B A D A N I A O P E R A C Y J N E I D E C Y Z J E

Nr 3 2008

Adam KUCHARSKI*

EFEKTYWNOŚĆ ALGORYTMU GENETYCZNEGO JAKO NARZĘDZIA PROGNOZOWANIA SZEREGÓW CZASOWYCH

Jedną z najbardziej charakterystycznych cech algorytmów genetycznych jest ich elastyczność zastosowań. Prowadzone od 2005 roku badania dowiodły, że i w przypadku predykcji wykonywanej na podstawie szeregów czasowych można sięgnąć po to narzędzie.

W niniejszej pracy zbadamy efektywność zmodyfikowanego klasycznego algorytmu genetycznego w zakresie wyżej wspomnianej tematyki. Dokonamy tego zarówno z punktu widzenia prognozowanego szeregu, sięgając po prognozy wygasłe, jak i od strony mechanizmów tworzących sam algorytm, a wpływających na szybkość i jakość otrzymywanych rozwiązań.

Słowa kluczowe: *algorytm genetyczny, prognozowanie, rynek kapitałowy*

Algorytmy genetyczne jako metoda optymalizacji charakteryzują się stosunkowo niewielkim stopniem sformalizowania, co pozwala na dużą swobodę w kwestii doboru obszaru zastosowań. Dowiodły tego wcześniejsze badania, przeprowadzone przez autora [6], [7]. Powstaje wszakże pytanie o efektywność algorytmu jako kolejnego narzędzia do otrzymywania prognoz na podstawie szeregów czasowych. Podstawową zaletą wspomnianej metody jest jej duża niezależność od dekompozycji szeregu, lecz otwarta pozostaje na przykład kwestia jakości prognoz.

Przyjrzymy się efektywności zmodyfikowanego klasycznego algorytmu genetycznego pod kątem wydajności obliczeń (analizując wewnętrzne mechanizmy algorytmu wpływające na czas potrzebny na znalezienie rozwiązania[1]) oraz jakości otrzymywanych prognoz.

* Społeczna Wyższa Szkoła Przedsiębiorczości i Zarządzania, ul. Sienkiewicza 9, 90-113 Łódź, e-mail: adamk@uni.lodz.pl

[1] Obliczenia przeprowadzono za pomocą programu napisanego w języku VBA, operującego w środowisku arkusza MS Excel.

1. Zastosowany algorytm genetyczny i jego modyfikacje

Zastosowany algorytm podobny jest w swoim działaniu do metod naiwnych, używanych w prognozowaniu niestrukturalnym. Wspomniana grupa modeli bazuje na założeniu, iż czynniki determinujące zjawisko do tej pory, zachowają swój wpływ w najbliższej przyszłości na dotychczasowym poziomie i w niezmieniony sposób [1], [2]. Skutkuje to tym, że np. prognoza *ex post* na dany okres równa jest wartości szeregu z okresu bezpośrednio poprzedzającego (jak w przypadku metody naiwnej prostej) lub wcześniejszego (metoda naiwna z sezonowością).

Prezentowany algorytm można traktować jako swoiste uogólnienie wspomnianego powyżej postępowania. Obie wymienione metody, będące inspiracją tworzonych chromosomów, podlegają silnym ograniczeniom, jeśli chodzi o dekompozycję szeregu czasowego. Zaleca się na przykład stosować je przy niewielkich wahaniach przypadkowych. Poza tym występuje wyraźne ograniczenie horyzontu czasowego danej tworzącej prognozę. Opisana w pracy metoda jest pod tym względem bardziej elastyczna, choć idea mechanizmu powstawania prognoz nie zmieniła się. Wśród danych z przeszłości poszukujemy wartości, która stanie się prognozą na bieżący okres. Wartość ta zostaje jednak uzmienniona i może zmieniać się w kolejnych prognozach. Nadal jednak predykcja *ex ante* dotyczy krótkiego horyzontu czasowego.

Algorytm genetyczny od metody naiwnej różni maksymalny, dopuszczalny numer opóźnienia danej rzeczywistej, będącej prognozą dla aktualnego okresu, który oznaczymy jako *tm*. Nie jest on bowiem stały, lecz przyjmuje wartości z pewnego przedziału. Parametr ów ustala się przed rozpoczęciem obliczeń i do ich końca pozostaje niezmieniony. Przyjmuje wartości z przedziału $<1, n - 2>$, gdzie n oznacza liczbę posiadanych obserwacji, co powoduje, iż pierwsza prognoza *ex post* powstaje dla trzeciego okresu. Nie wymagamy jednak, aby dana prognoza sięgała w przeszłość na maksymalną możliwą głębokość.

Zobrazujmy to następującym przykładem: Niech $\mathbf{y}_{(t)}$ oznacza pewien hipotetyczny szereg, dla którego liczba obserwacji $n = 5$. Wartość *tm* przyjmijmy na poziomie równym 3. Prognozowana wartość równa się jednej z obserwacji pochodzących sprzed $t - tm$ okresów, np. w okresie trzecim będzie to y_{t-1} lub y_{t-2}. Możliwe szeregi prognoz to m.in.:

1. $y_{t-1}, y_{t-1}, y_{t-2}$.
2. $y_{t-1}, y_{t-2}, y_{t-2}$.

Z uwagi na to, iż tworzonych w ten sposób potencjalnych szeregów prognoz przybywa lawinowo, należy użyć algorytmu genetycznego. Jego zadanie to znalezienie takiego zbioru opóźnionych wartości, który da jak najmniejszy błąd prognozy.

Ostatnia dana rzeczywista nie staje się prognozą *ex post*. Wykorzystywana jest dopiero w chwili wyznaczania prognoz *ex ante*, które tworzy się następująco:

1. Wyznaczamy prognozę *ex post* na podstawie kryterium, które stanowi wybrana miara błędów prognoz *ex post*[2].

2. Określamy wartość opóźnienia wskazującego, które dane rzeczywiste wezmą udział w tworzeniu prognoz *ex ante*. Do tego celu może posłużyć dominanta lub mediana wartości opóźnień wykorzystywanych podczas tworzenia prognoz *ex post*.

Odnosząc się do przedstawionego przykładu, w pierwszym przypadku mediana wynosi 1, a w drugim 2 okresy.

Dalsze postępowanie zależy od wyników dekompozycji szeregu czasowego. W razie wystąpienia stałego poziomu zmiennej prognozą *ex ante* może być wartość rzeczywista z ostatniego okresu lub któregoś z okresów wcześniejszych, wskazanych przez dominantę (względnie medianę) opóźnień. Ostatnia dana rzeczywista staje się ostatnią z prognoz *ex ante*.

Dla szeregów, w których stwierdzono trend liniowy wartość, o której była mowa wcześniej korygujemy o przyrost między ostatnią obserwacją, a tą wskazaną przez wybraną miarę statystyczną. Należy oczywiście pamiętać o krótkookresowym charakterze takich prognoz.

Algorytm genetyczny przeszedł modyfikacje stosowne do zaproponowanej metody tworzenia prognoz. Po pierwsze, zastosowaliśmy w nim kodowanie rzeczywiste. Pojedynczy chromosom reprezentuje jeden z wariantów predykcyjnych, zaś gen zawiera informację o opóźnieniu danej tworzącej prognozę. Na przykład chromosom dający pierwszą z prognoz w prezentowanym wcześniej przykładzie wyglądałby następująco: $Ch_1 = [2\ 3\ 3]$, a po rozkodowaniu otrzymamy wektor zawierający już konkretne prognozy *ex post*.

Odpowiada to jednej z możliwych prognoz *ex post* za okresy <3, *n*>.

Funkcję przystosowania, a tym samym kryterium oceny jakości prognoz *ex post*, stanowił błąd RMSPE:

$$\text{RMSPE} = \sqrt{\frac{1}{S}\sum_{i=1}^{S}\left(\frac{y_i - y_i^*}{y_i}\right)^2}\ , \tag{1}$$

gdzie:

y_i – wartość rzeczywista zmiennej y w okresie i,

y_i^* – prognoza *ex post* zmiennej y w okresie i,

S – liczba okresów prognozy *ex post*.

Jeżeli przez $F(x)$ oznaczymy funkcję przystosowania, to jej minimalizacja będzie równoważna z maksymalizacją pewnej funkcji $G(x)$ [8], czyli

$$\min F(x) = \max G(x) = \max (-F(x) + C), \tag{2}$$

[2] Sugerujemy użycie któregoś z błędów absolutnych np. MAPE lub kwadratowych np. RMSPE.

gdzie:

$F(x)$ – wartość funkcji przystosowania,

$G(x)$ – pewna funkcja, dla której zachodzi $G(x) = -F(x) + C$,

C – stała, określana *a priori* na wejściu.

Selekcja chromosomów do reprodukcji odbywa się metodą elitarną i wartości oczekiwanej [3]. Kiedy wybrane chromosomy zostają dobrane w pary, wymiana genów dokonuje się w sposób analogiczny jak w przypadku reprezentacji binarnej w klasycznym algorytmie genetycznym. Chromosomy potomków powstają w wyniku wymiany genów rodziców za punktem krzyżowania.

Z kolei w operatorze mutacji należy uwzględnić użycie parametru *tm*. Wylosowane i zamieniane miejscami dwa geny nie może dzielić w czasie więcej niż *tm* okresów.

Wydłużanie się szeregu i wzrost parametru *tm* gwałtownie powiększa zbiór rozwiązań. Z punktu widzenia algorytmu genetycznego oznacza to konieczność operowania dużymi populacjami, co wydłuża czas obliczeń.

Przed wyznaczeniem prognoz szeregi poddano analizie, mającej wskazać obserwacje nietypowe, które utrudniają uzyskiwanie dobrej jakości prognoz. W tym celu wykorzystano narzędzia znane z estymacji odpornej. Pierwszym z nich jest macierz rzutowania [5], która w naszym przypadku przyjęła postać

$$\mathbf{H} = \mathbf{y}(\mathbf{y}^T\mathbf{y})^{-1}\mathbf{y}^T. \tag{3}$$

Dla danego szeregu wyznaczyliśmy prognozy metodą średniej ruchomej prostej o stałej wygładzania równej 2. Na ich podstawie obliczyliśmy reszty z prognozy *ex post*, które następnie przekształciliśmy według formuły

$$e_t^* = \frac{e_t}{s\sqrt{1-h_t}}, \tag{4}$$

gdzie:

e_t – reszta z prognozy *ex post*,

s – odchylenie standardowe reszt z prognozy,

h_t – element przekątnej głównej macierzy $\mathbf{H}$, odpowiadający prognozie o numerze t.

Reszty wyznaczone według formuły (4) posłużyły do określenia, które obserwacje należy uznać za nietypowe. Jako kryterium potwierdzające tę tezę przyjęliśmy wartość $|e_t^*| > 2$.

Obserwację uznaną za nietypową należy z wektora danych wykluczyć. Ponieważ jednak mamy do czynienia z szeregami czasowymi, konieczne jest zastąpienie jej inną wielkością. W związku z tym nietypową obserwację zastąpiliśmy średnią arytmetyczną z obserwacji poprzedniej i następnej.

W przypadku dalszego występowania dużej zmienności wśród danych przydają się procedury zwiększające różnorodność populacji, a przy tej okazji bardziej wrażliwe

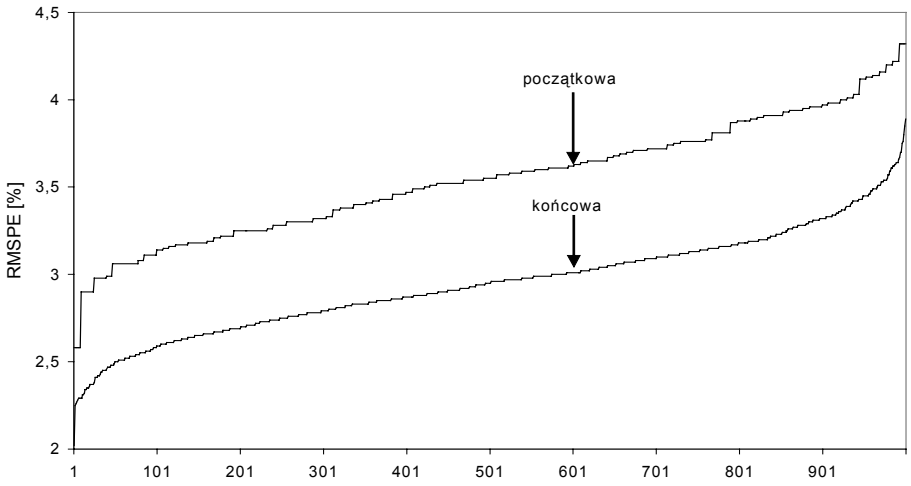
na ewentualność pojawiania się ekstremów lokalnych. Aby zwiększyć zróżnicowanie wśród chromosomów, wprowadziliśmy mechanizm preselekcji zaproponowany przez Cavicchio [4], który polega na tym, że potomek o lepszym przystosowaniu zastępuje w nowo tworzonej populacji rodzica o gorszym przystosowaniu. W przeciwnym wypadku to potomek jest kasowany, a jego miejsce zajmuje rodzic.

2. Prezentacja wyników obliczeń

Efektywność algorytmu genetycznego zbadamy sięgając po szeregi czasowe pochodzące z giełdy warszawskiej, obejmujące dzienne kursy zamknięcia i wolumeny obrotów za okres od 2.01.2007 do 22.02.2007 (38 obserwacji) dla następujących spółek: BZ WBK, KGHM, TP SA oraz indeksu WIG20. Prezentują one różne branże i odmienne warianty zachowań dekompozycji szeregów czasowych, przy czym w żadnym z analizowanych przypadków nie występowała sezonowość.

Parametry sterujące przebiegiem algorytmu przyjmują następujące wartości:

- pojedynczy chromosom składa się z 36 genów,
- populacja liczy 1000 chromosomów z liczbą pokoleń równą 50,
- prawdopodobieństwo krzyżowania wyniesie 0,5, zaś mutacji 0,1,
- parametr *tm* dla notowań równa się 3 lub 5 (zależnie od tego, która wartość dawała lepsze prognozy), zaś dla wolumenów obrotów wyniesie 10.



Rys. 1. Zmiana wartości funkcji przystosowania
między populacjami początkową i końcową dla BZWBK
Ź r ó d ł o: Opracowanie własne.

Na początek omówimy wnioski płynące z analizy różnic funkcji przystosowania dla chromosomów tworzących populację początkową i końcową. Rysunek 1 na przykładzie notowań BZWBK przedstawia sytuację, która okazała się typowa również dla pozostałych spółek. Znajdują się na nim wartości funkcji przystosowania dla wszystkich chromosomów w obu populacjach. Dla jasności obrazu, wartości RMSPE uporządkowaliśmy rosnąco – oddzielnie dla każdej populacji.

Dolna łamana pokazuje niższe, czyli lepsze wartości błędu RMSPE prognoz *ex post*, otrzymanych dla 1000 chromosomów populacji końcowej. Dla wszystkich pozostałych szeregów wyniki końcowe również okazały się wyraźnie lepsze od początkowych. Aby zilustrować wspomnianą poprawę wyników, w tabeli 1 pokazano średnie arytmetyczne różnic funkcji przystosowania między populacjami pierwszą i ostatnią.

Tabela 1. Średnia różnic wartości funkcji przystosowania między populacjami początkową i końcową [*p*%]

Spółka	Prognozy kursów	Prognozy wolumenów
TP S.A.	0,53	36,86
BZWBK	0,60	57,10
KGHM	1,08	33,07
WIG20	0,40	13,75

Ź r ó d ł o: Opracowanie własne.

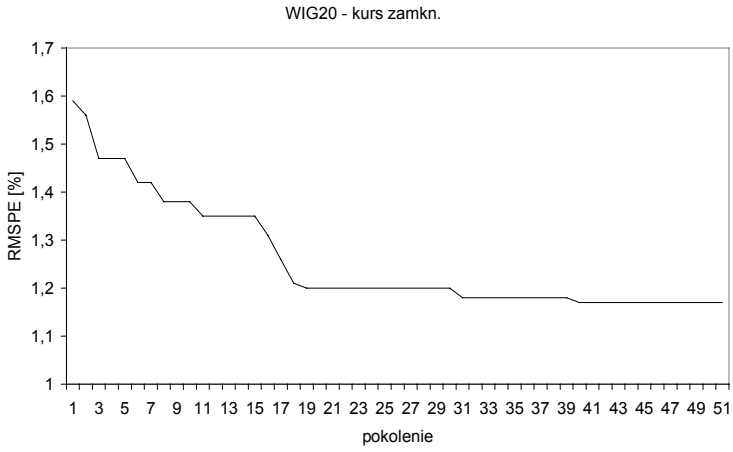
Znacznie wyższe wartości w ostatniej kolumnie tabeli 1 wynikają z faktu, że wolumeny obrotów charakteryzowały się dużą zmiennością, co przełożyło się na jakość prognoz *ex post*. Wolumen obrotów ze swej natury zachowuje się bardzo nieprzewidywalnie. Nie znaczy to jednak, że należy zrezygnować z prób oceny jego zachowania. Poza tym stanowi on źródło danych, przydatnych do testowania nowych metod predykcji.

Warto zwrócić uwagę, że wykres dla populacji końcowej jest bardziej gładki niż dla początkowej. Stanowi to wynik prowadzonych przez 50 pokoleń selekcji i krzyżowań chromosomów. Poza tym startowy zbiór prognoz powstaje w drodze losowania i zależy na przykład od rodzaju użytego generatora liczb pseudolosowych.

Obserwowane na rysunku 1 „schodki" świadczą o tym, że w danej populacji pojawiają się grupy chromosomów o takim samym przystosowaniu. Zastosowanie mechanizmu preselekcji spowodowało, że chromosomy gromadzą się w poszczególnych niszach. To samo podejście, zaproponowane przez Cavicchio, przyczyniło się do tak wyraźnego odsunięcia się od siebie populacji początkowej i końcowej.

W algorytmie zastosowano selekcję do reprodukcji, będącą połączeniem metody elitarnej i wartości oczekiwanej. W metodzie elitarnej wybiera się najlepszy element z danej populacji. W następnym pokoleniu sprawdza się, czy znalazł się on wśród chromosomów. Jeżeli nowa populacja nie zawiera tego łańcucha, to dołącza się go do niej jako kolejny element.

Wykres znajdujący się na rysunku 2 służy ilustracji efektywności metody elitarnej.



Rys. 2. Poprawa wartości funkcji przystosowania w kolejnych pokoleniach
Źródło: Opracowanie własne.

Uwagę zwraca początkowe szybkie poprawianie się wartości RMSPE, po którym następuje względna stabilizacja. Dla kursów zamknięcia przyjęcie mniej więcej stałej trajektorii zajmowało około 20 pokoleń, zaś dla wolumenów obrotów (z uwagi na znacznie wyższą zmienność danych) trwało niekiedy dłużej. Dalsza poprawa wyników przychodzi algorytmowi z większym trudem, co wiąże się z pojawianiem się w większej ilości lepiej przystosowanych chromosomów. Sama poprawa nie dokonuje się w każdym następnym pokoleniu, ponieważ dany chromosom może pozostawać najlepszy przez kilka przebiegów algorytmu.

Do zmian zachodzących w kolejnych pokoleniach można podejść od strony miar dynamiki. W tabeli 2 znalazły się wartości średniego tempa zmian, obliczonego według formuły

$$T = \bar{i}_G - 1, \tag{5}$$

gdzie $\bar{i}_G$ – średnia geometryczna indeksów zmian RMSPE obliczona dla 50 pokoleń.

Tabela 2. Średnie tempo zmian funkcji przystosowania

Spółka	Średnie tempo dla kursu [%]	Średnie tempo dla wolumenu [%]
TP SA	–0,4	–1,5
BZWBK	–0,6	–1,0
KGHM	–0,6	–1,4
WIG20	–0,6	–1,1

Źródło: Obliczenia własne.

Jak nietrudno zauważyć, przeciętne tempo poprawiania się wyników w następujących po sobie pokoleniach w przypadku wolumenów obrotów okazało się wyraźnie wyższe od tego dla kursów zamknięcia. Odpowiada za to różnica w zmienności szeregów kursów i wolumenów (wyższa dla tych drugich), która utrzymuje się nawet po usunięciu obserwacji nietypowych.

Dotychczasowe rozważania koncentrowały się na efektywności funkcjonowania algorytmu genetycznego jako takiego. Nie zapominajmy jednak, że po narzędzie to sięgnęliśmy, aby otrzymać prognozy. Dlatego jako kolejny sprawdzian wybraliśmy zbadanie dokładności prognoz wygasłych. W tym celu każdy z szeregów skróciliśmy o trzy obserwacje i ponownie wyznaczyliśmy prognozy *ex post* za pomocą algorytmu genetycznego. Następnie obliczyliśmy prognozy poza próbę, zgodnie z wcześniej opisaną w artykule metodologią.

W tabeli 3 zebraliśmy wartości uśrednionych błędów prognoz *ex post* i wygasłych. Aby porównanie wyników uczynić możliwym, w drugim przypadku również obliczono RMSPE.

Tabela 3. Błędy prognoz *ex post* i wygasłych

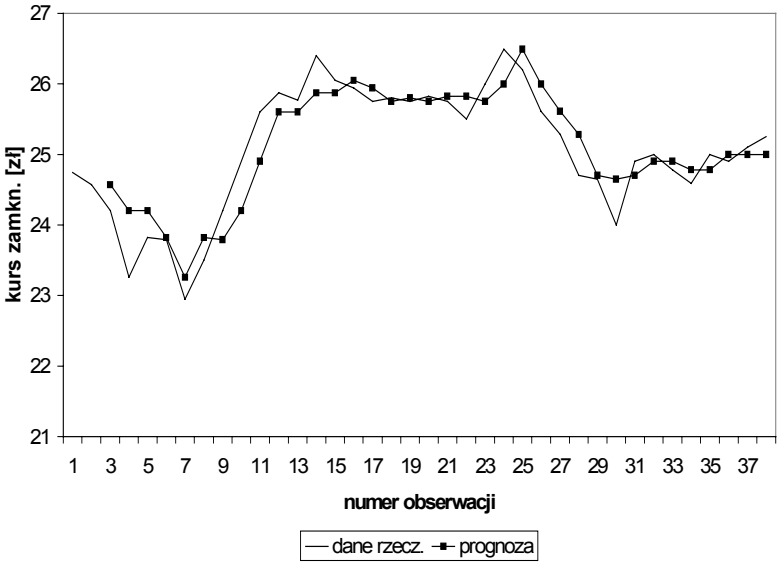
Spółka	RMSPE prognoz *ex post* [%]		RMSPE prognoz wygasłych [%]	
	Kurs zamkn.	Wolumen	Kurs zamkn.	Wolumen
TP SA	1,52	29,82	0,66	36,13
BZWBK	2,00	44,01	7,56	58,71
KGHM	2,02	24,97	2,96	*74,43*
WIG20	1,17	13,95	2,92	17,67

Źródło: Opracowanie własne.

Na początek wyjaśnienia wymaga wyróżniona wartość dla wolumenu obrotów KGHM w przypadku prognoz wygasłych. Otóż w trzecim i ostatnim okresie prognoza okazała się ujemna, co w oczywisty sposób nie może mieć miejsca. Dlatego w tym wypadku obliczyliśmy reszty tylko dla dwóch okresów. Rysunek 3 przedstawia szereg złożony z połączonych prognoz *ex post* i wygasłych, zestawiony ze wszystkimi danymi rzeczywistymi. Znalazł się na nim kurs akcji TP SA, ponieważ jako jedyna spółka miała ona niższy RMSPE dla prognoz wygasłych.

Na marginesie warto wspomnieć, że RMSPE dla pełnych, liczących 38 obserwacji szeregów kształtowały się na bardzo zbliżonym poziomie.

Generalnie błędy prognoz wygasłych przyjęły znacznie wyższe wartości w porównaniu z ich odpowiednikami *ex post*. Jest to szczególnie zauważalne w przypadku wolumenów obrotów i jeszcze raz podkreśla trudności, związane z prognozowaniem tego typu szeregów. Jeśli chodzi o kursy zamknięcia, to pomijając wyraźnie odstający pod tym względem BZWBK, RMSPE w okresie prognoz wygasłych kształtowały się wciąż na niskim poziomie.



Rys. 3. Prognozy kursu zamknięcia TP SA
Ź r ó d ł o: Opracowanie własne.

Wnioski

Zastosowaliśmy trzy metody, aby ocenić efektywność algorytmu genetycznego jako narzędzia prognozowania szeregów czasowych. Porównanie wartości funkcji przystosowania chromosomów populacji początkowej i końcowej pozwoliło stwierdzić, że wykorzystany mechanizm preselekcji przyczynił się do wyraźnej poprawy wyników. Procedura ta ma tę zaletę, że zwiększa zróżnicowanie populacji, co szczególnie przydaje się w naszej sytuacji, ponieważ zbiór możliwych prognoz jest bardzo obszerny, a nam zależy na możliwie dokładnym go przeszukaniu.

Porównanie przystosowania najlepszych chromosomów w kolejnych pokoleniach dowiodło, że najszybsza poprawa wartości RMSPE następowała na początku obliczeń. Później dochodzi do wyrównania się jakości prognoz, ponieważ najlepiej przystosowane chromosomy zaczynają dominować w ramach danej populacji.

Analiza wartości RMSPE obliczonych dla prognoz wygasłych pokazała, że w przypadku kursów zamknięcia pogorszenie jakości prognoz jest akceptowalne, lecz dla wolumenów obrotów średni błąd bardzo wyraźnie rośnie. Zresztą dla wszystkich instrumentów prognozowanie wolumenów sprawiało problemy, z uwagi na dużą zmienność występującą w tego typu szeregach.

Ogólnie rzecz biorąc algorytm genetyczny okazał się skuteczny jako narzędzie prognozowania. Można go używać zarówno w sytuacjach, kiedy sprawdzają się metody naiwne, jak i wtedy, gdy te ostatnie zawodzą z uwagi na wysokie wahania losowe. Nie zamyka to rzecz jasna drogi do dalszych badań tej tematyki.

Bibliografia

[1] CIEŚLAK M. (red.), *Prognozowanie gospodarcze. Metody i zastosowania*, PWN, 2001.
[2] GAJDA J.B., *Prognozowanie i symulacje a decyzje gospodarcze*, C.H. Beck, 2001.
[3] GOLDBERG D., *Algorytmy genetyczne i ich zastosowania*, WNT, Warszawa 1995.
[4] GWIAZDA T., *Algorytmy genetyczne – wstęp do teorii*, T.D.G. S. cyw., Warszawa 1995.
[5] KONARZEWSKA I., KARWACKI Z., *Planowanie i kontrola kosztów – wybrane problemy statystyczne*, 2001, s. 445–459.
[6] KUCHARSKI A., *O pewnym zastosowaniu algorytmów genetycznych do prognozowania szeregów czasowych*, 2005, Prace Naukowe Akademii Ekonomicznej we Wrocławiu, Wrocław 2007, s. 143–153.
[7] KUCHARSKI A., *Wykorzystanie algorytmów genetycznych do krótkookresowych prognoz na giełdzie papierów wartościowych*, konferencja naukowa „Rynek kapitałowy – skuteczne inwestowanie", Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, Szczecin 2007, s. 135–145.
[8] MICHALEWICZ Z., *Algorytmy genetyczne + struktury danych = programy ewolucyjne*, WNT, Warszawa 1996.

Efficiency of genetic algorithm as forecasting instrument of time series

The genetic algorithm is numbered among less formal methods which allows using it in different areas including forecasting. However, there is a question about efficiency of that instrument. We decided to check inner procedures of the algorithm affecting gaining speed and quality of solutions. We also tested exactness of expired forecasts. All the calculations were made using time series from the Stock Exchange.

Comparing all chromosomes from the initial and final populations leads to a conclusion that preselection contributed to the unquestionable improvement of forecasts. It is possible because of a bigger diversity between chromosomes.

RMSPE for the best chromosomes rapidly falls during the first twenty of all fifty generations. After that chromosomes with the best fitness dominate the whole population and the improvement of results is not so fast.

The expired forecasts of close price turned out much better than volume although we observed deterioration for all stock instruments. The reason was big volatility of those volumes which influenced the efficiency of the algorithm.

Keywords: *genetic algorithm, forecasting, stock market*