Raport Badawczy

Research Report

RB/58/2013

Zastosowanie ewolucji różnicowej do wyznaczania zależności statystycznych

O. Hryniewicz, K. Opara

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences

POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6

01-447 Warszawa

tel.: (+48) (22) 3810100

fax: (+48) (22) 3810105

Kierownik Zakładu zgłaszający pracę:
Prof. dr hab. inż. Olgierd Hryniewicz

Warszawa 2013

1. Zastosowanie ewolucji różnicowej do wyznaczania zależności statystycznych

1.1. Problem badawczy

Jednym z podstawowych zadań we wnioskowaniu statystycznym jest badanie i testowanie zależności. W tym celu najczęściej używany jest współczynnik korelacji τ Pearsona określający stopień zależności liniowej. Do analizy danych zależnych w sposób nieliniowy (bądź nieznany) można wykorzystać metody nieparametryczne wyznaczając współczynnik ρ Spearmana oraz τ Kendalla. Pierwszy z nich mówi o liniowości zależności pomiędzy rangami par obserwacji, drugi natomiast wyraża różnicę pomiędzy prawdopodobieństwem ułożenia się punktów w tym samym porządku i w porządku przeciwnym.

W praktyce często występują dane znane jedynie z dokładnością do przedziału $x \in [X_{i,L}, X_{i,U}]$. W analizach zazwyczaj są one zastępowane pojedynczymi liczbami, co skutkuje utratą informacji. Dokładniejszym podejściem byłoby bezpośrednie wykorzystanie danych przedziałowych lub rozmytych.

Problem testowania niezależności statystycznej dla danych rozmytych został poruszony w pracach [5, 6, 7]. Ich autorzy zwracają uwagę, że główną przeszkodę w bezpośrednim wykorzystaniu danych przedziałowych stanowią problemy obliczeniowe.

W tym rozdziale omówiono zastosowanie algorytmu ewolucji różnicowej do obliczania współczynnika korelacji Kendalla dla danych przedziałowych. Weryfikacja wyników oparta została o doświadczenia symulacyjne dla specjalnie wygenerowanych przedziałowych zbiorów danych.

1.1.1. Obliczanie τ Kendalla dla danych przedziałowych

Współczynnik τ Kendalla

Niech (X_i, Y_i) dla $i = 1, ..., n$ będzie próbą losową zawierającą n par obserwacji zależnych zmiennych losowych X i Y. Wówczas współczynnik τ Kendalla zdefiniowany jest jako

$$\tau = \mathrm{P}\left[(X_1 - X_2)(Y_1 - Y_2) > 0\right] - \mathrm{P}\left[(X_1 - X_2)(Y_1 - Y_2) < 0\right]. \qquad (1.1)$$

Genest [4] zaproponował następujący (nieklasyczny) wzór pozwalający na wyznaczanie τ Kendalla z próby

$$\tau_n = \frac{4}{n-1}\sum_{i=1}^{n-1} V_i - 1, \tag{1.2}$$

gdzie

$$V_i = \frac{card\{j : X_j < X_i, Y_j < Y_i\}}{n-2}, \text{ dla } i = 1, \ldots, n. \tag{1.3}$$

Współczynnik τ Kendalla zależy jedynie od porządku pomiędzy poszczególnymi obserwacjami, nie zaś od ich wartości. Z tego powodu jest on nieparametryczną (rangową) miarą zależności.

Współczynnik τ Kendalla dla danych przedziałowych

Załóżmy, że dane mają charakter przedziałowy. W praktyce sytuacja taka występuje dość często, ze względu na błędy pomiarowe, czy też dyskretyzację przy ich zbieraniu. Wówczas dane przybierają postać $(\mathbf{X}_i, \mathbf{Y}_i)$, $i = 1, \ldots, n$, gdzie $\mathbf{X}_i : [X_{i,L}, X_{i,U}]$ i $\mathbf{Y}_i : [Y_{i,L}, Y_{i,U}]$.

Dla danych przedziałowych współczynniki zależności nie są jednoznacznie określone, dlatego zaproponowano [3] wprowadzenie przedziałowego współczynnika Kendalla $\tau = [\tau_{n,L}, \tau_{n,U}]$

$$\tau_{n,L} = \frac{4}{n-1}\sum_{i=1}^{n-1} V_{i,L} - 1, \quad \tau_{n,U} = \frac{4}{n-1}\sum_{i=1}^{n-1} V_{i,U} - 1, \tag{1.4}$$

gdzie

$$V_{i,L} = \min_{\substack{X_j \in [X_{j,L}, X_{j,U}] \\ Y_j \in [Y_{j,L}, Y_{j,U}]}} \frac{card\{j : X_j < X_i, Y_j < Y_i\}}{n-2}, \tag{1.5}$$

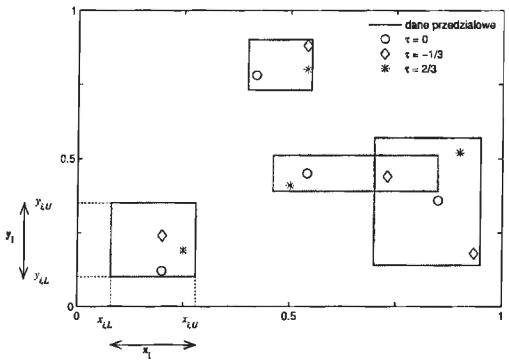
oraz

$$V_{i,U} = \max_{\substack{X_j \in [X_{j,L}, X_{j,U}] \\ Y_j \in [Y_{j,L}, Y_{j,U}]}} \frac{card\{j : X_j < X_i, Y_j < Y_i\}}{n-2}. \tag{1.6}$$

Jego wyznaczenie polega na znalezieniu takich położeń obserwacji wewnątrz odpowiadających im przedziałów, które minimalizują (1.5) bądź maksymalizują (1.6) wartość współczynnika Kendalla.

Przykład takiego zadania przedstawiono na rysunku 1.1. Czterem parom przedziałowych obserwacji odpowiadają cztery prostokąty. Zaznaczono również trzy możliwe realizacje obserwowanej zmiennej losowej dające różne wartości współczynnika korelacji $\tau \in \{-1/3, 0, 2/3\}$.

W pracach Hryniewicza i Opary [8, 9, 10] zaproponowano heurystyczne metody wyznaczania minimalnego i maksymalnego τ Kendalla dla danych przedziałowych. Zastosowano również kilka wariantów algorytmów optymalizacyjnych i porównano ich wyniki z rezultatami uzyskanymi za pomocą ewolucji różnicowej.



Rysunek 1.1. Niejednoznaczność określenia współczynnika zależności τ Kendalla dla przypadku czterech par danych przedziałowych [10]

1.2. Problem optymalizacyjny

Problem wyznaczania współczynnika τ Kendalla dla danych przedziałowych można sformułować jako zadanie w dziedzinie ciągłej bądź dyskretnej.

1.2.1. Ciągła wersja problemu

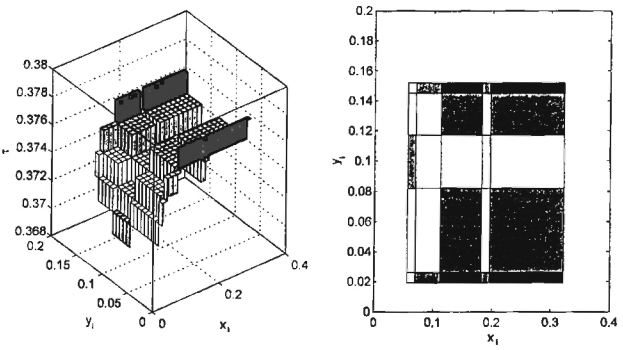
Problem optymalizacyjny zdefiniowany jest jako minimalizacja $\tau_{n,L}$ i maksymalizacja $\tau_{n,U}$ opisanych relacjami (1.4–1.6). Dziedziną funkcji celu f jest zatem hiperkostka w $\mathbb{R}^{2n}$

$$f\colon [X_{1,L}, X_{1,U}] \times \ldots \times [X_{n,L}, X_{n,U}] \times [Y_{1,L}, Y_{1,U}] \times \ldots \times [Y_{n,L}, Y_{n,U}] \to [-1, 1]. \quad (1.7)$$

Problem ten może być rozwiązany na kilka sposobów, takich jak:

— próbkowanie Monte Carlo,
— zastosowanie metod optymalizacji niewypukłej,
— wykorzystanie rozwiązań heurystycznych oparte na:
  — naśladowaniu silnych zależności,
  — przybliżaniu współczynnika ρ Pearsona.

Ponadto, rozwiązania heurystyczne można łączyć z procedurami optymalizacyjnymi.



Rysunek 1.2. Współczynnik τ Kendalla dla przekroju [9]

1.2.2. Dyskretna wersja problemu

Współczynnik τ Kendalla jest statystyką rangową i w związku z tym jest on niezmienniczy względem przekształceń zachowujących porządek punktów. Wykorzystując ten fakt można przedstawić problem optymalizacyjny (1.4–1.6) w skończonej dziedzinie możliwych porządków liniowych. W ten sposób otrzymuje się dyskretne zagadnienie, które w literaturze rozwiązywane jest poprzez:

— algorytm dokładny (efektywny tylko dla bardzo małych próbek),
— losowe generowanie rozszerzeń liniowych algorytmem Bubleya i Dyera [2] (do 30 obserwacji).

Efektywność powyższych metod jest silnie ograniczona przez liczność próbki. Z tego powodu zbadano, na ile i w jakich sytuacjach zastosowanie prostych metod heurystycznych lub ogólnych algorytmów optymalizacji ciągłej – takich jak ewolucja różnicowa – pozwala efektywnie wyznaczać przedziałowe wartości τ Kendalla.

1.2.3. Wizualizacja zadania optymalizacyjnego

Do porównania efektywności różnych metod optymalizacyjnych konieczne było stworzenie zbioru problemów testowych (benchmarku). W tym celu posłużono się ogólnym sposobem opisu zależności jakim są kopuły przedstawione szerzej na przykład w monografii [11].

Kopuły posłużyły do wygenerowania punktów o zadanym charakterze zależności, a następnie utworzenia na ich podstawie danych przedziałowych podobnych do tych, jakie przedstawiono na rysunku 1.1. Rozważono sytuację, w której wszystkie, oprócz i-tej, pary obserwacji są znane dokładnie. Obliczając τ Kendalla

dla różnych możliwych położeń i-tej pary w obrębie jej zbioru dopuszczalnego można stworzyć wykres wartości współczynnika korelacji dla dwuwymiarowego przekroju przez przestrzeń poszukiwań. Tego typu wykresy przedstawiono na rysunku 1.2. Ich analiza pozwala wskazać kilka cech rozwiązywanego problemu optymalizacji o istotnych implikacjach w kwestii wyboru właściwej metody optymalizacji:

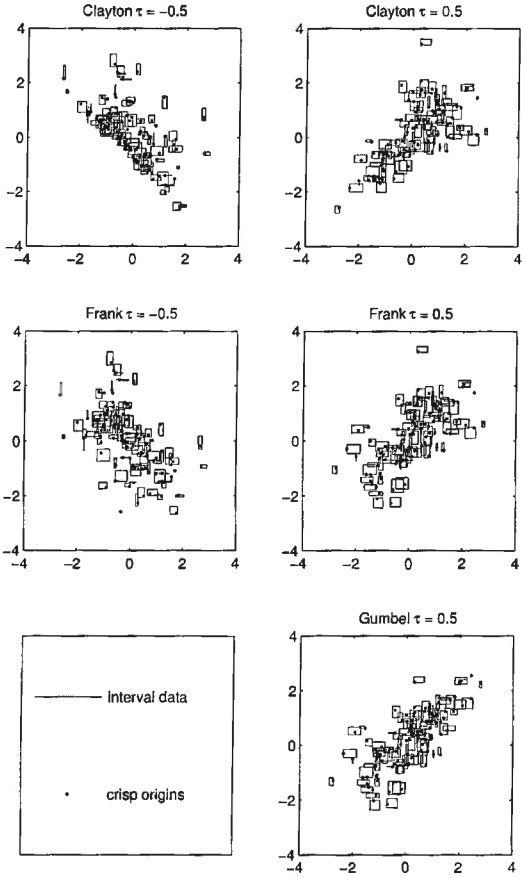
— Funkcja celu (τ Kendalla) ma charakter schodkowy, czyli przyjmuje stałe wartości dla pewnych kostek w $\mathbb{R}^{2n}$. Oznacza to również, że jego ciągłą wersję najlepiej rozwiązywać za pomocą bezgradientowych metod optymalizacji, gdyż gradient jest prawie wszędzie zerowy.
— Funkcja celu jest wielomodalna, a minima mogą występować na ograniczeniach. W związku z tym należy użyć algorytmów optymalizacji globalnej i zwrócić szczególną uwagę na metody uwzględniania ograniczeń, których rola wydaje się być często niedoceniana w analizie algorytmów [1].
— Zmiany wartości funkcji celu występują jedynie w kierunkach prostopadłych do osi układu współrzędnych. Tego typu właściwości są wykorzystywane na przykład przez algorytmy ewolucyjne z krzyżowaniem wymieniającym.
— Optymalna wartość współczynnika korelacji może być uzyskana dla kilku rozłącznych przedziałów.

1.3. Badanie symulacyjne

W celu sprawdzenia adekwatności oraz efektywności zaproponowanych metod poszukiwania przedziałowego współczynnika Kendalla przeprowadzono badania symulacyjne. Aby zbadać różne rodzaje zależności wykorzystano najczęściej stosowane kopuły: normalną, Claytona, Franka, FGM i Gumbela [11]. Sztuczne dane przedziałowe stworzono poprzez rozmycie punktów wygenerowanych dla każdej z kopuł. Badaniu poddano zależności silne ($\tau = \pm 0.9$), umiarkowane ($\tau = \pm 0.5$) i słabe ($\tau = \pm 0.1$). Dla każdego przypadku wygenerowano po 50 zestawów danych przedziałowych. Każdy z nich składał się ze 100 par przedziałów, zatem wymiar problemu optymalizacyjnego wynosił 200. Plan eksperymentu ograniczono do wersji przedstawionej w tabeli 1.1, gdyż kopuła FGM opisuje jedynie słabe zależności, a kopuła Franka tylko zależności dodatnie. Przykład wygenerowanych danych testowych zilustrowano na rysunku 1.3.

Przebadano pięć metod wyznaczania współczynnika τ Kendalla dla danych przedziałowych:

1. Heur. – zastosowanie 16 heurystyk [9]
2. DE – algorytm ewolucji różnicowej DE/rand/∞
3. HeurDE – algorytm DE/rand/∞ zainicjowany przy pomocy 16 heurystyk
4. MC – próbkowanie losowe w przestrzeni rzeczywistej



Rysunek 1.3. Dane przedziałowe wygenerowane dla umiarkowanych zależności opisanych różnymi kopułami

TABELA 1.1. Plan eksperymentu: liczba przebadanych zestawów danych dla zależności opisanych przez kopułę oraz oryginalny współczynnik τ

Kopuła	Oryginalny współczynnik τ					
	-0.9	-0.5	-0.1	0.1	0.5	0.9
normalna	50	50	50	50	50	50
Claytona	50	50	50	50	50	50
Franka	50	50	50	50	50	50
FGM			50	50		
Gumbela				50	50	50

5. BD – algorytm Bubleya i Dyera [2], błądzenie losowe w przestrzeni rozszerzeń liniowych

Porównania oparto na metodzie ustalonego kosztu obliczeniowego ustalonego na 10^4 wyznaczeń funkcji celu. Podejście to promuje algorytm Bubleya i Dyera, w którym występuje znaczny narzut obliczeniowy związany z losowym poszukiwaniem sąsiedniego rozszerzenia liniowego. Dla zniwelowania wpływu stochastycznego charakteru analizowanych metod optymalizacyjnych każdą z nich uruchomiono 7-krotnie dla każdego zestawu danych testowych aby wybrać najlepszy z otrzymanych wyników. Procedurę tę powtórzono dla każdego z 50 zestawów niezależnie wygenerowanych danych przedziałowych.

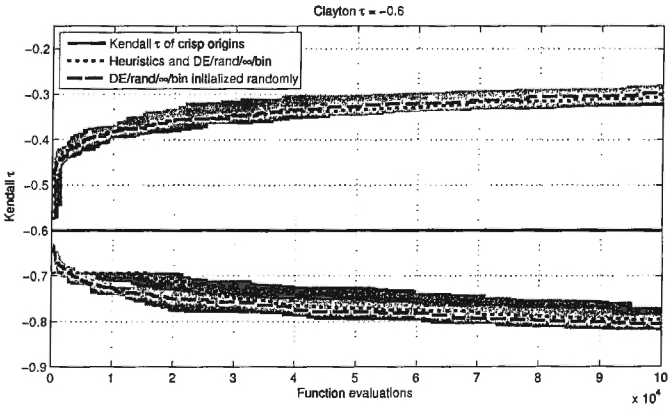
Rezultaty wyznaczania minimalnej i maksymalnej wartości współczynnika τ Kendalla zawarte są w tabelach A.1–A.6 zamieszczonych w załączniku. Najlepsze algorytmy wyróżniono tłustym drukiem. Zidentyfikowano je za pomocą sparowanego testu znaku z poprawką Bonferroniego. Okazuje się, że wybór najlepszego algorytmu nie zależy od rodzaju zależności a tylko od jej siły. Wyniki podsumowano w tabeli 1.2. Dla silnych zależności najlepsze wyniki daje zastosowanie heurystyk zaproponowanych w pracach [8, 9, 10] dla słabszych natomiast ewolucji różnicowej.

Algorytm ewolucji różnicowej pozwala polepszyć wyniki uzyskane heurystycznie i działa znacznie skuteczniej niż losowe próbkowanie przestrzeni poszukiwań. Inicjalizowanie algorytmu ewolucyjnego za pomocą heurystycznie uzyskanych, dobrych wyników okazało się zaś rozwiązaniem nieefektywnym, które może prowadzić bądź do przedwczesnej zbieżności, bądź do znacznego spowolnienia procesu poszukiwań. Ilustruje to rysunek 1.4, gdzie kropkowana niebieska linia, oznaczająca medianę błędów uzyskiwanych spośród 15 niezależnych uruchomień algorytmu, początkowo położna jest poziomo. Po kilku tysiącach obliczeń funkcji celu daje wręcz słabsze wyniki niż algorytm inicjalizowany losowo zaznaczony kreskowaną, czerwoną linią.

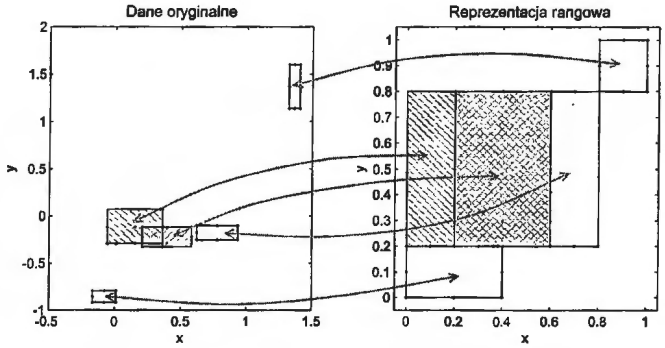
Szybkość zbieżności można znacznie poprawić uwzględniając wiedzę o rangowym charakterze problemu poprzez wzajemnie jednoznaczne przekształcenie przestrzeni poszukiwań uwzględniające rangowy charakter problemu [9, 8], co zo-

TABELA 1.2. Metody prowadzące do najlepszych rozwiązań dla problemu poszukiwania minimalnej i maksymalnej wartości współczynnika τ

zadanie	Kopuła	Oryginalny współczynnik τ					
		-0.9	-0.5	-0.1	0.1	0.5	0.9
min. τ^L	normalna Claytona Franka FGM Gumbela	Heur. HeurDE		DE HeurDE			
max. τ^U	normalna Claytona Franka FGM Gumbela	DE HeurDE				Heur. HeurDE	



Rysunek 1.4. Krzywe zbieżności przy poszukiwaniu minimalnej i maksymalnej wartości współczynnika τ Kendalla [9]



Rysunek 1.5. Przekształcenie oryginalnej przestrzeni poszukiwań z wykorzystaniem rang [9]

brazowano na rysunku 1.5. Powoduje to regulazyzację problemu ciągłego poprzez zapewnienie, że obszary poszczególnych „schodków" funkcji celu są porównywalnych rozmiarów. Umożliwia ono dodatkowo wprowadzenie kryterium zatrzymania algorytmu ewolucyjnego opartego na stopniu rozproszenia populacji.

TABELA 1.3. Zalecane metody poszukiwania maksymalnych i minimalnych wartości współczynnika τ Kendalla[9]

	Zależność ujemna			Zależność dodatnia		
	silna	umiarkowana	słaba	słaba	umiarkowana	silna
$\tau_{n,L}$	Heur.	Heur. i DE	DE.	DE	DE	DE
$\tau_{n,U}$	DE	DE	DE	DE	Heur. i DE	Heur.

Powyższe rozważania można podsumować za pomocą zaleceń zawartych w tablicy 1.3. Heurystyki okazują się bardzo efektywne dla silnych zależności i umiarkowanych zależności. W pozostałych przypadkach nie wpływają na efektywność optymalizacji [10].

Bibliografia

[1] Arabas, J., Szczepankiewicz, A. i Wroniak, T. Experimental comparison of methods to handle boundary constraints in Differential Evolution. W *Parallel Problem Solving from Nature PPSN XI*, tom 6239 z serii *Lecture Notes in Computer Science*, 411–420. 2010.

[2] Bubley, R. i Dyer, M. Faster random generation of linear extensions. W *Proc. 9th Annu. ACM-SIAM Symp. on Discrete Algorithms*, 175–186. 1998.

[3] Denœux, T., Masson, M.-H. i Hébert, P. Nonparametric rank-based statistics and signifficance tests for fuzzy data. *Fuzzy Sets and Systems*, 153(1) 1–28, 2005.

[4] Genest, C. i Rivest, L.-P. Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association*, 88(423) 1034–1043, 1993.

[5] Hébert, P.-A., Masson, M. i Denœux, T. Fuzzy rank correlation between fuzzy numbers. W *Proc. of IFSA World Congress*, 224–227. Istanbul, 2003.

[6] Hryniewicz, O. Goodman-Kruskal γ measure of dependence for fuzzy ordered categorical data. *Computational Statistics & Data Analysis*, 51(1) 323–334, 2006.

[7] Hryniewicz, O. Measures of association for fuzzy ordered categorical data. W Lopez-Diaz, M., Gil, M., Grzegorzewski, P., Hryniewicz, O. i Lawry, J., redaktorzy, *Soft Methodology and Random Information Systems*, 503–510. Springer Verlag, 2013.

[8] Hryniewicz, O. i Opara, K. Computation of the measures of dependence for imprecise data. W Atanassov, K., Baczynski, M., Drewniak, J., Kacprzyk, J., Krawczak, K., Szmidt, E., Wygralak, M. i Zadrozny, S., redaktorzy, *New Developments in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics Volume I: Foundations*, 99–112. SRI PAS, 2012.

[9] Hryniewicz, O. i Opara, K. Efficient calculation of kendall's τ for interval data. W Kruse, R., Berthold, M. R., Moewes, C., Gil, ., M, Grzegorzewski, P. i Hryniewicz, O., redaktorzy, *Synergies of Soft Computing and Statistics for Intelligent Data Analysis*, Advances in Intelligent Systems and Computing, 203–210. Springer, 2012.

[10] Hryniewicz, O. i Opara, K. On computational problems in the analysis of statistical dependence for imprecise data. Raport techniczny RB/13/2012, IBS PAN, 2012.

[11] Nelsen, R. *Introduction to Copulas*. Springer, 1999.

[12] Pratap, R. *Matlab 7 dla naukowców i inzynierów*. Mikom, Warszawa, 2007.