Autoreferat rozprawy doktorskiej

ADAPTACYJNE ALGORYTMY DETEKCJI ZDARZEŃ W SZEREGACH CZASOWYCH

mgr inż. Tomasz Pełech-Pilichowski

PROMOTOR: prof. dr hab. inż. Jan T. Duda – Akademia Górniczo-Hutnicza

RECENZENCI: prof. dr hab. inż. Adam Grzech – Politechnika Wrocławska
prof. dr hab. inż. Edward Nawarecki – Akademia Górniczo-Hutnicza

Rozprawa doktorska poświęcona jest konstrukcji algorytmów detekcji zdarzeń, ukierunkowanych na usprawnienie predykcji niestacjonarnych szeregów czasowych. Problem wczesnej sygnalizacji zdarzeń w szeregach ma bezpośredni wpływ na jakość prowadzonego przetwarzania danych. Związany jest on z minimalizowaniem opóźnienia detekcji bądź szybką estymacją parametrów statystycznych szeregu, co ma szczególne znaczenie dla uzyskiwania wiarygodnych prognoz. Autor podjął próbę poszukiwania możliwości usprawnień algorytmicznej analizy rozległych zasobów danych, ukierunkowanej na wykrywanie zdarzeń poprzedzających długoterminowe zmiany właściwości statystycznych szeregów czasowych, a także na rozproszoną analizę środowiska z wykorzystaniem zdywersyfikowanego zestawu detektorów. Praca zawiera opis algorytmu detekcji zdarzeń, inspirowanego działaniem naturalnych układów odpornościowych oraz komplementarnych detektorów krótkoterminowych zmian w szeregach (pojawiających się współbieżnie bądź ze zmiennym opóźnieniem). Przedstawiono wyniki badań skuteczności zaproponowanych mechanizmów detekcji zdarzeń, przeprowadzone na danych symulowanych i rzeczywistych (danych giełdowych, dla których wykazano zwiększenie skuteczności średnioterminowej predykcji ekstrapolacyjnej), pokazujące zasadność przyjętej w pracy koncepcji.

Wstęp

W zadaniach zarządzania, a także sterowania nadrzędnego procesami, kluczowe znaczenie ma prawidłowe podejmowanie decyzji eksperckich (operatorskich). Ich skuteczność zależy od dostępności informacji i jej sprawnej selekcji, w tym wczesnego sygnalizowania zdarzeń istotnych. Problem wczesnej sygnalizacji takich zdarzeń ma bezpośredni wpływ na jakość prowadzonego przetwarzania danych. Związany jest on z minimalizowaniem opóźnienia detekcji bądź, co ma szczególne znaczenie dla uzyskiwania wiarygodnych prognoz krótko i średnioterminowych, z szybką estymacją parametrów statystycznych szeregu.

Możliwości pozyskiwania, gromadzenia oraz udostępniania olbrzymich zasobów informacji stwarzają konieczność algorytmizacji selekcji informacji, w tym detekcji zdarzeń polegających na występowaniu specyficznych sekwencji próbek rejestrowanych szeregów. Poszerzanie zbioru przetwarzanych szeregów, nawet jeśli pochodzą z heterogenicznych źródeł, niekoniecznie implikuje istotny wzrost skuteczności detekcji, z uwagi na występujące często podobieństwo szeregów i koincydencję zdarzeń istotnych.

Cechą charakterystyczną środowiska reprezentowanego przez szeregi finansowe są częste, znaczące zmiany losowe ich parametrów statystycznych, w szczególności trendów średnioterminowych mogących być podstawą prognozowania krótko i średnioterminowego (w wielu pracach zakłada się wręcz, że szeregi finansowe są procesami błądzenia przypadkowego). Niemniej uzasadnione jest założenie, że zmiany istotne są poprzedzane pewnymi zdarzeniami symptomatycznymi, które są jednak trudne do jednoznacznego sprecyzowania i występują ze zmiennym wyprzedzeniem. Zdarzenia symptomatyczne mogą charakteryzować się różnymi cechami, co wymusza stosowanie zróżnicowanych metod ich detekcji. Selekcja względnie rzadkich zdarzeń istotnych wymaga analizy wielu zdarzeń występujących przypadkowo, co w połączeniu z ogromnym zestawem dostępnych szeregów uniemożliwia ich kompleksowe przetwarzanie. Sugeruje to potrzebę zastosowania rozproszonych mechanizmów detekcji, implementowanych w formie przekrojowych analiz krótkich segmentów szeregów.

Zagadnienia konstrukcji algorytmów detekcji zdarzeń w szeregach czasowych są ważnym i szybko rozwijającym się obszarem badawczym informatyki. Prowadzone w wielu ośrodkach naukowych badania związane z usprawnieniem jakości prognoz szeregów finansowych nie doprowadziły do znaczącej poprawy wiarygodności predykcji, z uwagi na niemożność opracowania efektywnego predyktora uniwersalnego.

Istotnym mankamentem predykcji szeregów finansowych metodami statystycznymi (ekstrapolacji składowych okresowych i długoterminowego trendu, predykcji przyrostów miesięcznych w oparciu o wieloczynnikowe modele ARMAX) jest konieczność wykorzystania długich ciągów danych do identyfikacji współczynników predyktora. Powoduje to zbyt wolną adaptację formuły prognostycznej do ustawicznie zmieniających się właściwości statystycznych szeregu, a w przypadku silnych zmian – długookresową utratę adekwatności predyktorów

Analiza literatury dotyczącej detekcji zdarzeń oraz predykcji szeregów czasowych wskazała na potrzebę podjęcia badań nad algorytmami automatycznie przetwarzającymi dane pochodzące ze źródeł heterogenicznych (rozległe zbiory szeregów), wykorzystującymi metody statystyczne oraz odpowiednie mechanizmy adaptacji, umożliwiające selekcję źródeł informacji i dostrajanie parametrów klasycznych procedur analizy statystycznej.

Cel, tezy oraz zakres pracy

Ideą przewodnią rozprawy było poszukiwanie sposobów stałej, samoczynnej adaptacji parametrycznej i strukturalnej predyktorów z wykorzystaniem informacji generowanych na podstawie analizy możliwie krótkich segmentów szeregów, prowadzonych w rozległym środowisku. Zadanie to ma wyraźną analogię do funkcji obronnych systemów immunologicznych. Względnie krótkie przedziały czasu, w których szeregi czasowe mogą być przewidywalne na podstawie stosunkowo prostych metod statystycznych (ekstrapolacja trendów, modele sygnałowe ARMAX), można postrzegać jako stan zdrowia środowiska (brak istotnej autokorelacji jednokrokowych przyrostów szeregu charakteryzowanych w szerokim oknie). Natomiast często nieokreślone czynniki zewnętrzne mogą być przyczyną utraty aktualności adekwatnych wcześniej predyktorów, co można postrzegać jako stan choroby.

Na podstawie powyższych przesłanek sformułowano następująca tezę główną rozprawy: *Zastosowanie podejścia immunologicznego do konstrukcji adaptacyjnych algorytmów wykrywania zdarzeń w rozległych zbiorach szeregów czasowych umożliwia usprawnienie średnioterminowej predykcji takich szeregów przez zmniejszenie opóźnienia adaptacji parametrów predyktora po silnych zakłóceniach w ich otoczeniu*. Sformułowano także sześć pomocniczych tez roboczych.

Celem rozprawy było opracowanie nowych algorytmów detekcji zdarzeń, ukierunkowanych na usprawnienie krótko i średnioterminowej predykcji niestacjonarnych szeregów finansowych poprzez algorytmiczną analizę rozległych zasobów danych, ukierunkowaną na wykrywanie zdarzeń krótkoterminowych, poprzedzających istotne, długoterminowe zmiany właściwości statystycznych szeregów. Jako główną koncepcję badawczą przyjęto, że podstawą do usprawnienia prognoz może być wieloaspektowa analiza zdarzeń w otoczeniu badanego szeregu, zwiastujących zmiany jego trendu. Do detekcji takich zdarzeń i ich algorytmicznej interpretacji wykorzystano paradygmat immunologiczny, zgodnie z którym uzyskanie wysokiej wiarygodności detekcji zdarzeń istotnych w zmiennym otoczeniu o słabo zdeterminowanej strukturze i właściwościach można osiągnąć przez odwzorowanie działania wybranych mechanizmów naturalnych systemów odpornościowych. Proponowane podejście opiera się na przetwarzaniu reprezentatywnego zestawu szeregów czasowych, połączonym z losowym doborem sygnałów do bieżącej analizy, umożliwiającym adaptację strukturalną środowiska poddawanego analizie.

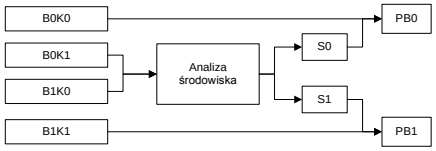
Do prognozowania średnioterminowego zastosowano ekstrapolację lokalnego trendu liniowego prognozowanego szeregu. W tym ujęciu składową cykliczną i trend długookresowy aproksymuje się funkcją odcinkowo-liniową z wykorzystaniem uogólnionej metody najmniejszych kwadratów. Detekcja załamań trendów średnio i długookresowych z wysoką skutecznością (stwierdzoną we wcześniejszych badaniach) i dopuszczalnie małym opóźnieniem prowadzona jest w oparciu o uogólnione testy największej wiarygodności (stosunku funkcji wiarygodności LR).

Zarys koncepcji detekcji zdarzeń z wykorzystaniem paradygmatu immunologicznego

W rozdziale 4 rozprawy zdefiniowano oryginalną koncepcję immunopodobnego algorytmu detekcji zdarzeń, przeznaczonego do wspomagania średnioterminowej predykcji. Scharakteryzowano podstawowe obiekty systemu oraz metody detekcji krótkoterminowych zmian. Zaproponowano nowe miary chwilowego podobieństwa szeregu i przeprowadzono obszerne analizy skuteczności detekcji na danych symulowanych oraz rzeczywistych (szczegółowe wyniki badań umieszczono w załączniku 4).

Immunopodobne przetwarzanie danych ukierunkowane jest na rozróżnienie pomiędzy stanem zdrowia i choroby, a następnie eliminację efektów choroby (dla szeregów czasowych jest to np. poprawna estymacja parametrów lokalnego trendu). W stanie choroby (przyrosty szeregów niestacjonarne lub wykazujące chwilową autokorelację) wykorzystanie do predykcji ekstrapolacyjnej wcześniej wyznaczonych parametrów trendu stwarza ryzyko wystąpienia nieakceptowalnie dużych błędów prognoz. Wydzielenie nowego segmentu szeregu o odpowiednio długim hipotetycznym czasie adekwatności można widzieć jako przywrócenie stanu zdrowia.

Nadrzędnym celem detekcji jest wykrycie zdarzenia zwiastującego zmiany parametrów trendu szeregu prognozowanego – szeregu bazowego B. Ze środowiska szeregów S wybiera się stosunkowo mało liczny podzbiór – otoczenie K, okresowo modyfikowane (adaptacja strukturalna). Podstawą detekcji jest równoczesna analiza par segmentów sygnałów o zadanej długości (przyjęto długość równą 22. próbki) – komórek systemu BK. Typowo, w czasie rzeczywistym analizowana jest względnie niewielka liczba komórek BK. Pierwszym etapem analizy jest detekcja zdarzeń w szeregach B i K (patrz rys. 1) w celu stwierdzenia sytuacji: B0K0 (brak zdarzeń); B1K1 (zdarzenia w obu szeregach); B0K1 (zdarzenie w K) lub B1K0 (zdarzenie w B). Zadaniem algorytmu jest stwierdzenie stanu zdrowia PB0 (parametry statystyczne znane, możliwość prowadzenia predykcji średnioterminowej na podstawie segmentacji LR) lub stanu choroby PB1 (potrzeba szybkiej identyfikacji zdarzenia w szeregu B: zmiana parametrów predyktorów) z wykorzystaniem dedykowanych metod badania podobieństwa.



Rys.1 Ogólny schemat detekcji zdarzeń w środowisku szeregu B.

W sytuacji nieokreślonej (B0K1 lub B1K0) konieczne jest poszerzenie zbioru BK w celu uzyskania ostatecznej diagnozy: potwierdzenia obecności zdarzenia (S1) lub jego braku (S0). Wykorzystuje się tutaj dodatkowe, bardziej złożone mechanizmy detekcji (adaptacja strukturalna). Dodatkowo prowadzone są analizy w tle dla komórek zawierających szeregi losowe w celu obliczenia referencyjnych wartości miar odległości. Zwiększenie szybkości oraz skuteczności wykrywania stanów PB0/PB1 winno być wspomagane wykorzystaniem pamięci rejestrującej informacje o momentach załamania trendu, opóźnieniu detekcji, zmianach korelacji pomiędzy sygnałami oraz zmianach chwilowego podobieństwa.

Badania prezentowane w rozprawie skoncentrowano na konstrukcji skutecznego algorytmu bezpośredniej detekcji stanu choroby (B1K1 $\rightarrow$ PB1).

Konstrukcja limfocytów

Autor zaproponował zastosowanie trzech rodzajów limfocytów, realizujących zadania detekcji krótkotrwałych zmian w szeregach (limfocyty L1 i L2) oraz prowadzących

analizę otoczenia (L3) – limfocyty typu L3 nie były celem obliczeń numerycznych prowadzonych w ramach rozprawy.

Limfocyt L1 aktywowany jest synchronicznie. Bada zmiany stacjonarności ciągu jednopróbkowych przyrostów szeregu przez porównywanie błędów jednokrokowej predykcji przyrostów z wykorzystaniem trójparametrowej metody Holta (e_H), dedykowanej do prognozowania ciągów niestacjonarnych, z błędami prognozy podtrzymania zerowego rzędu ZOH (e_{ZOH}), adekwatnej dla predykcji szeregów zbliżonych do procesu Wienera. Drugim mechanizmem L1 jest detekcja serii istotnych odchyłek od wartości średniej przyrostów sprawdzanych testem Studenta. Wykrycie zdarzenia następuje, gdy $e_{ZOH} > e_H$ i/lub wykryto serię istotnych odchyłek.

Limfocyt L2 jest aktywowany asynchronicznie po wykryciu zdarzenia przez L1. Realizuje zadania analizy podobieństwa zdarzeń w szeregach komórki, z zastosowaniem dodatkowych transformacji szeregów. Dla uzyskanych sygnałów diagnostycznych obliczany jest wskaźnik ich podobieństwa i potwierdzana jest istotność zdarzenia (B1K1) albo stwierdzana niezgodność sygnałów L1 lub anulowana wstępna diagnoza L1 (zmniejszenie prawdopodobieństwa fałszywych alarmów).

Autor zdefiniował uwarunkowania pracy limfocytu L3, przeznaczonego do adaptacji parametrów testów statystycznych LR, agregacji informacji przesyłanych ze wszystkich L2 oraz obliczania statystyk skuteczności detekcji. Efektem działania L3 jest sygnał blokujący akceptację prognoz lub zezwalający na kontynuację predykcji. W zależności od obciążenia czasowego systemu, L3 prowadzi retrospektywne analizy efektywności środowiska na podstawie wcześniej stwierdzonych zdarzeń oraz dokonuje eliminacji mało efektywnych limfocytów L2. Mogą być również modyfikowane na bieżąco kryteria reaktywacji prognozowania przez modyfikację długości ciągu próbek, wymaganej do wiarygodnej estymacji trendu w nowym segmencie.

Warunkiem skutecznej detekcji zdarzeń przez limfocyty L1 i L2 jest wykorzystanie w przetwarzaniu wielu informatywnych sygnałów diagnostycznych. Zbiór szeregów tworzących środowisko komórki (*zewnętrznych sygnałów diagnostycznych*) generowany jest na podstawie globalnych przekształceń szeregów oryginalnych. W pracy wykorzystano jednokrokowe przyrosty logarytmiczne, zwykłe, filtrację dolnoprzepustową szeregów oryginalnych (filtr MNK) oraz ciągi błędów jednokrokowych ekstrapolat długookresowego trendu i składowych cyklicznych (okres 2088 próbek).

Fragmenty szeregów, ujęte w oknie komórki BK, mogą być poddawane dalszym transformacjom w celu uzyskania *wewnętrznych sygnałów diagnostycznych*, stanowiących dane wejściowe w analizach podobieństwa, uwypuklających określone cechy zdarzeń. Transformacje te są dwojakiego rodzaju: (1) obligatoryjna uniformizacja sygnałów (uzyskanie scentrowanych ciągów próbek tego samego rzędu wielkości; ekstrakcja średniej lub detrending) oraz (2) opcjonalne transformacje uwypuklające lub maskujące określone cechy zunifikowanych szeregów komórki (podzielenie przez wartość referencyjną: odchylenie standardowe szeregu resztowego w komórce, wartość maksymalną modułów, stałą wartość będącą parametrem detektora). Zuniformizowane sygnały w oknie mogą być poddawane kolejnym *transformacjom wtórnym*, spośród których w pracy wykorzystano zaokrąglenie do wartości całkowitoliczbowych ciągów rezidualnych podzielonych przez dyspersję, zmianę znaków wartości ujemnych na dodatnie, rangowanie sygnałów rezidualnych.

Struktura limfocytu L2 jest określona przez iloczyn kartezjański (*MP* x *TS*), gdzie *MP* – zbiór detektorów, natomiast *TS* – zbiór typów sygnałów diagnostycznych uzyskiwanych w wyniku transformacji szeregów komórki. Na dostępny zbiór detektorów *MP* (przestrzeń detektorów różniących się zastosowanymi w nich miarami podobieństwa szeregów) składa się w praktyce przestrzeń dostępnych metod badania podobieństwa (*TM*) oraz wektor ich parametrów (*WP*). Zbiór typów sygnałów diagnostycznych (*TS*) jest iloczynem kartezjańskim zbioru zewnętrznych sygnałów diagnostycznych (*TSD*), przestrzeni przekształceń globalnych (*TGS*), zbioru transformacji unifikujących stosowanych w oknie komórki (*TSU*) oraz przestrzeni *MTS* przekształceń wyostrzających lub łagodzących cechy wewnętrznych sygnałów diagnostycznych. Otrzymuje się w ten sposób bardzo liczny, zdywersyfikowany zbiór detektorów *ZD*, określony jako $ZD = TM \times WP \times TSD \times TGS \times MTS \times TSU$. Zróżnicowanie detektorów umożliwia detekcję różnych, niekonieczne jednoznacznie sprecyzowanych zdarzeń w komórkach, co pozwala m.in. rozpraszanie zadania detekcji zdarzeń.

Miary odległości szeregów

W rozprawie zaproponowano koncepcję wykorzystania *miar odległości* pomiędzy segmentami szeregów tworzącymi komórkę, jako kryterium algorytmicznej kwalifikacji sygnałów środowiska do tworzenia otoczenia szeregu bazowego oraz do dalszej jego analizy (działanie limfocytów L2).

Z uwagi na niemożność wykorzystania do tego celu miar klasycznych (założenie dużego podobieństwa potencjalnych anomalii, niska efektywność w przypadku przesunięć między zdarzeniami) skonstruowano i przeanalizowano pięć dedykowanych miar odległości szeregów, ukierunkowanych na badanie występowania wzorców w szeregach, podobieństwa kształtu szeregów oraz współwystępowania podobnych sekwencji: odległość sygnałów zdeterminizowanych (**Sd**), odległość widm Fouriera ($\mathbf{S_F}$), miarę odległości uelastycznionych wzorców (**W**), miarę opartą na porównywaniu zliczonych zdarzeń (**Z**) oraz miarę opartą na porównaniu zunifikowanych wzorców (**U**).

Do pomiaru chwilowej odległości szeregów komórki zaproponowano zastosowanie odległości widm amplitudowych. Odległość Fouriera d_F. (określana w pracy symbolem $\mathbf{S_F}$) pozwala zredukować wpływ zmienności opóźnień reakcji badanego szeregu na zdarzenia w szeregach należących do jego otoczenia (nie zależy od wzajemnych opóźnień szeregów). Niech A_{xi}, A_{yi} dla $i = 1,...,k_F$ (rząd metody) oznaczają amplitudy k_F niższych harmonicznych szeregów o długości N danych (długość N jest rzędu 22. próbek). Odległości d_F definiuje wzór:

$$d_F = \sum_{i=1}^{k_F} \left| A_{xi}^2 - A_{yi}^2 \right| \qquad (1)$$

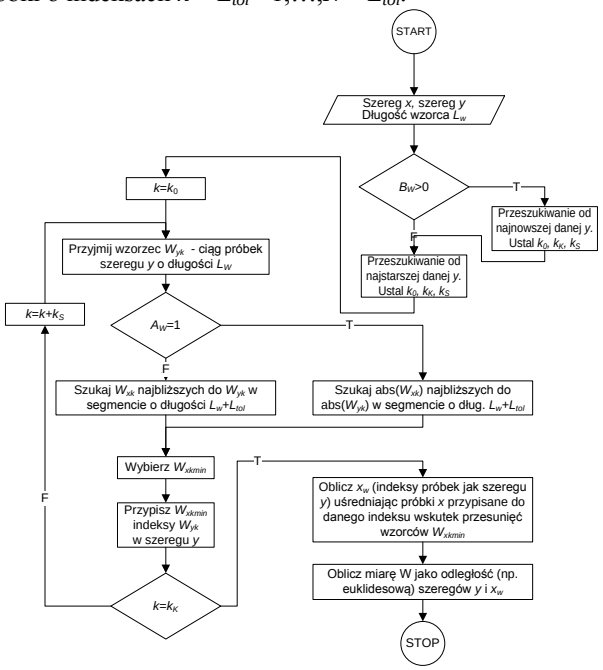
Istotą metody **Sd** (miara spektralna determinizowana) jest niwelowanie informacji o wzajemnym opóźnieniu zdarzeń w dwóch badanych szeregach, przez:

a) obliczenie transformat Fouriera obu badanych szeregów w oknie analizy o długości N,
b) obliczenie determinizowanych sygnałów o długości k_d próbek (k_d – rząd metody) przez zastosowanie transformaty odwrotnej dla widma amplitudowego (z wyzerowaniem przesunięcia fazowego wszystkich harmonicznych),

c) obliczenie odległości Minkowskiego dla uzyskanych w (b) sygnałów diagnostycznych.

Sygnał diagnostyczny (b) agreguje informację o kształcie i mocy dominujących serii odchyłek, pozwala wychwycić podobieństwo ze znaczną redukcją wpływu tła losowego i wzajemnych przesunięć serii.

Metoda pomiaru odległości typu **W** (scharakteryzowana na rys. 2) jest dedykowana do wykrywania w szeregach wzorców (sekwencji próbek) o określonej długości L_w. Badany jest szereg otoczenia x o długości $N + L_{tol}$ (L_{tol} – założona tolerancja) oraz szereg bazowy y o długości N, zawierający próbki o indeksach $n = L_{tol}+1,\ldots,N + L_{tol}$.



Rys.2 Schemat obliczania odległości typu W dla szeregów x oraz y (wersja podstawowa metody).

Głównymi parametrami metody W są: L_w (parametr podstawowy), A_W – binarna opcja sposobu dopasowywania wzorców ($A_W = 0$ – wzorce oryginalne; $A_W = 1$ – moduły wzorców) oraz kierunek analizy B_W. Miara W jest tolerancyjna na lokalne przesunięcia podobnych sekwencji występujących w obydwu szeregach. Nie jest tolerancyjna na różnice kształtu profili występujących w szeregu x, zatem opracowano bardziej rygorystyczną jej wersję, w której analizowane są tylko sekwencje W_x, których próbki nie były wcześniej przesuwane – przypisane do innych indeksów niż oryginalne ($W_{x\mathrm{kmin}}$ dopasowane do wcześniejszych wzorców W_{yk}). Zatem wynik analizy zależy od kierunku przeszukiwania wzorców. B_W przyjmuje jedną z trzech wartości: $B_W = 0$ (brak blokady wzorców), $B_W = –1$ (blokada z przeszukiwaniem od najstarszej danej y), $B_W = 1$ (blokada z przeszukiwaniem od najnowszej danej).

Metodę **U** ukierunkowano na detekcję zmian unikalnych, tj. analizę występowania serii skoków w szeregach komórki o amplitudzie przekraczającej arbitralnie ustaloną progową wartość ρ_U (główny parametr metody), o różnej długości serii (1,2,...,N) w porównywanych szeregach o długości N. Detekcja przebiega w następujących etapach:

a) Klasyfikacja serii istotnych odchyłek w szeregach: x oraz y wg długości serii i znaku odchyłek, z zapamiętaniem długości serii oraz jej liczności.

b) Zliczenie krotności występowania serii istotnych odchyłek o różnych długościach w szeregu x (L_{kx} – liczba serii o długości k) oraz y (odpowiednio – L_{ky}); $k = 1,2,\ldots,K_k$, K_k – największa długość wykrytej serii.

c) Obliczenie procentu zgodności w_{pzg} wykrytych serii skoków w szeregach, uwzględniającego krotność występowania poszczególnych serii oraz długość (L_{kx}, L_{ky}):

$$w_{pzg} = w_p \sum_{k=1}^{Kk} k \cdot \min(L_{kx}, L_{ky}) \tag{2}$$

gdzie

$$w_p = 1/\left(\sum_{k=1}^{K_k} k \cdot \max(L_{kx}, L_{ky})\right) \tag{3}$$

d) Końcowe obliczenie miary odległości $d_U = 1 – w_{pzg}$.

W metodzie **Z** dla szeregów x i y oblicza się średnie wartości bezwzględnych odchyłek dodatnich i ujemnych o module większym od założonej wartości progowej ρ_{zd} (dla odchyłek o module mniejszym niż ρ_{zd} przyjmuje się zero). Na postawie uzyskanych dwóch par wartości dodatnich (x_{dsr}, y_{dsr}) oraz ujemnych (x_{usr}, y_{usr}) oblicza się wskaźnik chwilowego podobieństwa: $d_Z = \sqrt{(x_{dsr} - y_{dsr})^2 + (x_{usr} - y_{usr})^2}$.

Zniwelowanie efektu zmiennego opóźnienia zdarzeń o charakterze przyczynowo-skutkowym uzyskano poprzez zastosowanie odpowiedniej *tolerancji* dla obliczanych wartości chwilowych miar odległości szeregów komórki. W badaniach przyjęto tolerancję $L_{tol} = 5$ próbek.

W celu wyeliminowania wpływu skali miar odległości skonstruowano wskaźnik odległości W_{Dz} (wzór 6), wykorzystany jako podstawowa informacja przesyłana przez L2. W_{Dz} rośnie ze wzrostem chwilowej miary odległości D_{BK} i równocześnie ją skaluje według uśrednionej wartości obliczonej dla losowych sygnałów nieskorelowanych o standardowym rozkładzie Gassa ($D_{BK_{B0K0}R_0}$):

$$W_{Dz} = \exp\left(-\frac{1}{2} \cdot \frac{D_{BK_{B0K0}R_0}}{D_{BK}}\right) \tag{4}$$

Dla ułatwienia analizy skuteczności detektorów zaprojektowano wskaźnik podobieństwa, który uwzględnia wartości referencyjne wyników badania podobieństwa w przypadku kontrolowanego umieszczenia w sygnałach skoków o określonej charakterystyce (w_{ref} przyjmuje wartości 0, gdy szeregi są podobne, a wartość 1 dla szeregów niepodobnych). Wyraża się on wzorami:

$$W_{Dt} = \frac{W_{Dz} - w_{pr}}{1 - w_{pr}}, \quad w_{pr} = \exp(-1) \tag{5}$$

$$W_{Dm} = 1 - \left|w_{ref} - \min(W_{Dt}, 0)\right| \tag{6}$$

W_{Dt} przyjmuje wartości 1 dla szeregów bardzo podobnych, a wartości ujemne dla szeregów mniej podobnych niż nieskorelowane ciągi losowe. Z kolei W_{Dm} jest miarą trafności rozpoznania podobieństwa szeregów, z zaniedbaniem zróżnicowania szeregów większego niż w przypadku szeregów losowych.

Zmniejszenie wrażliwości W_{Dm} na zróżnicowanie miary odległości szeregów bardzo bliskich i bardzo odległych osiągnięto przez wykorzystanie wskaźnika W_{Def}:

$$W_{Def} = \frac{1}{2}\left(1 + \sin\left((W_{Dm} - 0.5)\pi\right)\right) \tag{7}$$

W_{Def} zmniejsza zróżnicowanie miary odległości w pobliżu wartości ekstremalnych 0 i 1. Wysoka skuteczność danej metody pomiaru odległości oznacza wartość W_{Def} bliską 1, skuteczność zadowalająca osiągana będzie dla $W_{Def} > 0.5$.

Przedstawione w pracy wyniki analizy podobieństwa szeregów komórki przekształcano do postaci zestandaryzowanej wskaźnikiem (7) oraz (4) (symulacyjne badania skuteczności miar odległości) bądź tylko

wskaźnikiem (4) (badanie skuteczności algorytmu detekcji na danych empirycznych, tj. rzeczywistych szeregów czasowych).

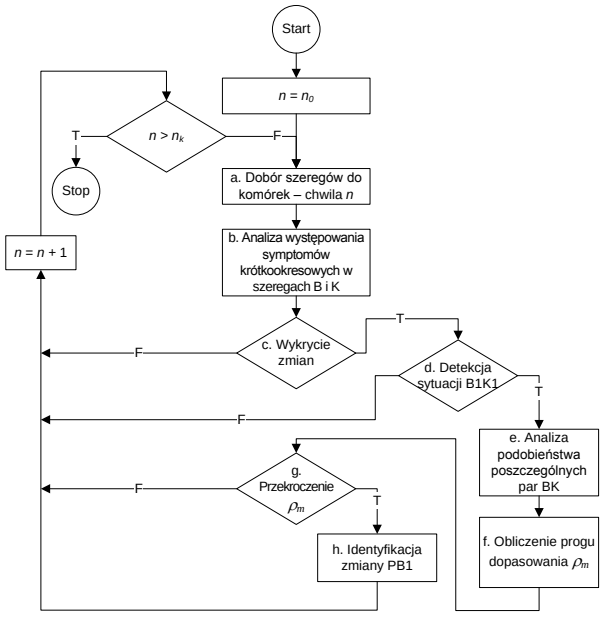
Obliczenia numeryczne prezentowane w rozprawie

W pierwszym etapie badań (rozdział 4.6.5) przeprowadzono analizy skuteczności detekcji zmian w szeregach wykorzystując zaproponowane miary odległości szeregów komórki dla sygnałów przekształconych z zastosowaniem sześciu transformacji unifikujących oraz z jedną transformacją unifikującą, uznaną za najkorzystniejszą. Zastosowano wewnętrzne obligatoryjne i dodatkowe przekształcenia sygnałów w oknie komórki, a także transformacje dodatkowe. Badania przeprowadzono na danych symulowanych $N(0,1)$ oraz dla przykładowych szeregów finansowych. Dla określonych próbek wprowadzono duże odchyłki (przekraczające 3σ) w różnych konfiguracjach.

Stwierdzono wysoką skuteczność miar Sd i S_F o niskim rzędzie – najwyższą dla zdarzenia obecnego w jednym szeregu. Miara W wykazała wysoką skuteczność szczególnie dla wzorców współbieżnych (wykrywanie niezgodności sygnałów zwiększa się przy założeniu blokowania wzorców). Wyniki wskazały, że miary Z oraz U wymagają ustawienia niskiego progu detekcji.

Ponadto, sprawdzono skuteczność detekcji dla różnych konfiguracji wzorców zdarzeń pojawiających się w szeregach BK. Szczególnie dobre rezultaty uzyskano metodą W oraz Sd. Wskazano na istotny wpływ zdarzeń występujących współbieżnie oraz zmiennych opóźnień na skuteczność detekcji. Wykazano brak zasadności stosowania dodatkowych przekształceń wtórnych oraz sygnałów przekształconych na wartości bezwzględne w dalszych analizach. Stwierdzono też ograniczoną możliwość skutecznej detekcji zmian obecnych w szeregach finansowych z wykorzystaniem klasycznych metod odległościowych.

W drugim etapie badań (szczegółowo opisanym w rozdziale 4.7 rozprawy) przeprowadzono analizy efektywności proponowanego algorytmu detekcji zdarzeń. Schemat działania algorytmu realizującego zadanie detekcji zmian typu B1K1 oraz identyfikację stanu PB1 przedstawiono na rysunku 3. Blok (b) reprezentuje działanie limfocytów L1, natomiast bloki (e), (f) – limfocytów L2.



Rys. 3 Schemat działania algorytmu detekcji zdarzeń typu B1K1.

Szeregi tworzące komórki dobierano z wykorzystaniem odległości City. Liczbę generowanych komórek n_K ustalano arbitralnie. Wstępną detekcję B1K1 prowadzono w oparciu o wykrycie przez L1 zmian w szeregach B i K. Działanie L2 oparto na zaproponowanych przez autora miarach odległości, przy założeniu losowości parametrów oraz typu sygnału diagnostycznego poddawanego analizie. Identyfikację stanu PB1 stwierdzano na podstawie analizy progu dopasowania ρ_m (frakcji komórek, w których potwierdzono detekcję zdarzenia).

Badania symulacyjne przeprowadzono dla dwóch losowych sygnałów $N(0,1)$, nieskorelowanych oraz z korelacją bliską 1, do których wprowadzono symulowane wzorce zdarzeń. Sygnały komórki poddano uniformizacji poprzez ekstrakcję trendu w oknie analizy i skalowanie wariancją przyrostów w okresie wcześniejszym. Zastosowano miary S_F (dla rzędu $k_F = \{2; 3\}$), Sd (dla rzędu równego długości analizowanego wzorca – $k_d = \{3; 4; 5\}$), W ($A_W = 1$, $B_W = 0$, $L_w = \{3; 4; 5\}$), Z oraz U (wartość progów ρ_U oraz ρ_Z: podwójna wartość odchylenia standardowego szeregów losowych). Obliczono także podobieństwo z wykorzystaniem odległości klasycznych: Euklidesa oraz Kendalla.

Wykazano efektywność zaproponowanych miar, zależną od sposobu transformacji wtórnej sygnału komórki, atrybutów wzorca oraz sposobu generowania sygnałów. Stosowanie transformacji do niektórych miar odległości (m.in. W oraz U) warunkuje prowadzenie wiarygodnych analiz.

Największą skuteczność detekcji osiągnięto dla miar W, Sd oraz S_F, dla wariantu zdarzeń w obydwu szeregach. Pokazano, że identyfikacja zmian w jednym szeregu wymaga stosowania zróżnicowanych metod detekcji oraz transformacji wtórnych. Zidentyfikowano miary szczególnie przydatne do wykrywania określonych wzorców. Potwierdzono komplementarność skonstruowanych detektorów.

W rozdziale 4.7.4 rozprawy opisano założenia oraz wyniki przeprowadzonego badania skuteczności algorytmu detekcji zdarzeń dla danych empirycznych (64. szeregów finansowych, pochodzących z heterogenicznych źródeł – 153. sygnałów diagnostycznych uzyskanych w wyniku przekształceń globalnych). Obliczenia wykonano dla 24. szeregów bazowych i powtórzono dla różnych wartości arbitralnie ustawianych parametrów: trzech wariantów liczebności komórek otoczenia szeregu B ($n_K = \{10;20;30\}$), pięciu przesunięć szeregu K względem B (0,5,10,15,20 próbek), trzech wartości progu dopasowania $\rho_m = \{0.3;0.4;05\}$. Informację przesyłaną przez L2 (W_{Dz}) zamieniono na postać binarną – zbadano dwie wartości progu podobieństwa $\rho_D = \{0.1;0.5\}$ (dla $W_{Dz} < \rho_D$ przyjmowano zerową odległość szeregów; w sytuacji odwrotnej – odległość jednostkową). Parametry metod oceny odległości przyjęto jak we wcześniejszych badaniach symulacyjnych.

Skuteczność detekcji mierzono poprzez odniesienie chwili wykrycia zdarzenia (potwierdzenia B1K1) do momentów załamania trendu, wykrytych retrospektywnie z wykorzystaniem rygorystycznego testu największej wiarygodności (czynnika Bayesa) o niskim prawdopodobieństwie fałszywego alarmu. Zastosowany algorytm detekcji poprawia jakość prognozy ekstrapolacyjnej w przypadku detekcji zmian w okresie poprzedzającym (do 22. próbek) moment faktycznej zmiany trendu (S_{zw} – suma detekcji zdarzeń w tym przedziale) lub chwilę wykrycia takiej zmiany (S_{LR} – suma aktywacji L2 w takim okresie). Odnosząc S_{zw} i S_{LR} do sumy S_A wszystkich wykrytych zmian B1K1, obliczano wskaźnik skuteczności detekcji:

$$Ws_{\text{ef}} = \frac{S_{LR} + S_{zw}}{S_A} \cdot 100\% \quad (8)$$

Na podstawie zestawień szczegółowych (umieszczonych w załączniku 4 pracy) obliczono zbiorcze, uśrednione wskaźniki efektywności W_{Sef} (tab. 1-3).

Przesunięcie		0		5		10		15		20	
ρ_D		0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5
ρ_m	0.3	56.00%	59.15%	55.03%	57.21%	55.89%	57.52%	57.01%	56.31%	57.59%	57.02%
	0.4	56.63%	59.08%	57.93%	58.01%	55.62%	58.51%	52.56%	58.25%	53.16%	57.32%
	0.5	53.33%	58.92%	59.82%	59.08%	56.45%	59.12%	50.85%	55.52%	51.67%	57.39%

Tab. 1 Zbiorcze wyniki pomiaru efektywności algorytmu detekcji zdarzeń (W_{Sef}) dla liczby generowanych komórek $n_K = 10$.

Przesunięcie		0		5		10		15		20	
ρ_D		0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5
ρ_m	0.3	56.04%	56.31%	58.17%	56.36%	56.54%	57.25%	56.90%	57.56%	58.74%	57.43%
	0.4	57.89%	56.28%	60.68%	56.44%	57.19%	57.01%	57.36%	58.23%	54.92%	56.11%
	0.5	54.69%	58.88%	58.73%	58.59%	65.00%	57.30%	63.41%	57.55%	47.22%	56.51%

Tab. 2 Zbiorcze wyniki pomiaru efektywności algorytmu detekcji zdarzeń (W_{Sef}) dla liczby generowanych komórek $n_K = 20$.

Przesunięcie		0		5		10		15		20	
ρ_D		0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5
ρ_m	0.3	55.16%	57.12%	56.36%	56.91%	55.40%	58.50%	55.43%	57.01%	56.17%	56.69%
	0.4	61.19%	56.31%	54.82%	55.86%	59.48%	57.35%	55.51%	58.11%	56.96%	55.36%
	0.5	62.50%	60.68%	60.87%	57.89%	72.00%	56.98%	54.84%	59.68%	56.67%	56.80%

Tab. 3 Zbiorcze wyniki pomiaru efektywności algorytmu detekcji zdarzeń (W_{Sef}) dla liczby generowanych komórek $n_K = 30$.

Wyniki obliczeń numerycznych potwierdziły zasadność stosowania proponowanej w pracy koncepcji detekcji zdarzeń. Dla większości przypadków wskaźnik W_{Sef} przyjmuje wartości powyżej 50% (połowa wykrytych zmian typu B1K1 jest trafnym zwiastunem zmian trendu).

Dla wybranych szeregów lokalnych oraz dla większości szeregów cen surowców wartości W_{Sef} dochodzą do 100%, co oznacza trafny sygnał zwiastujący lub krótkie opóźnienie detekcji dla danego szeregu bazowego. Dla szeregów zagregowanych (głównie SP500, DAX, FTSE, DJI) oraz walut (USD, GBP) uzyskano skuteczność, średnio rzędu około 30%. Wynika ona ze sposobu konstruowania takich wskaźników (informacja zagregowana). Niemal binarna zależność W_{Sef} od wartości parametrów algorytmu detekcji wskazuje na szybkie i pewne sygnalizowanie wystąpienia zdarzenia w szeregu.

Zwiększenie liczby generowanych komórek (n_K) powoduje wzrost efektywności detekcji (sięgający ponad 70% dla niektórych konfiguracji wartości parametrów), przy jednoczesnym zwiększeniu czasochłonności obliczeń. Znaczny przyrost czasu obliczeń (tab. 4) nie przekłada się na analogiczny przyrost skuteczności detekcji.

Przesunięcie	0		5		10		15		20	
ρ_D	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5	0.1	0.5
$n_K = 10$	74.69	94.90	95.95	95.14	96.50	95.05	97.80	95.65	103.54	95.59
$n_K = 20$	128.42	127.36	126.62	127.00	127.76	126.12	127.05	125.96	127.20	132.72
$n_K = 30$	184.81	183.80	184.56	175.90	183.72	182.67	184.14	182.14	185.10	182.79

Tab. 4 Uśrednione czasy (w sekundach) działania algorytmu dla pięciu wartości przesunięć oraz dwóch progów podobieństwa ρ_D.

Istotną rolę w zadaniu detekcji zmian B1K1 ma dobór opóźnienia (przesunięcia) analizowanych fragmentów szeregów wchodzących w skład komórek BK. Najlepsze rezultaty uzyskano dla przesunięcia wynoszącego co najmniej 5 próbek.

Podsumowanie

Przeprowadzone w ramach rozprawy badania potwierdziły ogólną koncepcję badawczą, że skuteczną drogą do usprawnienia predykcji średnioterminowej szeregów finansowych może być konstrukcja rozproszonych, immunopodobnych algorytmów detekcji zdarzeń zwiastujących załamania trendu, z wbudowanymi mechanizmami samoadaptacji parametrycznej i strukturalnej.

Wykazano skuteczność detekcji zmian krótkoterminowych w oparciu o identyfikację podobieństwa krótkich odcinków szeregów. Komplementarność detektorów warunkowana jest stosowaniem różnych przekształceń sygnałów diagnostycznych w komórce, a także niewrażliwością na losowe przesunięcia sygnałów.

Proponowane detektory zwiastują istotne załamania trendu z prawdopodobieństwem znacząco większym niż 0.5, co umożliwia między innymi skrócenie opóźnienia adaptacji parametrycznej predyktorów. Losowy dobór sygnałów, ich transformacji oraz metod detekcji chwilowych zdarzeń jest więc obiecującą drogą do skutecznego wykrywania zdarzeń zwiastujących istotne zmiany trendu w szeregach.

Wyniki uzyskane dla trudno prognozowalnych szeregów finansowych pozwalają przyjąć, że implementacja takich algorytmów może być celowa także do wykrywania uszkodzeń w systemach technicznych na podstawie kompleksowej analizy rezidualnej zmiennych procesowych.

W rozprawie podjęto tylko wybrane wątki ogólnej koncepcji badawczej, wskazując celowość przeprowadzenia dalszych badań, zarówno w obszarze usprawnienia działania detektorów zdarzeń krótkoterminowych, jak i implementacji adaptacyjnego predyktora. Problemy te będą przedmiotem dalszych, planowanych badań autora.

Publikacje doktoranta związane z rozprawą:

[1] Pełech-Pilichowski T., Duda J.T.: *Wykorzystanie podejścia immunologicznego do prognozowania szeregów czasowych*. Automatyka: półrocznik Akademii Górniczo-Hutniczej im. Stanisława Staszica w Krakowie; 2009

[2] Duda J.T., Pełech-Pilichowski T.: *Miary podobieństwa szeregów czasowych w detekcji zdarzeń*. [W:] Systemy wykrywające, analizujące i tolerujące usterki / red. Kowalczuk Z., Pomorskie Wydawnictwo Naukowo-Techniczne PWNT, 2009

[3] Pełech-Pilichowski T., Duda J.T.: *General Structure of T-Lymphocyte Applied to Immune-Based Event Detection in Financial Time Series*. [W:] Proceedings of the International Multiconference on Computer Science and Information Technology / eds.: M. Ganzha [et al.], November 6-10, 2006

[4] Pełech T., Duda J.: *Event Detection in Financial Time Series By Immune-Based Approach*. [W:] Intelligent Information Processing and Web Mining / eds. Kłopotek M.A., Wierzchoń S.T., Trojanowski K., Advances in Soft Computing , Springer-Verlag, 2006

[5] Duda J.T., Pełech T.: *Wykrywanie zdarzeń w szeregach finansowych z wykorzystaniem metod statystycznych*. [W:] Inżynieria wiedzy i systemy ekspertowe, T.2 / red. Grzech A., Oficyna Wydawnicza Politechniki Wrocławskiej, 2006

[6] Pełech T.: *Adaptive Holt's forecasting model based on immune paradigm*. [W:] Poleznye iskopaemye Rossii i ih osvoenie / red. Sinkov L.S., T.167, cz.2, *Zapiski Gornogo Instituta*, Sankt Petersburg, 2006

[7] Pełech T., Duda J.T.: *Immune algorithm of stock rates parallel monitoring*. [W:] Systemy informatyczne i metody obliczeniowe w zarządzaniu / red. Duda J.T., Uczelniane Wydawnictwa Naukowo-Dydaktyczne AGH, 2005

[8] Pełech T., Duda J.T.: *Zastosowanie paradygmatu immunologicznego do monitorowania wskaźników giełdowych.* [W:] Zastosowania teorii systemów / red. Kochan E., WIMiR AGH, 2005