PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 469 • 2017

Taksonomia 29
Klasyfikacja i analiza danych – teoria i zastosowania

ISSN 1899-3192
e-ISSN 2392-0041

Kamil Sapała, Marcin Piołun-Noyszewski, Marcin Weiss
Free Construction Sp. z o.o.
e-mail: data-science@freeconstruction.pl

PORÓWNANIE WYBRANYCH METOD STATYSTYCZNYCH I METOD SZTUCZNEJ INTELIGENCJI DO PRZEWIDYWANIA ZDARZEŃ W OPROGRAMOWANIU ZABEZPIECZAJĄCYM SYSTEMY PRZECHOWYWANIA DOKUMENTÓW CYFROWYCH, W TYM SYSTEMY KLASY *ENTERPRISE CONTENT MANAGEMENT*

A COMPARISON OF SOME STATISTICAL METHODS AND ARTIFICIAL INTELLIGENCE METHODS FOR PREDICTING EVENTS IN SOFTWARE PROTECTING DIGITAL DOCUMENTS REPOSITORIES, INCLUDING ENTERPRISE CONTENT MANAGEMENT

DOI: 10.15611/pn.2017.469.16
JEL Classification: C45, C53

Streszczenie: W ostatnich latach nastąpił wzrost zainteresowania wykorzystywaniem metod statystycznych do analizy zdarzeń z zakresu bezpieczeństwa teleinformatycznego. Coraz częściej moduły analityczne implementuje się w systemach chroniących przedsiębiorstwa przed zagrożeniami. Bardzo duże znaczenie ma w tej dziedzinie automatyzm i wykonywanie analiz bez nadzoru człowieka. W pracy opisane zostały efekty zastosowania działających automatycznie modułów eksperckich do przewidywania wartości szeregów czasowych, w sytuacji gdy nie były znane ich własności. Bez zastosowania właściwych metod przekształcenia szeregu i odpowiedniej parametryzacji tworzone modele mogą w wielu sytuacjach działać niepoprawnie. Natomiast w przypadku mających charakter cykliczny szeregów uzyskiwane prognozy mogą stanowić wartościową informację o potencjalnym zagrożeniu dla bezpieczeństwa przedsiębiorstwa.

Słowa kluczowe: sieci neuronowe, ARIMA, wyrównywanie wykładnicze, analiza w czasie rzeczywistym.

Summary: Recently statistical analysis of IT security events has been focusing more attention. Analytical modules have more often been implemented in systems protecting companies from security threats. In this field automation and analysis without human supervision are of great importance. The paper presents a performance of automatic expert modules applied to

predict time series, if its quantities were unknown. Created models without appropriate time series modification procedures and correct specification of parameters work only in a limited way. Nevertheless, the predictions of seasonal time series can provide valuable information about potential security threats to a company.

Keywords: neural networks, ARIMA, exponential smoothing, real-time analysis.

1. Wstęp

W ostatnich latach znacznie wzrosło zainteresowanie wykorzystywaniem metod predykcyjnych w rozwiązaniach biznesowych. Dotyczy to także dziedzin, w których dotychczas metody statystyczne nie miały szerokiego zastosowania. Sektorem, wykorzystującym w coraz większym stopniu techniki eksploracji i analizy danych jest bezpieczeństwo teleinformatyczne. Firmy tworzące oprogramowanie zabezpieczające przedsiębiorstwa przed zagrożeniami, chcąc udoskonalić swój produkt, poszukują nowych kierunków rozwoju. Jednym z nich jest implementacja modułów analitycznych pozwalających gromadzone informacje o zdarzeniach zachodzących w systemach bezpieczeństwa przedsiębiorstwa przetwarzać i analizować. Szczególne zainteresowanie wzbudzają w tej branży funkcjonujące bez nadzoru człowieka rozwiązania czasu rzeczywistego. Co do zasady modele statystyczne są uruchamiane w ten sposób, natomiast ich budowanie (dobór parametrów, transformacja danych) odbywa pod kontrolą człowieka. W praktyce oznacza to, że każda firma, która korzysta z systemu bezpieczeństwa, musi zatrudnić osobę posiadającą wiedzę specjalistyczną ze statystyki lub też korzystać z usług konsultanta spoza firmy [Cichowicz i in. 2012, s. 116]. Opisywane rozwiązanie jest z pewnością najbardziej pożądane, gwarantuje bowiem jakość analiz. Niestety, znaczna grupa przedsiębiorstw nie jest nim zainteresowana, na co wpływ ma szereg czynników, pośród których wymienić warto bezpieczeństwo danych i koszty, po pierwsze, związane z pracą analityka i po drugie, wynikające z konieczności dostosowania sytemu gromadzenia danych do potrzeb innych niż administratorzy użytkowników. Zdefiniowana została w ten sposób potrzeba konsumencka powstania systemu bezpieczeństwa teleinformatycznego, posiadającego moduły analityczne wykonywane na każdym etapie automatycznie i działającego u klientów samodzielnie. Autorzy tego opracowania postanowili sprawdzić, czy choćby w ograniczonym zakresie wykorzystanie dostępnych w wybranych programach statystycznych modułów eksperckich w połączeniu z własnymi kryteriami oceny rozwiązań może być odpowiedzią na potrzeby klientów. W tym celu zaprojektowano i zrealizowano badanie symulacyjne, w którym przewidywano wartości dwóch indeksów.

2. Cel badania

Jednoznaczne wskazanie, co nie było celem badania, pozwoli zrozumieć, co nim było. Nie była nim analiza danych w klasycznym rozumieniu, polegająca na wy-

konaniu przez analityka eksploracji danych (ich zrozumienia), dokonaniu właściwych transformacji, doborze optymalnych metod i ich parametrów, ocenie działania utworzonych modeli i wdrożeniu najlepszego. Już we wstępie zasygnalizowano, że w niniejszym opracowaniu zaprezentowane zostały wyniki eksperymentu, w którym opisane elementy procesu badawczego wykonywane przez człowieka zostały zastąpione przez ustawienia domyślne modułów eksperckich (bardzo zbliżone do dostępnych w wybranych pakietach statystycznych) i autorskie reguły/kryteria porównywania rozwiązań. Celem autorów było sprawdzenie, czy może to być użyteczna metoda przewidywania szeregów czasowych, w sytuacji gdy nie wiemy, jakie mogą mieć własności czy rozkład. Generowane w ten sposób prognozy będą z pewnością obarczone znacznie większym błędem, niż gdyby przygotował je doświadczony statystyk. Mogłyby jednak stanowić atrakcyjne dla klientów rozwiązanie w sytuacji, gdy nie ma możliwości nadzorowania przez człowieka procesu budowania modeli.

3. Metody badawcze i procedury automatyzujące prognozowanie

Wybór metod badawczych powinien być z całą pewnością następstwem poznania charakteru analizowanego zjawiska. W rozpatrywanej sytuacji konieczne stało się jednak odwrócenie kolejności i zaproponowanie metod bez praktycznie żadnej wiedzy o zjawisku.

Opracowana procedura zakłada wybór optymalnej w konkretnym przypadku metody prognostycznej spośród trzech uwzględnianych (sieci neuronowych typu perceptron, modeli ARIMA, modeli wyrównywania wykładniczego). O ile porównywanie prawidłowo skonstruowanej sieci neuronowej z modelami wyrównywania wykładniczego nie wydaje się zasadne – sieć pozwoli na dokładniejsze prognozy, o tyle w przypadku automatycznego tworzenia modeli tej pewności nie ma. Parametry wskazanych modeli dobierano automatycznie na podstawie częściowo zmodyfikowanych kryteriów wykorzystywanych w modułach eksperckich wybranych programów statystycznych[1]. W przypadku zintegrowanych autoregresyjnych modeli średniej ruchomej tworzono modele o zadanych wartościach (od 1 to k) parametrów określających: p – rząd autoregresji , d – rząd różnicy, q – rząd średniej ruchomej (opóźnienie zakłóceń losowych). Wybór optymalnego rozwiązania następował automatycznie. Dostępny zbiór dzielono na części, 10% najnowszych pomiarów stanowiło próbę testową, na której liczona była wartość średniego absolutnego błędu prognozy (MAE). Na podstawie tego kryterium wybierano też najlepiej dopasowany model wyrównywania wykładniczego spośród dostępnych niesezonowych (prostego, Holta, Browna, wygasającego) i sezonowych (prostego, addytywnego Wintersa, multiplikatywnego Wintersa).

[1] Tego typu modele eksperckie znajdują się w aplikacjach, takich jak np. SPSS czy R (funkcja auto.arima).

Dobór architektury sieci neuronowej typu wielowarstwowy perceptron następował na podstawie wskazanych ustawień. Wartości w szeregu sprowadzono do wspólnej skali [–1; 1] przy pomocy normalizacji[2] [Walesiak 2014, s. 365], stosowano trzy warstwy, w tym jedną ukrytą. Funkcją aktywacyjną dla warstwy wejściowej był tangens hiperboliczny, a dla warstwy wyjściowej funkcja liniowa. Liczbę neuronów w warstwie ukrytej określano przy pomocy mechanizmu usuwania neuronów do momentu zatrzymania istotnej poprawy jakości predykcji. Algorytmem optymalizującym wartości wag inicjacyjnych była propagacja wsteczna błędu, bazująca na minimalizacji sumy kwadratów błędu przy pomocy metody gradientowej (najszybszego spadku)[3]. Początkowo wagi losowano z przedziału [–0,5; 0,5], stosując przy tym mechanizm „symulowanego wyżarzania"[4]. Tworzono i oceniano modele uwzględniające liczbę opóźnień ze wskazanego przedziału (od 1 do k)[5].

4. Wyniki badania

W celu przetestowania funkcjonowania procedur automatyzujących prognozowanie przeprowadzono eksperyment, w którym zastosowano je do gromadzonych danych – dwóch indeksów, o których własnościach wiedza była bardzo ograniczona. Znany był jedynie przedział wartości, jakie mogły przyjmować zmienne, i interwał, w jakim będą trafiały do bazy. Kolejny pomiar pojawiał się w bazie w przypadku pierwszego indeksu co minutę, natomiast w przypadku drugiego co sekundę. Zadecydowano, aby tworzenie modeli predykcyjnych rozpoczęło się po 2 dwóch dniach roboczych pracy systemu, tzn. 16 godzinach. Oznacza to, że w momencie uruchomienia procesu budowania modeli w pierwszej bazie znajdowało się 960 odczytów, natomiast w drugiej 57 600.

Zbiór testowy liczył w przypadku pierwszego indeksu 96 obserwacji. Modele oceniane były na podstawie błędów prognoz o horyzoncie 1 obserwacji. W tym celu obliczano dla poszczególnych modeli autoregresyjnych średniej ruchomej, sieci neuronowych, wyrównywania wykładniczego wartości średniego absolutnego błędu prognozy[6]:

$$MAE = \frac{1}{r}\sum_{t=1}^{r}\left|y_t - y_{tp}\right|. \qquad (1)$$

[2] Znana była minimalna i maksymalna wartość, jaką przyjąć mogły oba indeksy, dlatego też zadecydowano o takiej metodzie sprowadzenia zmiennych do porównywalności.

[3] Opis metody [Morajda 2005, s. 95].

[4] Opis metody [Kowalik 2014, s. 218-220].

[5] W przypadku tego eksperymentu liczba ta wynosiła 1/20 obiektów znajdujących się w bazie danych.

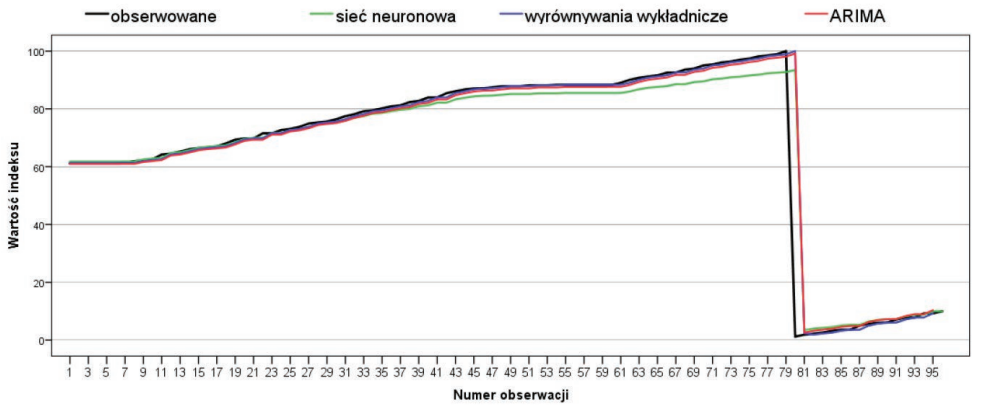
[6] Wzór podaję za Marią Szmukstą-Zawadzką i Janem Zawadzkim [2012, s. 213].

W przypadku modeli ARIMA dokonywano wyboru optymalnego spośród 720[7], w przypadku sieci neuronowych 48[8], natomiast w przypadku modeli wyrównywania wykładniczego 145[9]. Zamieszczenie w artykule tabel z wartościami średniego absolutnego błędu dla wszystkich modeli byłoby niemożliwe, dlatego też zamieszczono jedynie błędy dla najlepszych modeli (tabela 1). Prognozy uzyskiwane przy ich pomocy w horyzoncie 1 obserwacji znajdują się na rysunku 1.

Tabela 1. Średni absolutny błąd – wynik uzyskano prognozując pierwszy indeks

Metoda	Średni absolutny błąd
ARIMA (1, 0, 0)	1,43
Prosty model wyrównywania wykładniczego	1,54
Sieć neuronowa typu perceptron (rząd opóźnień = 27)	1,67

Źródło: opracowanie własne.



Rys. 1. Pierwszy indeks – wartości przewidywane i obserwowane w zbiorze testowym

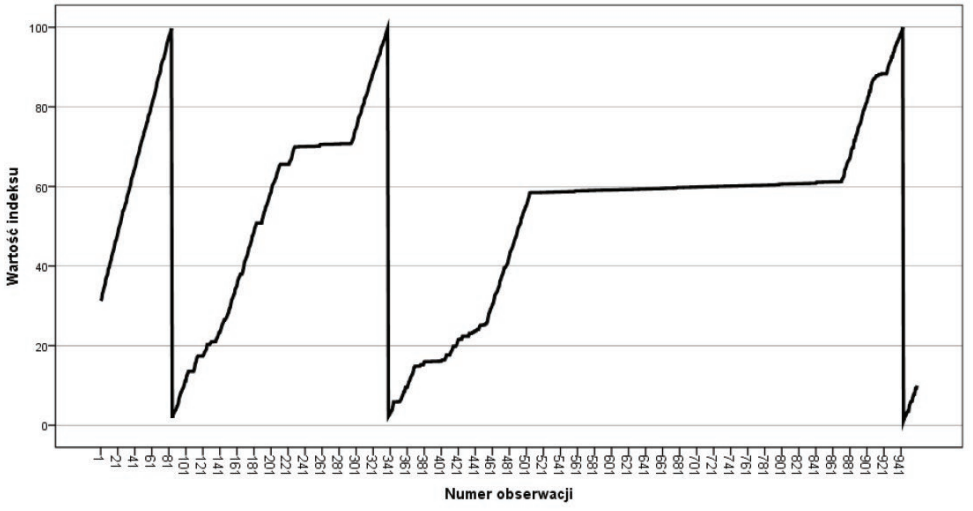
Źródło: opracowanie własne.

Na rysunku 2 zamieszczono cały szereg, który starano się przewidywać. Charakteryzuje się on występowaniem wielu niezależnych od siebie przebiegów. Jak widać, automatyzacja procesu porównywania i wyboru modeli bez zastosowania właściwych w konkretnym przypadku metod przekształcenia szeregu i poprawnej parametryzacji nie prowadzi do uzyskiwania dokładnych prognoz. Przy pomocy znacznie

[7] Porównywano modele o parametrach: p – od 1 do 48, d – od 0 do 2, q – od 0 do 5.

[8] Co zostało już wcześniej wskazane, jedynym zamienianym parametrem sieci był uwzględniany rząd opóźnień.

[9] Cztery modele niesezonowe (prosty, Holta, Browna, wygasający) i trzy sezonowe (prosty, addytywny Wintersa, multiplikatywny Wintersa), w których cykl wynosił od 2 do 48 obserwacji.



Rys. 2. Pierwszy indeks – wszystkie dostępne pomiary

Źródło: opracowanie własne.

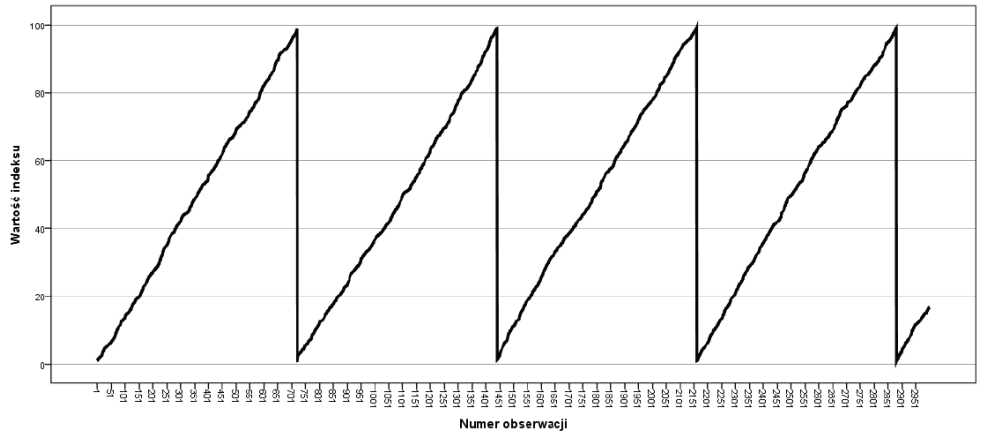
prostszej metody wyrównywania wykładniczego uzyskiwano mniejszy błąd prognozy niż po zastosowaniu sieci neuronowej typu perceptron. W rozpatrywanym przypadku lepsze rezultaty uzyskano by, gdyby skomplikowany schemat wyboru modeli zastąpić średnią kroczącą *k*-elementową obliczaną na przekształconym do różnicy 1 rzędu szeregu.

W przypadku drugiego indeksu część testowa liczyła 5760 obserwacji. Rozpatrywana zmienna miała charakter jednostajnego wzrostu w określonym cyklu, w związku z czym utworzone automatycznie modele pozwalały na dość dobre przewidywania zarówno w krótkim, jak i długim horyzoncie. Szereg ten można by przekształcić, eliminując występowanie cyklu, co z pewnością poprawiłoby dokładność prognoz, należy jednak zauważyć, że nawet w tej niedoskonałej formie predykcje mogą być użyteczne dla użytkowników systemu bezpieczeństwa, znacząca różnica (np. przekraczająca dwukrotnie średni błąd) pomiędzy wartością prognozowaną a obserwowaną mogłaby wskazywać nietypowe zdarzenie – potencjalne zagrożenie dla bezpieczeństwa przedsiębiorstwa.

Tabela 2. Średni absolutny błąd – wynik uzyskano prognozując drugi indeks

Metoda	Średni absolutny błąd
ARIMA (720, 0, 0)	2,52
Prosty model sezonowy wyrównywania wykładniczego	0,88
Sieć neuronowa typu perceptron (rząd opóźnień = 720)	1,5

Źródło: opracowanie własne.



Rys. 3. Drugi indeks

Źródło: opracowanie własne.

5. Zakończenie

Wykorzystywane mechanizmy automatyzujące proces prognozowania mogą być użyteczne w przypadku szeregów cyklicznych. W sytuacji gdy nie ma możliwości kontrolowania przez człowieka procesu analizy danych, pozwalają trafnie wskazać liczbę pomiarów składających się na cykl, umożliwiają też wykrywanie nietypowych wartości na podstawie różnic pomiędzy wartościami obserwowanymi i przewidywanymi. Należy jednak zauważyć, że wykonywane bez właściwych przekształceń i parametryzacji modeli analizy są niedoskonałe, dokładniejsze prognozy uzyskiwane są przy pomocy prostszych metod (wyrównywania wykładniczego), niewymagających skomplikowanej parametryzacji. Aby tego typu przewidywanie bez nadzoru człowieka miało charakter uniwersalny, mogło być efektywnie wykorzystywane do szeregów o różnych własnościach, konieczne wydaje się zautomatyzowanie procesu rozpoznawania charakterystyki szeregów i utworzenie reguł określających właściwy sposób transformacji w określonych przypadkach.

Literatura

Cichowicz T., Frankiewicz M., Rytwiński F., Wasilewski J., Zakrzewicz M., 2012, *Odkrywanie anomalii w szeregach czasowych pochodzących z monitoringu systemów teleinformatycznych*, Zeszyty Naukowe Wyższej Szkoły Bankowej w Poznaniu, nr 40, s. 115-130.

Kowalik S., 2014, *O Symulowane wyżarzanie w zastosowaniu do wyznaczania ekstremum globalnego funkcji o wielu ekstremach lokalnych daleko oddalonych od siebie lub bardzo zagęszczonych*, [w:]

Jastriebow A., Wowra K. (red.), *Współczesne technologie informatyczne i ich zastosowanie w teorii i praktyce*, Wydawnictwo Politechniki Radomskiej, Radom, s. 217-228.

Morajda J., 2005, *Sieci neuronowe i ich wykorzystanie w analizie danych ekonomicznych na przykładzie prognozowania sprzedaży energii elektrycznej*, Zeszyty Naukowe MWSE w Tarnowie, zeszyt 7, s. 87-100.

Stefanowski J., 2017, *Analiza szeregów czasowych*, http://www.cs.put.poznan.pl/jstefanowski/aed/TPtimeseries.pdf (1.03.2017).

Stopczyk M., 2005, *Symulowane wyżarzanie jako przykład algorytmu optymalizacji stochastycznej,* Mikroelektronika i Informatyka: Prace Naukowe, t. Z, nr 5, s. 139-142.

Szmuksta-Zawadzka M., Zawadzki J., 2012, *O miernikach dokładności prognoz ex post w prognozowaniu zmiennych o silnym natężeniu sezonowości*, Metody Ilościowe w Badaniach Ekonomicznych, t. 13, nr 1, s. 212-223.

Walesiak M., 2014, *Przegląd formuł normalizacji wartości zmiennych oraz ich własności w statystycznej analizie wielowymiarowej*, Przegląd Statystyczny, t. 61, nr 4, s. 363-372.

Wywiał J., Żądło T., 2008, *Prognozowanie szeregów czasowych za pomocą pakietu SPSS*, SPSS Polska.