Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663
www.czasopisma.uni.lodz.pl/foe/
4(337) 2018

DOI: http://dx.doi.org/10.18778/0208-6018.337.03

Kamil Sapała
Free Construction Sp. z o.o., kamil.sapala@freeconstruction.pl

Marcin Piołun-Noyszewski
Free Construction Sp. z o.o., marcin.noyszewski@freeconstruction.pl

Marcin Weiss
Free Construction Sp. z o.o., marcin.weiss@freeconstruction.pl

Automatyczne wykrywanie zagrożeń w systemach teleinformatycznych za pomocą metod Data Mining

Streszczenie: W pracy przedstawiono wybrane rozwiązania analityczne stosowane w czasie rzeczywistym w autorskim systemie bezpieczeństwa teleinformatycznego. Opisano metody wykorzystywane w celu wykrywania (bez nadzoru człowieka) potencjalnych zagrożeń – niepożądanych zdarzeń systemowych i zachowań użytkowników korzystających z dokumentów cyfrowych. Przestawiono procedury automatyzujące, stosowane w przypadku danych mających postać szeregów czasowych i dokumentów tekstowych. Poddane eksploracji dane pochodziły z testowego funkcjonowania oprogramowania zabezpieczającego systemy przechowywania dokumentów cyfrowych utworzonego przez firmę Free Construction.

Słowa kluczowe: systemy teleinformatyczne, dokumenty cyfrowe, przekształcenia, metody eksploracji danych

JEL: C02

1. Wprowadzenie

Bezpieczeństwo teleinformatyczne jest dziedziną, w której firmy – mimo gromadzenia ogromnych ilości informacji – nie wykorzystywały do analizy (lub robiły to w bardzo ograniczonym stopniu) metod Data Mining (więcej w Sapała, Piołun-Noyszewski, Weiss, 2017: 160). Było to spowodowane wieloma czynnikami, pośród których w pierwszej kolejności należy wymienić ograniczenia technologiczne i wielopoziomową, rozbudowaną strukturę gromadzonych informacji. Duże znaczenie miał również fakt, że dostępne aplikacje monitorujące systemy teleinformatyczne wymagały złożonej wstępnej konfiguracji i nieustannej asysty eksperckiej. Powiadomienia o możliwym zagrożeniu bezpieczeństwa miały ograniczony charakter, informując jedynie o przekroczeniu wskazanych wartości krytycznych rozpatrywanych indeksów (Cichowicz i wsp., 2012: 115–118). Przechowywanie danych historycznych w celu ich późniejszej pogłębionej analizy może pozwolić na zrozumienie przyczyn zachodzących zdarzeń, nie umożliwia jednak natychmiastowej reakcji. Niedoskonałość opisywanego narzędzia sprawiła, że podjęto prace zmierzające do stworzenia oprogramowania zabezpieczającego przedsiębiorstwa, działającego w pełni automatycznie i bez nadzoru człowieka. Zaimplementowano w nim wiele domyślnych metod przygotowywania danych i ich analizy (m.in. pobierania, grupowania, klasyfikacji, predykcji), uruchamianych na podstawie rozbudowanych zestawów reguł. Kluczowe znaczenie dla szerszego zastosowania metod Data Mining i ich praktycznej użyteczności w systemach działających w czasie rzeczywistym miało zaprojektowanie odpowiednich przekształceń i kryteriów oceny skuteczności ich zastosowania w danym przypadku. W dalszej części artykułu przedstawione zostały wybrane mechanizmy automatyzujące proces przekształcania danych i wykrywania zagrożeń w systemach IT.

2. Cel artykułu

Nadrzędnym celem implementowania modułów analitycznych w systemach bezpieczeństwa jest identyfikacja zdarzeń mogących zagrażać bezpieczeństwu firmy. Chodzi tu zarówno o wykrywanie niepożądanych zachowań indywidualnych (poszczególnych pracowników), jak i zdarzeń zbiorowych (grup pracowników).

Artykuł przedstawia studium przypadku autorskiego rozwiązania automatyzującego proces przekształcania danych i wykrywania zagrożeń w systemach IT. Skoncentrowano się na dwóch typach rozwiązań – detekcji odstępstw (*anomaly detection*) w aktywności pracowników w czasie oraz w zakresie tworzenia treści dokumentów cyfrowych. Przedstawione analizy są przykładowe, przez co nie dają pełnego obrazu możliwości zaproponowanego rozwiązania, jednakże pozostają dobrym przybliżeniem przebiegu i automatyzacji procesu analitycz-

nego w systemach bezpieczeństwa. Zaprezentowane autorskie rozwiązanie jest elastyczne: umożliwia zarówno analizy ilościowe (jaką jest analiza szeregów czasowych oraz detekcja odstępstw), jak i opierające się na bardziej jakościowej, dogłębnej analizie tekstu.

3. Analizowane dane

Analizy przeprowadzono na przykładzie danych zebranych podczas pilotażowego funkcjonowania systemu bezpieczeństwa u klienta. Analizowano zarówno aktywność pracowników, jak i treść tworzonych przez nich dokumentów. Dane dotyczące zdefiniowanych aktywności zbierane były co sekundę przez dwa miesiące. Analizie poddano bazę aktywności liczącą 604 800 obserwacji.

W przypadku dokumentów cyfrowych zgromadzono 680 dokumentów tekstowych w języku angielskim utworzonych w działach księgowości i zarządzania zasobami ludzkimi.

4. Automatyzacja procesu przekształcania danych

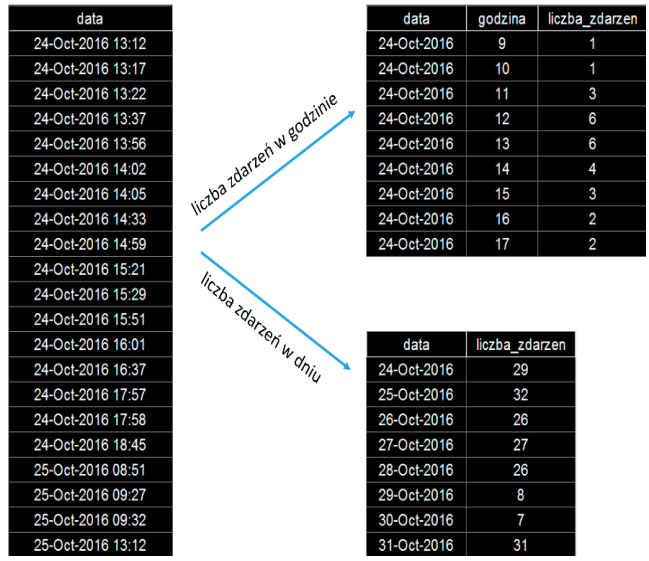
Przed uruchomieniem właściwego modułu analiz, czyli np. klasyfikacji czy predykcji, dane wymagają właściwego przygotowania, co może oznaczać konieczność wykonania transformacji polegającej na przykład na konwersji typu danych, kategoryzacji, normalizacji czy agregacji. Wyzwaniem w procesie automatyzacji jest dobór właściwych metod transformacji bez nadzoru człowieka. To, jak należy przekształcić konkretną krotkę danych czy zmienną, zależy od jej ustalonej początkowo roli, wartości odpowiednich statystyk wskazujących optymalne w danej sytuacji rozwiązanie oraz zestaw zdefiniowanych reguł. W przypadku analizy aktywności pracowników w czasie automatyzacja polega na przekształceniu danych do postaci szeregu czasowego, a następnie na:

1) wyborze sposobu agregacji jednostki czasu w celu zidentyfikowania regularności,
2) ustalaniu okresowości oraz
3) porównywaniu modeli szeregów czasowych w celu wyboru najlepiej dopasowanego do danych.

W artykule opisane zostały cztery mechanizmy poprzedzające prognozowanie wartości konkretnego indeksu mierzącego aktywność pracowników badanej firmy. Scharakteryzowano sposób przekształcania danych niemających początkowo formy szeregu czasowego do takiej właśnie postaci, agregowania danych w celu zidentyfikowania regularności, ustalania okresowości i porównywania proponowanych rozwiązań w celu wyboru najwłaściwszego.

4.1. Przekształcanie danych do postaci szeregów czasowych

Znaczna część gromadzonych przez systemy bezpieczeństwa informacji ma formę szeregu czasowego, wartość mierzonych atrybutów odczytywana jest w równym interwale czasowym. Jest jednak grupa informacji, które zapisywane są w bazie w momencie ich wystąpienia (zdarzeniowo) za pomocą zmiennej czasu wystąpienia (*timestamp*). W takiej sytuacji pierwszym etapem przygotowania danych jest przekształcenie polegające na zsumowaniu liczby zdarzeń we wskazanych jednostkach czasu, np. godzinie, dniu, tygodniu, miesiącu. Przykładowy rezultat przekształcenia danych wejściowych do szeregów czasowych zamieszczono na rysunku 1.



Rysunek 1. Schemat przekształcania zmiennej zawierającej datę wystąpienia zdarzeń do postaci szeregów czasowych

Źródło: opracowanie własne

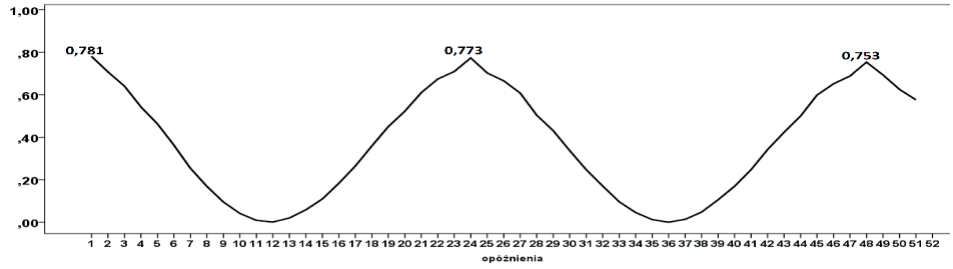
4.2. Odkrywanie regularności występowania zdarzeń

W przypadku wielu monitorowanych zmiennych standardowe interwały czasowe[1] agregacji nie umożliwiają wykrycia występującej, lecz prawie niedostrzegalnej powtarzalności. Konieczne stało się zatem zaprojektowanie bardziej elastycznych,

[1] W tym przypadku za standardową uznaje się agregację na poziomie godzin, dni, tygodni lub miesięcy.

w pełni automatycznych reguł, umożliwiających wybór optymalnej jednostki czasowej, wedle której powinna następować agregacja. Wybrana na potrzeby tego artykułu zmienna mierzyła co sekundę liczbę występujących w firmie zdefiniowanych aktywności użytkowników. W omawianym przykładzie zsumowano występującą liczbę zdarzeń co 1, 2, 3, 4, 5, 6, 8, 10, 12, 15 minut. Kolejny krok zakładał określenie liczby pomiarów składających się na cykl w każdym przypadku. Początkowo, wzorując się na dostępnych w niektórych programach analitycznych modelach eksperckich[2], planowano wykorzystać w tym celu statystyki dopasowania (stacjonarny R-kwadrat, znormalizowane Bayesowskie kryterium informacyjne) tworzone automatycznie dla autoregresyjnych zintegrowanych modeli średniej ruchomej (ARIMA)[3].

W trakcie testów udowodniono jednak zawodność tego podejścia. Na zamieszczonych poniżej wykresach (rys. 2, 3) przedstawiono wartości wskazanych statystyk w przypadku agregacji danych co 10 minut. Wyraźnie widać, że kierowanie się wartościami statystyk dopasowania i błędów *ex post* (rys. 4) może prowadzić do sprzecznych wniosków. Miary dopasowania sugerują, że optymalnym rozwiązaniem jest model o rzędzie autoregresji równym 1, natomiast wartości błędów *ex post* – o rzędzie autoregresji równym 48. W związku z odnotowaną dokładnością prognoz uzyskiwanych za pomocą modeli autoregresyjnych w tym konkretnym przypadku (rys. 5) zadecydowano, aby ustalanie okresowości odbywało się w przyszłości na podstawie wartości błędów prognoz *ex post*, ponieważ kryterium to wydaje się bardziej uniwersalne.

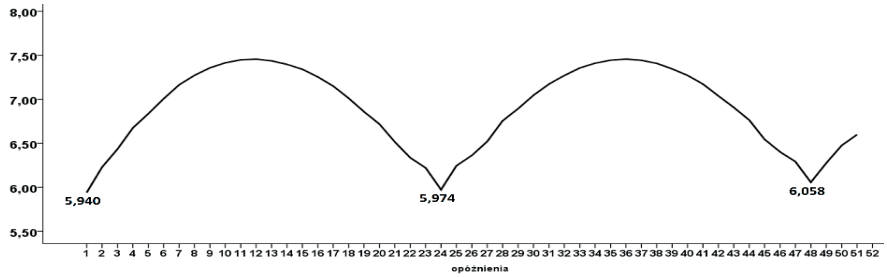


Rysunek 2. Wartość statystyki dopasowania (stacjonarnego R-kwadrat, oś OY) tworzonych modeli autoregresyjnych w zależności od parametru określającego rząd autoregresji (opóźnienie, oś OX)

Źródło: opracowanie własne

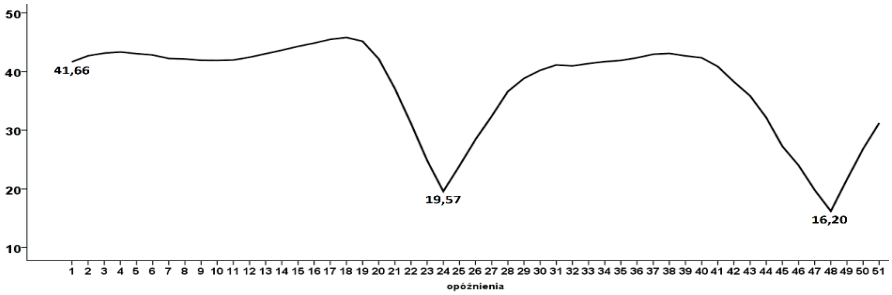
[2] Tryb ekspercki dobierający w sposób automatyczny parametry (p – rząd autoregresji, d – rząd różnic, q – rząd średniej ruchomej) modelu ARIMA posiada m.in. program IBM SPSS lub też pakiet *forecast* działający w środowisku R (więcej w Hyndman, Khandakar, 2007).

[3] Testowana liczba opóźnień (rząd autoregresji) była uzależniona od liczby obserwacji. Ocenie poddawano rozwiązania, w których rząd autoregresji wynosił od 1 do x, gdzie x stanowiło 1/20 liczby obiektów w bazie.



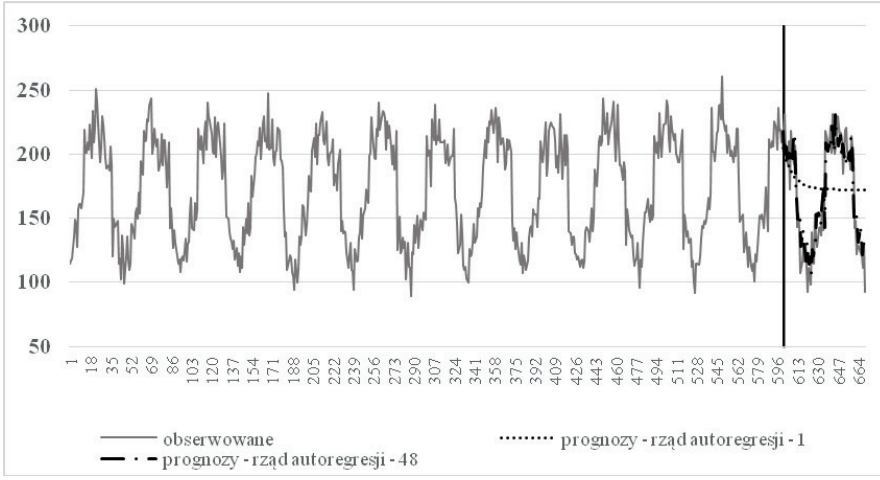
Rysunek 3. Wartość statystyki dopasowania (znormalizowanego Bayesowskiego kryterium informacyjnego, oś OY) tworzonych modeli autoregresyjnych w zależności od parametru określającego rząd autoregresji (opóźnienie, OX)

Źródło: opracowanie własne



Rysunek 4. Wartość błędu prognozy *ex post* (pierwiastka błędu średniokwadratowego, oś OY) tworzonych modeli autoregresyjnych w zależności od parametru określającego rząd autoregresji (opóźnienie, OX)

Źródło: opracowanie własne



Rysunek 5. Wartości prognozowane za pomocą modeli, w których rząd autoregresji ustalono na 1 i 48

Źródło: opracowanie własne

4.3. Kryteria wyboru optymalnego rozwiązania

Zastosowanie opisanej procedury umożliwiło agregację danych we wskazanych jednostkach czasowych i ustalenie liczby pomiarów składających się na cykl w każdym zaproponowanym podziale. Kolejnym wykonywanym automatycznie krokiem było dokonanie oceny proponowanych rozwiązań i wybór optymalnego. Co zrozumiałe, im większa jest jednostka czasowa, w której następuje agregacja, tym lepszą statystykę dopasowania (stacjonarny R-kwadrat) uzyskują tworzone modele autoregresyjne. Opisaną zależność można zaobserwować w tabeli 1. Wykorzystany wskaźnik dopasowania zaimplementowany w programie SPSS – stacjonarny R kwadrat (Harvey, 1990: 268) – jest odmianą klasycznego wskaźnika dobroci dopasowania dla obserwacji po różnicowaniu pierwszego rzędu. Wskaźnik można zapisać za pomocą wzoru (1):

$$R_D^2 = 1 - \frac{SSE}{\sum_{t=2}^{T} \overline{\Delta y_t - (\Delta y)}^2}, \tag{1}$$

gdzie: *SSE* – suma kwadratów błędów (*residual sum of squares*), $\overline{\Delta y}$ – średnia obserwacji po różnicowaniu pierwszego rzędu.

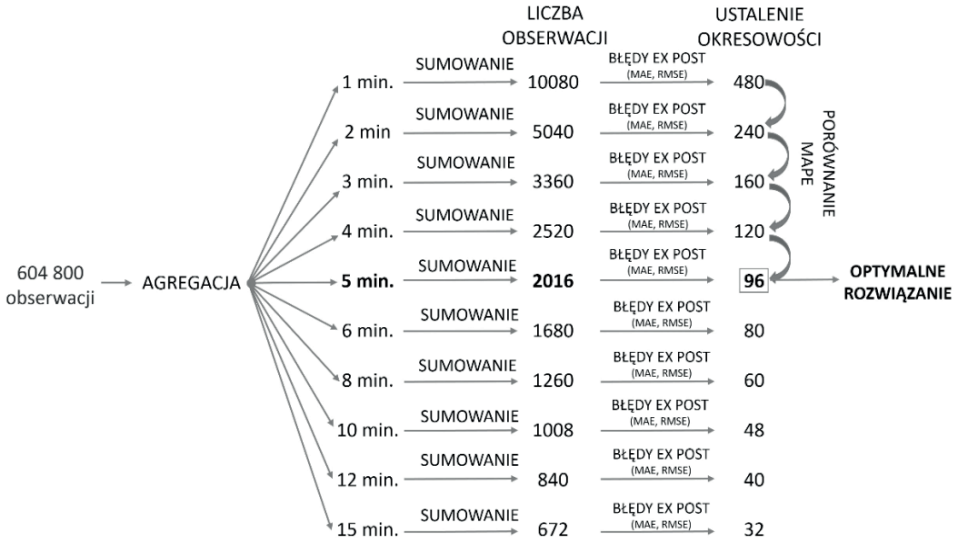
Tabela 1. Wartość statystyki dopasowania (stacjonarnego R-kwadrat) utworzonych modeli autoregresyjnych w zależności od jednostki czasowej, w której zagregowano dane wejściowe

Proponowane rozwiązanie numer	Jednostka agregacji (w minutach)	Liczba obserwacji w cyklu	Stacjonarny R-kwadrat
1	1	480	0,027
2	2	240	0,276
3	3	160	0,402
4	4	120	0,539
5	5	96	0,610
6	6	80	0,664
7	8	60	0,731
8	10	48	0,775
9	12	40	0,819
10	15	32	0,852

Źródło: opracowanie własne

Kluczowe znaczenie mają zatem reguły zatrzymania. W rozpatrywanym przypadku mechanizm zakładał posługiwanie się średnim procentowym błędem (*mean absolute percentage error* – MAPE) prognoz *ex post*. Jeśli w kolejnych dwóch krokach nie zmniejszał się łącznie o 2%, to konkretne rozwiązanie zostało uznane za najlepsze. Na podstawie tego kryterium w opisywanym przykładzie uznano, że liczba występujących zdarzeń powinna być sumowana co 5 minut.

Na rysunku 6 zaprezentowano podsumowanie analiz w postaci schematu ilustrującego automatyczny proces przygotowywania danych.



Rysunek 6. Schemat przedstawiający mechanizm wykorzystywany do automatycznego przekształcania zmiennych wymagających przed prognozowaniem agregacji

Źródło: opracowanie własne

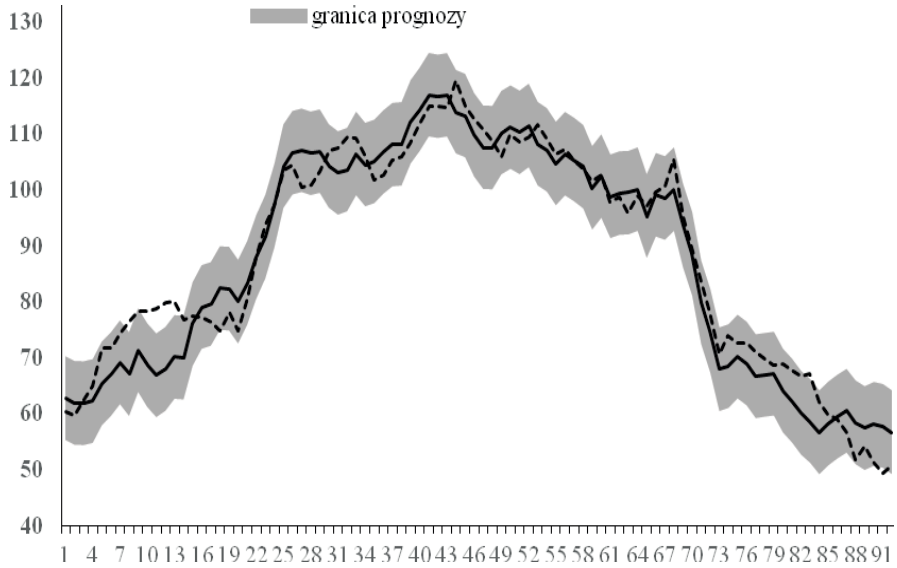
4.4. Prognozowanie odpowiednio przekształconych szeregów czasowych

Po dokonaniu odpowiednich transformacji w zależności od problemu badawczego i roli konkretnej zmiennej przeprowadzana jest analiza właściwa. W opisywanym przykładzie, po dokonaniu niezbędnych przekształceń, rozpatrywany atrybut posłużył do predykcji i identyfikacji potencjalnych zagrożeń na podstawie wykraczających poza normę błędów prognozy. Oprócz modelu autoregresyjnego weryfikowano dokładność uzyskiwaną za pomocą alternatywnej metody – sieci neuronowych typu wielowarstwowy perceptron. W literaturze (np. Zhang, 2003; Okasha, Yaseen, 2013; Mitrea, Lee, Wu, 2009) wskazuje się, że sieci neuronowe mogą być obiecującą alternatywą dla tradycyjnych modeli szeregów czasowych (gdzie parametry szacuje się metodą najmniejszych kwadratów), szczególnie w przypadku opisywania złożonych lub nieliniowych relacji w czasie.

Lepsza[4] spośród wskazanych metod stosowana jest do przewidywania wartości zmiennej i określenia przedziałów prognozy. Wykroczenie poza tak ustalone

[4] Lepsza metoda to taka, za pomocą której uzyskiwane są mniejsze błędy prognoz *ex post*.

granice będzie przez system bezpieczeństwa sygnalizowane. Budowany na podstawie domyślnych ustawień[5] model sieci neuronowych umożliwiał w rozpatrywanym przypadku dokładniejsze przewidywania, dlatego też za pomocą tej właśnie metody prognozowano wartość zmiennej. Na rysunku 7 ciągłą linią oznaczono wartości prognozowane wraz z przedziałem ufności, a przerywaną wartości rzeczywiste. Przekroczenie wskazanych granic stanowi informację o potencjalnym zagrożeniu dla bezpieczeństwa firmy. Generowane automatycznie przez system bezpieczeństwa powiadomienie powinno zawierać nie tylko informacje o wystąpieniu odstępstwa, ale także o jego skali[6], w celu opisania poziomu zagrożenia i podjęcia decyzji o uruchamianych procedurach alarmowych i ochronnych. Przykładowy komunikat mógłby mieć następującą postać: „W dniu 10.11.2016 wykryto nietypowe zdarzenie, indeks X przyjął o godzinie 11:30 nietypową wartość, stan ten utrzymywał się przez 5 kolejnych pomiarów, było to przeciętne odstępstwo".



Rysunek 7. Wartości obserwowane i przewidywane za pomocą modelu sieci neuronowych

Źródło: opracowanie własne

[5] Domyślnie ustawione parametry sieci zakładają standaryzację danych, trzy warstwy, funkcję aktywacji warstwy: wejściowej – tangens hiperboliczny, wyjściowej – liniowa, zastosowanie propagacji wstecznej błędu jako algorytmu optymalizacji wartości wag synaptycznych, liczbę neuronów w warstwie ukrytej ustalaną przy użyciu krokowej metody ich pomniejszania (zaczynając od liczby o 1 mniejszej niż w warstwie wejściowej) do momentu, w którym nie nastąpi istotna poprawa predykcji.

[6] Do określenia skali odstępstwa wykorzystać można takie statystyki, jak na przykład średni bezwzględny błąd (MAE), którego wartości mogą wyznaczać stopień anomalii. W opisywanym przykładzie: norma – do 1, nieznaczne odstępstwo – 1–1,5, przeciętne odstępstwo – 1,5–2, duże odstępstwo – 2–3, bardzo duże odstępstwo – 3–5, skrajnie duże odstępstwo – powyżej 5.

5. Klasyfikacja gromadzonych dokumentów tekstowych

W pierwszym rozpatrywanym przykładzie konieczne było zaprojektowanie własnych reguł automatyzujących proces przygotowania danych do analizy. W przypadku niektórych zadań można skorzystać z dostępnych algorytmów zaimplementowanych w standardowych pakietach statystycznych. Przykładem takiej analizy może być kontrola nowo powstałych, cyfrowych dokumentów tekstowych wraz z identyfikacją przypadków, kiedy treść jest nietypowa dla działu firmy, w którym on powstał. Integracja informacji o autorze dokumentu i występujących w tekście słowach może pomóc w wykryciu niestandardowych zachowań użytkowników.

W czasie próbnego funkcjonowania systemu u klienta zgromadzono 680 dokumentów tekstowych w języku angielskim, utworzonych w działach księgowości i zarządzania zasobami ludzkimi. Proces transformacji zakładał przeprowadzenie:

1) tokenizacji – przekształcenia występujących w dokumentach ciągów nieuporządkowanych znaków w ustrukturyzowany zbiór elementów; tekst dzielony jest po zdefiniowanych separatorach (spacja, myślnik, nawias, przecinek itp.); na tym etapie zamienia się też wszystkie wielkie litery na małe, czyści tekst z pozostałych znaków interpunkcyjnych i specjalnych oraz liczb (Lula, Wójcik, Tuchowski, 2016: 153–164);
2) filtrowania – usuwania z dokumentów słów funkcyjnych ze stop listy (*stop words*); wyrazy te poza funkcją gramatyczną nie niosą informacji o treści tekstu, więc w analizie dokumentów są zbędne; przykładem takich słów mogą być m.in. spójniki (*i, oraz, lub, ani, więc, aczkolwiek* itp.);
3) stemmingu – usunięcia końcówek fleksyjnych ze słów i pozostawienia tylko ich tematów (*stem*).

W najprostszym podejściu podstawową analizowaną jednostką jest słowo (unigram), natomiast w pewnych sytuacjach bardziej użyteczna może być wiedza o występowaniu w dokumencie danego zwrotu składającego się z dwóch (bigramy) lub trzech (trigramy) słów. Bardzo rzadko podstawową jednostką analizy jest zwrot składający się z więcej niż trzech wyrazów. W rozpatrywanym przykładzie przedmiotem analizy były pojedyncze słowa. Po przeprowadzeniu wstępnych transformacji należy wybrać metodę przeliczania wystąpień słów w dokumencie na konkretne wartości numeryczne, które umieszczone zostaną w macierzy częstości[7]. Najprostsza metoda polega na kodowaniu binar-

[7] Szczegółowy opis metod wykorzystywanych do transformacji macierzy częstości występowania słów w dokumentach można znaleźć w pracach Pawła Luli (2005) i Marcina Mirończuka (2012).

nym. Wystąpienie konkretnego słowa w danym dokumencie przyporządkowuje mu wartość 1. W rezultacie taką samą wartość otrzymuje dokument, w którym słowo pojawia się raz i dokument, w którym pojawia się wielokrotnie. Dlatego też w rozpatrywanym przypadku posługiwano się indeksem TF-IDF, w którym słowom nadawane są wagi w zależności od liczby ich wystąpień. TF-IDF (2) powstaje jako iloczyn indeksu TF (*term frequency*) i IDF (*inverse document frequency*) – zob. wzory (3) i (4):

$$(TF - IDF)_{i,j} = TF_{i,j} \times IDF_i, \tag{2}$$

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{3}$$

$$IDF_i = \log \frac{|D|}{|\{d : t_i \in d\}|}, \tag{4}$$

gdzie:
$ni_{,}j$ – liczba wystąpień konkretnego słowa w danym dokumencie,

$\sum_k n_{k,j}$ – liczba wystąpień wszystkich słów w danym dokumencie,
$|D|$ – liczba wszystkich dokumentów,
$|\{d : ti \in d\}|$ – liczba dokumentów, w których co najmniej raz występuje dane słowo.

Odpowiednio przygotowaną bazę wykorzystano do klasyfikacji dokumentów tekstowych do działów, w których powinny powstać. Wykorzystano naiwny klasyfikator Bayesa. Wybrano go z uwagi na prostotę obliczeń – umożliwia wykonanie analizy szybciej niż inne algorytmy klasyfikacyjne[8], ponadto jego skuteczność w porównaniu do innych algorytmów jest zadowalająca (Friedman, Geiger, Goldszmidt, 1997: 131–163). Utworzony model zweryfikowano, stosując prostą walidację dzielącą losowo zbiór na próbę uczącą (70%) i testową (30%). W tabeli 2 znajdują się wartości uzyskanych na zbiorze testowym wybranych miar oceny klasyfikacji (dokładności, czułości i precyzji). Za pomocą modelu klasyfikującego wykryto 8 dokumentów, których treść nie pasuje do działu, w którym powstały.

[8] W rozpatrywanym przypadku zebrano 680 dokumentów, co oznacza, że baza zawierała 680 wierszy, ale jednocześnie kilkadziesiąt tysięcy kolumn. Można potraktować je wszystkie jako zmienne objaśniające lub też uprzednio dokonać redukcji wielkości macierzy za pomocą techniki LSA (zob. Lula, 2005: 79–80). Biorąc pod uwagę fakt, że docelowo opisywany system bezpieczeństwa będzie gromadził od kilku do kilkunastu tysięcy dokumentów tygodniowo, warto implementować możliwie najmniej obciążające metody analityczne, stąd też decyzja o skorzystaniu z klasyfikatora Bayesa.

Tabela 2. Miary oceny klasyfikacji analizowanych dokumentów testowych

Dokładność: 96,08%			
	Prawidłowe przyporządkowania		**Precyzja**
	Księgowość	**Zarządzanie zasobami ludzkimi**	
Przewidywana klasa: księgowość	152	5	96,82%
Przewidywana klasa: zarządzanie zasobami ludzkimi	3	44	93,62%
Czułość	98,06%	89,90%	

Źródło: opracowanie własne

6. Podsumowanie

Na odnotowanie zasługuje fakt, że opisany przykład pochodzi z pilotażowego uruchomienia testowej wersji systemu wyposażonego w predykcyjne moduły analityczne. Opisany proces wykrywania potencjalnych naruszeń zasad w docelowym systemie monitorowania treści będzie znacznie bardziej rozbudowany, analizy uruchamiane będą cyklicznie, przy uwzględnieniu znacznie większego zbioru zmiennych i kategorii. Informacja o nieprawidłowym przyporządkowaniu dokumentu będzie wysyłana do oddzielnej bazy danych, zawierającej informacje o potencjalnie niebezpiecznych działaniach pracowników. Monitorowana będzie też liczba tego typu zdarzeń w określonych jednostkach czasu.

Automatyczne wykrywanie zagrożeń w systemach IT w czasie rzeczywistym jest możliwe, lecz wymaga zaprojektowania wielu bardzo skomplikowanych reguł wyznaczających kolejność transformacji zmiennych i analiz (wraz z regułami zatrzymania i kryteriami porównywania alternatywnych rozwiązań), a także przypisania wstępnie roli poszczególnym zmiennym oraz ustalenia, w jakich analizach i schematach badawczych mogą być wykorzystywane. Wymaga także przygotowania (zakodowania) alternatywnych metod doboru procedur, transformacji i analizy danych, w zależności od wartości statystyk mierzonych w czasie rzeczywistym. W artykule zaprezentowano przykład tworzenia tego typu mechanizmów automatyzujących transformację i prognozowanie szeregów czasowych.

System bezpieczeństwa teleinformatycznego posiadający automatyczne moduły analityczne jest atrakcyjny dla klientów, należy jednak pamiętać, że brak nadzoru człowieka może wiązać się z większą niedoskonałością budowanych modeli predykcyjnych czy klasyfikacyjnych.

Brak możliwości pełniejszego zrozumienia danych i modyfikacji domyślnych ustawień działania algorytmów sprawia, że wykonywana w tym trybie analiza będzie obarczona znacznie większym błędem, niż gdyby przygotował ją doświadczony analityk. Biorąc jednak pod uwagę ogromną ilość danych, liczbę monitoro-

wanych całą dobę zmiennych i fakt, że opisywane oprogramowanie musi działać u klientów samodzielnie, obrany kierunek rozwoju systemów bezpieczeństwa wykorzystujących metody Data Mining należy uznać za słuszny.

Bibliografia

Cichowicz T., Frankiewicz M., Rytwiński F., Wasilewski J., Zakrzewicz M. (2012), *Anomaly Detection in Time Series for System Monitoring*, „The Poznan School of Banking Research", nr 40, s. 115–130.

Friedman N., Geiger D., Goldszmidt M. (1997), *Bayesian network classifiers*, „Machine Learning", t. 29(2–3), s. 131–163.

Harvey A.C. (1990), *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press, New York.

Hyndman R.J., Khandakar Y. (2007), *Automatic time series for forecasting: the forecast package for R*. Working paper 06/07, Monash University, Department of Econometrics and Business Statistics, Melbourne.

Lula P. (2005), *Text mining jako narzędzie pozyskiwania informacji z dokumentów tekstowych*, StatSoft Polska, https://media.statsoft.pl/_old_dnn/downloads/text_mining_jako_narzedzie_pozyskiwania.pdf [dostęp: 22.11.2016].

Lula P., Wójcik K., Tuchowski J. (2016), *Feature-based sentiment analysis of opinions in polish*, „Research Papers of Wrocław University of Economics: Taxonomy 27. Classification and Data Analysis. Theory and Applications", s. 153–164.

Mirończuk M. (2012), *Review of methods and text data mining*, „Studies and Materials in Applied Computer Science", t. 4, nr 6, s. 25–42.

Mitrea C.A., Lee C.K.M., Wu Z. (2009), *A comparison between neural networks and traditional forecasting methods: A case study*, „International Journal of Engineering Business Management", t. 1, s. 19–24.

Okasha M.K., Yaseen A.A. (2013), *Comparison between ARIMA models and artificial neural networks in forecasting Al-Quds indices of Palestine stock exchange market*, The 25th Annual International Conference on Statistics and Modeling in Human and Social Sciences, Department of Statistics, Faculty of Economics and Political Science, Cairo University, Cairo.

Sapała K., Piołun-Noyszewski M., Weiss M. (2017), *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*, „Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu. Taksonomia 29. Klasyfikacja i analiza danych: teoria i zastosowania", s. 159–166.

Zhang G.P. (2003), *Time series forecasting using a hybrid* ARIMA *and neural network model*, „Neurocomputing", t. 50, s. 159–175.

Automatic Threat Detection in ICT Systems by Selected Data Mining Methods and Software

Abstract: The paper presents some real-time analytical solutions that work in a proprietary-designed system for IT security. It describes automatic methods of data transformations and analysis aiming at detection of potential threats (irregular system events, abnormal user behavior) both for time series and text documents without human supervision. Automation procedures used for time series and text documents are presented. Analyzed data was collected by Free Construction while protecting systems of electronic documents repositories (also including the Enterprise Content Management standards).

Keywords: ICT systems, electronic documents, transformations, data mining methods

JEL: C02

Received: 2016-12-28; verified: 2018-01-13. Accepted: 2018-04-23