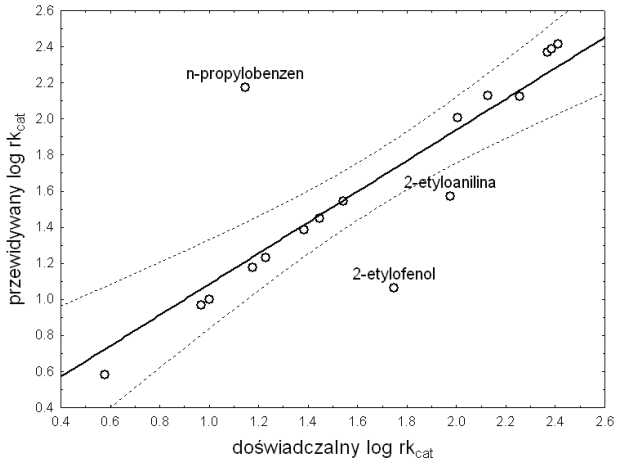


***Rysunek 2**. Testowana w badaniach sieć liniowa o strukturze 20:20-1:1.*

***Figure 2**. The architecture of 20:20-1:1 linear neural network tested in the study.*

W sieci liniowej stosowano oczywiście beziteracyjny algorytm uczenia metodą pseudoinwersji macierzy. Sieć ta, podobnie jak wszystkie inne rozważane w tym artykule, powstała w wyniku procedury optymalizacyjnej, jaką przeprowadził *Automatyczny projektant* pakietu *Statistica® Neural Networks*. W sieci liniowej jak wiadomo struktura oraz para-

metry są jednoznacznie wyznaczona przez zbiór uczący, dlatego optymalizacja jest ograniczona wyłącznie do optymalnego doboru danych wejściowych. Jednak ten właśnie efekt procesu optymalizacji zaobserwowano, mianowicie w wybranej sieci zamiast 24 neuronów wejściowych po optymalizacji przyjęto ich 20, co właśnie było skutkiem automatycznie przeprowadzonej optymalizacji. Zbudowana sieć działała oczywiście jako sieć regresyjna, to znaczy na wyjściu sieci oczekiwano wartości, które mogą być traktowane jako przewidywane przez sieć wartości aktywności badanych związków. Wartości te można bezpośrednio porównać z wartościami aktywności badanych związków uzyskanymi eksperymentalnie.



***Rysunek 3**. Porównanie wartości logarytmów względnej aktywności (rk_{cat}) badanych związków uzyskanych eksperymentalnie (oś pozioma) oraz wyznaczonych na drodze obliczeniowej za pomocą liniowej sieci neuronowej.*

***Figure 3.** The comparison of logarithmic values of experimental relative activity (rk_{cat}) of studied compounds with values calculated by linear neural network.*

Uzyskane wyniki były zaskakująco dobre. Na rysunku 3 pokazano wykres wskazujący na bardzo wysoki stopień korelacji wartości aktywności związków wyznaczonych empirycznie oraz uzyskiwanych na drodze obliczeniowej za pomocą sieci liniowej. Co prawda dla kilku związków (ich nazwy podano na rysunku 3) rozbieżność pomiędzy danymi obliczonymi i zmierzonymi eksperymentalnie były bardzo duże, ale **średnie** wartości błędów, podane w tablicy 2, są zadziwiająco dobre. Wartości prezentowane w tablicy 2 (jak i następnych tablicach dotyczących tego samego tematu) są błędami RMS (Root Mean Squared) sieci wyznaczonymi na podstawie wartości błędów jednostkowych obliczonych za pomocą przyjętej funkcji błędu pomiędzy otrzymaną i rzeczywistą wartością wyjścia. W przypadku sieci regresyjnych zastosowaną funkcją błędu była suma kwadratów różnic pomiędzy wartościami zadanymi i wartościami otrzymanymi na wyjściach każdego neuronu wyjściowego. W przypadku zadania klasyfikacyjnego zastosowano funkcję entropii wielokrotnej, która oblicza sumę iloczynów zadanych wartości oraz logarytmów błędów dla każdego neuronu wyjściowego.

***Tablica 2.** Średnie wartości błędów popełnianych przez sieć liniową.*

***Table 2.** The average error values of linear network.*

Błąd na zbiorze uczącym	Błąd na zbiorze walidacyjnym	Błąd na zbiorze testowym
$1.31 \cdot 10^{-15}$	0.399	0.308

Syntetyczną miarą jakości dopasowania działania sieci neuronowej liniowej do empirycznych danych może być współczynnik korelacji danych przewidywanych przez sieć i wyników uzyskanych empirycznie. Współczynnik ten wynosił dla najlepszej sieci liniowej r = 0.8467 i zapewniał znamienność związku statystycznego tych wartości na poziomie $p = 9 \cdot 10^{-7}$. Wynik uzyskany przez liniową sieć neuronową warto porównać z wynikiem uzyskanym za pomocą metody postępującej krokowej regresji wielorakiej MLR (*Multiple Linear Regression*). Przy zastosowaniu tej metodologii, niewątpliwie szerzej znanej i częściej stosowanej, w wyniku stworzenia dla tych samych danych najlepszego z możliwych modelu MLR powstało równanie oparte o cztery ilościowe zmienne. Oczywiście warunkiem wprowadzenia zmiennych do równania MLR było stwierdzenie, że pozostają one istotnie statystycznie. Przewidywana przez model regresyjny aktywność koreluje z wynikami eksperymentalnymi na poziomie r = 0.8704 (co odpowiada znamienności związku statystycznego tych wartości na poziomie p = 0.00059).

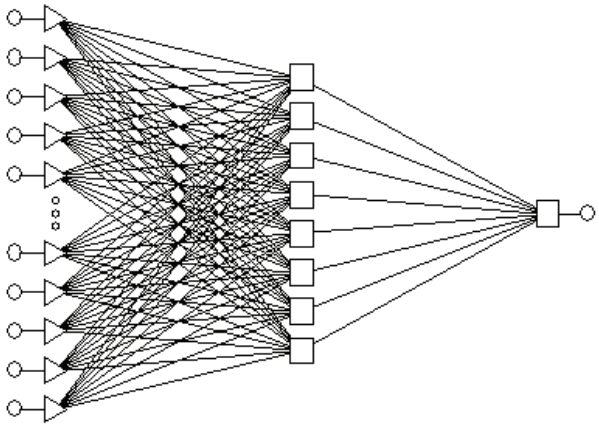
3.2. Sieć nieliniowa typu MLP

Przyjmując wyżej podane wartości jako poziom odniesienia przystąpiono do poszukiwania optymalnej **nieliniowej** sieci – na początku ograniczając się do sieci typu MLP (*Multi-Layer Perceptron*) (Liang 2005). W tym przypadku *Automatyczny projektant* sieci musiał przeanalizować działanie około dwóch tysięcy modeli, w których zmianom podlegały wybierane do zasilenia sieci sygnały wejściowe oraz liczba neuronów ukrytych. Przyjęto, że optymalizowana sieć będzie miała tylko jedną warstwę ukrytą. Badano wprawdzie także sieci z dwoma warstwami

ukrytymi, jednak w sytuacji posiadania bardzo ograniczonej liczby przypadków w zbiorze uczącym sieci takie (o dwóch warstwach ukrytych) nie dawały się skutecznie wytrenować, gdyż posiadany zasób danych nie wystarczał dla wystarczająco skutecznego zdeterminowania wartości wszystkich współczynników wagowych występujących w takiej dużej sieci.

Po przeprowadzeniu około dwóch tysięcy prób znaleziono sieć, której strukturę przedstawia rysunek 4. Sieć ta posiada ograniczoną liczbę elementów wejściowych (procedura optymalizacji sieci doprowadziła do wyboru 20 spośród 24 danych wejściowych) oraz stosunkowo niewielką warstwę ukrytą (8 neuronów). Sieć o podanej strukturze uczona była początkowo (przez 15 pierwszych epok procesu uczenia) z użyciem algorytmu szybkiej wstecznej propagacji błędów (quick propagation), potem w ciągu kolejnych 74 epok wykorzystywano algorytm Quasi-Newton z włączonym czynnikiem momentum o wartości 0,3. Podczas uczenia do sygnałów podawanych do sieci dodawano szum gaussowski o amplitudzie 0,5 celem uniknięcia zjawiska zatrzymywania procesu uczenia w lokalnych minimach.

Doprowadziło to do wyników, które można ocenić jako zdecydowanie zadowalające. Jakość działania sieci przedstawionej na rysunku 4 charakteryzuje talica 3.



Rysunek 4. *Najlepsza spośród znalezionych sieci klasy MLP o strukturze 20:20-8-1:1.*

Figure 4. *The architecture (20:20-8-1:1) of the best network from MLP class.*

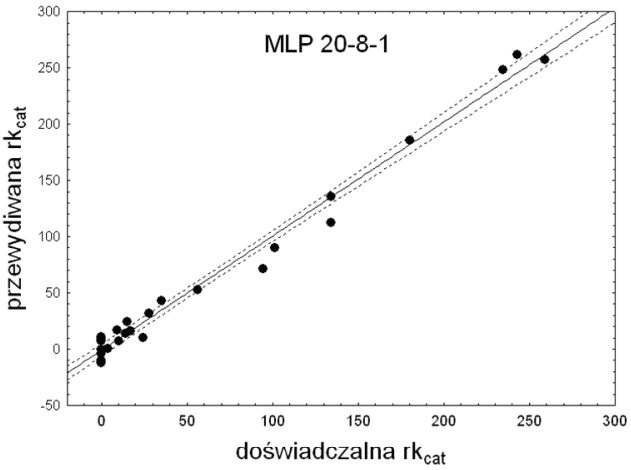
Jak widać błędy popełniane przez wytrenowaną sieć MLP są przynajmniej o rząd wielkości mniejsze, niż analogiczne błędy w przypadku sieci liniowej. Dotyczy to błędu ocenionego na zbiorze walidacyjnym i testowym, bo tylko one są naprawdę miarodajne (mogą stanowić dobre oszacowanie błędów popełnianych przez sieć w trakcie normalnej, roboczej eksploatacji, to znaczy przy ustalaniu właściwości nowych, całkowicie jeszcze nieznanych lub nie przebadanych związków chemicznych). Jak widać, w tym zakresie jakość najlepszej sieci MLP jest **znacząco** lepsza, niż jakość najlepszej sieci liniowej.

Tablica 3. *Średnie wartości błędów popełnianych przez sieć nieliniową klasy MLP.*

Table 3. *The average error values of non-linear MLP network.*

Błąd na zbiorze uczącym	Błąd na zbiorze walidacyjnym	Błąd na zbiorze testowym
0.028075	0.026944	0.049206

Warto może na chwilę skupić się na sprawie wartości błędu popełnianego przez sieć na zbiorze uczącym, który w sieci liniowej uzyskał wyjątkowo małą wartość (praktycznie był zerowy). Otóż jest to skutek użycia do uczenia sieci liniowej precyzyjnego algorytmu optymalizacyjnego zamiast iteracyjnej techniki adaptacyjnej. Nota bene gdyby nie przerywać procesu uczenia sieci MLP na podstawie informacji pochodzących z obserwacji zmienności zachowania sieci podczas przetwarzania danych pochodzących ze zbioru walidacyjnego – to także bez trudu można by było uzyskać dowolnie dobre dopasowanie zachowania tej sieci do zbioru danych uczących, przy równoczesnym pogorszeniu jej działania dla danych, na których sieć uczona nie była (w tym przypadku dałoby się to zaobserwować dla danych walidacyjnych oraz testowych). Jednak taka sytuacja jest uważana powszechnie za niepożądaną, jako że wiąże się z utratą (lub ograniczeniem) zdolności sieci do generalizacji wyników procesu uczenia.



Rysunek 5. *Wykres zgodności danych eksperymentalnych oraz danych przewidywanych przez wybraną najlepszą sieć MLP.*

Figure 5. *The correlation plot of experimental data with results of prediction of the best MLP network..*

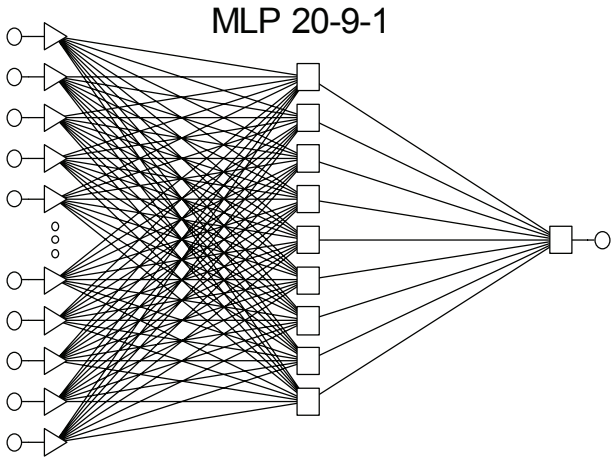
Dobrą jakość wyselekcjonowanej sieci MLP potwierdza także wykres (analogiczny do rysunku 3) ilustrujący związek pomiędzy przewidywaną a sprawdzoną empirycznie wartością aktywności rozważanej grupy wzorców, przedstawiony na rysunku 5. Warto zwrócić uwagę, jak wąski zakres ma na tym wykresie obszar rozrzutu (zaznaczony przerywanymi liniami), dopuszczalnego przy założeniu 95% przedziału ufności. Analogiczny obszar na rysunku 3 jest wielokrotnie większy. Oznacza to, że związek pomiędzy rzeczywistymi danymi a danymi uzyskiwanymi przy użyciu prognozowania aktywności chemicznej badanych związków za pomocą sieci neuronowej – jest bardzo silny i dobrze zdeterminowany. Na to samo wskazują uzyskane dla tej sieci wartości współczynnika korelacji oraz miary jego istotności: $r = 0.9925$; $p = 1.95*10^{-23}$. Wynik ten potwierdza dane przytaczane w licznych pracach (patrz na przykład Liang 2005, Raji et al. 2005) wskazujące, że w podobnych studiach porównawczych sieci typu MLP zwykle uzyskują bardzo dobre wyniki.

3.3. Wariant sieci typu MLP z liniowym neuronem wyjściowym

Sieci MLP rozpatrywane są w literaturze w dwóch wariantach (Raji et al. 2005). Pierwszy z nich, zastosowany w eksperymentach przedstawionych w podrozdziale 3.2, dotyczy sieci mającej wszystkie neurony nieliniowe (najczęściej o charakterystyce sigmoidalnej, chociaż dyskutowane są także sieci prezentujące struktury z neuronami mającymi inne charakterystyki nieliniowe, na przykład wykorzystujące funkcję tangens hiperboliczny). Natomiast w niektórych doniesieniach literaturowych przedstawiane są sugestie, że w zadaniach typu regresyjnego, w których na wyjściu sieci oczekiwana jest przewidywana przez model **wartość**, a nie **decyzja,** jako neurony **warstwy wyjściowej** mogą być używane neurony **liniowe**. Niekiedy spotyka się nawet pogląd, że neurony liniowe są wręcz **zalecane** w sieci tego typu, ponieważ to daje większą swobodę doboru zachowania neuronowego modelu (wyjście nie podlega restrykcjom wynikającym z nasycenia sigmoidy na poziomie 0 oraz 1).

Dysponując dobrze zebranymi danymi uczącymi na temat rozważanego problemu analizy aktywności związków chemicznych oraz mając sprawne narzędzie do optymalizacji struktury sieci (*Automatyczny projektant* pakietu *Statistica® Neural Networks*), postanowiono sprawdzić, jak się w tym zdaniu zachowa sieć typu MLP z liniowym neuronem wejściowym.

W tym celu przy użyciu *Automatycznego projektanta* przeprowadzono badania wstępne, w wyniku których uzyskana została sieć o strukturze pokazanej na rysunku 6.



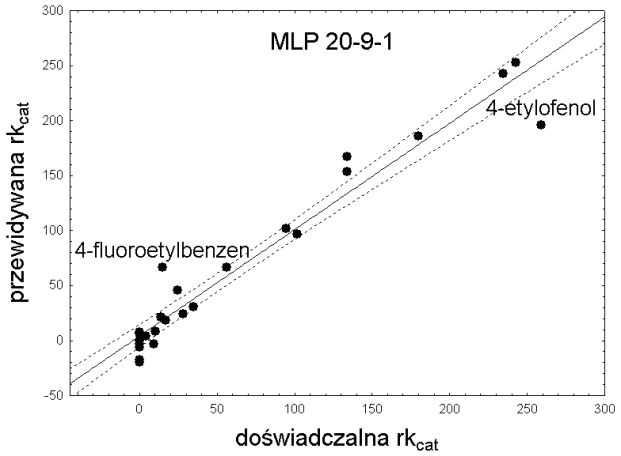
Rysunek 6. *Struktura znalezionej w wyniku badań optymalnej sieci typu MLP o strukturze 20:20-9-1:1 z liniowym neuronem w warstwie wyjściowej.*

Figure 6. *The optimal architecture of MLP network with linear output neuron: 20:20-9-1:1.*

Łatwo zauważyć, że struktura ta jest podobna do tej, jaką ustalono (w odrębnych i niezależnych badaniach) dla typowej, w pełni nieliniowej sieci MLP, opisanej w poprzednim podrozdziale. Należy jednak podkreślić, że wybrany na drodze optymalizacji zbiór danych wejściowych częściowo różnił się od tego użytego w sieci w pełni nieliniowej. Jedyna zauważalna różnica strukturalna polega na nieco większej warstwie ukrytej, jaką wybrał w wyniku wstępnych prób *Automatyczny projektant* dla sieci z elementem liniowym na wyjściu, co dowodzi, że poprawne wyuczenie sieci z elementem liniowym okazuje się zadaniem **trudniejszym** niż zadanie uczenia sieci w pełni nieliniowej, więc koniecznych jest więcej danych pośrednich (sygnałów produkowanych przez warstwę ukrytą), żeby uzyskać dobre funkcjonowanie sieci.

Całkowicie rozczarowujące okazały się jednak wyniki, jakie udało się uzyskać przy użyciu tego typu sieci. Pierwszy sygnał ostrzegawczy w tym zakresie uzyskano analizując przytoczony na rysunku 7 wykres zgodności danych eksperymentalnych oraz danych przewidywanych przez wybraną sieć. Okazało się, że korelacja była w tym przypadku znacznie gorsza, niż dla całkowicie nieliniowej sieci

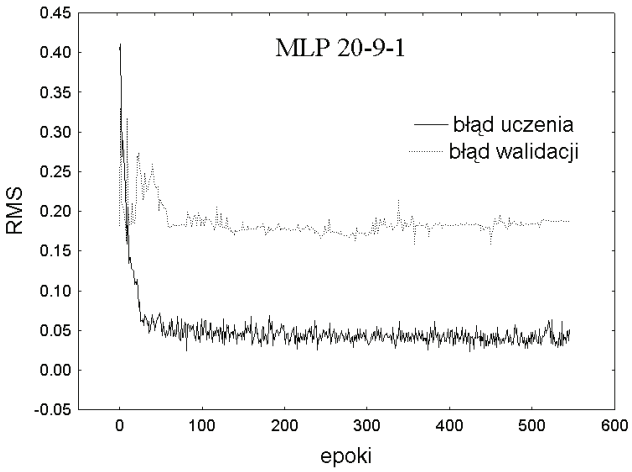
MLP (r = 0.9717; p = $1.5 \cdot 10^{-16}$), co dowodzi, że jakość modelu pozostawia w tym przypadku wiele do życzenia.



Rysunek 7. *Wykres zgodności danych eksperymentalnych oraz danych przewidywanych przez wybraną najlepszą sieć MLP z elementem liniowym na wyjściu.*

Figure 7. *The correlation plot of experimental data with results of prediction of the MLP with linear output neuron.*

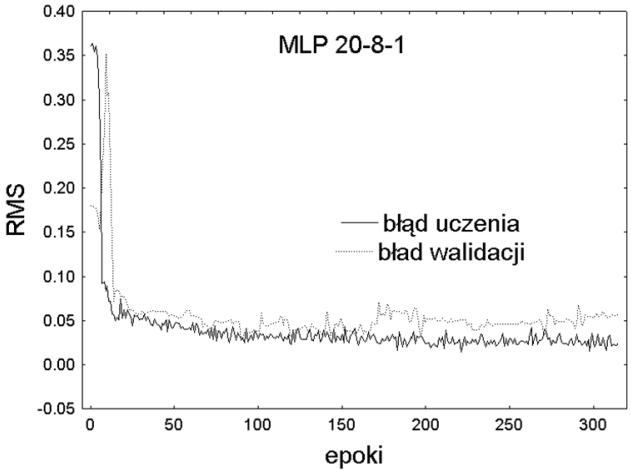
Jeszcze bardziej niepokojące objawy stwierdzono jednak obserwując (patrz rysunek 8) zmiany wartości błędu popełnianego przez sieć dla zbioru uczącego (linia ciągła) oraz dla zbioru walidacyjnego (linia przerywana).



Rysunek 8. *Przebieg uczenia optymalnej sieci MLP z elementem liniowym na wyjściu.*

Figure 8. *Learning curve of optimal MLP network with linear output neuron.*

Na rysunku tym widać wyraźnie, że badana sieć uczy się dosyć sprawnie, ma jednak spore trudności z uogólnieniem wyników procesu uczenia, gdyż linia błędów uzyskanych dla zbioru walidacyjnego przebiega znacząco powyżej linii błędów uzyskanych dla zbioru uczącego. Dla kontrastu warto obejrzeć na rysunku 9 analogiczne wykresy, które rejestrują zjawiska, jakie miały miejsce przy uczeniu sieci MLP całkowicie nieliniowej (opisanej w podrozdziale 3.2).



Rysunek 9. *Przebieg uczenia optymalnej sieci MLP.*

Figure 9. *Learning curve of optimal MLP network with non-linear output neuron.*

Jak widać w tym poprzednim przypadku poprawianie jakości działania sieci dla zbioru uczącego było ściśle skorelowane z poprawianiem jakości jej działania dla zbioru walidacyjnego. Co więcej, dokładna analiza wykresy podanego na rysunku 9 może wskazywać na to, że mniej więcej w okolicach 170 epoki proces uczenia wszedł w fazę, w której zaznaczył się (bardzo delikatnie) efekt „przeuczenia" – dalszemu opadaniu średniego trendu linii błędu notowanej dla zbioru uczącego towarzyszyło zauważalne dla specjalisty, pogorszenie działania sieci ocenianego na zbiorze walidacyjnym. Efekt przeuczenia, mimo że w tym przypadku subtelny, został wykryty przez automat kontrolujący jakość sieci. W wyniku tego wstecznie, po arbitralnym przerwaniu uczenia, została odzyskana sieć z wagami z 84 epoki (15 epok szybkiej propagacji i 74 epoki algorytmu Quasi-Newtona).

Wracając do sprawy sieci MLP z elementem liniowym na wyjściu należy stwierdzić, że przebieg uczenia zasygnalizowany na rysunku 8 wyraźnie wskazuje, że sieć przy przewidywaniu aktywności chemicznej poszczególnych związków chemicznych uzyskuje znacząco lepsze wyniki dla tych cząsteczek, których opis był jej podany w trakcie procesu uczenia, niż dla tych, na których usiłujemy sprawdzić jakość jej działania (w zbiorze walidacyjnym oraz w zbiorze testowym). Potwierdzają to wartości odpowiednich błędów zestawione w tablicy 4 (którą warto porównać z tablicą 3), a także dokładniejsze porównanie wartości dostarczanych przez modelowaną sieć oraz wartości rzeczywistych dla poszcze-

gólnych konkretnych związków, przedstawione w tablicy 5.

Tablica 4. *Średnie wartości błędów popełnianych przez sieć nieliniową klasy MLP z elementem liniowym na wyjściu.*

Table 4. *The average error values of non-linear MLP network.*

Błąd na zbiorze uczącym	Błąd na zbiorze walidacyjnym	Błąd na zbiorze testowym
0.032292	0.158581	0.114655

Tablica 5. *Porównanie wartości aktywności rzeczywistych oraz obliczonych za pomocą modelu z liniową warstwą wyjściową dla poszczególnych związków chemicznych z uwidocznieniem faktu, do którego zbioru (uczącego, walidacyjnego lub testowego) należał dany związek.*

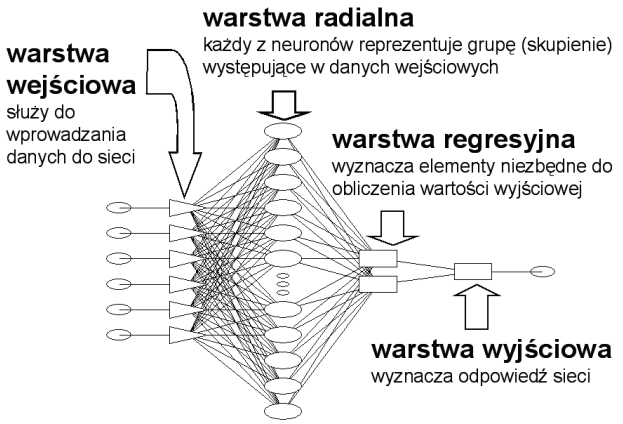
Table 5. *The comparison of experimental activities for each chemical compound with values predicted by neural model with linear output layer along with cases assignment to learning, validation or testing subsets.*

Nazwa związku	Zbiór	Wartość rzeczywista	Wartość przewidywana
1,2-dietylobenzen	Uczący	0.0	7.1
1,4-dietylobenzen	Uczący	35.0	30.5
1-etylonapftalen	Uczący	0.0	-5.9
2-etyloanilina	Uczący	94.5	101.9
2-etylofuran	Walidacyjny	133.9	166.7
2-etylonaftalen	Uczący	9.3	-3.5
2-etylofenol	Uczący	56.1	66.5
2-etylopirydyna	Testowy	0.0	-19.5
2-etylopirol	Uczący	234.6	242.5
2-etylotiofen	Uczący	242.9	252.8
2-etylotoluen	Testowy	3.8	4.0
2-metylofuran	Uczący	0.0	-17.8
2-metylopirol	Uczący	0.0	6.4
2-metylotiofen	Uczący	0.0	-0.2
3-etylofenol	Walidacyjny	24.3	45.5
3-etylopirydyna	Uczący	16.9	18.4
3-etylotoluen	Uczący	10.0	8.4
4-etylofenol	Walidacyjny	259.0	195.6
4-etylopirydyna	Uczący	0.0	7.6
4-etylotoluen	Uczący	28.0	24.1
4-fluoroetylobenzen	Testowy	15.0	66.9
etylobenzen	Uczący	100	96.7
n-propylobenzen	Uczący	14.0	21.2
toluen	Uczący	0.0	-3.0
4-etyloanilina	Walidacyjny	134.0	153.6
4-propylofenol	Testowy	180.0	185.9

Podsumowując tę część eksperymentów można stwierdzić, że hipoteza mówiąca o korzystnym wpływie liniowej charakterystyki wyjściowego neuronu w sieci MLP **nie** została potwierdzona. Być może przyczyna tego zjawiska leżała po stronie *Automatycznego projektanta,* który dla sieci MLP z liniowym wyjściem wybrał (jako optymalną) sieć mającą 9 neuronów ukrytych, podczas gdy optymalizacja dla sieci w pełni nieliniowej zakończyła się wskazaniem struktury korzystającej z 8 neuronów ukrytych. Czasem takie dodanie **jednego** neuronu do warstwy ukrytej sieci może „przerzucić" sieć z trybu uczenia z uogólnianiem do trybu uczenia na pamięć – szczególnie w przypadku sieci mającej stosunkowo dużą liczbę wejść (w rozważanym przypadku – 20) i uczonej na niezbyt licznym zbiorze uczącym (w rozważanym przypadku do dyspozycji było zaledwie 18 przypadków uczących, 4 przypadki walidacyjne oraz cztery testowe). Być może przyczyny należy szukać w innym wektorze wyjściowym, (choć został on wybrany do tego typu sieci na drodze eksperymentalnej spośród tysięcy innych sieci). Niemniej wynik, jaki zaobserwowano w tych badaniach, wydaje się wart tego, żeby go w tej pracy zaprezentować i omówić.

3.4. Sieci nieliniowe typu GRNN

Model pozwalający przewidywać właściwości (aktywność chemiczną) jeszcze nieprzebadanych związków chemicznych na podstawie topologicznych i kwantowo-chemicznych właściwości ich cząsteczki jest niewątpliwie jednym z trudniejszych i bardziej skomplikowanych modeli, jakie budowano przy wykorzystaniu sieci neuronowych w Laboratorium Biocybernetyki AGH. Dlatego budując te modele wykorzystano (eksperymentalnie) różne typy sieci neuronowych, w tym między innymi sieci GRNN (uogólnionej regresji), które uchodzą w literaturze za szczególnie predestynowane do budowy takich właśnie złożonych modeli (Cigizoglu & Alp 2006) – co prawda stosowanych do innego typu systemów (Zhang & Sun 2004). Przykładowa architektura sieci GRNN, przedstawiona na rysunku 10, uwidacznia większą złożoność takiej sieci w stosunku do sieci MLP (nie wspominając już o liniowych).



Rysunek 10. *Ogólna struktura sieci GRNN i rola poszczególnych jej elementów.*

Figure 10. *General structure of GRNN and role of individual components.*

Proces budowy modelu w sieci GRNN jest podzielony na dwa etapy. W pierwszym z nich w przestrzeni sygnałów wejściowych wydzielane są grupy sygnałów tworzące (na skutek wzajemnego podobieństwa) dające się wyodrębnić skupiska. Etap ten realizowany jest przy pomocy warstwy radialnej sieci GRNN. W etapie drugim tworzy się regresyjną aproksymację poszukiwanej zależności opierając się na wcześniej wyznaczonych decyzjach warstwy radialnej, określających stopień podobieństwa aktualnie rozważanego sygnału wejściowego do poszczególnych wcześniej wyodrębnionych klas (albo grup sygnałów).

Nawet ten szkicowy i skrótowy opis działania sieci GRNN wskazuje na to, że sieć tego rodzaju powinna w rozważanym tu zadaniu wykazywać szczególnie duży stopień skuteczności, bowiem droga rozumowania eksperta (człowieka) przy próbie przewidzenia właściwości nieznanego związku chemicznego biegnie zwykle poprzez próbę ustalenia, do jakich znanych związków podobna jest cząsteczka nieznanej substancji, a następnie odwołuje się do wnioskowania przez analogię – czego swoisty model oferuje właśnie sieć GRNN. Z tego powodu w opisywanych tu pracach przywiązywano początkowo spore nadzieje właśnie do możliwości zastosowania sieci GRNN jako neuronowego modelu przeznaczonego do przewidywania aktywności chemicznej nieznanych związków.

Dodatkową okolicznością skłaniającą do skupienia uwagi na sieciach GRNN był fakt, że we wszystkich rozważanych w tym artykule sieciach mieliśmy do czynienia z jednym (tylko) wyjściem, co jest warunkiem koniecznym przy próbach stosowania sieci GRNN.

Z tego powodu przebadano wstępnie (z pomocą *Automatycznego projektanta*) około dwa tysiące wariantów sieci GRNN i do dalszych szczegółowych badań wybrano nie jedną, ale trzy struktury sieci tego typu. Wytypowane do badań sieci miały struktury przedstawione w tablicy 6.

Tablica 6. *Struktury sieci GRNN wytypowane do dokładnych badań.*

Table 6. *The architecture of GRNN networks that were considered in the study.*

Numer porządkowy badanej sieci	Struktura
GRNN 40	24:24-14-2-1:1
GRNN 46	23:23-14-2-1:1
GRNN 58	22:22-14-2-1:1

Jak widać *Automatyczny projektant* konsekwentnie wskazywał jako najkorzystniejszą strukturę sieci zawierającą 14 neuronów w warstwie radialnej i 2 neurony w warstwie regresyjnej, natomiast zróżnicowanie sieci wytypowanych jako najkorzystniejsze polegało głównie na doborze różnych podzbiorów zbioru danych wejściowych, branych pod uwagę przy tworzeniu neuronowego modelu.

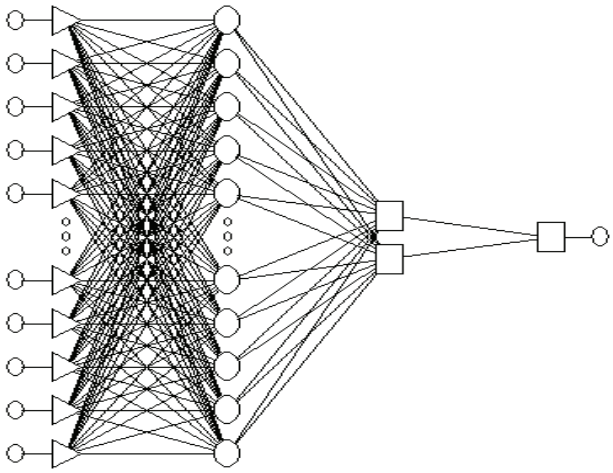
Struktura przykładowej badanej sieci GRNN przedstawiona jest na rysunku 11. Sieci wymienione w tablicy 6 poddano uczeniu, w wyniku którego uzyskano wyniki, przedstawione w tablicy 7.

Tablica 7. *Zestawienie typów najlepszych sieci GRNN wraz z błędami dla zbioru uczącego, walidacyjnego i testowego.*

Table 7. *Summary of the best GRNN networks along with errors for learning, validation and testing subsets.*

Typ sieci	Błąd na zbiorze uczącym	Błąd na zbiorze walidacyjnym	Błąd na zbiorze testowym
GRNN 24:24-14-2-1:1	0.857	0.739	1.189
GRNN 23:23-14-2-1:1	0.964	0.589	1.213
GRNN 22:22-14-2-1:1	1.037	0.481	1.137

Analiza tablicy 7 i jej porównanie z danymi podawanymi wcześniej w podobnych tablicach dla sieci MLP, a nawet dla sieci liniowej nie pozostawia żadnych złudzeń: sieci GRNN w rozważanym przykładzie nie okazały się „konkurencyjne" dla żadnej w z wcześniej rozważanych sieci.



Rysunek 11. *Struktura 24:24-14-2-1:1 jednej z badanych sieci typu GRNN.*

Figure 11. *The architecture 24:24-14-2-1:1 – one of studied network of GRNN type.*

Wniosek ten potwierdza także analiza wartości współczynników korelacji oraz ich istotności. Odpowiednie wartości były następujące: dla GRNN 24:24-14-2-1:1 r = 0.9108, p = 0.0000002, dla sieci GRNN 23:23-14-2-1:1 r = 0.8893, p = 0.0000008 oraz dla sieci GRNN 22:22-14-2-1:1 r = 0.8524, p = 0.000007.

3.5. Sieć hybrydowa wykorzystująca strukturę GRNN i wejście MLP

Znaczna rozbieżność pomiędzy kiepskimi wynikami uzyskanymi przez sieci GRNN oraz zdecydowanie dobrym wynikiem uzyskanym w sieci MLP, wykazana powyżej, skłoniła do poszukiwania przyczyny tego stanu rzeczy. Analiza porównawcza właściwości i struktur badanych sieci (znajdowanych każdorazowo w wyniku procesu optymalizacji prowadzonego z wykorzystaniem narzędzia, jakim jest *Automatyczny projektant*) była utrudniona przez fakt, że wyłonione w następstwie tej optymalizacji sieci GRNN korzystały z innych podzbiorów danych wejściowych, niż optymalna sieć MLP. Korzystając z sugestii jednego z recenzentów tej pracy postanowiono spróbować stworzyć sieć hybrydową, o strukturze sieci GRNN, ale korzystającą ze zdeterminowanego podzbioru danych wejściowych (o liczebności 20 elementów) zdeterminowanego uprzednio podczas optymalizacji sieci MLP.

Wyniki okazały się bardzo ciekawe. Po optymalizacji struktury sieci, w której zaangażowany był oczywiście *Automatyczny projektant*, otrzymaliśmy sieć o strukturze GRNN 20:20-18-2-1:1. Sieć ta po treningu dała **uśrednione** wyniki jeszcze lepsze, niż najlepsza z wcześniej omawianych (w pełni nieliniowa sieć MLP). Ilustruje to tablica 8.

Tablica 8. *Wyniki dla sieci hybrydowej.*

Table 8. *Results for hybryd network.*

Błąd na zbiorze uczącym	Błąd na zbiorze walidacyjnym	Błąd na zbiorze testowym
0.000203	0.019030	0.016527

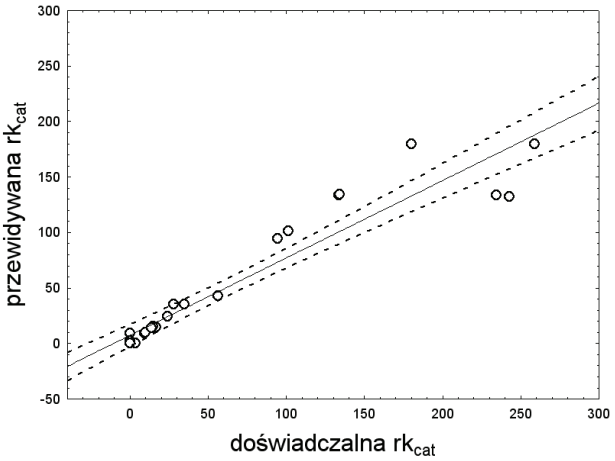
Pozornie więc w pełni uzasadniony byłby wniosek, że oto odkryliśmy najlepszą możliwą sieć i że taka mieszana metodologia odrębnego optymalizowania zbioru danych wejściowych oraz oddzielnego trenowania wybranego typu sieci (w tym przypadku – sieci GRNN) jest godnym polecenia sposobem tworzenia modeli neuronowych złożonych zjawisk lub procesów. Jednak bliższa analiza uzyskanych wyników ujawniła, że sprawa nie jest taka prosta ani oczywista.

Uzyskana w opisany wyżej sposób sieć bardzo dobrze dopasowuje się do większości posiadanych danych (zarówno uczących, jak i walidacyjnych oraz testowych). Z drugiej jednak strony sieć ta doprowadziła do powstania **bardzo dużych** błędów w przypadku niektórych danych (wyłącznie w zbiorze walidacyjnym i testowym). Efekt ten ze względu na relatywnie niewielką ilość błędnie ocenionych przypadków (3 znaczące błędy na 26 przypadków) może zostać niedostrzeżony gdyby ograniczyć się tylno do analizie miary dopasowania opartej na współczynnikach korelacji (uzyskano r = 0,9494; $p = 1{,}45 \cdot 10^{-13}$), które wydają się wskazywać na doskonały wynik! Wystarczy jednakowoż przedstawić tę zależność graficznie (rysunek 12) by jasno zobaczyć, jak od zbiorowości punktów dobrze dopasowanych odskakują na dużą odległość niektóre punkty ze zbioru danych testowych i walidacyjnych.

Opisane zjawisko można interpretować na dwa sposoby.

Z jednej strony można przypuszczać, że duża liczba neuronów radialnych w warstwie ukrytej sieci GRNN radialnych doprowadziła do „wysycenia" przestrzeni sygnałów wejściowych. Istotnie, automatyczna analiza wyłoniła jako najlepszą sieć GRNN strukturę o 18 neuronach radialnych przy zbiorze uczącym liczącym 18 przykładów. Oznacza to, że sieć w warstwie radialnej nie skupiała grup sygnałów wejściowych w pewne kategorie tylko po prostu mogła „sfotografować" cały zbiór uczący. Tego rodzaju uproszczone „uczenie" warstwy radialnej przy stosunkowo ubogiej warstwie regresyjnej doprowadziło do tego, że rozważana sieć dobrze wy-

pełniała zadania bardzo podobne do tych, z jakimi miała do czynienia w zbiorze uczącym, ale totalnie zawodziła w przypadku problemów walidacyjnych lub testowych, których sposób rozwiązania odbiega znacząco od tych właśnie „sfotografowanych" danych uczących.



Rysunek 12. *Wykres zgodności danych eksperymentalnych oraz danych przewidywanych przez wybraną sieć hybrydową.*

Figure 12. *The correlation plot of experimental data with results of prediction of the chosen hybrid network (GRNN with optimal MLP input vector).*

Z drugiej strony można sformułować przypuszczenie, że badane zjawisko nie podlega jednemu ogólnemu modelowi, tylko dla części danych powinno się rozważać inny model. Jeśli bowiem model teoretyczny dopasowuje się (za pomocą przyjętej neuronowej metodyki) niemal idealnie do znaczącej większości obserwowanych danych – i to zarówno uczących, jak i testowych oraz walidacyjnych – ale bardzo istotnie nie zgadza się z pewnymi danymi (dla pewnych wybranych związków chemicznych) – to być może dla tych innych związków konieczne byłoby poszukanie innego modelu, a cały problem poszukiwania relacji pomiędzy elementami strukturalnymi cząsteczki chemicznej a właściwościami (aktywnością) rozważanego związku należałoby przedefiniować z odwołaniem do dwóch (lub większej liczby) modeli, dobieranych do tego zadania zależnie od jakiegoś dodatkowego kryterium. Hipotezy tej jednak nie da się zweryfikować na tym etapie opisywanych tu badań, bo zestawy danych nie pasujące do opisanego wyżej hybrydowego modelu są zbyt mało liczne, żeby w oparciu o nie można było próbować stworzyć jakikolwiek model.

Niezależnie od tego, jakie są przyczyny opisanego wyżej zachowania modelu hybrydowego jego działanie trzeba uznać za niezadowalające. Model, który dla niektórych danych testowych wykazuje bardzo duże błędy nie jest narzędziem godnym zaufania, jako że w przypadku wykorzystywania go do predykcji właściwości jakiegoś związku chemicznego przed eksperymentalną weryfikacją, nigdy nie będzie wiadomo, czy odpowiedź uzyskana na podstawie neuronowych obliczeń należy do tej **większości** wyników, dla których sieć uzyskuje bardzo dobrą zgodność przewidywań teoretycznych z obserwacją empiryczną, czy też trafiliśmy na jeden z tych **nielicznych** przypadków, w których rozbieżność przewidywań neuronowych i rzeczywistości jest bardzo duża. Dlatego pozostawiając kwestię dalszego doskonalenia modelu hybrydowego do czasu, aż będzie można go przebadać na bazie większej liczby danych doświadczalnych – na obecnym etapie nie rekomendujemy jego użycia.

4. WYBÓR SPOSOBU REPREZENTACJI WYNIKU NA WYJŚCIU SIECI

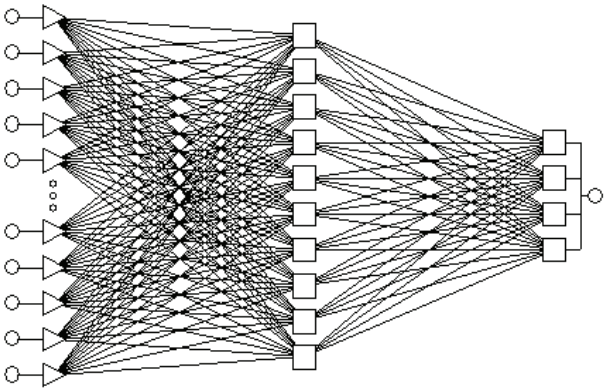
Po dokonaniu wyżej opisanych badań, których przedmiotem była struktura i rozmiar sieci rozwiązującej postawione zadanie, podjęto kolejną próbę uzyskania jeszcze lepszych rezultatów, tym razem sięgając do zmiany sposobu reprezentacji wyników pracy sieci na jej wyjściu. Jak wiadomo z wcześniej przedstawionej dyskusji, przedmiotem badań jest ocena aktywności chemicznej badanych związków, przy czym aktywność ta daje się wyrażać ilościowo i konfrontować z danymi doświadczalnymi. Jednak dla potrzeb praktyki może być ona również wyrażona w postaci ustalenia przynależności związku do określonej klasy aktywności, o czym wspomniano już we wstępie do tego artykułu. W przypadku przyjęcia **klas** aktywności związków (w miejsce **wartości** ich aktywności) – zadanie, które ma rozwiązać sieć zostaje zmienione z zadania regresyjnego do postaci zadania klasyfikacyjnego. Otóż podjęto próbę sformułowania i rozwiązania takiego zadania klasyfikacyjnego w kontekście rozważanych problemów, wprowadzając na wyjściu sieci jedną zmienną **jakościową** (zamiast ilościowej) o czterech możliwych kategoriach podanych w tablicy 9.

Tablica 9. *Sposób kodowania aktywności różnych związków w zadaniu klasyfikacji.*

Table 9. *Encoding of activity in classification task.*

Kod	Poziom aktywności
-1	inhibitory
0	aktywność do 50%,
1	51-150 %,
2	powyżej 150 %

Przy takim założeniu oraz przy przyjęciu typowej w takich przypadkach formy reprezentacji danych w postaci odwzorowania 1-z-N uruchomiono ponownie *Automatycznego projektanta* w celu uzyskania struktury optymalnej sieci. 1000 przebadanych przez *Automatycznego projektanta* sieci były szkolone takim samym algorytmem, tzn. przez 100 pierwszych epok procesu uczenia z użyciem algorytmu wstecznej propagacji błędów, potem w ciągu kolejnych 25 epok wykorzystywano algorytm gradientów sprzężonych, a na koniec w ciągu 25 epok korzystano z techniki gradientów sprzężonych z włączonym czynnikiem momentum. Najlepsza sieć, która została wybrana przedstawiona jest na rysunku 13, poddana została następnie ręcznemu uczeniu. Warto zauważyć, że praca *Automatycznego projektanta* tym razem doprowadziła do znaczącej redukcji wejściowego zbioru danych (do używania w sieci zakwalifikowano jedynie 15 zmiennych spośród 24 możliwych).



Rysunek 13. *Optymalna sieć klasyfikacyjna, 15:15–10–4:1 przyjęta w opisywanych badaniach.*

Figure 13. *The optimal classification network used in the study.*

Sieć o podanej strukturze 15:15–10–4:1 uczona była początkowo przez 100 epok z użyciem algorytmu wstecznej propagacji błędów a następnie przez 19 epok wykorzystywano algorytm Quasi-Newton z włączonym czynnikiem momentum o wartości 0,3. Podobnie jak w przypadku uczenia regresyjnych MLP do sygnałów podawanych do sieci dodawano szum gaussowski o amplitudzie 0,5 celem uniknięcia zjawiska zatrzymywania procesu uczenia w lokalnych minimach.

Wyniki przekroczyły najśmielsze oczekiwania. W tablicy 10 pokazano formalnie obliczone wartości błędów, które były zdecydowanie najlepsze ze wszystkich, uzyskiwanych w opisywanych tu (a także pominiętych z braku miejsca) doświadczeniach. Co więcej, wyniki te, zinterpretowane w kategoriach zaliczenia poszczególnych związków do odpowiednich klas, pokryły się całkowicie z klasyfikacją tych samych danych przeprowadzoną przez eksperta. Dotyczyło to zarówno danych ze zbioru uczącego, jak i danych walidacyjnych i testowych. Zestawienie odpowiednich klasyfikacji przedstawia tablica 11. Wynik ten, jak się wydaje, nie wymaga komentarza.

Tablica 10. *Średnie wartości błędów popełnianych przez sieć nieliniową klasy MLP z elementem decyzyjnym (klasyfikującym) na wyjściu*

Table 10. *The average error values in non-linear classification MPL.*

Błąd na zbiorze uczącym	Błąd na zbiorze walidacyjnym	Błąd na zbiorze testowym
0.107042	0.009889	0.000605

Tablica 11. *Porównanie klasyfikacji dokonanej przez eksperta i podanej przez sieć.*

Table 11. *The comparison of the classification performed by expert and classification neural network.*

Związek chemiczny	Zbiór	Klasyfikacja prawidłowa	Klasyfikacja podana przez sieć
1,2-dietylobenzen	Walidacyjny	-1	-1
1,4-dietylobenzen	Uczący	0	0
1-etylonaftalen	Walidacyjny	-1	-1
2-etyloanilina	Walidacyjny	1	1
2-etylofuran	Uczący	1	1
2-etylonaftalen	Uczący	0	0
2-etylofenol	Uczący	1	1
2-etylopirydyna	Uczący	-1	-1
2-etylopirol	Uczący	2	2
2-etylotiofen	Testowy	2	2
2-etylotoluen	Uczący	0	0
2-metylofuran	Uczący	-1	-1
2-metylopirol	Uczący	-1	-1
2-metylotiofen	Uczący	-1	-1
3-etylofenol	Testowy	0	0
3-etylopirydyna	Uczący	0	0
3-etylotoluen	Uczący	0	0
4-etylofenol	Uczący	2	2
4-etylopirydyna	Uczący	-1	-1
4-etylotoluen	Uczący	0	0
4-fluoroetylobenzen	Uczący	0	0
etylobenzen	Uczący	1	1
n-propylobenzen	Testowy	0	0
toluen	Uczący	-1	-1
4-etyloanilina	Walidacyjny	1	1
4-propylofenol	Testowy	2	2

5. PODSUMOWANIE I WNIOSKI

W pracy podjęto próbę przedstawienia wybranych problemów doboru i optymalizacji modelu w postaci sieci neuronowej do posiadanego zestawu danych empirycznych. Przykładowy problem dotyczył próby zbudowania sieci przewidującej na podstawie danych topologicznych i kwantowo-chemicznych aktywność pewnej klasy związków chemicznych, jednak szczegóły chemiczne rozwiązywanego problemu nie miały tu zasadniczego znaczenia. Z punktu widzenia tej pracy ważne było, że rozwiązywany problem był trudny, a ponadto charakteryzował się szeregiem cech, które wyjątkowo często pojawiają się przy praktycznym stosowaniu sieci neuronowych:

- zadanie wymagało brania pod uwagę (w charakterze potencjalnych sygnałów wejściowych) **dużej** liczby danych o zróżnicowanym charakterze (były tam dane ilościowe i jakościowe);
- istniało uzasadnione podejrzenie, że nie wszystkie dane wejściowe są równie przydatne przy rozwiązywaniu postawionego zadania, ale brak było dokładnych przesłanek, żeby dokonać ich wstępnej selekcji przed zbudowaniem neuronowego modelu;
- liczba przykładów, na bazie których można było sieć uczyć, a także dokonywać walidacji i testowania jej działania, była bardzo ograniczona.

Przy takich założeniach przebadano przydatność szeregu różnych struktur i zasad działania sieci, uzyskując następujące ważniejsze wyniki:

- stwierdzono, że jakość modelu uzyskiwanego przy zastosowaniu liniowej sieci neuronowej jest bardzo istotnie gorsza, niż jakość najlepszego uzyskanego modelu nieliniowego w postaci sieci MLP;
- wykazano, że wprowadzenie do sieci MLP liniowego neuronu w jej wyjściowej warstwie nie polepsza jakości uzyskiwanego rozwiązania;
- zaobserwowano, że niewielkie zwiększenie liczby neuronów warstwy ukrytej (o jedne neuron!) może prowadzić do tego, że sieć zamiast budować model problemu, nadający się do rozwiązywania całej klasy zadań podobnego typu, zaczyna uczyć się „na pamięć" zbioru uczącego i traci zdolność do generalizacji;
- stwierdzono, że użycie sieci GRNN nie przynosi dobrych rezultatów – uzyskiwane wyniki były nie tylko gorsze od tych, jakie wykazywała najlepsze sieć MLP, ale co gorsza – plasowały się poniżej wyników, jakie dawała sieć liniowa;
- próba stworzenia sieci hybrydowej, wykorzystującej doświadczenia zdobyte podczas optymalizacji struktury MLP oraz zalety struktury GRNN doprowadziła do powstania modelu dobrze dopasowanego do większości posiadanych danych, ale w sposób zasadniczy odbiegającego od rzeczywistości w przypadku pewnej liczby danych doświadczalnych, co uznano za przesłankę do nie używania tego modelu w dalszych pracach;
- wykazano, że transformacja zadania regresyjnego (wymagającego, by sieć obliczyła i podała określoną **wartość** sygnału wyjściowego) do postaci zadania klasyfikacyjnego, w którym odpowiedź sieci może być interpretowana jako **decyzja**, pozwala uzyskać najlepsze rezultaty, które w rozważanym problemie osiągnęły poziom 100% zgodności zachowania sieci z wymaganiami wynikającymi z natury rozwiązywanego zadania.

Przytoczone wyżej spostrzeżenia i sformułowane na ich podstawie wnioski z całą pewnością są silnie uwarunkowane właściwościami konkretnego rozwiązywanego tu zdania. Jednak można sądzić, że z dużym prawdopodobieństwem podobne prawidłowości będą wiązały się z innymi zastosowaniami techniki sieci neuronowych – dlatego zebrano te spostrzeżenia i przedstawiono w tej publikacji w celu ułatwienia pracy innym badaczom, którzy zdecydują się w swojej pracy na użycie sieci neuronowych jako narzędzia modelowania zjawisk rzeczywistego świata.

6. PODZIĘKOWANIE

Badania zostały sfinansowane przez Ministerstwo Nauki i Szkolnictwa Wyższego w ramach sieci naukowej EKO-KAT i grantu obliczeniowego KBN/SGI2800/PAN/037/2003. Maciej Szaleniec składa podziękowania za przyznanie stypendium doktoranckiego Polskiej Akademii Nauk.

BIBLIOGRAFIA

Cigizoglu, H. K. Alp, M., 2006, Generalized regression neural network in modelling river sediment yield, *Adv. Eng. Soft.,* 37, 63 – 68.

Fogelman, S., Blumenstein, M., Zhao, H. 2006, Estimation of chemical oxygen demand by ultraviolet spectroscopic profiling and artificial neural networks, *Neural Comput. Appl*., 15, 197 – 203.

Hartono, P.; Hashimoto, S., 2005, Learning with ensemble of linear perceptrons. Artificial Neural Networks: Formal Models and their Applications ICANN 2005 15th International Conference, Proceedings, *Part II Lec. Notes*

Comp. Sci., eds, Duch, W., Kacprzyk, J., Oja, E., Zadrozny, S., Springer-Verlag, Berlin, 3697, 115 - 120.

Hong, L., Chunsheng, Y., 2006, Specifying distributed multi agent systems in chemical reaction metaphor, *Appl. Intell*., 24, 155 – 168.

Lang, B., 2005, Monotonic multi layer perceptron networks as universal approximators. Artificial Neural Networks: Formal Models and their Applications ICANN 2005 15th International Conference, Proceedings, *Part II Lec. Notes Comp. Sci*., eds, Duch, W., Kacprzyk, J., Oja, E., Zadrozny, S., Springer-Verlag, Berlin, 3697, 31 – 37.

Mei, H., Zhou, Y., Liang, G., Zhiliang, L. 2005, Support vector machine applied in QSAR modeling, *Chinese Sci. Bull*., 50, 2291 –2296.

Pianese, C., Arsie, I., Sorrentino, M., 2006, A procedure to enhance identification of recurrent neural networks for simulating air fuel ratio dynamics in SI engines, *Eng. Appl. Artif*. Intel., 19, 65 – 77.

Plewczynski, D., Spieser, S.A.H., Koch, U., 2006, Assessing different classification methods for virtual screening, *J. Chem. Inf. Mod*., 46, 1098 – 1106.

Raji, U. , Mashor, M. Y., Ali, A. N., Adom, A. H., Sadullah, A. F., 2005, HMLP, MLP and recurrent networks for carbon monoxide concentrations forecasting: a comparison studies, *WSEAS Transactions on Systems*, 4, 812 – 820.

Szaleniec M., Goclon J., Witko M., Tadeusiewicz R., 2006a, Application of artificial neural networks and DFT-based parameters for prediction of reaction kinetics of ethylbenzene dehydrogenase, J. Comput., *Aid. Mol. Des*., 20, 1573-4951.

Szaleniec M., Witko M., Tadeusiewicz R., 2006b, Prediction of the ethylbenzene dehydrogenase reaction kinetics by comparative molecular field analysis (COMFA) and artificial neural network's systems basing on DFT parameters. *Materiały XXXVIII Ogólnopolskiego Kolokwium Katalitycznego*, ed, Mokrzycki L., Instytut Katalizy i Fizykochemii Powierzchni PAN, Kraków, 215–217.

Weiss, Y., Schölkopf B.**,** Platt J., eds, 2006, *Advances in Neural Information Processing Systems 18*, The MIT Press, Cambridge.

Yixing, L., Yuzhang, W., Shilie, W., Yonghong, W., 2006, Application of artificial neural network in countercurrent spray saturator, Advances in Neural Networks ISNN 2006. Proc. 3rd Int. Symp. on Neural Networks, *Lec. Notes Comp. Sci*., eds, Wang, J., Yi, Z., Zurada, J.M., Lu, B. L., Yin, H., Springer-Verlag, Berlin, 3973, 1277 – 1282.

Zhang, J., Sun, J., 2004, Automatic classification of MRI images for three dimensional volume reconstruction by using general regression neural networks, Conf. Rec. 2003 *IEEE Nuc. Sci. Sym*., 5, 3188 -3189.

Received: August 30, 2006
Received in a revised form: October 2, 2006
Accepted: October 4, 2006