Zeszyty Naukowe
Wyższej Szkoły Bankowej w Poznaniu
Nr 40/2012

Tomasz Cichowicz, Michał Frankiewicz, Filip Rytwiński, Jacek Wasilewski, Maciej Zakrzewicz

Politechnika Poznańska

Odkrywanie anomalii w szeregach czasowych pochodzących z monitoringu systemów teleinformatycznych

Streszczenie. *Zautomatyzowana analiza szeregów czasowych pochodzących z monitoringu systemów teleinformatycznych jest odpowiedzią na rosnącą złożoność topologiczną i techniczną współczesnych systemów. Jednym z trudniejszych zagadnień z zakresu analizy szeregów czasowych jest wykrywanie anomalii, sygnalizujących awarię lub niewłaściwe użycie systemu teleinformatycznego. W artykule omówiono kontekst wykrywania anomalii w szeregach czasowych pochodzących z monitoringu systemów teleinformatycznych, dokonano przeglądu dotychczasowych metod i algorytmów, zaproponowano dwie nowe metody wykrywania anomalii oraz zaprezentowano wyniki złożonych badań eksperymentalnych.*

Słowa kluczowe*: systemy teleinformatyczne, analiza szeregów czasowych*

1. Wprowadzenie

Złożoność topologiczna i techniczna współczesnych systemów teleinformatycznych wymaga ciągłego monitorowania sprawności i efektywności ich funkcjonowania. Każdy ze składników systemu teleinformatycznego – aktywne urządzenie sieciowe, serwer bazy danych, serwer aplikacji, urządzenie pamięci masowej, aplikacja itp. – dokonuje automatycznych obserwacji swojego stanu pracy, a wyniki tych obserwacji udostępnia zewnętrznym aplikacjom narzędziowym (*Network Monitoring Tools*). Bieżąca analiza wskaźników raportowanych przez składniki systemu teleinformatycznego umożliwia wczesne wykrywanie awarii, identyfikację działań „podejrzanych" (np. ataków typu *Denial of Service*, prób włamania), optymalizację wydajności i utylizacji systemów. Ze względu na rozmiary obecnych systemów i mnogość raportowanych przez nie wskaźników monitorowanie pracy systemów teleinformatycznych bezwzględnie wymaga daleko idącej automatyzacji.

Obecnie wykorzystywane aplikacje narzędziowe służące do monitorowania pracy systemów teleinformatycznych (np. IBM Tivoli, HP Network Node Manager, WhatsUpGold, Solar Winds Orion, Zenoss, Oracle Enterprise Manager) realizują zaledwie podstawowy funkcjonalny poziom automatyzacji. Skupiają się na wizualizacji obserwowanych wielkości w formie interakcyjnych wykresów graficznych, gromadzą odczyty historyczne w celu późniejszej analizy, automatycznie sygnalizują fakt przekroczenia statycznych poziomów alarmowych oraz znaczące odstępstwa od typowego poziomu wartości. Pomimo tych funkcjonalności wymienione narzędzia wymagają jednak zarówno wykonania złożonej wstępnej konfiguracji, jak i późniejszej ciągłej asysty ze strony doświadczonego administratora lub operatora systemu. W przypadku systemów bardzo dużych, w których liczba monitorowanych wielkości osiąga rząd tysięcy i dziesiątków tysięcy, nieprzekraczalną granicą stają się możliwości człowieka w zakresie jednoczesnego śledzenia wielu niezależnych zdarzeń.

Wśród odbiorców i użytkowników aplikacji narzędziowych służących do monitorowania pracy systemów teleinformatycznych coraz wyraźniej artykułowana jest potrzeba stosowania wysoce zautomatyzowanych algorytmów zarządzania, umożliwiających nienadzorowane wykrywanie zjawisk „nietypowych" na podstawie obserwacji wybranych cech statystycznych monitorowanych wielkości. Tak formułowany problem nazywany jest w piśmiennictwie wykrywaniem anomalii (*Anomaly Detection*), a w przedmiotowym obszarze zastosowań – wykrywaniem anomalii w szeregach czasowych (*Time Series Anomaly Detection*). Całkowicie zautomatyzowane wykrywanie anomalii jest zadaniem bardzo trudnym, wymagającym zapożyczeń z takich dziedzin naukowych, jak uczenie maszynowe, sztuczna inteligencja i eksploracja danych. Oczekuje się, że w niedalekiej przyszłości aplikacje narzędziowe służące do monitorowania pracy systemów teleinformatycznych będą wyposażane w takie rozwiązania.

W ogólności, problemu wykrywania anomalii nie zawęża się wyłącznie do analizy danych syntetycznych, generowanych przez urządzenia lub aplikacje komputerowe. Analogiczne wyzwania pojawiają się np. w epidemiologii, gdzie na podstawie statystyk zachorowań podejmuje się decyzje o ogłoszeniu epidemii, czy w sejsmologii, gdzie na podstawie zapisów amplitudy drgań gruntu w czasie przewiduje się nadejścia fal sejsmicznych. Zwykle jednak algorytmów wykrywania anomalii opracowanych dla innych dziedzin nauki nie można przenieść wprost na grunt monitorowania systemów teleinformatycznych – ze względu na inną charakterystykę szeregów czasowych lub inny minimalny czas reakcji (np. w epidemiologii – dni, w monitoringu systemów – sekundy).

W niniejszej pracy przedstawiono wyniki doświadczeń dotyczących zastosowania algorytmów wykrywania anomalii dla szeregów czasowych pochodzących z monitoringu systemów teleinformatycznych. W wyniku przeprowadzonych studiów literaturowych wyłoniono podzbiór algorytmów przystających do zdefinio-

wanych wymagań, a ponadto opracowano kilka specjalizowanych algorytmów własnych. Następnie zrealizowano złożony eksperyment obliczeniowy, w ramach którego badano skuteczność algorytmów wykrywania anomalii. W badaniach eksperymentalnych wykorzystywano rzeczywiste szeregi czasowe opisujące funkcjonowanie systemu klasy *e-commerce*.

2. Specyfika szeregów czasowych pochodzących z monitoringu systemów teleinformatycznych

Cechy charakterystyczne (zmienność, kształt, częstotliwości składowe itp.) szeregu czasowego pochodzącego z monitoringu są w znacznej mierze zależne od typu urządzenia, którego stan pracy podlega obserwacji. Podczas prowadzonych badań dokonano następującej klasyfikacji typów urządzeń/elementów:

a) **urządzenia sieciowe**, dokonujące retransmisji strumieni danych przysyłanych do systemu lub wysyłanych przez system do odbiorców zewnętrznych: karty sieciowe, routery, zapory sieciowe (*Firewall*), punkty dostępowe (*Access Point*). Mierzone wielkości to m.in. gęstość strumienia (rys. 1),

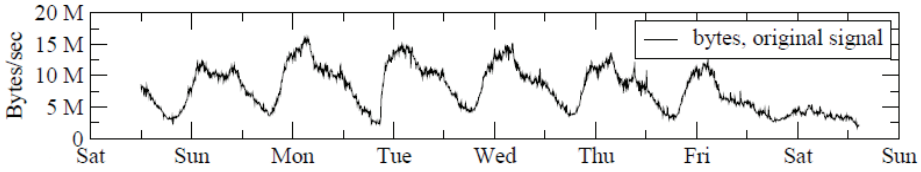
b) **urządzenia pamięci masowej**, zapisujące i odczytujące pliki użytkowników i aplikacji: macierze dyskowe, serwery plików, dyski sieciowe. Mierzone wielkości to m.in.: wykorzystana pojemność, średni czas dostępu, ilość odczytywanych/zapisywanych danych,

c) **oprogramowanie systemowe**, przetwarzające żądania użytkowników i aplikacji: systemy operacyjne, serwery baz danych, serwery aplikacji. Mierzone wielkości to m.in.: liczba równoczesnych sesji/połączeń, wskaźniki wydajnościowe (np. współczynniki trafień w bufory, liczby zdarzeń oczekiwania na zwolnienie blokady), liczba udanych i nieudanych logowań, zajętość pamięci operacyjnej, liczba procesów, liczba współbieżnych wątków,

d) **aplikacje biznesowe**, implementujące funkcje użytkowe: aplikacje ERP, aplikacje CRM, aplikacje *e-commerce* itp. Mierzone wielkości to m.in.: liczba żądań wykonania funkcji biznesowych, liczba równoczesnych sesji, średni czas odpowiedzi,

e) **procesory** wykonujące kod programowy aplikacji biznesowych i oprogramowania systemowego. Ta klasa urządzeń obejmuje zarówno procesory fizyczne, jak i wirtualne. Mierzone wielkości to m.in. chwilowe obciążenie procesora.

W zależności od typu urządzenia/elementu różnego znaczenia nabiera pojęcie anomalii w szeregu czasowym. O wystąpieniu anomalii w gęstości strumienia dostarczanego do karty sieciowej może świadczyć wzrost powyżej poziomu typowego (przeciążenie), ale też spadek do poziomu bliskiego zeru (awaria urządzenia poprzedzającego). Anomalią dla wykorzystanej pojemności macierzy dyskowej będzie zbliżenie się do poziomu 100% zapełnienia. Z kolei 100%



Rys. 1. Przykładowy szereg czasowy pochodzący z monitoringu karty sieciowej (gęstość strumienia wychodzącego, w bajtach na sekundę)

Ź r ó d ł o: P. Barford, J. Kline, D. Plonka, A. Ron, *A signal analysis of network trafic anomalies*, w: *IMW'02 Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurement*, ACM, New York 2002, s. 71-82.

obciążenie nie będzie traktowane jako anomalia w przypadku procesora, o ile nie będzie utrzymywać się przez dłuższy czas. Dla serwera aplikacji przejawem anomalii może być zbyt szybki wzrost liczby sesji (atak typu *Denial of Service*). W wielu przypadkach przejawem anomalii może być inny rozkład szeregu czasowego w skali dnia/tygodnia w odniesieniu do analogicznego okresu w przeszłości. Na przykład liczba 50 nieudanych prób logowania do aplikacji ERP może być naturalna podczas dnia roboczego, ale może świadczyć o anomalii, gdy zostanie odnotowana w niedzielę (próba włamania).

Dla wielu systemów informatycznych typowy jest powolny, stopniowy wzrost obciążenia i rozmiaru wynikający ze zwiększania się liczby użytkowników, klientów, obiektów przechowywanych w bazie danych itp. Zwykle skutkuje to pojawieniem się trendu wznoszącego w szeregu czasowym. Zjawisko takie powinno być uwzględniane przez algorytmy wykrywania anomalii – poprzez wznoszenie stosownych poziomów alarmowych. Podobnego traktowania wymagają zjawiska sezonowości (przejściowy wzrost obciążenia systemów *e-commerce* w okresach przedświątecznych).

W związku ze stosowanymi metodami pomiaru wielkości obserwowanych, opartymi na próbkowaniu czasowym, w analizowanych szeregach czasowych często występują składowe wysokiej częstotliwości, mogące zaburzać działanie algorytmów wykrywania anomalii. W celu eliminacji niepożądanych efektów stosuje się wstępne wygładzanie (*Smoothing*) szeregu czasowego. Intensywność wygładzania musi być jednak dostosowana do specyfiki badanego szeregu, aby nie spowodowała uszkodzenia informacji o występującej anomalii.

3. Dotychczasowe badania

3.1. Badane algorytmy wygładzania szeregu czasowego

Prowadzone badania objęły obserwację skuteczności działania opisanych w literaturze wybranych algorytmów wygładzania szeregu czasowego, takich jak:

średnia krocząca, wygładzanie wykładnicze, podwójne wygładzanie wykładnicze, potrójne wygładzanie wykładnicze, wygładzanie cepstralne.

Najprostszym algorytmem wygładzania jest **średnia krocząca**[1], powszechnie stosowana w finansach i analizie technicznej – średnia arytmetyczna wartości z ostatnich *n* próbek. Występuje w wielu odmianach, jak: ważona średnia krocząca, wykładnicza średnia krocząca, średnia krocząca poprawiona o wolumen, trójkątna średnia krocząca. Może służyć również jako metoda odcinania wartości skrajnych.

Wygładzanie wykładnicze ma na celu zmniejszenie wariancji źródłowego szeregu czasowego za pomocą ważonej średniej kroczącej z przeszłych wartości[2]. Wagi średniej maleją wykładniczo wraz z upływem czasu. Wygładzanie wykładnicze może być zastosowane w usuwaniu szumów oraz prognozowaniu przebiegów czasowych, gdzie stosunek między sygnałem a szumem jest niewielki oraz dane nie wykazują wyraźnego trendu i wahań sezonowych.

Podwójne wygładzanie wykładnicze[3], znane również jako wygładzanie wykładnicze Holta, jest udoskonaleniem modelu zwykłego wygładzania wykładniczego. Uwzględnia występowanie tendencji rozwojowych (trendów), jak i wahania przypadkowe. Opiera się na liniowym modelu Holta:

$$F_{t-1} = \alpha y_{t-1} + (1-\alpha)(F_{t-2} + S_{t-2})$$
$$S_{t-1} = \beta(F_{t-1} - F_{t-2}) + (1-\beta)S_{t-2}$$

gdzie: F_{t-1} – wygładzona wartość zmiennej prognozowanej na moment $t - 1$; S_{t-1} – wygładzona wartość przyrostu trendu na moment $t - 1$; α, β – parametry modelu o wartościach z przedziału [0, 1].

W przypadku niewystępowania trendu lub sezonowości najlepsze rezultaty wygładzania dawała metoda zwykłego wygładzania wykładniczego, natomiast w sytuacji, gdy w szeregach źródłowych pojawiał się trend wznoszący lub opadający, zwykłe wygładzanie wykładnicze wykazywało skłonność do opóźniania wygładzania. Jednakże mimo dobrych wyników wygładzania dla szeregów czasowych wykazujących trend wygładzanie Holta nie sprawdzało się przy szeregach czasowych ujawniających cechy sezonowości. Dla takich przebiegów czasowych bardziej atrakcyjną metodą było potrójne wygładzanie wykładnicze.

Potrójne wygładzanie wykładnicze[4], często nazywane wygładzaniem Holta-Wintersa, uwzględnia sezonowość w szeregu czasowym. Występuje w dwóch

---

[1] J. Durbin, *Efficient estimation of parameters in moving-average models*, „Biometrika" 1959, nr 3.

[2] *Averaging and exponential smoothing models*, www.duke.edu/~rnau/411avg.htm [01.2012].

[3] Ibidem; *OpenForecastAPI*, http://openforecast.sourceforge.net/docs [01.2012].

[4] *Averaging and exponential smoothing models*, op. cit.; *Triple exponential smoothing*, www.itl.nist.gov/div898/handbook/pmc/section4/pmc435.htm [01.2012]; *OpenForecastAPI*, http://openforecast.sourceforge.net/docs [01.2012].

wersjach modelu: addytywnej i multiplikatywnej. Dla wersji addytywnej równania modelu przedstawiają się następująco:

$$F_{t-1} = \alpha(y_{t-1} - C_{t-1-r}) + (1-\alpha)(F_{t-2} + S_{t-2})$$

$$S_{t-1} = \beta(F_{t-1} - F_{t-2}) + (1-\beta)S_{t-2}$$

$$C_{t-1} = \gamma(y_{t-1} - F_{t-1}) + (1-\gamma)C_{t-1-r}$$

natomiast dla wersji multiplikatywnej:

$$F_{t-1} = \alpha\frac{y_{t-1}}{C_{t-1-r}} + (1-\alpha)(F_{t-2} + S_{t-2})$$

$$S_{t-1} = \beta(F_{t-1} - F_{t-2}) + (1-\beta)S_{t-2}$$

$$C_{t-1} = \gamma\frac{y_{t-1}}{F_{t-1}} + (1-\gamma)C_{t-1-r}$$

gdzie: F_{t-1} – wygładzona wartość zmiennej prognozowanej na moment $t - 1$ po eliminacji wahań sezonowych; S_{t-1} ocena przyrostu trendu na moment $t - 1$; C_{t-1} – ocena wskaźnika sezonowości na moment $t - 1$; r – liczba okresów; α, β, γ – parametry modelu o wartościach z przedziału [0, 1].

Wadą tej metody jest wymagalność przynajmniej jednego zakończonego szeregu czasowego sezonowego do wyznaczenia początkowych estymat wskaźników sezonowości C. Kompletne dane sezonowe składają się z r okresów, ponieważ wymagana jest estymacja współczynnika trendu przy przejściu z jednego okresu do kolejnego. Zalecane jest wykorzystywanie dwóch zakończonych, kompletnych sezonów, tzn. $2r$ okresów – a w praktyce 5-6, gdyż umożliwia to modelowi skuteczniejszą adaptację do danych, a nie ślepe typowanie wartości lub poprawną estymację jedynie dla początkowych elementów. W badaniach wykorzystano model multiplikatywny oraz wymagano dwóch kompletnych cykli danych do inicjalizacji modelu.

Wygładzanie współczynnikami cepstralnymi bazuje na transformacie Fouriera przeniesionej w dziedzinę decybelowę (cepstrum) i na oknie dolnoprzepustowym[5]. Kroki algorytmu są następujące:

1. Wykonywana jest szybka dyskretna transformata Fouriera na źródłowym szeregu czasowym.

2. Otrzymany wynik przekształcany jest tak, aby stał się widmem, w którym amplituda wyrażona jest w decybelach.

[5] J.O. Smith III, *MUS421/EE367B applications lecture b: Cross synthesis using cepstral smoothing or linear prediction for spectral envelopes*, https://ccrma.stanford.edu/~jos/SpecEnv/SpecEnv.pdf [01.2012]; *Cepstral smoothing*, https://ccrma.stanford.edu/~jos/SpecEnv/Cepstral_Smoothing.html [01.2012].

3. Za pomocą funkcji okna dolnoprzepustowego obcinane są mało znaczące składowe periodyczne, które z założenia zaburzają sygnał.

4. Wynik transformowany jest odwrotnie w szereg czasowy za pomocą odwrotnej transformaty Fouriera.

3.2. Badane algorytmy wykrywania wystąpienia anomalii

Prowadzone badania objęły obserwację skuteczności działania opisanych w literaturze wybranych algorytmów wykrywania wystąpienia anomalii, m.in.: metod finansowej analizy technicznej, metod ekstrakcji składowych sygnału, metod WSARE, metod opartych na klasyfikatorach decyzyjnych oraz metod trzech sigm.

Badaniom poddano trzy najpopularniejsze **modele analizy technicznej**: MACD (*Moving Average Convergence/Divergence*), Momentum (wskaźnik zmiany ROC) oraz wstęgę Bollingera. MACD[6] jest wskaźnikiem badającym zbieżność i rozbieżność średnich kroczących. Reprezentuje różnice wartości długoterminowej i krótkoterminowej średniej wykładniczej. Produktem tego modelu są dwie linie – MACD oraz linia sygnału (średnia z linii MACD). Moment przecięcia linii sygnału z linią MACD oznacza zmianę trendu, interpretowaną w badaniach jako prawdopodobne wystąpienie anomalii. Momentum[7] oznacza procentową zmianę wartości pomiędzy stanem aktualnym a stanem sprzed *k* punktów czasowych. Osiąganie ekstremum przez ten wskaźnik może być interpretowane jako wzmocnienie trendu – np. anomalia ataku przyrostowego na system lub anomalia zwiększenia wykorzystania systemu. Metoda wstęgi Bollingera[8] zakłada, że zmienność wartości obserwowanego parametru jest dynamiczna, a nie statyczna. Wstęga Bollingera składa się z: (1) wstęgi środkowej, będącej *n*-okresową średnią kroczącą, (2) wstęgi górnej, będącej *k*-krotnością *n*-okresowego odchylenia standardowego powyżej wstęgi środkowej, (3) wstęgi dolnej, która jest *k*-krotnością *n*-okresowego odchylenia standardowego poniżej wstęgi środkowej. Wstęga Bollingera tworzy swoisty korytarz, którego opuszczenie jest traktowane jako anomalia.

Metody ekstrakcji składowych sygnału traktują szereg czasowy jako opis próbek okresowego sygnału ciągłego, w stosunku do którego stosować można

---

[6] J.J. Murphy, *Technical analysis of the financial markets*, „Pennsylvania Dental Journal" 1999, nr 77(2); *Encyklopedia analizy technicznej*, www.wdsoftware.com/pl/encyklopedia-at/index.html [01.2012].

[7] *Encyklopedia analizy technicznej*, op. cit.; T. Fawcett, *An introduction to roc analysis*, „Pattern Recogn. Lett." 2006, nr 27, s. 861-874.

[8] *Encyklopedia analizy technicznej*, op. cit.; J. Bollinger, *Bollinger on Bollinger bands*, Mc-Graw-Hill, 2001.

techniki wyodrębniania składowych (np. sinusoidalnych). Przy wykorzystaniu takiego modelu przewidywane są przyszłe wartości sygnału, a następnie odnoszone do wartości faktycznie mierzonych – duże odstępstwo wskazuje na wystąpienie anomalii. Badane podejścia obejmowały: szybką transformację Fouriera, falki (*Wavelets*) i analizę głównych składowych (*Principal Component Analysis* – PCA)[9].

Interesującym podejściem jest WSARE (*What's Strange About Recent Events*), pierwotnie opracowane w celu wczesnego wykrywania zagrożeń epidemiologicznych na podstawie danych pochodzących z różnych źródeł, takich jak: przychodnie, szpitale, stacje meteorologiczne, dane o migracji ludności, ruchu ulicznym itp.[10] Głównym założeniem WSARE jest operowanie na dyskretnym, wielowymiarowym zbiorze danych i porównywanie wektora wartości teraźniejszych do danych historycznych, np. w postaci statystyk. W związku z docelowym zastosowaniem WSARE projektowano tak, aby bez względu na konkretne zastosowane algorytmy wykrywanie anomalii odbywało się szybko, a ogólna złożoność obliczeniowa była stała lub liniowa względem rozmiaru historii. Opublikowano trzy oficjalne implementacje (wersje) WSARE: 2.0, 2.5 i 3.0.

Metody wnioskowania probabilistycznego przy wykorzystaniu **klasyfikatorów decyzyjnych** opierają się na przewidywaniu prawdopodobieństwa wystąpienia określonej przyszłej wartości próbki w szeregu czasowym, a następnie porównania wartości przewidywanej z wartością faktycznie odnotowaną. Najbardziej rozpowszechnionym modelem pozwalającym określać prawdopodobieństwo zajścia pewnego ciągu zdarzeń są sieci Bayesa. Modelują one zależności przyczynowe poprzez tworzenie acyklicznego grafu skierowanego. Wierzchołki tego grafu reprezentują zdarzenia. W kontekście wykrywania anomalii w szeregu czasowym wierzchołkiem może być wartość badanej funkcji w ustalonym momencie czasu. Łuki natomiast modelują związki przyczynowe między zdarzeniami. Dzięki tak stworzonej sieci w łatwy sposób można wyznaczyć prawdopodobieństwo warunkowe zajścia konkretnych zdarzeń w systemie.

Naiwny klasyfikator Bayesa, będący obecnie jednym z popularniejszych klasyfikatorów, jest oparty na regule Bayesa, pozwalającej obliczać prawdopodobieństwo warunkowe zajścia zdarzenia[11]. Wykorzystuje on upraszczające obli-

[9] *Factor analysis*, www.psych.cornell.edu/Darlington/factor.htm [23.01.2012]; W.J. Krzanowski, *Principles of multivariate analysis: a user's perspective*, „Oxford statistical science series", Oxford University Press, Oxford 2000.

[10] W.-K. Wong, A. Moore, G. Cooper, M. Wagner, *What's Strange About Recent Events*, „Journal of Urban Health", czerwiec 2003, Supplement 1; W.-K. Wong, A. Moore, G. Cooper, M. Wagner, *What's Strange About Recent Events (WSARE): An algorithm for the early detection of disease outbreaks*, „Journal of Machine Learning Research" 2005, nr 6.

[11] S. Thrun, P. Norvig, *Online introduction to artificial intelligence*, www.ai-class.com/course/topic/6 [01.2012].

czenia założenie o wzajemnej warunkowej niezależności atrybutów opisujących próbkę względem zmiennej decyzyjnej. Mimo takiego uproszczenia modelowanej rzeczywistości algorytm daje w praktyce bardzo dobre rezultaty, m.in. przy wykrywaniu spamu.

Metoda trzech sigm opiera się na założeniu, że wartości szeregu czasowego przyjmują rozkład zbliżony do krzywej Gaussa. Zgodnie z tą metodą pojawienie się nowej próbki oznacza uaktualnienie średniej wartości dotychczasowej historii oraz odchylenia standardowego w obrębie tej historii. Następnie aktualna próbka jest porównywana z obliczoną średnią i w przypadku różnicy większej niż ustalona wielokrotność odchylenia standardowego (w badanym rozwiązaniu: trzykrotność) licznik sygnału anomalii jest zwiększany o 1. W przeciwnym przypadku licznik jest zerowany. Algorytm sygnalizuje anomalię z chwilą, gdy licznik przekroczy ustaloną wartość.

4. Propozycje nowych rozwiązań

4.1. Profile szeregów czasowych

Motywacją do opracowania metody profili była obserwacja cykliczności i wewnętrznego podobieństwa szeregów czasowych generowanych w zbliżonych warunkach, np. obciążenie serwera poczty elektronicznej we wtorek wykazuje podobną charakterystykę jak obciążenie tego samego serwera w poniedziałek (wysoka aktywność w godzinach pracy biura, niska aktywność w godzinach nocnych). Termin „profil szeregu czasowego" reprezentuje typowy kształt szeregu czasowego zobrazowanego w formie wykresu, wykorzystywany do analizy zachowania się systemu w analogicznych oknach czasowych w przyszłości. Znaczące odstępstwo od kształtu profilu jest sygnałem wystąpienia anomalii.

Profile szeregów czasowych generowano z wykorzystaniem algorytmu dwufazowego, obejmującego wygładzenie szeregu w oknie czasowym, a następnie ekstrakcję cech szeregu. Do wygładzania szeregu zastosowano algorytmy średniej kroczącej i wygładzania wykładniczego. Ekstrakcji cech dokonywano za pomocą funkcji: pochodnej (jako miary szybkości zmian wartości w szeregu czasowym), całki (jako sumy wartości szeregu czasowego), transformaty Fouriera (jako dekompozycji na częstotliwości składowe). Do wykrywania rozbieżności pomiędzy profilem historycznym a profilem bieżącym wykorzystano algorytm trzech sigm.

4.2. Tablice znamionowe

Metoda tablic znamionowych polega na generowaniu zestawów parametrów charakteryzujących zachowanie się szeregu czasowego w historycznych wąskich oknach czasowych, a następnie na porównywaniu tych parametrów z cechami szeregu aktualnego. Strukturę tablicy znamionowej przedstawiono w tabeli 1.

Tabela 1. Struktura tablicy znamionowej szeregu czasowego

Nazwa parametru	Opis
valueMin	minimalna wartość sygnału w oknie
valueMax	maksymalna wartość sygnału w oknie
valueAvg	średnia wartość sygnału w oknie
valueMed	mediana wartości sygnału w oknie
valueStd	odchylenie standardowe wartości sygnału w oknie
derivativeMin	minimalna wartość z pochodnej sygnału w oknie
derivativeMax	maksymalna wartość z pochodnej sygnału w oknie
derivativeAvg	średnia wartość pochodnej sygnału w oknie
derivativeMed	mediana wartości pochodnej sygnału w oknie
derivativeStd	odchylenie standardowe wartości pochodnej sygnału w oknie
fourierArea	pole pod wykresem widma fourierowskiego sygnału

Ź r ó d ł o: opracowanie własne.

5. Eksperymenty obliczeniowe

W celu oceny użyteczności i dokładności metod wykrywania anomalii w szeregach czasowych pochodzących z monitoringu systemów teleinformatycznych przeprowadzono eksperyment obliczeniowy z wykorzystaniem rzeczywistych danych pomiarowych, opisujących działanie m.in.: serwerów Microsoft Active Directory, serwerów poczty elektronicznej, serwerów baz danych, serwerów aplikacji i routerów Cisco. Dane pomiarowe prezentowały wartości takich wskaźników, jak: zużycie dysku, obciążenie procesora, ruch sieciowy przychodzący i wychodzący, zużycie pamięci operacyjnej maszyn wirtualnych Java, zużycie pamięci operacyjnej serwera. Badane algorytmy zostały zaimplementowane w formie wielowątkowej, modułowej aplikacji Java EE uruchomionej na platformie serwera aplikacji Glassfish 3.1. Dane źródłowe były utrwalone w bazie danych Oracle Database 11g. Całość środowiska obliczeniowego była osadzona na platformie Linux CentOS, wyposażonej w 8 rdzeni procesorów i 32 GB pamięci operacyjnej.

Podczas eksperymentów wyszukiwane były zarówno anomalie naturalne, jak i anomalie sztuczne dodane do rzeczywistych danych pomiarowych. Anomalie sztuczne były generowane jako trapezoidalne: (1) punktowe, (2) trójkątne krótkotrwałe, (3) długotrwałe. Na podstawie każdego eksperymentu odczytywano wartości czterech wskaźników:

1) liczba poprawnie wykrytych anomalii (TP – *True Positive*),
2) liczba niepoprawnie wykrytych anomalii (FP – *False Positive*),
3) liczba niewykrytych anomalii (FN – *False Negative*),
4) liczba poprawnie przeanalizowanych próbek bez anomalii (TN – *True Negative*).

Wskaźniki te posłużyły do wyprowadzenia wartości współczynników jakościowych: *czułości* i *swoistości*. Czułość (*Sensitivity*) przedstawia procent anomalii, które zostały poprawnie wykryte, wyrażony za pomocą formuły:

$$Sens = \frac{TP}{TP + FN}$$

Swoistość (*Specificity*) zdefiniowano jako procent poprawnie zdiagnozowanych przypadków, które nie były anomaliami:

$$Spec = \frac{TN}{TN + FP}$$

W celu ułatwienia prezentacji wyników eksperymentów czułość i swoistość złożyły się na *F*-wartość (*F-score*):

$$F = 2 \cdot \frac{Spec \cdot Sens}{Spec + Sens}$$

5.1. Profile szeregów czasowych

Wykonano ponad 60 000 testów opartych na permutacjach zestawu parametrów: 26 szeregów czasowych dla różnych rodzajów urządzeń, 8 rodzajów badanych anomalii, 5 długości okna (1, 2, 4, 12, 24 godziny), rodzaj algorytmu wygładzania (średnia krocząca, wygładzanie wykładnicze), 6 parametrów algorytmu wygładzania (okno o rozmiarze 2, 4, 6, współczynnik wygładzania wykładniczego 0,5, 0,6, 0,7), 2 rodzaje algorytmu ekstrakcji cech (pochodna, całka), 2 rodzaje algorytmu oceny wyniku (trzy sigmy, tolerancja procentowa), 4 poziomy tolerancji procentowej (10, 25, 50, 75).

Przykładowe najlepsze wyniki pomiarów przedstawiono w tabeli 2. Eksperyment został przeprowadzony z wykorzystaniem danych opisujących sieciowy ruch wychodzący z routera Cisco.

Tabela 2. Przykładowe najlepsze wyniki pomiarów dla metody profili szeregów czasowych – dane źródłowe opisujące ruch sieciowy wychodzący

Lp.	A	B	C	D	E	F	H	I
1	1,0000	0,8607	0,9251	wygł. wykł.	całka	alg. 3-sigma	1	0,6
2	1,0000	0,8607	0,9251	śr. krocząca	całka	alg. 3-sigma	1	2
3	1,0000	0,8607	0,9251	wygł. wykł.	całka	alg. 3-sigma	1	0,7
4	1,0000	0,8607	0,9251	śr. krocząca	całka	alg. 3-sigma	1	4
5	1,0000	0,8607	0,9251	wygł. wykł.	pochodna	alg. 3-sigma	1	0,6
6	1,0000	0,8607	0,9251	śr. krocząca	pochodna	alg. 3-sigma	1	4
7	1,0000	0,8607	0,9251	wygł. wykł.	całka	alg. 3-sigma	1	0,5
8	1,0000	0,8607	0,9251	wygł. wykł.	pochodna	alg. 3-sigma	1	0,7
9	1,0000	0,8607	0,9251	wygł. wykł.	pochodna	alg. 3-sigma	1	0,5
10	1,0000	0,8607	0,9251	śr. krocząca	pochodna	alg. 3-sigma	1	2

Objaśnienia: A – czułość, B – swoistość, C – *F*-wartość, D – algorytm wygładzania, E – algorytm ekstrakcji cech, F – algorytm wykrywania anomalii, H – szerokość okna, I – współczynnik wygładzania/długość okna wygładzania

Ź r ó d ł o: badania własne.

Na podstawie przeprowadzonych eksperymentów stwierdzono, że w zakresie wykrywania anomalii za pomocą metody profili szeregów czasowych najskuteczniejsze okazały się następujące kombinacje parametrów algorytmu:

a) badanie szeregu czasowego z użyciem długich okien (od 12 do 24 godzin), ekstrakcja cechy za pomocą funkcji pochodnej, a następnie porównywanie otrzymanych profili procentowo, z przyjętym progiem tolerancji: wysokim (50-75%) w przypadku szeregów czasowych obrazujących specyficzny ruch sieciowy, niskim (10-25%) w przypadku szeregów czasowych o charakterze podobnym do Apache Tomcat NonHeapMemoryUsage,

b) badanie szeregu czasowego z użyciem krótkich okien (godzinnych), ekstrakcja cechy za pomocą funkcji całki lub pochodnej, a następnie wykrywanie anomalii na różnicy otrzymanych profili przy użyciu algorytmu trzech sigm.

5.2. Tablice znamionowe

Wykonano ok. 5000 testów opartych na permutacjach zestawu parametrów: 26 szeregów czasowych dla różnych rodzajów urządzeń, 8 rodzajów anomalii, 6 długości okna (1, 2, 4, 8, 12, 24 godziny), 4 poziomy tolerancji procentowej (10, 25, 50, 75), 4 wartości współczynnika wygładzania dla metody średniej kroczącej (2, 4, 6, 10).

Przykładowe najlepsze wyniki pomiarów przedstawiono w tabeli 3. Eksperyment został przeprowadzony z wykorzystaniem danych opisujących obciążenie procesora.

Tabela 3. Przykładowe najlepsze wyniki pomiarów dla metody tablic znamionowych – dane źródłowe opisujące obciążenie procesora

Lp.	A	B	C	D	E	F
1	1,0000	0,3885	0,5596	0,75	1	4
2	1,0000	0,3869	0,5579	0,75	1	2
3	1,0000	0,3841	0,5550	0,75	1	4
4	1,0000	0,3836	0,5545	0,5	1	2
5	1,0000	0,3836	0,5545	0,5	1	4
6	1,0000	0,3825	0,5533	0,75	1	2
7	1,0000	0,3807	0,5515	0,75	1	4
8	1,0000	0,3793	0,5499	0,5	1	2
9	1,0000	0,3793	0,5499	0,5	1	4
10	1,0000	0,3791	0,5498	0,75	1	2

Objaśnienia: A – czułość, B – swoistość, C – *F*-wartość, D – próg tolerancji, E – szerokość okna, F – współczynnik wygładzania

Ź r ó d ł o: badania własne.

Na podstawie przeprowadzonych eksperymentów stwierdzono, że w zakresie wykrywania anomalii za pomocą metody tablic znamionowych dla szeregów czasowych najskuteczniejsze okazały się kombinacje parametrów: 75% próg tolerancji, okno o rozmiarze 1 godziny, współczynniki wygładzania o wartościach 2 i 4.

5.3. Naiwny klasyfikator Bayesa

Wykonano testy oparte na permutacjach zestawu parametrów: 26 szeregów czasowych dla różnych rodzajów urządzeń, 8 rodzajów anomalii, 8 długości okna uczącego (0,5, 1, 2, 4, 8, 12, 24 godziny, 7 dni).

Przykładowe najlepsze wyniki pomiarów przedstawiono w tabeli 4. Eksperyment został przeprowadzony z wykorzystaniem danych opisujących zużycie pamięci Apache Tomcat HeapMemoryUsage. Najwyższa skuteczność detekcji anomalii została odnotowana dla okna o szerokości 30 minut oraz 7 dni.

Tabela 4. Przykładowe najlepsze wyniki pomiarów dla metody naiwnego klasyfikatora Bayesa – dane źródłowe opisujące zużycie pamięci Apache Tomcat HeapMemoryUsage

Lp.	A	B	C	D
1	0,9500	1,0000	0,9744	10080
2	0,9500	1,0000	0,9744	10080
3	0,9500	0,9994	0,9741	30
4	0,9500	0,9994	0,9741	30
5	0,9500	0,9994	0,9741	30
6	0,9500	0,9994	0,9741	30
7	0,9500	0,9994	0,9741	30
8	0,9500	0,9994	0,9741	30
9	0,9500	0,9991	0,9739	30
10	0,9500	0,9991	0,9739	30

Objaśnienia: A – czułość, B – swoistość, C – *F*-wartość, D – szerokość okna (minuty)

Ź r ó d ł o: badania własne.

5.4. Wnioski końcowe z eksperymentów

Przeprowadzone eksperymenty pozwoliły wskazać, które z algorytmów wykrywania anomalii zapewniają najwyższą dokładność działania w odniesieniu do różnych kategorii szeregów czasowych. Wnioski końcowe były następujące:

– najwyższą skutecznością w wykrywaniu anomalii w szeregach czasowych opisujących zużycie pamięci operacyjnej serwera Apache Tomcat oraz zużycie pamięci fizycznej systemu operacyjnego cechowała się metoda profili szeregu czasowego,

– najwyższą skutecznością w wykrywaniu anomalii w szeregach czasowych opisujących ruch sieciowy wychodzący z serwera i zużycie dysku cechowała się metoda oparta na naiwnym klasyfikatorze Bayesa,

– metoda tablic znamionowych oferowała przeciętną jakość we wszystkich badanych kategoriach.

6. Podsumowanie

W artykule przedstawiono problem wykrywania anomalii w szeregach czasowych pochodzących z monitoringu systemów teleinformatycznych. Główną motywacją dla prowadzonych badań było zapotrzebowanie rynkowe na automatyczne algorytmy wspomagające pracę operatora/administratora dużych syste-

mów teleinformatycznych. Dokonano przeglądu istniejących metod wykrywania anomalii oraz zaproponowano dwie nowe metody: profili szeregu czasowego oraz tablic znamionowych profilu czasowego. W ramach eksperymentalnej ewaluacji potwierdzono ich skuteczność oraz wskazano najbardziej przystające obszary zastosowań. Dalsze planowane prace badawcze obejmą wykorzystanie korelacji pomiędzy różnymi szeregami czasowymi w celu wykrywania zachowań nietypowych.

Literatura

Averaging and exponential smoothing models, www.duke.edu/~rnau/411avg.htm [01.2012].

Barford P., Kline J., Plonka D., Ron A., *A signal analysis of network trafic anomalies*, w: *IMW'02 Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurement, ACM*, New York 2002, s. 71-82.

Bertsekas D., Tsitsiklis J., *Probabilistic systems analysis and applied probability*, http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-041-probabilistic-systems-analysis-and-applied-probability-fall-2010/lecture-notes [01.2012].

Bollinger J., *Bollinger on Bollinger bands*, McGraw-Hill, 2001.

Cepstral smoothing, https://ccrma.stanford.edu/~jos/SpecEnv/Cepstral_Smoothing.html [01.2012].

Chandola V., Banerjee A., Kumar V., *Anomaly detection. A survey. ACM*, „Comput. Surv.", lipiec 2009.

Durbin J., *Efficient estimation of parameters in moving-average models*, „Biometrika" 1959, nr 3.

Encyklopedia analizy technicznej, www.wdsoftware.com/pl/encyklopedia-at/index.html [01.2012].

Factor analysis, www.psych.cornell.edu/Darlington/factor.htm [23.01.2012].

Fawcett T., *An introduction to roc analysis*, „Pattern Recogn. Lett." 2006, nr 27, s. 861-874.

Gao J., Hu G., Yao X., Chang R.K.C., *Anomaly detection of network traffic based on wavelet packet*, APCC '06. Asia-Pacific Conference on Communications, 2006.

Generating mechanical forecasts from statistical models, www.mrp3.com/fcst_models.html [01.2012].

Krzanowski W.J., *Principles of multivariate analysis: a user's perspective*, „Oxford statistical science series", Oxford University Press, Oxford 2000.

Kumar N., Lolla N., Keogh E., Lonardi S., Ratanamahatana Ch.A., *Time-series bitmaps: a practical visualization tool for working with large time series databases*, SIAM 2005 Data Mining Conference, SIAM, 2005, s. 531-535.

Lin J., Keogh E., Lonardi S., Chiu B., *A symbolic representation of time series, with implications for streaming algorithms*, Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, ACM Press, 2003.

Lo A.W., Mamaysky H., Wang J., *Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation*, „The Journal of Finance" 2000, nr 55(4), s. 1705-1770.

Murphy J.J., *Technical analysis of the financial markets*, „Pennsylvania Dental Journal" 1999, nr 77(2).

Naiwny klasyfikator Bayesa, www.statsoft.com.pl/textbook/stnaiveb.html [01.2012].

Ng A., *Machine learning*, www.ml-class.org/course/auth/welcome [01.2012].

OpenForecastAPI, http://openforecast.sourceforge.net/docs [01.2012].

Smith III J.O., *MUS421/EE367B applications lecture b: Cross synthesis using cepstral smoothing or linear prediction for spectral envelopes*, https://ccrma.stanford.edu/~jos/SpecEnv/SpecEnv.pdf [01.2012].

Stefanowski J., *Analiza szeregów czasowych*, www.cs.put.poznan.pl/jstefanowski/aed/TPtimeseries.pdf [01.2012].

Thrun S., Norvig P., *Online introduction to artificial intelligence*, www.ai-class.com/course/topic/6 [01.2012].

Triple exponential smoothing, www.itl.nist.gov/div898/handbook/pmc/section4/pmc435.htm [01.2012].

Wei L., Kumar N., Lolla V., Keogh E., Lonardi S., Ratanamahatana Ch.A., *Assumption-free anomaly detection in time series*, Proceedings of the 17th International Conference on Scientic and Statistical Database Management 2005, s. 237-242.

Wong W.-K., Moore A., Cooper G., Wagner M., *Bayesian network anomaly pattern detection for disease outbreaks*, Proceedings of the Twentieth International Conference on Machine Learning, Menlo Park, California, lipiec 2003, AAAI Press, s. 808-815.

Wong W.-K., Moore A., Cooper G., Wagner M., *What's Strange About Recent Events*. „Journal of Urban Health", czerwiec 2003, Supplement 1.

Wong W.-K., Moore A., Cooper G., Wagner M., *What's Strange About Recent Events (WSARE): An algorithm for the early detection of disease outbreaks*, „Journal of Machine Learning Research" 2005, nr 6.