

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

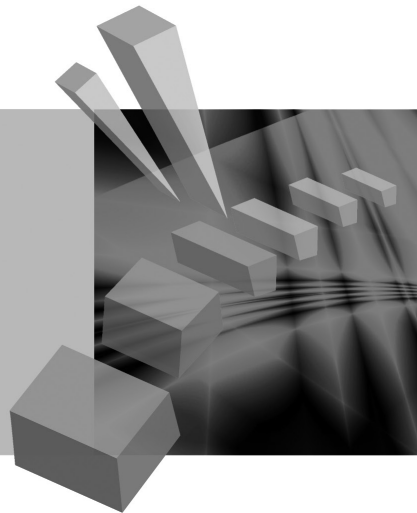
RESEARCH PAPERS

of Wrocław University of Economics

279

Taksonomia 21

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi

Krzysztof Jajuga

Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:

www.ibuk.pl, www.ebscohost.com,

The Central and Eastern European Online Library www.ceeol.com,

a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon

http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się

na stronie internetowej Wydawnictwa

www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego

oraz ze środków Sekcji Klasyfikacji i Analizy danych PTS

Kopiowanie i powielanie w jakiegokolwiek formie

wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu

Wrocław 2013

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)

ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp	9
Sabina Denkowska, Kamil Fijorek, Marcin Salamaga, Andrzej Sokolowski: Sejm VI kadencji – maszynka do głosowania	11
Barbara Pawelek, Adam Sagan: Zmienne ukryte w modelach ekonomicznych – respecyfikacja modelu Kleina I	19
Jan Paradysz: Nowe możliwości badania koniunktury na rynku pracy	29
Krzysztof Najman: Samouczące się sieci GNG w grupowaniu dynamicznym zbiorów o wysokim wymiarze	41
Kamila Migdał-Najman: Zastosowanie jednowymiarowej sieci SOM do wyboru cech zmiennych w grupowaniu dynamicznym	48
Aleksandra Matuszewska-Janica, Dorota Witkowska: Zróżnicowanie płac ze względu na płeć: zastosowanie drzew klasyfikacyjnych	58
Iwona Foryś, Ewa Putek-Szeląg: Przestrzenna klasyfikacja gmin ze względu na sprzedaż użytków gruntowych zbywanych przez ANR w województwie zachodniopomorskim	67
Joanna Banaś, Małgorzata Machowska-Szewczyk: Klasyfikacja internetowych rachunków bankowych z uwzględnieniem zmiennych symbolicznych.....	77
Marta Jaročka: Wpływ metody doboru cech diagnostycznych na wynik porządkowania liniowego na przykładzie rankingu polskich uczelni	85
Anna Zamojska: Badanie zgodności rankingów wyznaczonych według różnych wskaźników efektywności zarządzania portfelem na przykładzie funduszy inwestycyjnych.....	95
Dorota Rozmus: Porównanie dokładności taksonomicznej metody propagacji podobieństwa oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i>	106
Ewa Wędrowska: Wrażliwość miar dywergencji jako mierników niepodobieństwa struktur.....	115
Katarzyna Wójcik, Janusz Tuchowski: Wpływ automatycznego tłumaczenia na wyniki automatycznej identyfikacji charakteru opinii konsumenckich ...	124
Małgorzata Misztal: Ocena wpływu wybranych metod imputacji na wyniki klasyfikacji obiektów w modelach drzew klasyfikacyjnych.....	135
Anna Czapkiewicz, Beata Basiura: Badanie wpływu wyboru współczynnika zależności na grupowanie szeregów czasowych	146
Tomasz Szubert: Czynniki różnicujące poziom zadowolenia z życia oraz wartości życiowe osób sprawnych i niepełnosprawnych w świetle badań „Diagnozy społecznej”	154

Marcin Szymkowiak: Konstrukcja estymatorów kalibracyjnych wartości globalnej dla różnych funkcji odległości	164
Wojciech Roszka: Szacowanie łącznych charakterystyk cech nieobserwowanych łącznie	174
Justyna Brzezińska: Metody wizualizacji danych jakościowych w programie R	182
Agata Sielska: Regionalne zróżnicowanie potencjału konkurencyjnego polskich gospodarstw rolnych w województwach po akcesji do Unii Europejskiej	191
Mariusz Kubus: Liniowy model prawdopodobieństwa z regularyzacją jako metoda doboru zmiennych	201
Beata Basiura: Metoda Warda w zastosowaniu klasyfikacji województw Polski z różnymi miarami odległości	209
Katarzyna Wardzińska: Wykorzystanie metody obwiedni danych w procesie klasyfikacji przedsiębiorstw	217
Katarzyna Dębowska: Modelowanie upadłości przedsiębiorstw oparte na próbach niezbilansowanych	226
Danuta Tarka: Wpływ metody doboru cech diagnostycznych na wyniki klasyfikacji obiektów na przykładzie danych dotyczących ochrony środowiska ..	235
Artur Czech: Zastosowanie wybranych metod doboru zmiennych diagnostycznych w badaniach konsumpcji w ujęciu pośrednim	246
Beata Bal-Domańska: Ocena relacji zachodzących między inteligentnym rozwojem a spójnością ekonomiczną w wymiarze regionalnym z wykorzystaniem modeli panelowych	255
Mariola Chrzanowska: <i>Ordinary kriging</i> i <i>inverse distance weighting</i> jako metody szacowania cen nieruchomości na przykładzie warszawskiego rynku	264
Adam Depta: Zastosowanie analizy wariancji w badaniu jakości życia na podstawie kwestionariusza SF-36v2	272
Maciej Beręsewicz, Tomasz Klimanek: Wykorzystanie estymacji pośredniej uwzględniającej korelację przestrzenną w badaniach cen mieszkań	281
Karolina Paradysz: Benchmarkowa analiza estymacji dla małych obszarów na lokalnych rynkach pracy	291
Anna Gryko-Nikitin: Dobór parametrów w równoległych algorytmach genetycznych dla problemu plecakowego	301
Tomasz Ząbkowski, Piotr Jałowiecki: Zastosowanie reguł asocjacyjnych do analizy danych ankietowych w wybranych obszarach logistyki przedsiębiorstw przetwórstwa rolno-spożywczego	311
Agnieszka Przedborska, Małgorzata Misztal: Zastosowanie metod statystyki wielowymiarowej do oceny wydolności stawów kolanowych u pacjentów z chorobą zwyrodnieniową leczonych operacyjnie	321
Dorota Perło: Rozwój zrównoważony w wymiarze gospodarczym, społecznym i środowiskowym – analiza przestrzenna	331

Ewa Putek-Szeląg, Urszula Gieraltowska, Analiza i diagnoza wielkości produkcji energii odnawialnej w Polsce na tle krajów Unii Europejskiej..	342
--	-----

Summaries

Sabina Denkowska, Kamil Fijorek, Marcin Salamaga, Andrzej Sokolowski: VIth-term Sejm – a voting machine	18
Barbara Pawelek, Adam Sagan: Latent variables in econometric models – respecification of Klein I model	28
Jan Paradysz: New possibilities for studying the situation on the labour market	40
Krzysztof Najman: Self-learning neural network of GNG type in the dynamic clustering of high-dimensional data.....	47
Kamila Migdał-Najman: Applying the one-dimensional SOM network to select variables in dynamic clustering	57
Aleksandra Matuszewska-Janica, Dorota Witkowska: Gender wage gap: application of classification trees.....	66
Iwona Foryś, Ewa Putek-Szeląg: Spatial classification of communes by usable land traded by the APA in the Zachodniopomorskie voivodeship...	76
Joanna Banaś, Małgorzata Machowska-Szewczyk: Classification of Internet banking accounts including symbolic variables	84
Marta Jarocka: The impact of the method of the selection of diagnostic variables on the result of linear ordering on the example of ranking of universities in Poland.....	94
Anna Zamojska: Empirical analysis of the consistency of mutual fund ranking for different portfolio performance measures.....	105
Dorota Rozmus: Comparison of accuracy of affinity propagation clustering and cluster ensembles based on bagging idea.....	114
Ewa Wędrowska: Sensitivity of divergence measures as structure dissimilarity measurements	123
Katarzyna Wójcik, Janusz Tuchowski: Machine translation impact on the results of the sentiment analysis	134
Małgorzata Misztal: Assessment of the influence of selected imputation methods on the results of object classification using classification trees ...	145
Anna Czapkiewicz, Beata Basiura: Simulation study of the selection of coefficient depending on the clustering time series.....	153
Tomasz Szubert: Factors differentiating the level of satisfaction with life and the life's values of people with and without disabilities in the light of the "Social Diagnosis" survey	162
Marcin Szymkowiak: Construction of calibration estimators of totals for different distance measures	173

Wojciech Roszka: Joint characteristics' estimation of variables not jointly observed.....	181
Justyna Brzezińska: Visualizing categorical data in \mathbf{R}	190
Agata Sielska: Regional diversity of competitiveness potential of Polish farms after the accession to the European Union	200
Mariusz Kubus: Regularized linear probability model as a filter	208
Beata Basiura: The Ward method in the application for classification of Polish voivodeships with different distances.....	216
Katarzyna Wardzińska: Application of Data Envelopment Analysis in company classification process.....	225
Katarzyna Dębowska: Modeling corporate bankruptcy based on unbalanced samples	234
Danuta Tarka: Influence of the features selection method on the results of objects classification using environmental data.....	245
Artur Czech: Application of chosen methods for the selection of diagnostic variables in indirect consumption research.....	254
Beata Bal-Domańska: Assessment of relations occurring between smart growth and economic cohesion in regional dimension using panel models	263
Mariola Chrzanowska: Ordinary kriging and inverse distance weighting as methods of estimating prices based on Warsaw real estate market	271
Adam Depta: Application of analysis of variance in the study of the quality of life based on questionnaire SF-36v2	280
Maciej Beręsewicz, Tomasz Klimanek: Using indirect estimation with spatial autocorrelation in dwelling price surveys.....	290
Karolina Paradysz: Benchmark analysis of small area estimation on local labor markets	300
Anna Gryko-Nikitin: Selection of various parameters of parallel evolutionary algorithm for knapsack problems	310
Tomasz Ząbkowski, Piotr Jałowiecki: Application of association rules for the survey of data analysis in the selected areas of logistics in food processing companies	320
Agnieszka Przedborska, Małgorzata Misztal: Using multivariate statistical methods to assess the capacity of the knee joint among the patients treated surgically for osteoarthritis	330
Dorota Perło: Sustainable development in the economic, social and environmental dimensions – spatial analysis.....	341
Ewa Putek-Szeląg, Urszula Gieraltowska: Analysis and diagnosis of the volume of renewable energy production in Poland compared to EU countries	352

Anna Gryko-Nikitin

Politechnika Białostocka

DOBÓR PARAMETRÓW W RÓWNOLEGLYCH ALGORYTMACH GENETYCZNYCH DLA PROBLEMU PLECAKOWEGO

Streszczenie: Celem artykułu jest prezentacja metodyki doboru parametrów równoległego algorytmu ewolucyjnego. Dobór parametrów zaprezentowany zostanie na przykładzie problemu plecakowego, którym może być przybliżony np. problem doboru akcji do koszyka inwestycyjnego. Dyskretny problem plecakowy należy do zadań optymalizacyjnych *NP*-trudnych o złożoności obliczeniowej określanej jako $O(n^2)$. Do zbioru metod rekomendowanych do rozwiązywania dyskretnego problemu plecakowego należą między innymi: algorytmy zachłanne, programowanie dynamiczne oraz wspomniane algorytmy ewolucyjne w wersji równoległej.

Słowa kluczowe: problem plecakowy, równoległe algorytmy ewolucyjne, parametry w algorytmach ewolucyjnych.

1. Wstęp

Dyskretny problem plecakowy należy do zadań optymalizacyjnych *NP*-trudnych. Zadanie w problemie plecakowym polega na wyborze z *N*-elementowego zbioru jak największej liczby przedmiotów, o jak największej wartości, przy czym wybrane przedmioty muszą się zmieścić do plecaka o zadanym rozmiarze. Problem plecakowy znajduje zastosowanie w wielu praktycznych zagadnieniach m.in.: z zakresu informatyki i zarządzania [Spillman 1995; Taheri i in. 2012]. Do zbioru metod rekomendowanych do rozwiązywania dyskretnego problemu plecakowego należą między innymi: algorytmy zachłanne, programowanie dynamiczne oraz algorytmy genetyczne [Kumar, Banerjee 2006; Taheri i in. 2012].

W kontekście algorytmów genetycznych ważnym zagadnieniem jest dobór parametrów kontrolnych [Grefenstette 1986]. Proces poszukiwań najlepszego rozwiązania w algorytmach genetycznych jest kontrolowany przez parametry tego algorytmu. Zastosowany zestaw parametrów algorytmu genetycznego wpływa na jakość otrzymanego wyniku oraz na czas potrzebny na jego otrzymanie [Fernandez-Prieto i in. 2011].

W pracy przedstawiono strategię postępowania w doborze wybranych parametrów równoległego algorytmu genetycznego. Metoda doboru parametrów kontrolnych równoległych algorytmów genetycznych została zainspirowana pracami [Fernandez-Prieto i in. 2011; Sakurai i in. 2010]. Dobór parametrów przeprowadzony został na przykładzie problemu plecakowego, którym może być przybliżony np. problem wyboru akcji do koszyka inwestycyjnego.

Cel aplikacyjny artykułu został zdefiniowany jako opracowanie równoległego algorytmu genetycznego dla problemu doboru akcji do portfela inwestycyjnego.

2. Algorytmy genetyczne

Zasada działania klasycznych algorytmów genetycznych została zaczerpnięta z natury i opiera się na podstawowej zasadzie darwinowskiej ewolucji połączonej z dziedziczeniem, w myśl której proces dochodzenia do rozwiązania odbywa się na drodze ewolucji grupy początkowych, mało wartościowych, propozycji rozwiązań. Jakość reprezentowanego rozwiązania opisana jest wartością liczbową, zwaną przystosowaniem osobnika. Wyselekcjonowane osobniki podlegają w poszczególnych pokoleniach przemianom (krzyżowaniu, mutacji), doprowadzając ostatecznie do otrzymania najlepszego osobnika (optymalnego rozwiązania). Algorytm działa w środowisku, które opisuje się przy użyciu funkcji przystosowania [Arabas 2001].

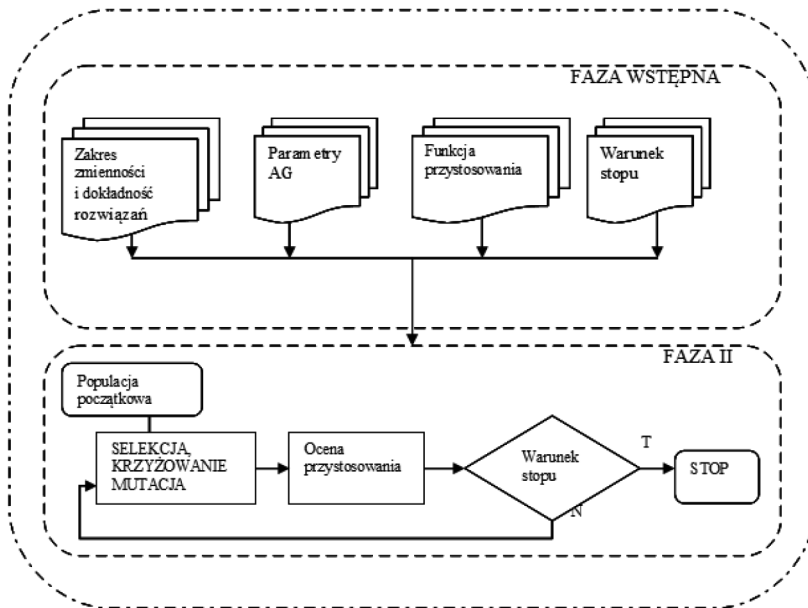
Zadaniem algorytmu genetycznego jest ciągła poprawa średniej wartości funkcji dopasowania całej populacji w iteracjach zmierzających do koncentracji kolejnych pokoleń osobników wokół pewnego, zbliżonego do optymalnego, rozwiązania.

Pracę z algorytmami genetycznymi można podzielić na dwie fazy (rys. 1). Fazę I rozpoczyna się od dokładnego sprecyzowania problemu (przyjęcie określonej reprezentacji problemu), po czym określa się parametry populacji i parametry algorytmu genetycznego. W następnej kolejności określa się funkcję przystosowania i warunek stopu. Na fazę II składają się wyznaczanie wartości funkcji dopasowania oraz operacje genetyczne, tj. selekcja, krzyżowanie i mutacja. Działanie algorytmu powtarza się aż do osiągnięcia założonego kryterium stopu.

Proces poszukiwań najlepszego rozwiązania w algorytmach genetycznych kontrolowany jest przez parametry tego algorytmu [Grefenstette 1986]. Parametry algorytmów genetycznych, tj.: rozmiar populacji, prawdopodobieństwo krzyżowania, prawdopodobieństwo mutacji, wpływają na jakość otrzymanego wyniku oraz na czas potrzebny na jego otrzymanie [Fernandez-Prieto i in. 2011].

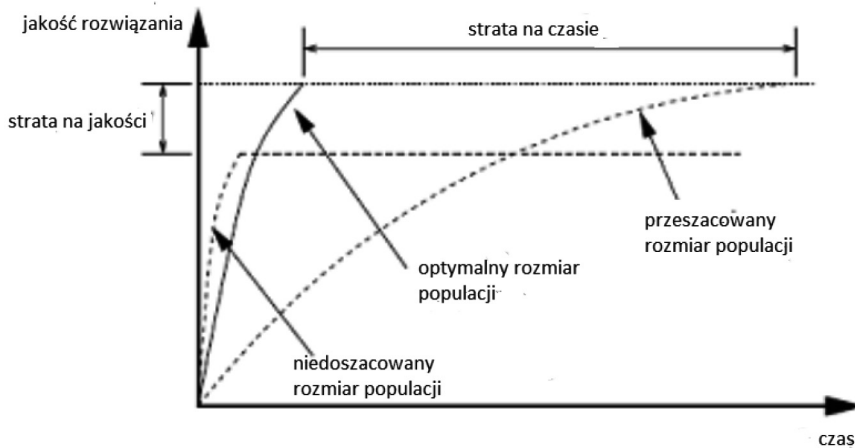
Rysunek 2 przedstawia wpływ rozmiaru populacji na jakość rozwiązania. Podczas określania tego parametru należy rozważyć dwie kwestie. Mianowicie niedoszacowana populacja startowa może skutkować poważnymi stratami w jakości rozwiązania. Sytuacja odwrotna powoduje znaczne wydłużenie czasu potrzebnego na znalezienie rozwiązania optymalnego.

Na zbieżność algorytmu genetycznego wpływ mają zastosowane operatory genetyczne. Napór genetyczny warunkowany jest metodą selekcji osobników do puli



Rys. 1. Schemat działania w algorytmach genetycznych

Źródło: opracowanie własne na podstawie [Chodak, Kwaśnicki 2002].



Rys. 2. Wpływ rozmiaru populacji na jakość rozwiązania i czas potrzebny na jego otrzymanie

Źródło: opracowanie na podstawie [Lobo, Goldberg 2004].

rodzicielskiej. Powszechną zasadą przy ustalaniu wartości prawdopodobieństwa krzyżowania i mutacji jest zasada zaczerpnięta z natury. W naturze mutacja zach-

dzi stosunkowo rzadko, zaś krzyżowanie osobników jest podstawą do zachowania gatunku. Poprzez operację krzyżowania następuje też wymiana materiału genetycznego, co skutkuje utworzeniem wielu osobników o wyższym przystosowaniu do środowiska, niż wykazywały osobniki rodzicielskie. Ostatnie eksperymenty dopuszczają przyjęcie 16 różnych wartości parametru p_c , (od 0,25 do 1,00 w krokach co 0,05) [Grefenstette 1986]. Mutacja, umożliwiająca przywrócenie utraconego materiału genetycznego, traktowana jest jako operacja drugorzędna [Aguirre i in. 1999]. Najczęściej stosowane wartości prawdopodobieństwa mutacji zawierają się w przedziale $\langle 0,001; 0,01 \rangle$ [Grefenstette 1986].

W zakresie doboru parametrów kontrolnych algorytmów genetycznych prowadzono wiele badań [Goldberg 1995; Sakurai i in. 2010; Grefenstette 1986]. W pracy [Michalewicz 1999] testowano zestawy różnych kombinacji parametrów. Deb i Agrawal [1999] badali m.in. interakcje pomiędzy parametrami. Powstało też wiele koncepcji na ten temat. Autorzy byli jednak zgodni co do tego, że dobór parametrów jest zadaniem trudnym, wymagającym przeprowadzenia wielu testów. Do chwili obecnej zaproponowano dwa podstawowe podejścia w kwestii wyznaczania wartości parametrów. Według jednego nurtu, parametry kontrolne są ustalane na początku eksperymentów (np. testuje się różne zestawy parametrów i wybiera najlepszy z nich) [Chan i in. 2002]. Drugi sposób polega na dostrajaniu parametrów kontrolnych w trakcie działania algorytmu [Sakurai i in. 2010; Fernandez-Prieto i in. 2011]. W pracy wykorzystano podejście z dostrajaniem.

3. Równoległy algorytm genetyczny dla problemu plecakowego

Opracowanie równoległego algorytmu genetycznego dla problemu plecakowego zostanie zaprezentowane na przykładzie wyboru akcji do koszyka inwestycyjnego (portfela inwestycyjnego).

Zadanie w problemie plecakowym polega na wyborze z N -elementowego zbioru $\{x_1, x_2, \dots, x_N\}$ jak największej liczby przedmiotów o jak największej wartości, przy czym wybrane przedmioty muszą się zmieścić do plecaka o zadanym rozmiarze B . Funkcja wartości plecaka:

$$f(x) = \sum_{j=0}^N c_j x_j,$$

$$j = 1, 2, \dots, N \quad x_j \in (0, 1)$$

przy ograniczeniu

$$\sum_{j=0}^N w_j x_j \leq B,$$

gdzie: c_j – wartość j -tego elementu,
 w_j – wielkość j -tego elementu,
 B – rozmiar plecaka.

W kontekście analizy portfelowej [Tarczyński 1997] problem może być opisany jako maksymalizacja oczekiwanej stopy zwrotu portfela przy określonej wariancji.

Wartość j -tego elementu (c_j) należy więc rozumieć jako oczekiwaną stopę zwrotu z akcji wyrażoną wzorem:

$$R = \frac{\sum_{t=1}^N R_t}{N},$$

gdzie: R – ocena oczekiwanej stopy zwrotu z papieru wartościowego,
 N – liczba wszystkich analizowanych stóp zwrotu,
 R_t – empiryczna stopa zwrotu wyrażona wzorem:

$$R_t = \frac{P_t - P_{t-1} + D_t}{P_{t-1}},$$

gdzie: P_t – cena papieru wartościowego w okresie t ,
 P_{t-1} – cena papieru wartościowego w okresie $t-1$,
 D_t – dywidenda wypłacona w t -tym okresie.

Przez ciężar j -tego przedmiotu (w_j) należy rozumieć ryzyko inwestycyjne mierzone odchyleniem standardowym stopy zwrotu papieru wartościowego. Odchylenie standardowe wyznacza się z następującego wzoru:

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (R_i - R)^2}.$$

Adaptując algorytm genetyczny na potrzeby rozwiązania konkretnego problemu, należy rozpocząć od wyboru reprezentacji poszczególnych osobników. Na podstawie przeglądu literatury [Arabas 2001; Zhao i in. 2008; Spillman 1995] dla problemu plecakowego przyjęto binarną reprezentację chromosomów.

Kodowanie binarne polega na przypisaniu poszczególnym genom wartości „0” lub „1”. Osobnikami są wówczas wektory binarne długości N . Osobnikiem w badanym zagadnieniu będzie pojedynczy plecak (koszt inwestycyjny) składający się z N genów (papierów wartościowych). Przyjmując, że plecak składa się z N przedmiotów, k -tą populację chromosomów można opisać następująco:

$$P_k = \{X_{k1}, X_{k2}, X_{k3}, \dots, X_{ki}, \dots, X_{kM}\} \quad i = 1, 2, \dots, M, \quad k = 1, 2, \dots, q,$$

gdzie: X_{ki} – i -ty chromosom reprezentujący potencjalne rozwiązanie w k -tej populacji;

M – liczba chromosomów w populacji;

k – liczba generacji.

Zakładając liczbę przedmiotów równą 20, pojedynczy chromosom można opisać następująco: $X_k = [0\ 1\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 1\ 1]$, co należy interpretować następująco: akcje oznaczone numerami 1, 3, 4, 5, 8, 9, 10, 11, 17 i 18 nie weszły do koszyka inwestycyjnego (plecaka).

Działanie algorytmów genetycznych rozpoczyna się od utworzenia populacji początkowej P^0 . Zazwyczaj generuje się ją losowo [Goldberg 1995]. W przypadku, gdy istnieją uzasadnione przesłanki, populacja startowa może być tworzona przy zachowaniu wybranego kryterium [Michalewicz 1999]. Bazując na teorii portfela efektywnego Markowitza, zdecydowano, że populację startową będą tworzyć portfele składające się z aktywów nieskorelowanych [Feder-Sempach 2011]. Podejście to pozwoli zmniejszyć ryzyko inwestycji.

Miarą zależności między badanymi akcjami jest współczynnik korelacji wyrażony wzorem [Tarczyński 1997]:

$$\rho_{12} = \frac{\sum_{i=1}^n (R_{1i} - R_1)(R_{2i} - R_2)}{(n - 1)S_1S_2},$$

gdzie: ρ_{12} – unormowana miara korelacji (współczynnik korelacji dwóch papierów wartościowych),

R_{1i} – możliwe stopu zwrotu pierwszej akcji,

R_1 – oczekiwana stopa zwrotu pierwszej akcji,

R_{2i} – możliwe stopu zwrotu drugiej akcji,

R_2 – oczekiwana stopa zwrotu drugiej akcji,

S_1 – odchylenie standardowe pierwszej akcji,

S_2 – odchylenie standardowe drugiej akcji,

n – liczba wszystkich badanych stóp zwrotu.

Celem opracowania efektywnego algorytmu rozwiązującego problem plecakowy dla koszyka inwestycyjnego wyróżniono fazy, w których obliczenia zajmują najwięcej czasu i mocy obliczeniowej. Jest to m.in. faza, podczas której następuje tworzenie populacji startowej.

W zagadnieniu zrównoleglenia wykorzystano schemat komunikacji *master-slaves*. W podejściu tym wyróżnia się jeden procesor jako główny (*master*), a reszta procesorów, tzw. podwładnych (*slaves*) – odpowiada za wykonanie obliczeń. Procesor główny jako nadzorca ma całą wiedzę o stanie algorytmu i kontroluje kolejność wykonywania obliczeń przez poszczególne procesory. Procesor główny rozsyła do

gdzie: c_j – wartość j -tego elementu,
 w_j – wielkość j -tego elementu,
 B – rozmiar plecaka.

W kontekście analizy portfelowej [Tarczyński 1997] problem może być opisany jako maksymalizacja oczekiwanej stopy zwrotu portfela przy określonej wariancji.

Wartość j -tego elementu (c_j) należy więc rozumieć jako oczekiwaną stopę zwrotu z akcji wyrażoną wzorem:

$$R = \frac{\sum_{t=1}^N R_t}{N},$$

gdzie: R – ocena oczekiwanej stopy zwrotu z papieru wartościowego,
 N – liczba wszystkich analizowanych stóp zwrotu,
 R_t – empiryczna stopa zwrotu wyrażona wzorem:

$$R_t = \frac{P_t - P_{t-1} + D_t}{P_{t-1}},$$

gdzie: P_t – cena papieru wartościowego w okresie t ,
 P_{t-1} – cena papieru wartościowego w okresie $t-1$,
 D_t – dywidenda wypłacona w t -tym okresie.

Przez ciężar j -tego przedmiotu (w_j) należy rozumieć ryzyko inwestycyjne mierzone odchyleniem standardowym stopy zwrotu papieru wartościowego. Odchylenie standardowe wyznacza się z następującego wzoru:

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (R_i - R)^2}.$$

Adaptując algorytm genetyczny na potrzeby rozwiązania konkretnego problemu, należy rozpocząć od wyboru reprezentacji poszczególnych osobników. Na podstawie przeglądu literatury [Arabas 2001; Zhao i in. 2008; Spillman 1995] dla problemu plecakowego przyjęto binarną reprezentację chromosomów.

Kodowanie binarne polega na przypisaniu poszczególnym genom wartości „0” lub „1”. Osobnikami są wówczas wektory binarne długości N . Osobnikiem w badanym zagadnieniu będzie pojedynczy plecak (koszt inwestycyjny) składający się z N genów (papierów wartościowych). Przyjmując, że plecak składa się z N przedmiotów, k -tą populację chromosomów można opisać następująco:

$$P_k = \{X_{k1}, X_{k2}, X_{k3}, \dots, X_{ki}, \dots, X_{kM}\} \quad i = 1, 2, \dots, M, \quad k = 1, 2, \dots, q,$$

Procesory podwładne obliczają dodatkowo podstawowe statystyki populacji. Wyznaczone statystyki, które nie mieszczą się w dopuszczalnym zakresie, są przekazywane procesorowi głównemu. Na podstawie informacji o bieżącej generacji procesor główny podejmuje decyzję o dostrojeniu poszczególnych parametrów. Dostrojenie polega m.in. na zmianie prawdopodobieństwa mutacji i prawdopodobieństwa krzyżowania.

Tabela 3. Algorytm dla węzła podrzędnego

<p>FAZA I. Konstrukcja populacji startowej</p> <p>KROK 1. Odebranie danych z procesora głównego</p> <p>KROK 2. Wyznaczenie poprawnych chromosomów</p> <p>KROK 3. Wyznaczenie funkcji dopasowania</p> <p>KROK 4. Wyznaczenie statystyk bieżącej populacji</p>
<p>Wyzeruj zmienne $licz_chrom, i, j, liczba_el$</p> <p>Dopóki nie ustalono wymaganej liczby chromosomów ($licz_chrom$) w podpopulacji startowej, powtarzaj</p> <p>zwiększ zmienną $licz_chrom$ o jeden</p> <p style="padding-left: 2em;">Dopóki nie ustalono wymaganej liczby genów ($liczba_el$) w chromosomie, powtarzaj</p> <p style="padding-left: 4em;">wylosuj 0 lub 1</p> <p style="padding-left: 4em;">zwiększ zmienną $liczba_el$ o jeden</p> <p style="padding-left: 4em;">wstaw wylosowaną pozycję do i-tego chromosomu na pozycję j-tą</p> <p style="padding-left: 4em;">zwiększ zmienną j o jeden</p> <p>Sprawdź, czy chromosom jest prawidłowy</p> <p style="padding-left: 2em;">dopóki chromosom nie jest prawidłowy, wywołuj procedurę naprawy</p>
<p>Wyznacz wartość funkcji dopasowania chromosomu</p>
<p>Wyznacz statystyki populacji</p> <p><i>Jeśli wartości statystyk wykraczają poza dopuszczalny zakres, prześlij informację do procesora głównego</i></p>
<p>Wyjście: $licz_chrom$- elementowa populacja chromosomów z obliczoną wartością funkcji dopasowania oraz statystyka populacji</p>

Źródło: opracowanie własne.

Druga faza omawianego algorytmu to faza ewolucji populacji chromosomów. Operacja ta wykonywana jest przez procesory podwładne według schematu zawartego w tab. 4.

Tabela 4. Algorytm dla węzła podrzędnego

FAZA II Ewolucja populacji
Dopóki nie osiągnięto warunku stopu, powtarzaj KROK I – KROK VI
KROK I Wyznaczenie funkcji dopasowania
KROK II Selekcja osobników {Wybór rodziców}
KROK III Krzyżowanie
KROK IV Mutacja
KROK V Podmiana osobników w populacji
KROK VI Wyznaczenie statystyk bieżącej populacji
<i>Jeśli statystyki nie mieszczą się w dopuszczalnych granicach wyślij informację do procesora głównego</i>
KROK VII Wysłanie populacji do procesora głównego

Źródło: opracowanie własne na podstawie [Sakurai i in. 2010; Fernandez-Prieto i in. 2011].

4. Uwagi końcowe

W artykule przedstawiono strategię postępowania w doborze wybranych parametrów równoległego algorytmu genetycznego, uwzględniając jednocześnie ich znaczenie dla badanego zjawiska. Metoda została zaprezentowana na przykładzie wyboru akcji do koszyka inwestycyjnego (problem plecakowy). Propozycja obliczeń równoległych oparta została na jednym z czterech modeli równoległych algorytmów genetycznych, tj.: modelu synchronicznym scentralizowanym.

Literatura

- Aguirre H., Tanaka K., Sugimura T., *Cooperative model for genetic operators to improve GAs*, Proc. IEEE ICIS 1999, pp. 98-109.
- Arabas J., *Wykłady z algorytmów ewolucyjnych*, Wyd. Naukowo-Techniczne, Warszawa 2001.
- Chan M.C., Wong C.C., Cheung B.K., Tang G.Y., *Genetic algorithms in multi-stage portfolio optimization system. In proceedings of the eighth international conference of the Society for Computational Economics*, Computing in Economics and Finance, Aix-en-Provence, France 2002.
- Chodak G., Kwaśnicki W., *Zastosowanie algorytmów genetycznych w prognozowaniu popytu*, „Gospodarka Materialowa & Logistyka” 2002, nr 4.
- Deb K., Agrawal S., *Understanding interactions among genetic algorithm parameters*, “Foundations of Genetic Algorithms” 1999, pp. 265-286.
- Eklund S.E., *A massively parallel architecture for distributed genetic algorithms*, “Parallel Computing” 2004, vol. 30.

- Feder-Sempach E., *Ryzyko inwestycyjne. Analiza polskiego rynku akcji*, CeDeWu, Warszawa 2011.
- Fernandez-Prieto J.A., Canada-Bago J., Gadeo-Martos M.A., Velasco J.R., *Optimisation of control parameters for genetic algorithms to test computer networks under realistic traffic loads*, "Applied Soft Computing" 2011, 11(4), pp. 3744-3752.
- Goldberg D.E., *Algorytmy genetyczne i ich zastosowanie*, Wydawnictwo-Naukowo Techniczne, Warszawa 1995.
- Grefenstette J.J., *Optimization of control parameters for genetic algorithms*, Ieee Transactions on Systems, Man, and Cybernetics 1986, vol. SMC-16, no. 1.
- Kumar R., Banerjee N., *Analysis of a multiobjective evolutionary algorithm on the 0-1 knapsack problem*, "Theoretical Computer Science" 2006, 358, pp. 104-120.
- Lobo F.G., Goldberg D.E., *The parameter-less genetic algorithm in practice*, "Information Sciences" 2004, vol. 167.
- Michalewicz Z., *Algorytmy genetyczne + struktury danych = programy ewolucyjne*, Wyd. Naukowo-Techniczne, Warszawa 1999.
- Sakurai Y., Takada K., Kawabe T., Tsuruta S., *A metod to Control Parameters of Evolutionary Algorithms by using Reinforcement Learning*, 2010 Sixth International Conference on Signal-Image Technology and Internet Based Systems, 2010 IEEE.
- Spillman R., *Solving Large Knapsack Problems with a Genetic Algorithm, Systems, Man and Cybernetics, 1995. Intelligent Systems for the 21st Century*, IEEE International Conference on, 1995, vol. 1, pp. 632-637.
- Taheri J., Sharif S., Xing P., Zomaya A.Y., *Paralleled Genetic Algorithm for Solving the Knapsack Problem in the Cloud, P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)*, 2012 Seventh International Conference on, 2012, pp. 303-308.
- Tarczyński W., *Rynki kapitałowe. Metody ilościowe. Vol II*, Agencja Wydawnicza „Placet”, Warszawa 1997.
- Zhao T., Yang L., Man Z., *A MSM-PGA based on multi-agent for solving 0-1 knapsack problem*, [in:] Computer Science and Information Technology, 2008. ICCSIT'08. International Conference on. IEEE, 2008, pp. 898-902.

SELECTION OF VARIOUS PARAMETERS OF PARALLEL EVOLUTIONARY ALGORITHM FOR KNAPSACK PROBLEMS

Summary: The aim of the paper is the presentation of the methodology for selection of various parameters of parallel evolutionary algorithm. The selection process will be presented on the example of knapsack problem, which can be used for example for the problem of selection of investment shares to the cart. Discreet knapsack problem is one of the optimization tasks of NP-hard with computational complexity known as $O(n^2)$. The methods which are recommended for solving discrete knapsack problem are: greedy algorithms, dynamic programming, and mentioned above evolutionary algorithms in the parallel version.

Keywords: parallel evolutionary algorithms, evolutionary algorithms, parameters, knapsack problems.