

MODELOWANIE KONIUNKTURY GOSPODARCZEJ Z WYKORZYSTANIEM DANYCH TEKSTOWYCH

Szymon Chojnacki

Zakład Wspomagania i Analizy Decyzji, Szkoła Główna Handlowa, Warszawa

1 WPROWADZENIE

Gospodarka krajów rozwiniętych podlega fluktuacjom. Po okresie szybkiego rozwoju następuje okres spowolnienia i stagnacji. Podczas fazy dobrej koniunktury spada bezrobocie i rosną dochody [1]. Z kolei słaby stan koniunktury związany jest ze spadkiem popytu i obniżeniem indeksów giełdowych.

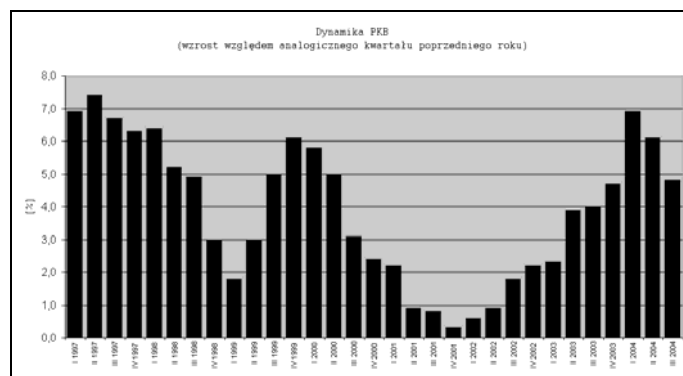
Umiejętność oceny aktualnego stanu gospodarki jest ważna dla prezesa banku centralnego i ministra finansów, którzy poprzez dostępne narzędzia starają się niwelować odchylenia gospodarki od długookresowego trendu wzrostowego. Również przedsiębiorstwa i gospodarstwa domowe zainteresowane są stanem koniunktury. Dostosowują one swoje inwestycje i zakupy do oczekiwanych przychodów w przyszłości.

Wśród narzędzi służących do prognozowania koniunktury szczególną popularnością cieszą się barometry koniunktury i modele ekonometryczne [2]. Pierwsze pozwalają budować wskaźniki diagnozujące aktualną koniunkturę, drugie dopasowują wielkości agregatów gospodarczych do równań opisujących prawa ekonomiczne i pokazują kierunek rozwoju gospodarki.

W pracy zostały użyte dane tekstowe do modelowania koniunktury. Podobne próby podejmowano wcześniej w celu wyznaczenia zależności pomiędzy komunikatami prasowymi a kształtowaniem się indeksów giełdowych. Cho [3] zbadał, jaki wpływ na kurs akcji na giełdzie w Hong Kongu ma pojawienie się w komunikatach giełdowych jednego z około 300 predefiniowanych zwrotów ekonomicznych. Lavrenko [6] skonstruował program, który przewiduje zmianę trendu 130 walorów na giełdzie

w Nowym Jorku, wykorzystując komunikaty prasowe i na tej podstawie podejmuje opłacalne decyzje inwestycyjne. Kroha [6] sprawdził, jakie słowa są charakterystyczne dla okresów wzrostu i spadku niemieckiego indeksu DAX 30. Istnieje szereg analogii pomiędzy modelowaniem indeksów giełdowych i modelowaniem koniunktury. Różnica pomiędzy obydwojema podejściami jest techniczna i polega na konieczności zawężenia tekstów służących do analizy giełdy do tych, które dotyczą spółek budujących dany indeks.

Do oceny stanu koniunktury wykorzystano wielkość dynamiki PKB. W latach od 1997 do 2004 mieliśmy dwie fazy spowolnienia i dwie fazy ożywienia gospodarczego, widoczne na rys. 1.



Rys. 1. Dynamika PKB względem analogicznego kwartału poprzedniego roku.

Jako źródło tekstów wykorzystano stronę internetową dziennika „Rzeczpospolita”. Do przetworzenia tekstów użyto makra w aplikacjach Excel i Access. Analizy ilościowe i budowa modeli prognostycznych została przeprowadzona w programie *STATISTICA*.

2 CEL PRACY

Podstawowym celem pracy jest weryfikacja hipotezy o istnieniu związku pomiędzy słowami występującymi w tytułach artykułów ekonomicznych a stanem koniunktury. W celu sprawdzenia, czy taki związek ma miejsce, wyznaczono słowa, których częstość wystąpień w okresach z dobrą koniunkturą istotnie różni się od częstości wystąpień w okresach złej koniunktury.

Drugim celem pracy jest próba zbudowania modeli prognozujących koniunkturę, które można interpretować i konfrontować z intuicją. Zbudowane modele to drzewo klasyfikacyjne oraz funkcja dyskryminacyjna.

Trzecim celem pracy jest sprawdzenie, czy można na części dostępnych danych zbudować model klasyfikacyjny, który z wysoką dokładnością będzie potrafił przewidzieć koniunkturę w nowych miesiącach. Zbudowany model prognostyczny jest siecią neuronową.

3 PRZETWORZENIE TEKSTÓW

Celem poniższego rozdziału jest opisanie procedury pozyskania tytułów wiadomości ekonomicznych z Internetu oraz ich przetworzenia do postaci umożliwiającej ilościową analizę.

Do modelowania koniunktury wykorzystano tytuły wiadomości gospodarczych umieszczonych na stronach internetowych dziennika „Rzeczpospolita” w latach 1997-2004. Najpierw ściągnięto 244 895 tytułów, które obejmowały 126 kategorii tematycznych. Następnie wydzielono 27 551 tytułów zawartych w dziale Ekonomia¹⁸. W kolejnym kroku usunięto tytuły, które wystąpiły w bazie ponad dziesięć razy, co mogło być związane z ich stałym miejscem na łamach dziennika (tj. notowania giełdowe). W rezultacie uzyskano 26 310 tytułów, które zagregowano względem 95 miesięcy od lutego 1997 do grudnia 2004.

Przykładowa strona internetowa pokazana jest na rys. 2. Znalezienie tytułów na stronie archiwalnej było zadaniem prostym w porównaniu z wyszukiwaniem tytułów w czasie rzeczy-

wistym. Problem ten jest szczegółowo opisany przez K. Nørvåg i R. Øyri [7] i związany jest z poszukiwaniem w kodzie html strony charakterystycznych wzorców (na przykład URL-Tytuł-URL, , Tytuł). W naszym przypadku tytuły zaczynały się zawsze w tym samym miejscu na stronie.

40. dokumentów [] Strona 1 z 2

RZECZPOSPOLITA PŁATNE SERWISY

RZECZPOSPOLITA Spis treści wydania 030. z dnia 03.02.1997 [40 artykułów w bezpłatnych w tym wydaniu]
szukaj

dodatek/Media, Internet

1. Departament stanu USA o wolności słowa w Polsce

dodatek/Moja Kariera (Praca, Specjaliści)

1. Japoński styl pracy
2. Praca a dla specjalisty
3. Umowa o pracę i zakaz konkurencji

dodatek/Moje Podróże

1. Hongkong
2. Kotlina Kłodzka

gazeta/Ekonomia

1. 1000. sesja giełdowa
2. Jak Kołodko PZU oddawał
3. Liberalizm bez wzajemności
4. Notowania - 06.02.1997
5. Notowania giełdowe - 06.02.1997
6. Prawo co dnia - spis treści

gazeta/Kraj

1. Czego SdRP spodziewa się po ministrze finansów
2. Nie wydać Mandułęgi
3. Prawo co dnia - spis treści
4. SLD chce postawić Milczanowskiego przed Trybunałem Stanu
5. TVP zagra komercja i politycy

gazeta/Kultura

1. Muzeum Narodowe w Warszawie
2. Pamela Anderson w Polsce
3. XVIII Przegląd Piosenki Aktorskiej

gazeta/Nauka i Technika

1. Komputerowa panorama ziemi
2. Nowa taryfa celna na komputery

gazeta/Prawo

1. Brak tytułu

http://www.rzeczpospolita.pl/szukaj/spis.pl?t=1997020519970205 2005-03-23

Rys. 2. Kopia strony internetowej ze spisem artykułów z 5 lutego 1997 r. pobrana z witryny www.rzeczpospolita.pl/szukaj/spis.pl?t=1997020519970205 w dniu 23.03.2005.

Tytuły wiadomości ekonomicznych zidentyfikowane na stronie archiwum zostały oczyszczone z cyfr, znaków interpunkcyjnych i innych znaków (na przykład procent, tylda, cudzysłów, nawias okrągły, nawias kwadratowy, amperсанд, hash, plus). Następnie zastąpiono odmienne formy wyrazów przez formy rdzenne (na przykład słowo „wzrosło” zamieniono na „wzrastać”). Tabela 1 pokazuje co stało się w wyniku tej operacji z tytułami z rys. 2. Do znalezienia form podstawowych wyrazów wykorzystano dwa słowniki bezpłatnie dostępne w internecie: słownik ortograficzny języka polskiego PWN, słownik odmian wyrazów TIP.

¹⁸ Najbardziej popularnymi kategoriami artykułów były: doda-tek/Notowania (50 259 wystąpień), gazeta/Ekonomia (27 551), gazeta/Prawo (25 483), gazeta/Kraj (17 091), gazeta/Świat (15 055), gazeta/Sport (14 426), gazeta/Publicystyka (14 000), gazeta/Kultura (12 673).

Czasami dana forma odmiany związana była z różnymi słowami pierwotnymi (na przykład słowo „mam” może być dopełniaczem rzeczownika „mama”, bądź formą czynną pierwszej osoby czasu teraźniejszego czasownika „mieć”). W takich sytuacjach zawsze następowała zmiana na jedno i to samo słowo pierwotne, jednakże słowo to było wybierane w sposób losowy.

Tabela 1. Tytuły z działu Ekonomia z dnia 5 lutego 1997, po przetworzeniu.

Data	Tytuł
05.02.1997	sesja giełda
05.02.1997	Jak Kołodko PZU oddawać
05.02.1997	Liberalizm bez wzajemność
05.02.1997	Notowanie
05.02.1997	Notowanie giełda
05.02.1997	Bezrobocie spadać
05.02.1997	Prawo co dzień spis treści

W dalszym etapie przetwarzania danych tekstowych do postaci umożliwiającej zastosowanie algorytmów ilościowych rozdzielono tytuły na nieuporządkowane zbiory słów w każdym miesiącu. W tym momencie pojawiły się liczne wątpliwości, czy tak znaczące przetworzenie danych tekstowych pozwoli nam zachować treść związaną z tytułami. Wydaje się, że poprzez sprowadzenie słów do formy rdzennej (*ang. stemming*) oraz rozbięcie słów na nieuporządkowany zbiór słów (*ang. bag of words*) traci się bezpośrednie znaczenie zdań. Jednocześnie liczne eksperymenty psycholingwistyczne [5] dowodzą, że udaje się zachować głębokie odczucia związane z czytaniem tych słów bez względu na kontekst i ich formę w zdaniu [4]. A zatem, jeżeli stan koniunktury jest opisywany w tytułach poprzez słowa o innym wydźwięku emocjonalnym w okresach ożywienia i stagnacji, to eksperymenty statystyczne z następnego rozdziału mają szansę wykryć taki związek.

4 ANALIZY STATYSTYCZNE

Tabela z danymi do analizy składała się z 95 wierszy odpowiadających miesiącom od lutego 1997 do grudnia 2004 oraz 6 614 kolumn odpowiadających formom podstawowym słów. Na przecięciu słów i miesięcy znajdują się unormowane częstości występowania słowa w danym miesiącu. Do każdego miesiąca została dodana etykieta określająca, czy koniunktura

w miesiącu była dobra czy zła. Dane w takiej postaci były już gotowe do analiz statystycznych. Ponieważ jednak liczba cech była zbyt duża pozostawiono tylko te słowa, które występowały w badanym okresie co najmniej 95 razy, co pozwoliło zredukować liczbę cech do 156. Jednakże nadal liczba ta była większa od liczby obiektów. Z tego powodu przeprowadzono analizę czynnikową do redukcji liczby cech. Pomimo zastosowania pięciu różnych metod wyznaczania czynników, nie udało się przekroczyć 22% wyjaśnionej wariancji przez pierwsze dziesięć czynników w żadnej metodzie. Z tego powodu konieczne było zastosowanie metody redukcji liczby cech, która korzysta z wiedzy o klasie, do jakiej należał każdy obiekt. Jedną z takich metod jest analiza wariancji.

4.1 Analiza wariancji

Aby ograniczyć liczbę cech (słów) opisujących analizowane miesiące, sprawdzono, które słowa istotnie różnicują miesiące z dobrą i złą koniunkturą. Do porównania zróżnicowania użyto średniej unormowanej częstości wystąpienia słowa w miesiącach z dobrą i złą koniunkturą.

Do weryfikacji hipotezy zerowej o równości średnich w obu grupach obliczono dla każdego słowa wartość statystyki F jako ważony iloraz zmienności między- i wewnątrzgrupowej:

$$F = \frac{\sum_{i=1}^k (\bar{x}_i - \bar{x})^2 n_i}{\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2} \cdot \frac{n - k}{k - 1},$$

gdzie:

n – liczba obserwacji (95),

k – liczba grup (2),

x_{ij} – unormowana częstość wystąpień danego słowa w j -tym miesiącu i -tej grupy,

\bar{x} – średnia unormowana częstość wystąpień danego słowa we wszystkich miesiącach,

\bar{x}_i – średnia unormowana częstość wystąpień danego słowa w i -tej grupie.

Jeżeli wartość statystyki F jest większa od wartości progowej, wówczas odrzuca się hipotezę zerową o braku różnic średnich pomiędzy grupami miesięcy z dobrą i złą koniunkturą. Wartość progowa jest pobierana z rozkładu F-Snedecora o $(k-1)$ i $(n-k)$ stopniach swobody.

Zależna Zm.	Test SS dla pełnego modelu względem SS dla reszt (Do statistica)										
	Wielokr. R	Wielokr. R2	Skorygow R2	SS Model	df Model	MS Model	SS Reszta	df Reszta	MS Reszta	F	p
ożywienie	0,451442	0,203800	0,195238	19,15717	1	19,15717	74,84283	93	0,804762	23,80477	0,000004
polski	0,307829	0,094759	0,085025	8,90731	1	8,90731	85,09269	93	0,914975	9,73503	0,002408
cena	0,307566	0,094597	0,084862	8,89212	1	8,89212	85,10788	93	0,915138	9,71670	0,002430
nadal	0,287683	0,082762	0,072899	7,77960	1	7,77960	86,22040	93	0,927101	8,39132	0,004700
klient	0,282780	0,079965	0,070072	7,51669	1	7,51669	86,48331	93	0,929928	8,08308	0,005493
rosnąć	0,281002	0,078962	0,069059	7,42247	1	7,42247	86,57753	93	0,930941	7,97308	0,005808
bycie	0,277799	0,077173	0,067250	7,25422	1	7,25422	86,74578	93	0,932750	7,77724	0,006417
prywatyzacja	0,272738	0,074386	0,064433	6,99227	1	6,99227	87,00773	93	0,935567	7,47383	0,007495
euro	0,260338	0,067776	0,057752	6,37091	1	6,37091	87,62909	93	0,942248	6,76140	0,010835
wolno	0,258210	0,066672	0,056636	6,26718	1	6,26718	87,73282	93	0,943364	6,64344	0,011524
kapitał	0,244016	0,059544	0,049431	5,59713	1	5,59713	88,40287	93	0,950569	5,88819	0,017171
czy	0,242056	0,058591	0,048469	5,50758	1	5,50758	88,49242	93	0,951531	5,78813	0,018113
ale	0,238897	0,057072	0,046933	5,36475	1	5,36475	88,63525	93	0,953067	5,62893	0,019726
droga	0,229579	0,052707	0,042521	4,95443	1	4,95443	89,04557	93	0,957479	5,17446	0,025219
poprawa	0,222880	0,049676	0,039457	4,66951	1	4,66951	89,33049	93	0,960543	4,86133	0,029931
inflacja	0,218883	0,047910	0,037672	4,50350	1	4,50350	89,49650	93	0,962328	4,67980	0,033083
podwyżka	0,212518	0,045164	0,034897	4,24542	1	4,24542	89,75458	93	0,965103	4,39893	0,038675
rząd	0,210743	0,044413	0,034138	4,17479	1	4,17479	89,82521	93	0,965862	4,32234	0,040368
na	0,209646	0,043951	0,033671	4,13144	1	4,13144	89,86856	93	0,966329	4,27539	0,041446
europejski	0,206818	0,042774	0,032481	4,02074	1	4,02074	89,97926	93	0,967519	4,15573	0,044333
rynek	0,205593	0,042269	0,031970	3,97325	1	3,97325	90,02675	93	0,968030	4,10447	0,045635
reforma	0,203529	0,041424	0,031117	3,89386	1	3,89386	90,10614	93	0,968883	4,01892	0,047900
lepszy	0,201466	0,040588	0,030272	3,81531	1	3,81531	90,18469	93	0,969728	3,93442	0,050257
mniej	0,201424	0,040572	0,030255	3,81374	1	3,81374	90,18626	93	0,969745	3,93272	0,050305
miliard	0,200981	0,040393	0,030075	3,79698	1	3,79698	90,20302	93	0,969925	3,91472	0,050823
mało	0,200974	0,040390	0,030072	3,79670	1	3,79670	90,20330	93	0,969928	3,91441	0,050832
obligacja	0,197677	0,039076	0,028744	3,67316	1	3,67316	90,32684	93	0,971256	3,78186	0,054832
strata	0,195649	0,038279	0,027938	3,59819	1	3,59819	90,40181	93	0,972062	3,70161	0,057418

Rys. 3. Słowa z największą wartością statystyki F. Na podstawie modułu ANOVA i procedury jednoczynnikowej.

Na rys. 3 znajdują się wyniki analizy wariancji dla słów z najwyższą wartością statystyki F. Dla analizowanych danych w przypadku 22 słów wartość statystyki F przekroczyła wartość progową (por. rys. 3). Wartość p informująca o ryzyku odrzucenia hipotezy zerowej w przypadku, gdy jest ona prawdziwa, znalazła się dla tych słów poniżej poziomu 5%. W rezultacie różnice w występowaniu tych słów w miesiącach z dobrą i złą koniunkturą są istotne. Oznacza to, że istnieją istotne związki pomiędzy wystąpieniami słów (ożywienie, polski, cena, nadal, klient, rosnać, bycie, prywatyzacja, euro, wolno, kapitał, czy, ale, droga, poprawa, inflacja, podwyżka, rząd, na, europejski, rynek, reforma) w tytułach a stanem koniunktury.

Wyniki analizy wariancji mogą być mylące w przypadku naruszenia założeń o normalności rozkładów zmiennych w grupach oraz w sytuacji, gdy wariancja jest różna pomiędzy grupami. W module ANOVA sprawdzono, że dla 22 wybranych słów założenie o normalności rozkładu jest w większości przypadków naruszone. Jednocześnie dla 15 słów wartość testu Barletta nie przekroczyła wartości krytycznej. Oznacza to, że dla tych słów nie ma podstaw do

odrzucenia hipotezy zerowej o równości wariancji pomiędzy grupami. Biorąc pod uwagę fakt, że test F jest mocno odporny na naruszenie założeń o normalności i równości wariancji, uznano, że dalsza analiza danych zostanie ograniczona do wytypowanych 22 słów.

4.2 Interpretacja modeli klasyfikacyjnych

W poniższym podrozdziale zbudowano drzewo klasyfikacyjne i funkcję dyskryminacyjną do poznania bezpośrednich związków pomiędzy słowami występującymi w miesiącach a stanem koniunktury. W celu wykrycia związków w całej dostępnej populacji uczenie modeli przeprowadzono na pełnym zbiorze obiektów. W rezultacie utworzone modele można wykorzystać do lepszego zrozumienia analizowanych danych. Jednocześnie trzeba mieć na uwadze, że dokładność klasyfikacyjna tych modeli na nieobserwowalnych danych jest z reguły niższa niż na danych uczących.

Liniowa analiza dyskryminacyjna, przeprowadzana dla zmiennej objaśnianej przyjmującej dwie wartości, polega na wyznaczeniu takich

współczynników funkcji klasyfikacyjnych, które zapewniają wysoką jakość klasyfikacji.

$$Y_j = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_{22} X_{22},$$

W powyższym wzorze Y_j oznacza zmienną grupującą związaną z klasą j (koniunkturą), natomiast X_i odpowiada zmiennym objaśniającym (słowom).

Algorytm poszukuje takich wartości współczynników, które maksymalizują iloraz zmienności międzygrupowej wartości funkcji klasyfikacyjnej przez zmienność wewnątrzgrupową wartości funkcji klasyfikacyjnej. A zatem kryterium wyboru parametrów funkcji klasyfikacyjnej bazuje na analizie wariancji opisanej wcześniej.

Tabela 2. Współczynniki przy zmiennych w funkcji klasyfikacyjnej, na podstawie modułu analiza klasyfikacyjna.

Zmienna	Stan koniunktury	
	zły	dobry
prywatyzacja	0,399245	-0,65432
polski	0,334194	-0,54771
ale	0,322457	-0,52850
nadal	0,292504	-0,47938
rynek	0,179061	-0,29346
cena	0,175337	-0,28736
reformacja	0,144642	-0,23705
podwyżka	0,101121	-0,16573
inflacja	0,038001	-0,06228
czy	0,033923	-0,05560
wolno	-0,051350	0,08416
europowski	-0,053216	0,08722
klient	-0,056391	0,09242
na	-0,108434	0,17771
euro	-0,231335	0,37913
droga	-0,292610	0,47021
poprawa	-0,293012	0,48021
rząd	-0,344058	0,56387
kapitał	-0,375775	0,61585
rosnąć	-0,430220	0,70508
bycie	-0,481952	0,78987
ożywienie	-0,559530	0,91701
Stała	-0,989090	-2,34758

Tabela 2 zawiera współczynniki stojące przy zmiennych w funkcjach klasyfikacyjnych. Przyjmuje się, że obiekt należy do tej klasy, dla której funkcja klasyfikacyjna przyjęła większą wartość. Dokładność klasyfikacyjna tak skonstruowanego modelu wynosi 90%. Siła dyskryminacji zmiennych diagnostycznych mierzona współczynnikiem korelacji kanonicznej wyniosła 0,79, a zatem była znaczna. Natomiast stopień zakłóceń modelu innymi czynnikami mierzony przy

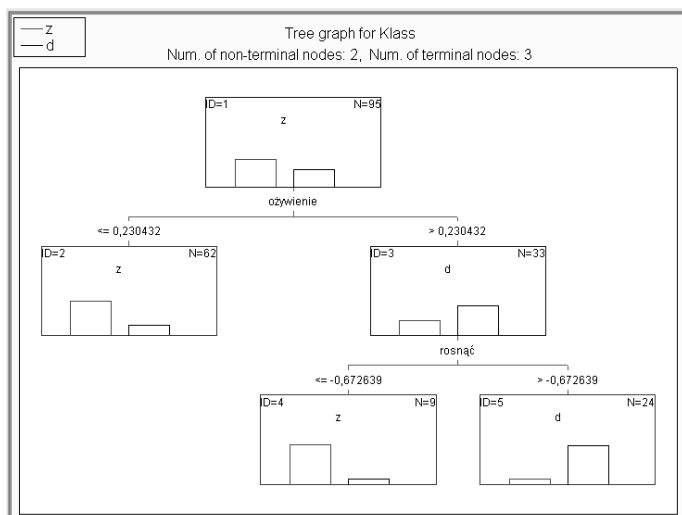
pomocy wskaźnika λ Wilksa wyniósł jedynie 0,36. Statystyka chi-kwadrat badająca hipotezę zerową: $H_0: \lambda = 1$ wyniosła 81,95, a związany z nią poziom istotności p wyniósł 0,0. Co oznacza, że ryzyko popełnienia błędu przy odrzuceniu hipotezy zerowej na rzecz hipotezy alternatywnej: $H_1: \lambda < 1$ jest znikome.

Ponieważ dane wykorzystane do budowy modelu zostały wcześniej wystandaryzowane, zatem współczynniki stojące przy zmiennych są porównywalne. Widzimy, że z dobrą koniunkturą związane są słowa „ożywienie”, „rosnąć”, „poprawa”. Natomiast słowa charakterystyczne dla złej koniunktury to „ale”, „nadal”, „czy”. Jednocześnie związek słów „cena”, „podwyżka” i „inflacja” ze złą koniunkturą jest sprzeczny z intuicją ekonomiczną, ponieważ w okresach złej koniunktury ceny spadają i pisze się o nich rzadziej niż w okresach dobrej koniunktury.

Drzewa klasyfikacyjne buduje się w celu wyznaczenia kryteriów podziału obiektów na jednorodne grupy. Kryterium podziału obiektów może być cięcie wielowymiarowe lub cięcie jednowymiarowe. Jeżeli jeden warunek podziału nie wystarczy do wyznaczenia jednorodnych grup obiektów, to dla wyznaczonych grup powtarza się rekurencyjne poszukiwanie najlepszego kryterium podziału. Czynność tę powtarza się do czasu wyznaczenia zadowolających podgrup obiektów.

Do oceny jakości warunku znajdującego się w tzw. węzłach drzewa klasyfikacyjnego korzysta się z heurystycznych miar różnicowania rozkładu (np. współczynnik Giniego, statystyka chi-kwadrat, entropia, rozróżnialność). W rezultacie szybko można otrzymać kryteria podziału obiektów na jednorodne grupy.

Rys. 4 przedstawia drzewo zbudowane w module *Drzewa interakcyjne* pakietu *STATISTICA* z wykorzystaniem domyślnych ustawień dla metody CHAID. Drzewo to poprawnie klasyfikuje 81% obiektów. Informuje ono nas, że jeżeli w danym miesiącu słowo „ożywienie” występuje rzadziej niż pewien poziom, to mamy złą koniunkturę. Jeżeli słowo to występuje wystarczająco często, wówczas aby określić, czy mamy dobrą koniunkturę, potrzebna nam jest dodatkowo informacja o wystąpieniach słowa „rosnąć”. Jeżeli oba słowa występują wystarczająco często, wówczas z dużą pewnością możemy przyjąć, że w miesiącu była dobra koniunktura.



Rys. 4. Drzewo klasyfikacyjne utworzone przez algorytm CHAID, na podstawie modułu *Drzewa interakcyjne*.

Sprawdzono, że trzy drzewa zbudowane przy pomocy algorytmu C&RT mają w korzeniu również słowo „ożywienie”. Jednakże kolejny warunek był oparty na słowach „klient” lub „prywatyzacja”. Jakość klasyfikacyjna tych drzew przy głębokości takiej jak na rys. 4 była średnio niższa o 5% od jakości algorytmu CHAID. Również algorytm exhaustive-CHAID w korzeniu umieścił słowo „ożywienie”, lecz wybrane kryterium podziału składało się z dwóch cięć.

4.3 Klasyfikacja nowych miesięcy

Oba modele z poprzedniego podrozdziału dają się w łatwy sposób interpretować. Cechuje je również zadowalająca skuteczność klasyfikacyjna na danych treningowych. Jeżeli jednak chcielibyśmy poznać potencjał naszych danych w prognozowaniu koniunktury w nowych miesiącach, należy podzielić zbiór dostępnych miesięcy na niezależne części: uczącą i testującą oraz zastosować bardziej złożone modele, tj. sieci neuronowe.

Sztuczne sieci neuronowe są klasą algorytmów prognostycznych, klasyfikacyjnych i grupujących, których działanie inspirowane jest działaniem mózgu. Szacuje się, że w mózgu człowieka znajduje się około 10 miliardów neuronów. Do przeciętnego neuronu dochodzi kilka tysięcy połączeń od innych neuronów (dendrytów) i wychodzi z niego jedno włókno (akson) rozgałęziające się do wielu neuronów. Elektrochemiczne sygnały wchodzące do neuronu są przetwarzane i przesyłane dalej lub ulegają wygaśnięciu. Neurony są połączone ze sobą

w sieć o skomplikowanej topologii, jednakże można określić wejścia do sieci (nerwy sensoryczne) oraz wyjścia z sieci (nerwy motoryczne). Rzeczywista sieć neuronowa jest w dużym stopniu odporna na uszkodzenia i ma zdolność do szybkiej nauki.

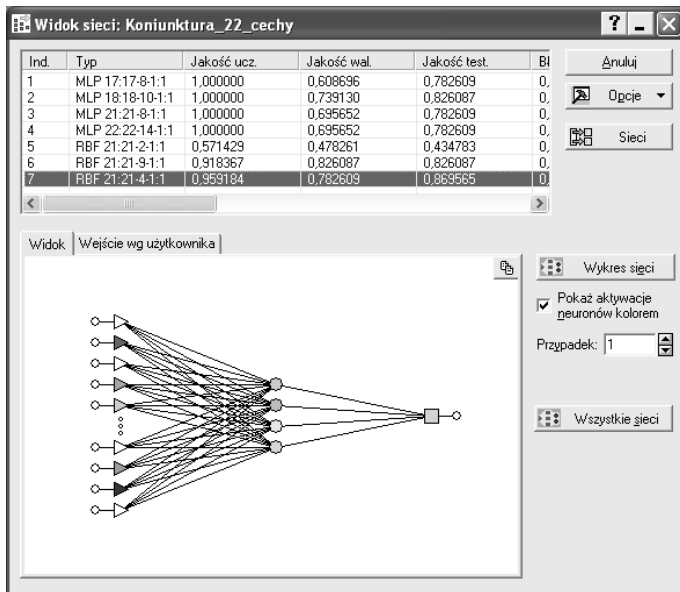
Sztuczne sieci neuronowe zostaną wykorzystane do sprawdzenia, jak przy ich pomocy można prognozować koniunkturę. W celu zapewnienia reprezentatywności uzyskanych wyników podzielono zbiór wszystkich obserwacji na trzy części. Pierwsza część składała się z obiektów uczących sieć. Druga grupa została wykorzystana do walidacji uczącej się sieci, czyli do określenia, kiedy sieć powinna przestać się uczyć, bo następuje jej przeuczenie. Trzeci zbiór obiektów ma na celu sprawdzenie, jak dokładnie nauczona sieć prognozuje koniunkturę obiektów niebiorących udziału w uczeniu i walidacji.

Najbardziej popularnym, a zarazem wykorzystanym w pracy modelem sztucznej sieci neuronowej jest perceptron wielowarstwowy (ang. *MultiLayer perceptron*) uczący się poprzez wsteczną propagację błędów. Sieć MLP składa się z jednej warstwy neuronów wejściowych, co najmniej jednej warstwy neuronów ukrytych oraz neuronu wyjściowego. A zatem kształt sieci MLP stanowi znaczne uproszczenie rzeczywistej sieci neuronowej. Neurony w kolejnych warstwach są połączone na zasadzie każdy z każdym. Natomiast sygnały wchodzące do neuronów są mnożone przez wagi dendrytów i sumowane. Na wyjściu następuje transformacja iloczynu skalarne go z wejścia przy pomocy funkcji przejścia (aktywacji). W badaniu stosowano logistyczną funkcję przejścia:

$$f(x) = \frac{1}{1 + e^{-x}}$$

Kształt funkcji $f(x)$ przypomina literę S, a sama funkcja przekształca wartość wejściową w liczbę z przedziału (0;1). Do najważniejszych cech dobrej funkcji przejścia należy łatwość w obliczaniu jej pochodnej, co ma duże znaczenie w procesie uczenia sieci. Uczenie przebiega w epokach. Podczas jednej epoki wprowadzane są do neuronów wejściowych wartości cech dla każdego obiektu i porównywana jest wartość neuronu wyjściowego z klasą obiektu uczącego sieć. Po każdej epoce następuje zamiana wag przy wejściach do neuronów na wagi najlepiej lokalnie poprawiające jakość klasyfikacyjną

całej sieci. Najczęściej stosowanym kryterium do zakończenia uczenia sieci jest spadek jakości klasyfikacyjnej na zbiorze walidacyjnym.



Rys. 5. Jakość sztucznych sieci neuronowych na zbiorach uczącym, walidacyjnym i testowym, na podstawie modułu *Automatyczny projektant*.

Drugim modelem sieci neuronowej wykorzystanym w niniejszej pracy jest sieć o radialnych funkcjach bazowych (ang. *Radial Basis Function*). Podstawowa różnica pomiędzy siecią MLP i RBF polega na innym sposobie działania neuronów. W sieci MLP wartości wejściowe są mnożone przez wagi, dodawane i przekształcane przez funkcję przejścia. W sieci RBF wagi związane z neuronem reprezentują współrzędne punktu. Interpretując wartości wejść do neuronu jako współrzędne drugiego punktu, można obliczyć odległość pomiędzy obydwooma punktami. Przyjmując, że punkt określony przez wagi jest centrum funkcji o wielowymiarowym rozkładzie normalnym, można dokonać nieliniowej transformacji odległości punktu wejściowego od centrum. W ten sposób otrzymuje się wyjście neuronu w sieci RBF. W rezultacie obszar decyzyjny zostaje podzielony przy pomocy okręgów (hipersfer), podczas gdy w sieciach MLP podział następuje przy pomocy prostych (płaszczyzn). Uczenie sieci RBF następuje z reguły szybciej niż analogicznej sieci MLP, jednakże działa ona wolniej i wymaga większych zasobów pamięci.

Do budowy sieci neuronowych wykorzystano *Automatycznego projektanta sieci* w programie *STATISTICA*. Do prezentacji wybrano siedem

sieci o najwyższej jakości klasyfikacji na zbiorze walidacyjnym, por. rys. 5. Sieci te zostały automatycznie wybrane przez program spośród sieci RBF i MLP. Wszystkie analizowane sieci składały się z jednej warstwy ukrytej, sieci RBF miały w niej nie więcej niż 24 neurony, a sieci MLP nie więcej niż 15 neuronów. Najwyższą dokładność na zbiorze testującym uzyskała sieć RBF z czterema neuronami w warstwie ukrytej. Najniższą dokładność miała sieć RBF z dwoma ukrytymi neuronami. Sieci MLP charakteryzują się wyższą stabilnością wyników niż sieci RBF. Dokładność sieci MLP była zbliżona do 4/5.

Wysoka jakość klasyfikacyjna zbudowanych sieci na zbiorach testujących pokazuje, że przy ich pomocy można z wysoką dokładnością prognozować stan koniunktury.

5 UWAGI KOŃCOWE

W pracy opisano, jak można modelować pojęcia gospodarcze poprzez wiadomości internetowe. W pierwszym kroku następuje pozyskanie danych tekstowych z Internetu i przetworzenie ich do postaci umożliwiającej ilościową analizę. W drugim kroku można skorzystać z technik opisu statystycznego do zrozumienia dostępnych danych, bądź bezpośrednio przejść do konstrukcji modeli prognostycznych.

Proponowany sposób działania został zobrazowany na przykładzie modelowania koniunktury gospodarczej z wykorzystaniem tytułów artykułów ekonomicznych dziennika „Rzeczpospolita” w latach 1997-2004. Jednakże postępując analogicznie, można spróbować zbadać inne pojęcia ekonomiczne (tj. bezrobocie, inflacja, kurs walutowy, indeks giełdowy).

Jednocześnie warto mieć na uwadze, że postrzeganie stanu koniunktury jako „dobra” lub „zła” jest podejściem klasyfikacyjnym. Jeżeli poprzez stan koniunktury będziemy rozumieć pewną liczbę na osi, to należałoby stosować podejście prognostyczne.

Przeprowadzone analizy statystyczne miały na celu odkrycie nowych relacji w danych oraz sprawdzenie, czy teksty są dobrym źródłem danych w modelowaniu koniunktury. Na wstępie spróbowano zmniejszyć liczbę słów opisujących miesiące. Ze względu na niskie współczynniki korelacji pomiędzy słowami, analiza czynnikowa nie pozwoliła zmniejszyć liczby cech. Z tego powodu zastosowano analizę wariancji

do wyznaczenia słów występujących istotnie częściej w okresach z jednym typem koniunktury. Przy 5% poziomie istotności znaleziono 22 słowa znacząco różnicowane przez koniunkturę. Bardzo ciekawe wyniki przedstawiają zbudowane drzewo klasyfikacyjne i funkcja dyskryminacyjna. Współczynniki stojące przy zmiennych w funkcji dyskryminacyjnej pokazują, w jaki sposób poszczególne słowa wpływają na koniunkturę. Natomiast utworzone drzewo klasyfikacyjne pozwoliło podzielić zbiór miesięcy ze względu na koniunkturę z dokładnością ponad 80% przy pomocy jedynie dwóch słów („ożywienie”, „rosnąć”). Następnie porównano jakość sieci neuronowych przy prognozowaniu koniunktury na zbiorze miesięcy niebiorącym udziału w uczeniu sieci. Średnia dokładność 7 sieci o najwyższej dokładności klasyfikacyjnej na zbiorze walidacyjnym wyniosła na zbiorze testującym ponad $\frac{3}{4}$, co wskazuje na potencjalne możliwości prognozowania koniunktury poprzez słowa z tytułów prasowych.

Barierą w rozwoju proponowanej procedury działania pozostaje fakt, że przekształcając zdania w nieuporządkowane zbiory słów, zachowujemy jedynie pewne głębokie odczucia związane z czytaniem tych słów, a tracimy praktycznie całą treść zdań.

BIBLIOGRAFIA

- 1) Burda M., Wyplosz C.,(2000): Makroekonomia – podręcznik europejski PWE, Warszawa.
- 2) Barczyk R.,(2004): Teoria i praktyka polityki antycyklicznej, AE, Poznań.
- 3) Cho V., Wutrich B., Zhang J., (1998): Text processing for classification, Technical report, The Hong Kong University of Science and Technology.
- 4) Gleason J.B., Ratner N.B., (2005): Psycholingwistyka, GWP, Gdańsk.
- 5) Kroha P., Baeza-Yates R.,(2004): Classification of Stock Exchange News, Technical Report, Engineering School, Universidad de Chile.
- 6) Lavrenko V., Schmill M., Lawrie D., Ogilvie P., Jensen D., Allan J.,(2000): Language models for financial news recommendation, Proceedings of the 9th ICIKM.
- 7) Norvag K., Oyri R.,(2005): News Item Extraction for Text Mining in Web Newspapers, Proceedings of International Workshop on Challenges in Web Information Retrieval and Integration.