PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

Julita Stańczuk, Patrycja Trojczak-Golonka
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

WPŁYW ZRÓŻNICOWANIA ZBIORÓW ATRYBUTÓW I PROCESU WALIDACJI NA EFEKTYWNOŚĆ KLASYFIKACJI PRZEDSIĘBIORSTW PRZY WYKORZYSTANIU SIECI NEURONOWYCH

Streszczenie: Celem artykułu jest przedstawienie wybranego aspektu z paroletnich badań nad wielostanową klasyfikacją spółek notowanych na GPW w Warszawie. W artykule zaprezentowano analizę wrażliwości algorytmów klasyfikacji na dobór różnych kombinacji atrybutów opisowych oraz wpływ różnej struktury prób atrybutów uczących, testujących oraz walidacyjnych na efektywność klasyfikacji. Problem ten jest często pomijany w publikacjach naukowych, choć sam algorytm procedury badawczej może mieć wpływ na otrzymywane wyniki. W badaniach wykorzystano sztuczne sieci neuronowe. O oryginalności pracy decyduje również dobór samej próby badawczej (dane do badania pochodzą ze sprawozdań finansowych rzeczywistych 286 przedsiębiorstw polskich z lat 2006-2007).

1. Wstęp

Celem artykułu jest prezentacja wpływu na jakość klasyfikacji selekcji atrybutów opisowych oraz sposobu konstrukcji zbiorów uczących, testujących i walidacyjnych wykorzystywanych podczas budowy klasyfikatora.

Jako materiał badawczy wykorzystano dane finansowe 286 przedsiębiorstw, które mają obowiązek publikowania swoich sprawozdań finansowych. Próba badawcza obejmuje wszystkie (niebędące instytucjami o charakterze finansowym) przedsiębiorstwa notowane na GPW w Warszawie w roku zarówno 2006, jak i 2007. Do analizy brano pod uwagę dane pochodzące ze sprawozdań finansowych za lata 2006-2007. Na ich podstawie skonstruowano 17 wskaźników finansowych z zakresu zadłużenia, rentowności, płynności oraz sprawności działania. W sumie powstały 34 atrybuty opisowe, gdzie W1 to wskaźnik za rok 2006, a W2 to ten sam wskaźnik za rok 2007 itd.

2. Metody badań

Klasyfikacja przedsiębiorstw ma charakter klasyfikacji wielostanowej, gdyż firmy są oceniane w międzynarodowej skali ratingowej. W tym przypadku rating kredytowy został wygenerowany przez bankowy system automatycznej oceny

przedsiębiorstw zgodnie z pięciostopniową skalą austriacką, gdzie klasa pierwsza oznacza najlepszą kondycję finansową, a w klasie piątej znajdują się spółki bankrutujące[1].

Do najczęściej stosowanych miar oceny zdolności klasyfikacyjnej modelu należy macierz pomyłek (*confusion matrix*), w której wiersze macierzy odpowiadają poprawnym klasom decyzyjnym, natomiast kolumny – decyzjom przewidywanym przez klasyfikator [Stoch 2007]. Prostą miarą jakości klasyfikacji, którą łatwo jest skonstruować na podstawie macierzy pomyłek, jest trafność (*accuracy*). Wybrano ją ze względu na arbitralny charakter i możliwość porównania z wcześniejszymi wynikami innych badaczy zjawiska.

$$acc = \frac{\sum_{i=1}^{5} T_i}{\sum_{i=1}^{5} T_i + \sum_{i=1}^{5} F_i} = \frac{\sum_{i=1}^{5} T_i}{N},$$

gdzie: T_i – liczba obiektów z klasy i-tej zaklasyfikowana prawidłowo,
F_i – liczba obiektów z klasy i-tej zaklasyfikowana do niewłaściwej klasy,
$i = 1, \ldots, c$, gdzie $c = 5$,
N – liczba obiektów klasyfikowanych.

Do badań wykorzystano sieć jednokierunkową wielowarstwową, zwaną perceptronem wielowarstwowym. Wybrana została taka metoda klasyfikacji z powodu wielu zalet, m.in. zdolności do „uczenia się", dzięki której sieć neuronowa potrafi nauczyć się prawidłowych reakcji na określony zestaw bodźców [Witkowska 2002], zdolności do generalizacji, która polega na tym, że na podstawie zdobytego doświadczenia sieć potrafi wygenerować właściwe odpowiedzi dla nowych wzorców wejściowych, tzn. takich, na których nie była uczona, oraz odporności na uszkodzenia i błędy [Markowska-Kaczmar 2006].

Sieć uczono metodą pod nadzorem (z nauczycielem) ze względu na to, że znano pożądaną odpowiedź w postaci ratingu przyporządkowanego każdej spółce. Sieć miała jedną warstwę ukrytą, natomiast liczbę neuronów w warstwie tej dobierano eksperymentalnie z zakresu od 1 do 50. Zbudowano modele, w których losowanie próbek do uczenia i testowania było niezależne od ratingu (losowe), również eksperymentalnie określono najlepszą kombinację procentowej liczby atrybutów uczonych do testujących spośród najczęściej wykorzystywanych w literaturze (80:20 lub 60:20:20 przy próbie walidacyjnej). Jako funkcję błędu wykorzystano funkcję entropii wzajemnej[2]. W badaniu uwzględniono liniową funkcję

[1] Przykładową skalę pięciostopniową przedstawiono w artykule B. Lepczyńskiego [2001].

[2] Funkcja błędu wykorzystywana jest w trakcie uczenia sieci, jak również przy określaniu błędu w trakcie jej działania.

aktywacji do neuronów ukrytych i funkcję softmax, która jest szczególnie skuteczna w zadaniach klasyfikacji dla neuronów wyjściowych. Jako algorytm uczący wykorzystano algorytm BFGS[3].

3. Analiza istotności doboru atrybutów

Każdy atrybut wejściowy wymaga osobnego neuronu wejściowego, a z kolei wzrost ich liczby komplikuje strukturę całej sieci neuronowej. Wymagana liczba punktów uczących rośnie wraz z wymiarem przestrzeni sygnałów wejściowych, a szacunki określają, że minimalna liczba punktów, która jest potrzebna do wiernego odwzorowania cech modelu, równa jest 2^n, gdzie n to liczba atrybutów wejściowych. Oznacza to, że zwiększanie wymiarowości wektora wejściowego musi skutkować powiększaniem liczebności zbioru obserwacji [*Wprowadzenie*… 2001].

Co prawda większość sieci neuronowych (w tym MLP) jest mniej podatna na problemy z wymiarowością modelu niż inne metody klasyfikacji. Spowodowane to jest zdolnością sieci neuronowych do samodzielnego wyboru podzbioru atrybutów opisowych. Owo „przycinanie" przestrzeni opisowej odbywa się przez wyzerowanie wag neuronów wejściowych przypisanych do niechcianych cech opisowych. Niemniej problem zbyt dużej wymiarowości ma wpływ na działanie sieci neuronowych i zapewne można polepszyć efekt ich pracy przez wcześniejszą eliminację nieistotnych zmiennych. W przypadku przedstawionej bazy badawczej mamy 34 atrybuty opisowe i 286 punktów badawczych (przedsiębiorstw). Ponieważ nie ma możliwości zwiększenia liczebności zbioru uczącego, niezbędna jest eliminacja części zmiennych. Optymalna liczba atrybutów, zalecana przez źródła literaturowe, w tym przypadku to 8, ponieważ $2^8 = 256$ obserwacji [*Wprowadzenie*… 2001].

Wybór najistotniejszych cech powinien uwzględniać zarówno stopień powiązania atrybutu wejściowego ze zmienną objaśnianą, jak i wzajemne powiązania pomiędzy zmiennymi tak, aby unikać niepotrzebnej redundancji danych. W pierwszym etapie można posłużyć się analizą wrażliwości. Pozwala ona na odróżnienie ważnych zmiennych od takich, które niewiele wnoszą do wyniku działania sieci (te ostatnie można odrzucić). Analiza wrażliwości wykazuje stratę, jaką ponosimy, odrzucając konkretną zmienną, a jej podstawową miarą jest iloraz błędu uzyskanego przy uruchomieniu sieci dla zbioru danych bez jednej zmiennej i błędu uzyskanego z kompletem zmiennych. Im większy błąd po odrzuceniu zmiennej w stosunku do pierwotnego błędu, tym bardziej wrażliwa jest sieć na brak tej zmiennej. W efekcie zmienne można uszeregować pod względem ważności, co zostało przedstawione w tab. 1 (dla wybranej najlepszej sieci MLP z liczbą neuronów 34-38-5) [Trojczak-Golonka, Stańczuk 2009].

[3] Algorytm ten jest zaawansowaną metodą uczenia perceptronów wielowarstwowych, należy do metod gradientowych, a dokładniej *quasi*-newtonowskich.

Tabela 1. Analiza wrażliwości

Pozycja	Wskaźnik	Wartość ilorazu	Pozycja	Wskaźnik	Wartość ilorazu
1	w8	31,45947	**18**	w29	1,014811
2	w32	29,74717	**19**	w2	1,012442
3	w31	3,050426	**20**	w3	1,007699
4	w7	2,533516	**21**	w19	1,005769
5	w15	1,886779	**22**	w28	1,005306
6	w9	1,803530	**23**	w26	1,003841
7	w16	1,499950	**24**	w20	1,003704
8	w10	1,353835	**25**	w23	1,003279
9	w11	1,276151	**26**	w12	1,002267
10	w25	1,258118	**27**	w24	1,001828
11	w34	1,084451	**28**	w21	1,001652
12	w18	1,079487	**29**	w27	1,001311
13	w30	1,066771	**30**	w1	1,001225
14	w22	1,044461	**31**	w13	0,981461
15	w33	1,031691	**32**	w5	0,977781
16	w17	1,029619	**33**	w14	0,958402
17	w4	1,020874	**34**	w6	0,926678

Źródło: opracowanie własne.

Analiza wrażliwości wykazała, że najistotniejszym atrybutem jest wskaźnik W8, czyli jeden ze wskaźników rentowności majątku własnego za rok 2007. Kolejnym jest inny wskaźnik rentowności majątku przedsiębiorstwa za rok 2007 oraz ten sam wskaźnik za rok 2006.

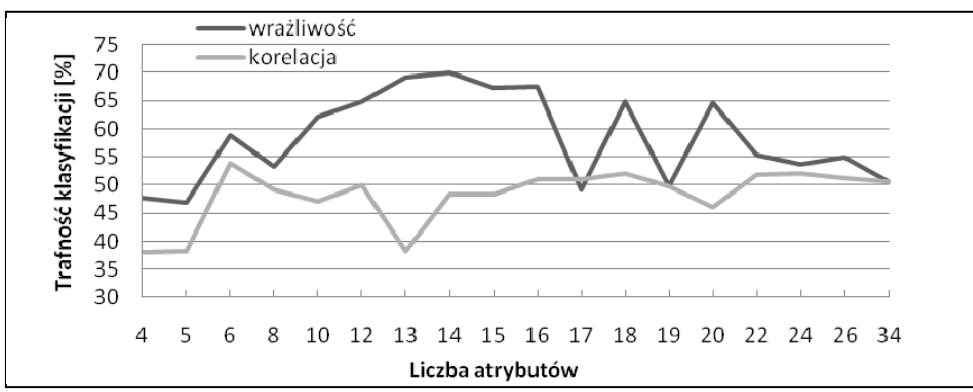
W drugim etapie należy usunąć ze zbioru zmiennych atrybuty wzajemnie skorelowane. Do pomiaru siły korelacji użyto współczynnik rang Spearmana, który ma – w stosunku do współczynnika korelacji – takie zalety, iż może zostać użyty przy braku założeń co do rozkładów badanych cech, jak też daje możliwość badania asocjacji wartości ilościowych i jakościowych jednocześnie [Aczel 2002]. W wyniku przeprowadzonego badania okazało się, że zgodnie z przewidywaniami część zmiennych wykazała silną korelację (nawet powyżej 0,9). Ponieważ należy usunąć zmienne najmniej istotne dla badanego obiektu, a jednocześnie pozostawione zmienne powinny być jak najbardziej różnorodne, zaproponowano algorytm selekcji, który uzależnia wartość wrażliwości zmiennych (czyli ich istotność) od korelacji z innymi zmiennymi ze zbioru zmiennych wybranych do dalszych analiz. W każdym kroku wybierana jest taka zmienna, której wartość wrażliwości pomniejszona o korelację z dotychczas wybranymi zmiennymi jest największa. Efektem pracy algorytmu powinien być zbiór atrybutów najistotniejszych z punktu widzenia klasyfikacji i jednocześnie najmniej ze sobą skorelowanych.

```
Start
W[34] - wektor wskaźników
W2[8] - wektor wskaźników posortowany względem wartości wrażliwo-
ści dla klasyfikacji z uwzględnieniem wzajemnej korelacji wskaź-
ników
```

```
WW[34] – liczbowa wartość wrażliwości dla poszczególnych wskaźni-
ków
K[34][34]- tablica korelacji wzajemnych
i,j,x,z – licznik naturalny

1: for i=1 to 8
2:  {
3:         x= z, dla którego WW[z]=max
4:         W2[i]=W[x]
5:         for j=1 to 34
6:                WW[j]=WW[j]*(1-abs(K[j][x]))
7: }
Stop
```

W rezultacie otrzymano taki oto zbiór ośmiu atrybutów: W8 (rentowność majątku), W11 i W12 (płynność bieżąca 2006 i 2007), W15 (efektywność majątku 2006), W22 (wskaźnik rentowności sprzedaży), W24 (obrotowość zapasów 2007), W27 (wskaźnik wystarczalności gotówki 2006), W31(rentowność majątku 2006). W drugim etapie zbadano, jak na jakość klasyfikacji wpływa selekcja atrybutów opisowych, w tym także selekcja z uwzględnieniem korelacji.



Rys. 1. Trafność klasyfikacji dla różnej liczby atrybutów opisowych

Źródło: opracowanie własne.

Na rysunku 1 pokazano, w jaki sposób ograniczanie liczby atrybutów wpływa na trafność klasyfikacji przedsiębiorstw. Seria „wrażliwość" to zmienne wyselekcjonowane jedynie na podstawie analizy wrażliwości, gdzie np. 4 atrybuty oznaczają cztery cechy o najwyższym poziomie wrażliwości. Seria „korelacja" oznacza zmienne wyselekcjonowane ze względu na poziom wrażliwości jednocześnie z uwzględnieniem wzajemnej korelacji według przedstawionego wcześniej algorytmu.

Wcześniejsze przypuszczenie o optymalnej liczbie 8 atrybutów nie znalazło potwierdzenia w badaniach. Okazało się, że najlepsze wyniki sieć osiąga dla 14 atrybutów (skuteczność blisko 70%). Korelację pomiędzy nimi przedstawia tab. 2.

Tabela 2. Korelacja wzajemna Spearmana

	w3	w8	w11	w12	w15	w16	w20	w21	w22	w23	w24	w27	w28	w31
w3	1,000	0,018	0,698	0,264	–0,403	–0,417	0,122	0,239	0,156	–0,056	–0,071	0,307	0,139	0,131
w8	0,018	1,000	0,046	0,174	0,200	0,233	0,695	–0,083	–0,165	–0,095	–0,043	0,233	0,496	0,563
w11	0,698	0,046	1,000	0,488	–0,172	–0,206	0,070	0,169	0,096	0,016	–0,016	0,343	0,185	0,162
w12	0,264	0,174	0,488	1,000	0,033	–0,151	0,128	0,077	0,037	0,038	0,071	0,188	0,448	0,296
w15	–0,403	0,200	–0,172	0,033	1,000	0,742	–0,060	–0,290	–0,241	0,069	0,119	–0,049	0,163	0,252
w16	–0,417	0,233	–0,206	–0,151	0,742	1,000	–0,141	–0,191	–0,142	0,123	0,118	–0,109	–0,015	0,038
w20	0,122	0,695	0,070	0,128	–0,060	–0,141	1,000	–0,043	–0,133	–0,087	–0,009	0,293	0,363	0,540
w21	0,239	–0,083	0,169	0,077	–0,290	–0,191	–0,043	1,000	0,898	–0,056	–0,089	–0,101	–0,048	–0,193
w22	0,156	–0,165	0,096	0,037	–0,241	–0,142	–0,133	0,898	1,000	–0,021	–0,068	–0,135	–0,109	–0,208
w23	–0,056	–0,095	0,016	0,038	0,069	0,123	–0,087	–0,056	–0,021	1,000	0,838	–0,113	–0,105	–0,100
w24	–0,071	–0,043	–0,016	0,071	0,119	0,118	–0,009	–0,089	–0,068	0,838	1,000	–0,118	–0,054	–0,047
w27	0,307	0,233	0,343	0,188	–0,049	–0,109	0,293	–0,101	–0,135	–0,113	–0,118	1,000	0,272	0,450
w28	0,139	0,496	0,185	0,448	0,163	–0,015	0,363	–0,048	–0,109	–0,105	–0,054	0,272	1,000	0,390
w31	0,131	0,563	0,162	0,296	0,252	0,038	0,540	–0,193	–0,208	–0,100	–0,047	0,450	0,390	1,000

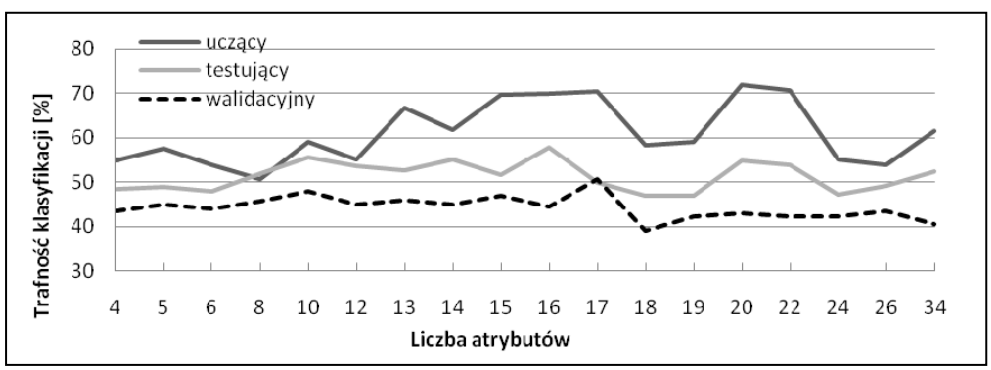
Źródło: opracowanie własne.

Faktycznie korelacja pomiędzy poszczególnymi wskaźnikami pozostaje na niskim poziomie. Z jeszcze mniejszej liczby atrybutów stosunkowo dobre wyniki osiąga się dla 6 cech. Nie potwierdziła się również w badaniach teza dotycząca szkodliwego wpływu na wyniki klasyfikacji redundancji informacji na skutek wzajemnej korelacji cech. Dla selekcji cech z uwzględnieniem korelacji prawie w każdym przypadku osiągnięto gorsze efekty. Dlatego dalsze badania wykonano z uwzględnieniem wyłącznie wrażliwości.

4. Wpływ walidacji na efektywność klasyfikacji

Istotnym problemem podczas tworzenia modelu sieci neuronowej jest możliwość zbytniego dopasowania się tego modelu do danych uczących. Powoduje to, że klasyfikacja odbywa się z dużą skutecznością, ale tylko dla danych podobnych do danych uczących. Natomiast na pewno model taki nie odzwierciedla rzeczywistego obrazu modelowanego zjawiska. Aby zapobiec temu negatywnemu zjawisku, stosuje się walidację. Większa sieć jest zdolna do modelowania złożonych zjawisk, ale może mieć również skłonność do dopasowania się do danych. Z kolei sieć o prostej strukturze może generować wyniki o zbyt ogólnym i niewystarczającym charakterze. I właśnie walidacja to proces, który ma pozwolić na wyznaczenie właściwego momentu przerwania procesu uczenia sieci. Dane wejściowe dzielone są

na trzy grupy: dane uczące, które biorą bezpośredni udział w budowie sieci, dane walidacyjne, które służą do bezpośredniej kontroli algorytmu uczenia, oraz zbiór testujący, którym weryfikuje się gotową sieć. Podczas tworzenia sieci jej jakość testowana jest na zbiorze uczącym i walidacyjnym. Początkowo oba te błędy (wartość funkcji błędu) są wysokie, później w miarę wzrostu skomplikowania sieci wartość obu błędów powinna szybko spadać. Po początkowym spadku obserwuje się szybszy spadek błędu dla danych uczących i wolniejszy dla grupy walidacyjnej. Przełomowym momentem podczas tworzenia sieci jest chwila, w której spadek błędu próby walidacyjnej zatrzyma się lub zacznie rosnąć. Jest to sygnał, że sieć zaczęła się zbytnio dopasowywać do danych uczących. Fakt ten oznacza przerwanie procesu uczenia i „cofnięcie" sieci do momentu, gdy błąd na próbie walidacyjnej był najmniejszy. Efektem takiego procesu uczenia powinna być sieć o uniwersalnych zdolnościach klasyfikacyjnych. Końcowa postać sieci jest testowana na zbiorze testowym, aby ostatecznie potwierdzić zdolności sieci na przypadkach, które zupełnie nie brały udziału w procesie uczenia, i wykluczyć jakiekolwiek wątpliwości na temat jakości i udziału zbiorów uczącego i walidacyjnego [*Internetowy podręcznik statystyki*... 2009].

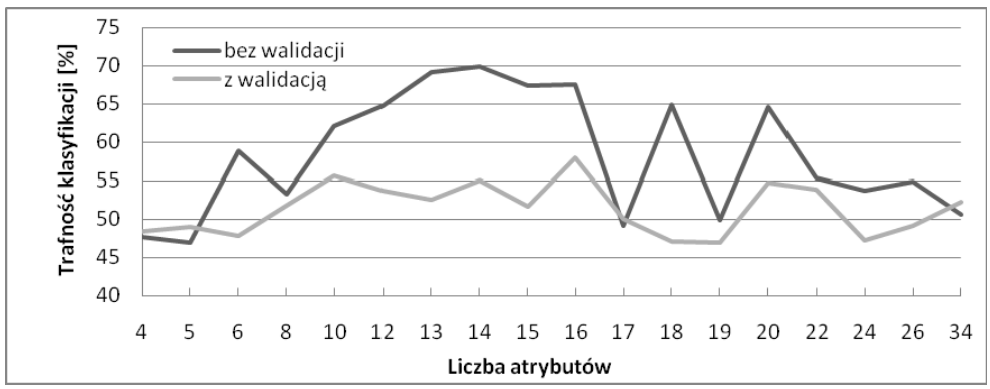


Rys. 2. Trafność klasyfikacji dla różnych zbiorów danych

Źródło: opracowanie własne.

Na rysunku 2 pokazano wyniki trafności dla zbioru uczącego, testującego i walidacyjnego, proporcje wszystkich zbiorów wynoszą odpowiednio 60:20:20. Najlepsze wyniki osiągnięto dla zbioru uczącego, ponieważ na zbiorze tym sieć się bezpośrednio uczy. Stosunkowo dobre efekty uzyskano dla zbioru testowego, zawierającego zupełnie nowe, nieznane sieci przypadki.

Na rysunku 3 zaprezentowano skuteczność sieci dla zbiorów testujących. Przeciętnie wyniki dla sieci bez próby walidacyjnej są lepsze, jednak skuteczność tych sieci jest bardziej zróżnicowana i zależna od liczby atrybutów. W drugim wariancie (z próbą walidacyjną) wyniki są słabsze, jednak bardziej jednorodne. Skutecz-



Rys. 3. Trafność klasyfikacji dla różnych zbiorów danych

Źródło: opracowanie własne.

ność tej sieci oscyluje w granicach 50-55% (przy klasyfikacji do 5 klas) i jest niezależna od liczby atrybutów. Można się również spodziewać, że sieć ta będzie sobie lepiej radzić z nowymi i nietypowymi obiektami. W badaniach tych próba ucząca i testująca są do siebie podobne (wynika to z techniki wykonania badania – losowanie z jednego zbioru) i pewnie dlatego sieć „przeuczona" (zbyt mocno dopasowana do próby uczącej) może osiągać lepsze wyniki. Odrębną kwestią jest fakt, że z przyjętej metody badań wynikła sytuacja, w której liczebność zbioru uczącego w obu wariantach jest różna.

5. Podsumowanie

Z przeprowadzonych badań wynikają następujące wnioski o charakterze ogólnym:

- ograniczanie liczby atrybutów opisowych do pewnego optymalnego poziomu zawsze prowadzi do zwiększenia trafności klasyfikacji,
- sieć neuronowa doskonale radzi sobie z wzajemną korelacją cech opisowych i nie wymaga wcześniejszej interwencji w tej kwestii,
- proces walidacji powoduje, że sieć ma ograniczone możliwości dopasowania się do danych uczących. W sytuacji gdy nowe przypadki będą podobne do przypadków uczących, lepiej radzi sobie sieć uczona bez walidacji. Natomiast gdy nowe przypadki będą stosunkowo inne od przypadków uczących, wówczas lepsze wyniki osiąga sieć uczona z walidacją, ze względu na jej bardziej ogólny charakter działania.

Na koniec należy zwrócić uwagę na to, aby przy porównywaniu wyników zbiorów testowych przeanalizować stosowaną technikę badania, a szczególnie proporcje poszczególnych zbiorów uczących i testujących. Temat ten zostanie poruszony w następnej publikacji.

Literatura

Aczel A.D., *Statystyka w zarządzaniu*, Wydawnictwo Naukowe PWN, Warszawa 2007.
Internetowy podręcznik statystyki, Statsoft Polska, dostępny w Internecie: http://www.statsoft.com.pl/textbook/stathome_stat.html?http://www.statsoft.com.pl/textbook/stneunet.html, 10.10.2009.
Lepczyński B., *Banki w ocenie agencji ratingowych*, „Bank" 2001 nr 7-8.
Markowska-Kaczmar U., *Ekstrakcja reguł z sieci neuronowych – podejście ewolucyjne*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław 2006.
Stoch P., *Zastosowanie narzędzi statystycznych i matematycznych metod sztucznej inteligencji do predykcji wystąpienia dysplazji oskrzelowo-płucnej u noworodków*, praca doktorska, Akademia Górniczo-Hutnicza, Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki, Kraków 2007.
Trojczak-Golonka P., Stańczuk J., *Zastosowanie sieci neuronowych w klasyfikacji obiektów ekonomicznych*, [w:] *Inżynieria wiedzy i systemy ekspertowe*, red. A. Grzech, K. Juszczyszyn, H. Kwaśnicka, Ngoc Than Nguye, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2009.
Witkowska D., *Sztuczne sieci neuronowe i metody statystyczne – wybrane zagadnienia finansowe*, Wydawnictwo C.H. Beck, Warszawa 2002.
Wprowadzenie do sieci neuronowych, Statsoft Polska, Kraków 2001.

THE INFLUENCE OF DIFFERING SETS OF ATTRIBUTES AND THE PROCESS OF VALIDATION ON THE EFFECTIVENESS OF COMPANIES CLASSIFICATION USING NEURAL NETWORKS

Summary: This aim of the article is presenting a selected aspect of a few years of research on multistate classification of companies listed on the Warsaw Stock Exchange. In the article, a sensitivity analysis of classification algorithms for the selection of different combinations of descriptive attributes and the impact of various attributes of the structure of learning trials, testing and validation on the effectiveness of classification are presented. This problem is often overlooked in scientific publications, although the algorithm itself of the test procedure may influence the results obtained. In the studies, artificial neural networks are used. The originality of the work determines the selection of the sample test (the test data derived from actual accounts of 286 Polish companies from the years 2006-2007).