Dwa nowe algorytmy klasyfikacji farmakokinetycznych szeregów czasowych

Two novel classification algorithms of the pharmacokinetic time series

Piotr Wilczek[1]

Streszczenie: Przedkładana praca zawiera propozycję dwóch nowych miar odległości pomiędzy szeregami czasowymi. Mianowicie, wprowadzamy tzw. rekurencyjno-kanoniczną miarę odległości oraz jej krzyżową formę. Nasze podejście opieramy na bezprogowych (krzyżowych) macierzach rekurencyjnych, których koncepcje zostały rozwinięte w ramach (krzyżowej) analizy rekurencyjnej szeregów czasowych oraz na kanonicznej mierze odległości pomiędzy dwoma wielowymiarowymi zbiorami danych, której koncepcja została rozwinięta w obrębie chemometrii. Wyczerpujące symulacje komputerowe przeprowadzone na 27 farmakokinetycznych szeregach czasowych pokazały, iż algorytmy bazujące na nowo opracowanych funkcjach odległości są bardziej efektywne niż algorytmy bazujące na klasycznych miarach $L2$ i DTW oraz na ich modyfikacjach.

Abstract: In the present contribution, we proposed two novel distance measures between time series. Namely, we introduced the so-called recurrence-canonical measure of distance as well as its cross form. Our approach is based on the notion of the unthresholded (cross-)recurrence matrices developed in the field of the (Cross-)Recurrence Quantification Analysis and the notion of the canonical measure of distance between multidimensional datasets developed in the field of chemometrics. The extensive computer simulations carried out on 27 pharmacokinetic time series showcased that the algorithms based on the newly designed distance functions outperform the protocols based on the classical $L2$ and DTW functions and on their modifications.

Słowa kluczowe: szereg czasowy, (krzyżowa) analiza rekurencyjna, kanoniczna miara odległości, farmakokinetyka

Keywords: time series, (cross-)recurrence quantification analysis, canonical measure of distance, farmacokinetics

[1] Computer Laboratory, 61–163 Poznań, e-mail: piotr.wilczek.net@onet.pl, ORCID 0000-0003-2758-343X.

1. Wprowadzenie

W ostatnich latach można zaobserwować wzrost popularności analiz opartych na tzw. „danych temporalnych". Mianowicie, biorąc pod uwagę powszechne wykorzystanie nowoczesnych technologii cyfrowych, możliwa jest rejestracja dużej liczby pomiarów czasowych w obrębie nauk podstawowych (np. w fizyce, chemii, biologii czy też geologii), jak również nauk stosowanych (np. w farmakologii, medycynie, robotyce czy też automatyce). W związku z tym nastąpił dramatyczny wzrost zainteresowania gromadzeniem oraz eksploracją danych temporalnych, co z kolei zaowocowało dużą liczbą artykułów naukowych proponujących nowatorskie techniki indeksacji, klasyfikacji, grupowania, aproksymacji oraz prognozowania szeregów czasowych. W szczególności zaproponowano wiele nowych miar odległości pomiędzy danymi temporalnymi. Przedkładana praca ma na celu wprowadzenie dwóch nowych funkcji odległości pomiędzy szeregami czasowymi. Efektywność nowo zaproponowanych miar temporalnych zostanie przetestowana na 27 farmakokinetycznych przykładowych zbiorach danych.

1. Podstawy teoretyczne i terminologia

Z matematycznego punktu widzenia sygnałem nazywamy zmianę jednej (lub kilku) wielkości fizycznie mierzalnych w zależności od zmian innej wielkości. W powszechnym użyciu pojęcie sygnału stosuje się do zmian pewnych fizycznie mierzalnych wielkości (np. chemicznych, biologicznych czy też biochemicznych) w funkcji czasu. Taki sygnał nazywamy szeregiem czasowym. Przypomnijmy, iż szereg czasowy to ciąg pomiarów (obserwacji) uporządkowany w czasie. W prezentowanej pracy przyjmujemy, iż zmienna niezależna (tj. zmienna czasowa) jest dyskretna. A więc, z formalnego punktu widzenia, szereg czasowy T to ciąg par uporządkowanych o postaci: $T = [(t_1, x_1), (t_2, x_2), \dots, (t_i, x_i), \dots, (t_n, x_n)]$ dla $t_1 < t_2 < \cdots < t_i < \cdots < t_n$ oraz dla zbioru indeksów $I = \{1,\ 2, \dots, i, \dots, n\}$, gdzie każdy wyraz x_i to wynik pomiaru (obserwacji) zmiennej zależnej w d-wymiarowej przestrzeni cech (ang. *feature space*) oraz każdy element t_i to punkt (krok) czasowy (ang. *timestamp*), w którym dany pomiar (dana obserwacja) został zarejestrowany. W prezentowanej pracy zakładamy, iż wszystkie analizowane dane temporalne są jednowymiarowe (ang. *univariate*), a więc $d = 1$ dla każdego z rozpatrywanych szeregów czasowych. Ponadto przyjmujemy, iż każdy badany ciąg temporalny jest regularny, a więc punkty czasowe t_i wszystkich rozpatrywanych szeregów czasowych są równomiernie rozmieszczone. Liczba zarejestrowanych pomiarów (obserwacji), (n), to długość szeregu czasowego T. Z kolei, i-ty wyraz szeregu czasowego T oznaczamy

przez $T(i)$. Etykietowany temporalny zbiór danych $\boldsymbol{TD}$ to kolekcja $\boldsymbol{T} = \{T_g\}_{g=1}^{m}$ szeregów czasowych z których każdy ma długość n razem z określonym (ang. *predefined*) dyskretnym wektorem[2] etykiet $\boldsymbol{C}$ o długości m. Konkludując, etykietowany temporalny zbiór danych to struktura o postaci $\boldsymbol{TD} = (\boldsymbol{T}, \boldsymbol{C}) = (T_g, C^{T_g})$, gdzie $T_g \in \boldsymbol{T}$ to g-ty szereg czasowy, natomiast $C^{T_g} \in \boldsymbol{C}$ to jego etykieta. W przedkładanej pracy przyjmujemy, iż dwie etykiety $C^{T_g}, C^{T_h} \in \boldsymbol{C}$ są równoważne ($\cong$) wtedy i tylko wtedy gdy są identyczne, tj. $C^{T_g} \cong C^{T_h} \leftrightarrow C^{T_g} = C^{T_h}$, gdzie $\leftrightarrow$ to logiczny spójnik równoważności międzyzdaniowej. Dla szeregu czasowego T o długości n, jego pierwsza dyskretna pochodna T' to nowy szereg czasowy o długości $n-1$, którego wyrazy dane są zależnością[3]: $T'(i) \coloneqq T(i+1) - T(i)$, gdzie $i = 1, 2, \ldots, n-1$.

3. Nowe algorytmy obliczania odległości pomiędzy szeregami czasowymi

Proponowane w tej części pracy nowe techniki obliczania odległości pomiędzy danymi temporalnymi oparte są na analizie rekurencyjnej (ang. *recurrence quantification analysis*) szeregów czasowych RQA[4], jej *krzyżowej* formie (ang. *cross-recurrence quantification analysis*) $CRQA$ oraz na metodzie obliczania odległości pomiędzy wielowymiarowymi zbiorami danych opracowanej przez Roberta Todeschiniego oraz jego współpracowników[5]. Techniki te składają się z następujących kroków:

1. Dla każdego szeregu czasowego w analizowanej bazie danych, tj. $T_g \in \boldsymbol{T}$ obliczamy jego bezprogową macierz rekurencyjną (ang. *unthresholded recurrence matrix*) UR^g lub jej krzyżową formę (ang. *unthresholded cross-recurrence matrix*) UCR^g zgodnie z następującymi formułami[6]:

$$UR_{ij}^{g} \coloneqq \left\| T_g(i) - T_g(j) \right\|_2$$

oraz

[2] Z matematycznego punktu widzenia wektor $\boldsymbol{C}$ to multizbiór.
[3] T. Górecki, M. Łuczak, *Multivariate time series classification with parametric derivative dynamic time warping*, „Expert Systems with Applications" 2015, nr 42, s. 2307.
[4] J.–P. Eckmann, S. O. Kamphorst, D. Ruelle, *Recurrence plots of dynamical systems*, „Europhysics Letters" 1987, nr 5, s. 973–977; N. Marwan, M. C. Romano, M. Thiel, J. Kurths, *Recurrence plots for the analysis of complex systems*, „Physics Reports" 2007, nr 438, s. 237–329.
[5] R. Todeschini, D. Ballabio, V. Consonni, A. Manganaro, A. Mauri, *Canonical measure of correlation (CMC) and canonical measure of distance (CMD) between sets of data. Part 1. Theory and simple chemometric applications*, „Analytica Chimica Acta" 2009, nr 648, s. 45–51.
[6] N. Marwan *et al.*, *op. cit*.

$$UCR_{ij}^{g} \coloneqq \left\| T_g(i) - T'_g(j) \right\|_2,$$

gdzie $\|\cdot\|_2$ to norma *euklidesowa*, a $T'_g(j)$ to j-ty wyraz pierwszej dyskretnej pochodnej T'_g szeregu temporalnego T_g. Macierz UR^g to symetryczna względem głównej przekątnej macierz kwadratowa o wymiarach $n \times n$, natomiast macierz UCR^g to prostokątna macierz o wymiarach $n \times (n-1)$[7].

2. W drugim kroku wyrazy otrzymanych macierzy, tj. UR_{ij}^{g} oraz UCR_{ij}^{g}, transformujemy według zależności[8]:

$$TUR_{ij}^{g} \coloneqq e^{-UR_{ij}^{g}}$$

oraz

$$TUCR_{ij}^{g} \coloneqq e^{-UCR_{ij}^{g}}$$

otrzymując transformowaną bezprogową macierz rekurencyjną (ang. *transformed unthresholded recurrence matrix*) TUR^g oraz jej krzyżową wersję (ang. *transformed unthresholded cross-recurrence matrix*) $TUCR^g$.

3. W trzecim kroku zapisujemy macierze TUR^g oraz $TUCR^g$ w formie kolumnowej[9]:

$$TUR^{g} = \begin{bmatrix} TUR_{11}^{g} & TUR_{12}^{g} & \dots & TUR_{1n}^{g} \\ TUR_{21}^{g} & TUR_{22}^{g} & \dots & TUR_{2n}^{g} \\ \cdots & \dots & \ddots & \dots \\ \dots & \dots & \ddots & \dots \\ TUR_{n1}^{g} & TUR_{n2}^{g} & \dots & TUR_{nn}^{g} \end{bmatrix} = \left[TUR_{\cdot 1}^{g}, TUR_{\cdot 2}^{g}, \dots, TUR_{\cdot n}^{g} \right]$$

oraz

[7] Przypomnijmy, iż macierze rekurencyjne (progowe lub bezprogowe) zawsze są symetrycznymi macierzami kwadratowymi, mającymi na celu porównywanie odległości pomiędzy wyrazami tego samego szeregu czasowego. Natomiast, krzyżowe macierze rekurencyjne progowe lub bezprogowe zawsze służą do porównywania odległości pomiędzy wyrazami dwóch różnych szeregów czasowych i tylko w wyjątkowych przypadkach, gdy długości obu szeregów są równe, są one kwadratowe. Por. *Ibidem*. Bezprogowe krzyżowe macierze rekurencyjne prównujące odległości pomiędzy wyrazami oryginalnego (wejściowego) szeregu czasowego i jego pierwszej dyskretnej pochodnej rozpatrywane były w: P. Wilczek, *An application of the local binary pattern algorithm and its uniform variant to improve the recurrence and cross-recurrence quantification analyses of the pharmacologically important time series* [w:] *Recent Advances in Computational Oncology and Personalized Medicine*. *Vol. 2. The Challenges of the Future !*, red. K. Krukiewicz, M. Marczyk, M. Bugdol, S. Bajkacz, Z. Ostrowski, Gliwice 2022, s. 128–152.

[8] D. Eroglu, T. K. DM. Peron, N. Marwan, F. A. Rodigues, L. da F. Costa, M. Sebek, I. Z. Kiss, J. Kurths, *Entropy of weighted recurrence plots*, „Physical Review E" 2014, nr 90, s. 042919.

[9] Por. formuły 3 i 4 w R. Todeschini *et al.*, *op. cit*.

$$TUCR^g = \begin{bmatrix} TUCR_{11}^g & TUCR_{12}^g & \dots & TUCR_{1(n-1)}^g \\ TUCR_{21}^g & TUCR_{22}^g & \dots & TUCR_{2(n-1)}^g \\ \dots & \dots & \ddots & \dots \\ \dots & \dots & \ddots & \dots \\ TUCR_{n1}^g & TUCR_{n2}^g & \dots & TUCR_{n(n-1)}^g \end{bmatrix} = \left[TUCR_{\cdot 1}^g, TUCR_{\cdot 2}^g, \dots, TUCR_{\cdot (n-1)}^g \right].$$

W powyższym zapisie symbole $TUR_{\cdot i}^g$ oraz $TUCR_{\cdot i}^g$ oznaczają i-tą kolumnę, odpowiednio macierzy TUR^g oraz macierzy $TUCR^g$.

4. W czwartym kroku obliczamy macierz współczynników korelacji Pearsona r pomiędzy dwiema macierzami TUR^g oraz TUR^h, odpowiadającymi dwóm szeregom czasowym $T_g, T_h \in \boldsymbol{T}$, pomiędzy którymi chcemy obliczyć odległość rekurencyjno-kanoniczną, lub pomiędzy dwiema macierzami $TUCR^g$ oraz $TUCR^h$, odpowiadającymi dwóm szeregom czasowym $T_g, T_h \in \boldsymbol{T}$, pomiędzy którymi chcemy obliczyć krzyżową odległość rekurencyjno-kanoniczną. W pierwszym przypadku macierz korelacji oznaczona jest przez r_{TUR^g,TUR^h}, a w drugim przez $r_{TUCR^g,TUCR^h}$. Mają one postać:

$$r_{TUR^g,TUR^h} = \begin{bmatrix} r\left(TUR_{\cdot 1}^g, TUR_{\cdot 1}^h\right) & r\left(TUR_{\cdot 1}^g, TUR_{\cdot 2}^h\right) & \dots & r\left(TUR_{\cdot 1}^g, TUR_{\cdot n}^h\right) \\ r\left(TUR_{\cdot 2}^g, TUR_{\cdot 1}^h\right) & r\left(TUR_{\cdot 2}^g, TUR_{\cdot 2}^h\right) & \dots & r\left(TUR_{\cdot 2}^g, TUR_{\cdot n}^h\right) \\ \dots & \dots & \ddots & \dots \\ \dots & \dots & \ddots & \dots \\ r\left(TUR_{\cdot n}^g, TUR_{\cdot 1}^h\right) & r\left(TUR_{\cdot n}^g, TUR_{\cdot 2}^h\right) & \dots & r\left(TUR_{\cdot n}^g, TUR_{\cdot n}^h\right) \end{bmatrix}$$

oraz

$$r_{TUCR^g,TUCR^h} = \begin{bmatrix} r\left(TUCR_{\cdot 1}^g, TUCR_{\cdot 1}^h\right) & r\left(TUCR_{\cdot 1}^g, TUCR_{\cdot 2}^h\right) & \dots & r\left(TUCR_{\cdot 1}^g, TUCR_{\cdot (n-1)}^h\right) \\ r\left(TUCR_{\cdot 2}^g, TUCR_{\cdot 1}^h\right) & r\left(TUCR_{\cdot 2}^g, TUCR_{\cdot 2}^h\right) & \dots & r\left(TUCR_{\cdot 2}^g, TUCR_{\cdot (n-1)}^h\right) \\ \dots & \dots & \ddots & \dots \\ \dots & \dots & \ddots & \dots \\ r\left(TUCR_{\cdot (n-1)}^g, TUCR_{\cdot 1}^h\right) & r\left(TUCR_{\cdot (n-1)}^g, TUCR_{\cdot 2}^h\right) & \dots & r\left(TUCR_{\cdot (n-1)}^g, TUCR_{\cdot (n-1)}^h\right) \end{bmatrix}.$$

Wyraz na pozycji (i, j) macierzy r_{TUR^g,TUR^h} to współczynnik korelacji liniowej Pearsona pomiędzy i-tą kolumną macierzy TUR^g a j-tą kolumną macierzy TUR^h. A więc w ogólnym przypadku ma on postać $r\left(TUR_{\cdot i}^g, TUR_{\cdot j}^h\right)$[10]. Ponieważ obie macierze mają n kolumn, macierz r_{TUR^g,TUR^h} ma wymiary $n \times n$. Podobnie, wyraz na pozycji (i, j) macierzy $r_{TUCR^g,TUCR^h}$

[10] Por. formuła 5 w *Ibidem*.

to współczynnik korelacji liniowej Pearsona pomiędzy i-tą kolumną macierzy $TUCR^g$ a j-tą kolumną macierzy $TUCR^h$. Wtedy ma on postać $r\left(TUCR^g_{\cdot i}, TUCR^h_{\cdot j}\right)$. Ponieważ obie macierze mają $n-1$ kolumn, macierz ich współczynników korelacji $r_{TUCR^g,TUCR^h}$ ma wymiary $(n-1)\times(n-1)$. Zauważmy, iż w ogólnym przypadku $r_{A,B} \neq r_{B,A}$, gdzie A oraz B to dowolne macierze mające równą ilość rzędów.

5. W piątym kroku obie macierze współczynników korelacji liniowych są symetryzowane względem głównej przekątnej za pomocą mnożenia macierzowego[11]:

$$Q_{RCMD} = r_{TUR^g,TUR^h} \times r_{TUR^h,TUR^g}$$

lub

$$Q^*_{RCMD} = r_{TUR^h,TUR^g} \times r_{TUR^g,TUR^h},$$

gdzie $\left(r_{TUR^g,TUR^h}\right)^T = r_{TUR^h,TUR^g}$

oraz

$$Q_{CRCMD} = r_{TUCR^g,TUCR^h} \times r_{TUCR^h,TUCR^g}$$

lub

$$Q^*_{CRCMD} = r_{TUCR^h,TUCR^g} \times r_{TUCR^g,TUCR^h},$$

gdzie $\left(r_{TUCR^g,TUCR^h}\right)^T = r_{TUCR^h,TUCR^g}$. Działanie $(\cdot)^T$ to transpozycja macierzowa. Macierze produktowe Q_{RCMD} oraz Q^*_{RCMD} mają wymiary $n \times n$, natomiast macierze Q_{CRCMD} oraz Q^*_{CRCMD} mają wymiary $(n-1)\times(n-1)$.

6. W szóstym kroku obliczmy wartości własne macierzy produktowych Q_{RCMD} (lub Q^*_{RCMD}) oraz Q_{CRCMD} (lub Q^*_{CRCMD}). Macierz produktowa Q_{RCMD} ma takie same niezerowe wartości własne jak macierz Q^*_{RCMD}. Podobnie, macierz produktowa Q_{CRCMD} ma takie same niezerowe wartości własne jak macierz Q^*_{CRCMD}.

7. W siódmym kroku obliczamy wartości odległości rekurencyjno-kanonicznej pomiędzy szeregami czasowymi $T_g, T_h \in \boldsymbol{T}$ według wzoru[12]:

[11] Por. Formuła 6 w *Ibidem*.

[12] W oryginalnej pracy Todeschiniego i współautorów kanoniczna miara odległości $CMD(A,B)$ pomiędzy dwoma wielowymiarowymi zbiorami danych A i B wyrażona jest formułą: $CMD(A,B) \coloneqq p_A + p_B - 2\sum_j^k \sqrt{\lambda_j}$, gdzie p_A oraz p_B to, odpowiednio liczba zmiennych (kolumn) w zbiorze A i w zbiorze B, λ_i to wartości własne

$$RCMD(T_g, T_h) \coloneqq 2n - 2\sum_{j=1}^{k}\sqrt{\lambda_j^{RCMD}},$$

gdzie n to długość obu szeregów, λ_j^{RCMD} to wartości własne macierzy Q_{RCMD} (lub Q_{RCMD}^*), a indeks k równy jest ilości niezerowych wartości własnych λ_j^{RCMD}. Przy tych założeniach zachodzi, iż dla dowolnych $T_g, T_h \in \boldsymbol{T}$, $0 \leq RCMD(T_g, T_h) \leq 2n$. Z kolei krzyżowa odległość rekurencyjno-kanoniczna pomiędzy szeregami $T_g, T_h \in \boldsymbol{T}$ dana jest zależnością:

$$CRCMD(T_g, T_h) \coloneqq 2(n-1) - 2\sum_{j=1}^{k}\sqrt{\lambda_j^{CRCMD}},$$

gdzie n to długość obu szeregów, λ_j^{CRCMD} to wartości własne macierzy Q_{CRCMD} (lub Q_{CRCMD}^*), a indeks k równy jest ilości niezerowych wartości własnych λ_j^{CRCMD}. W tym przypadku zachodzi, iż dla dowolnych $T_g, T_h \in \boldsymbol{T}$, $0 \leq CRCMD(T_g, T_h) \leq 2(n-1)$.

Aby zilustrować powyższy algorytm, rozpatrzmy dwa czteroelementowe szeregi czasowe T_1 oraz T_2, tj. $T_1 = \langle 1, 3, 8, 5\rangle$ oraz $T_2 = \langle 2, 0, 9, 7\rangle$. Ich bezprogowe macierze rekurencyjne oraz ich transponowane formy mają postać:

$$UR^1 = \begin{bmatrix} 0 & 2 & 7 & 4 \\ 2 & 0 & 5 & 2 \\ 7 & 5 & 0 & 3 \\ 4 & 2 & 3 & 0 \end{bmatrix}, UR^2 = \begin{bmatrix} 0 & 2 & 7 & 5 \\ 2 & 0 & 9 & 7 \\ 7 & 9 & 0 & 2 \\ 5 & 7 & 2 & 0 \end{bmatrix}$$

oraz

$$TUR^1 = \begin{bmatrix} 1 & 0{,}1353 & 0{,}0009 & 0{,}0183 \\ 0{,}1353 & 1 & 0{,}0067 & 0{,}1353 \\ 0{,}0009 & 0{,}0067 & 1 & 0{,}0498 \\ 0{,}0183 & 0{,}1353 & 0{,}0498 & 1 \end{bmatrix},$$

$$TUR^2 = \begin{bmatrix} 1 & 0{,}1353 & 0{,}0009 & 0{,}0067 \\ 0{,}1353 & 1 & 0{,}0001 & 0{,}0009 \\ 0{,}0009 & 0{,}0001 & 1 & 0{,}1353 \\ 0{,}0067 & 0{,}0009 & 0{,}1353 & 1 \end{bmatrix}.$$

Z kolei macierze korelacji liniowych Pearsona mają postać:

odpowiednich macierzy produktowych, a indeks k równy jest ilości niezerowych wartości własnych λ_i. Por. formuła 7 w *Ibidem*.

$$r_{TUR^1,TUR^2} = \begin{bmatrix} 0{,}9999 & -0{,}083 & -0{,}4688 & -0{,}4424 \\ -0{,}1417 & 0{,}9913 & -0{,}5105 & -0{,}3441 \\ -0{,}4229 & -0{,}4137 & 0{,}9961 & -0{,}1655 \\ -0{,}445 & -0{,}3001 & -0{,}2315 & 0{,}9821 \end{bmatrix}$$

oraz

$$\left(r_{TUR^1,TUR^2}\right)^T = r_{TUR^2,TUR^1} = \begin{bmatrix} 0{,}9999 & -0{,}1417 & -0{,}4229 & -0{,}445 \\ -0{,}083 & 0{,}9913 & -0{,}4137 & -0{,}3001 \\ -0{,}4688 & -0{,}5105 & 0{,}9961 & -0{,}2315 \\ -0{,}4424 & -0{,}3441 & -0{,}1655 & 0{,}9821 \end{bmatrix}$$

Obie macierze produktowe dane są poniżej:

$$Q_{RCMD} = r_{TUR^1,TUR^2} \times r_{TUR^2,TUR^1} = \begin{bmatrix} 1{,}4223 & 0{,}1676 & -0{,}7823 & -0{,}746 \\ 0{,}1676 & 1{,}3819 & -0{,}8018 & -0{,}4542 \\ -0{,}7823 & -0{,}8018 & 1{,}3696 & -0{,}0808 \\ -0{,}746 & -0{,}4542 & -0{,}0808 & 1{,}3061 \end{bmatrix}$$

oraz

$$Q^*_{RCMD} = r_{TUR^2,TUR^1} \times r_{TUR^1,TUR^2} = \begin{bmatrix} 1{,}3968 & 0{,}0851 & -0{,}7147 & -0{,}7606 \\ 0{,}0851 & 1{,}2509 & -0{,}8099 & -0{,}5306 \\ -0{,}7147 & -0{,}8099 & 1{,}5262 & -0{,}0091 \\ -0{,}7606 & -0{,}5306 & -0{,}0091 & 1{,}306 \end{bmatrix}$$

Wartości własne obu macierzy produktowych Q_{RCMD} oraz Q^*_{RCMD} są równe 2,838209; 1,513631; 1,128044; 0. Trzy pierwsze z nich służą do obliczenia wartości odległości $RCMD(T_1, T_2)$. Wynosi ona 0,04582047. Odległość $CRCMD(T_1, T_2)$ obliczana jest w sposób podobny. Wynosi ona 2,256611.

Nowo zaproponowane protokoły obliczania odległości pomiędzy danymi temporalnymi będą testowane w przykładowych zadaniach klasyfikacyjnych w części czwartej poniższej pracy. Przypomnijmy, iż dla kolekcji $\boldsymbol{T} = \{T_g\}_{g=1}^{m}$ szeregów czasowych i dyskretnego wektora ich etykiet $\boldsymbol{C}$ (w naszym przypadku jest to wektor binarny) zagadnienie klasyfikacji polega na aproksymacji funkcji o postaci: $c: \boldsymbol{T} \rightarrow \boldsymbol{C}$ (gdzie $c(T_g) = C^{T_g}$ dla $\forall T_g \in \boldsymbol{T}$) funkcją $\hat{c}: \boldsymbol{T} \rightarrow \boldsymbol{C}$, tak aby spełniony był warunek: $\hat{c}(T_g) = C^{T_g}$ dla $\forall T_g \in \boldsymbol{T}$. Funkcja c to klasyfikator aprioryczny (ang. *predefined*), natomiast funkcja $\hat{c}$ to klasyfikator aposterioryczny[13]. A więc, jeżeli:

[13] Oczywiście pojęcia klasyfikatora apriorycznego oraz aposteriorycznego mają sens tylko w odniesieniu do klasyfikacji formalnej (matematycznej). Oznacza to, iż chociaż elementy dyskretnego wektora etykiet C mogły zostać przyporządkowane poszczególnym danym temporalnym na podstawie procedur empirycznych, to podczas

$c(T_g) = \hat{c}(T_g)$ dla $\forall T_g \in \boldsymbol{T}$, to mówimy, że klasyfikator aposterioryczny w pełni (perfekcyjnie, doskonale) aproksymuje klasyfikator aposterioryczny. Natomiast jeżeli tylko dla większości $T_g \in \boldsymbol{T}$ zachodzi warunek $c(T_g) = \hat{c}(T_g)$, to mówimy, iż klasyfikator aposterioryczny tylko częściowo przybliża klasyfikator aprioryczny. Problem klasyfikacji polega na znalezieniu jak najdokładniejszego klasyfikatora aposteriorycznego. Tym samym wydajność klasyfikatora aposteriorycznego może być mierzona w procentach poprawnie sklasyfikowanych szeregów czasowych w analizowanym temporalnym zbiorze danych. Postępując zgodnie z sugestią Eamonna Keogha oraz ShrutiKasetty[14], w części czwartej poniższej pracy nowo zaproponowane funkcje odległości pomiędzy danymi temporalnymi będą testowane przez pryzmat klasycznego klasyfikatora jednego najbliższego sąsiada (ang. *1-nearest neighbor classifier* (1NN)) na 27 przykładowych farmakokinetycznych zbiorach danych.

4. Porównawcze algorytmy obliczania odległości pomiędzy szeregami czasowymi

Zgodnie z ogólnie przyjętą metodologią walidacji nowo zaproponowanych miar odległości pomiędzy danymi temporalnymi[15] w części czwartej prezentowanej pracy porównamy efektywność nowych technik klasyfikacji szeregów czasowych z efektywnością technik bazujących na klasycznych miarach odległości oraz na niedawno wprowadzonych funkcjach odległości. Mianowicie, wydajność w przykładowych zadaniach klasyfikacyjnych nowych algorytmów będzie porównana z wydajnością algorytmu euklidesowego ($L2$), jego trzech modyfikacji ($DL2$, $PDL2$ oraz $CIL2$), algorytmu DTW oraz jego trzech ulepszeń ($DDTW$, $PDDTW$ oraz $CIDTW$). Przypomnijmy, iż dla dwóch szeregów czasowych T_g oraz T_h, których wyrazy indeksowane są tym samym zbiorem I, algorytm euklidesowy ma postać $L2(T_g, T_h) \coloneqq \sqrt{\sum_{i=1}^{n}\left[T_g(i) - T_h(i)\right]^2}$. Z kolei miara DTW obliczana jest według procedury zaimplementowanej w pakietach obliczeniowych języka *R*, tj. w pakietach *dtw*[16] oraz *TSdist*[17]. Dla dwóch szeregów czasowych T_g oraz T_h, których wyrazy indeksowane są tym samym zbiorem I, niech $M(T_g, T_h)$ będzie kwadratową $n \times n$ dodatnią macierzą odległości (tzw.

matematycznego procesu klasyfikacji owo przyporządkowanie jest czymś z góry ustalonym (ang. *predefined*), a więc z wewnętrznego punktu widzenia klasyfikacji formalnej jest to przyporządkowanie aprioryczne.

[14] E. Keogh, S. Kasetty, *On the need for time series data mining benchmarks: a survey and empirical demonstration* [w:] *Proceedings of the eight ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York 2002, s. 102–111.

[15] *Ibidem*, a także P. Wilczek, *op. cit.*

[16] T. Giorgino, *Computing and visualizing dynamic time warping alignments in R: The dtw package*, „Journal of Statistical Software" 2009, nr (7), s. 1–24.

[17] U. Mori, A. Mendiburu, J. A. Lozano, *Distance measures for time series in R: The TSdist Package*, „The R Journal" 2016, nr 2, s. 451–459.

macierzą kosztów lokalnych), której wyraz M_{ij} dany jest zależnością: $M_{ij} \coloneqq \left|T_g(i) - T_g(j)\right|$. Wtedy tak zwana krzywa (ang. *warping path*) P o postaci: $P \coloneqq [(e_1, f_1), (e_2, f_2), \dots, (e_i, f_i), \dots, (e_s, f_s)]$ to ciąg punktów (tj. par indeksów) definiujących trawersalę macierzy $M(T_g, T_h)$. W definicji odległości DTW zakładamy, iż krzywa P musi spełniać następujące warunki: $(e_1, f_1) = M_{11}$ oraz $(e_s, f_s) = M_{nn}$, jak również $0 \leq e_{i+1} - e_i \leq 1$ oraz $0 \leq f_{i+1} - f_i \leq 1$ dla $\forall i < n$. A więc krzywa P to ścieżka biegnąca od elementu M_{11} macierzy $M(T_g, T_h)$ do elementu M_{nn} tejże macierzy, przy czym przy przejściu od wyrazu M_{11} do wyrazu M_{nn} dozwolone są tylko kroki: (0,1), (1,0) oraz (1,1). Niech $p_i = \gamma M_{e_i f_i}$ będzie odległością globalną (tj. kumulacyjną) pomiędzy elementem w pozycji e_i szeregu T_g a elementem w pozycji f_i szeregu T_h dla i-tej pary punktów (tj. indeksów) na krzywej P. Wtedy odległość pomiędzy szeregami czasowymi T_g i T_h wzdłuż krzywej P, D_P, dana jest zależnością $D_P(T_g, T_h) \coloneqq \sum_{i=1}^{s} p_i$. Niech $\mathcal{P}$ będzie przestrzenią wszystkich możliwych krzywych P. Wtedy ścieżka DTW, P^*, to krzywa minimalizująca odległość D_P, tj. $P^* \coloneqq \min_{P \in \mathcal{P}} D_P(T_g, T_h)$. Odległość DTW pomiędzy dwoma ciągami temporalnymi T_g oraz T_h to odległość wzdłuż ścieżki P^*[18]. Można ją znaleźć za pomocą procedur programowania dynamicznego[19] na podstawie następującego warunku rekurencyjnego, umożliwiającego obliczenie globalnej odległości $\gamma M_{e_i f_i}$ pomiędzy elementami szeregów T_g oraz T_h na podstawie znajomości wartości kosztów lokalnych: $\gamma M_{e_1 f_1} = M_{11}$ oraz dla $i, j \neq 1$ $\gamma M_{e_i f_i} = M_{ij} + \min\{\gamma M_{e_{i-1} f_{i-1}}, \gamma M_{e_{i-1} f_i}, \gamma M_{e_i f_{i-1}}\}$. A więc obliczenie kumulacyjne odległości $\gamma M_{e_i f_i}$ pomiędzy elementem w pozycji e_i szeregu T_g a elementem w pozycji f_i szeregu T_h dla i-tej pary indeksów na krzywej P redukuje się do znajomości odległości lokalnej bieżącego elementu M_{ij} oraz kumulacyjnych odległości elementów przyległych (ang. *adjacent*), tj. odległości $\gamma M_{e_{i-1} f_{i-1}}$, $\gamma M_{e_{i-1} f_i}$ oraz $\gamma M_{e_i f_{i-1}}$. Dla dwóch szeregów czasowych T_g oraz T_h, których wyrazy indeksowane są tym samym zbiorem I, oraz funkcji $DF \in \{L2, DTW\}$, tzw. algorytm odległości pochodnej DDF ma postać $DDF(T_g, T_h) \coloneqq DF(T_g', T_h')$, gdzie T_g' oraz T_h'

[18] D. J. Berndt, J. Clifford, *Using dynamic time warping to find patterns in time series* [w:] *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, 1994, s. 359–370.

[19] Wyczerpujące informacje na temat procedur programowania dynamicznego czytelnik znajdzie w następujących pozycjach: T. Giorgino, *op. cit*., E. Keogh, C. A. Ratanamahatana, *Exact indexing of dynamic time warping*, „Knowledge and Information Systems" 2005, nr 7, s. 358–386, L. Rabiner, B.–H. Juang, *Fundamentals of Speech Recognition*, New Jersey 1993, D. Sankoff, J. Kruskal, *Time Warps, String Edits, and Macromolecules. The Theory and Practice of Sequence Comparison*, Stanford 1999.

to pierwsze dyskretne pochodne szeregów T_g i T_h[20]. Z kolei dla dwóch szeregów czasowych T_g oraz T_h, których wyrazy indeksowane są tym samym zbiorem I, oraz funkcji $DF \in \{L2, DTW\}$, tzw. algorytm pochodnej odległości parametrycznej $PDDF$ to ważona wypukła kombinacja o postaci $PDDF(T_g, T_h) \coloneqq aDF(T_g, T_h) + bDDF(T_g, T_h)$, gdzie DDF to algorytm odległości pochodnej oraz $b = 1 - a$, gdzie $a \in [0,1]$ to parametry rzeczywiste[21]. W części czwartej poniższej pracy algorytm $PDDF$ obliczany jest względem parametrów równych: $a = b = cos\alpha = sin\alpha = 0{,}7071068$, gdzie $\alpha = \frac{\frac{\pi}{2}}{2}$. Natomiast dla dwóch szeregów czasowych T_g oraz T_h, których wyrazy indeksowane są tym samym zbiorem I, oraz funkcji $DF \in \{L2, DTW\}$, tzw. algorytm niezmienniczy ze względu na złożoność szeregów czasowych (ang. *complexity-invariant*) ma postać $CIDF(T_g, T_h) \coloneqq DF(T_g, T_h) \times CF(T_g, T_h)$, gdzie $CF(T_g, T_h)$ to tzw. czynnik korygujący złożoność (ang. *complexity correction factor*), obliczany dla T_g oraz T_h według zależności: $CF(T_g, T_h) \coloneqq \frac{max\{CE(T_g), CE(T_h)\}}{min\{CE(T_g), CE(T_h)\}}$. W powyższej zależności $CE(T)$ (dla $T \in \{T_g, T_h\}$) to tzw. wyznacznik złożoności (ang. *complexity estimate*) szeregu T. Ma on postać: $CE(T) \coloneqq \sqrt{\sum_{i=1}^{n-1}[T(i) - T(i-1)]^2}$ dla $i = 1, 2, \dots, n - 1$[22].

5. Materiały i metody

Nowo zaproponowane algorytmy obliczania funkcji odległości pomiędzy szeregami czasowymi zostaną przetestowane na części publicznie dostępnego zbioru danych HTS007[23]. Zbiór ten został wygenerowany w laboratorium *High Throughput Screening* Uniwersytetu Vanderbilta[24]. Obejmuje on pięciodniowe pomiary proliferacji komórek przeprowadzone na 8 liniach komórkowych (BT20, HCC1143, MCF10A-HMS, MCF10A-VU, MDAMB231, MDAMB453, MDAMB468 oraz SUM149) traktowanych 27 lekami przeciwnowotworowymi o różnym stężeniu (próba testowa) oraz nietraktowanych żadnym lekiem (próba kontrolna). Do naszych symulacji komputerowych użyliśmy tylko szeregów czasowych, będących wynikami

[20] E. J. Keogh, M. J. Pazzani, *Dynamic time warping with higher order features* [w:] *Proceedings of the 2001 SIAM International Conference on Data Mining*, red. V. Kumar, R. Grossman, 2001, s. 1–11.
[21] T. Górecki, M. Łuczak, *Using derivatives in time series classification*, „Data Mining and Knowledge Discovery" 2013, nr 2, s. 310–331.
[22] G. E. A. P. A. Batista, X. Wang, E. J. Keogh, *A complexity-invariant distance measure for time series* [w:] *Proceedings of the 2011 SIAM International Conference on Data Mining*, red. B. Liu, H. Liu, C. Clifton, T. Washio, C. Kamath, 2011, s. 699–710.
[23] Thunor, https://www.thunor.net, (on-line 16.02.2024).
[24] A. L. R. Lubbock, L. A. Harris, V. Quaranta, D. R. Tyson, C. F. Lopez, *Thunor: visualization and analysis of high-throughput dose-response datasets*, „Nucleic Acids Research" 2021, nr 46, s. w633–w640.

pomiarów przeprowadzonych na linii BT20. Nazwy poszczególnych zbiorów danych pomiarowych odpowiadają nazwą leków testowanych na komórkach, których te pomiary dotyczą. Są one wyszczególnione w Tabeli 1. Kolumny Test oraz Kontrola odpowiadają kolejno liczbie testowych oraz kontrolnych szeregów czasowych w każdym ze zbiorów danych.

Lp.	***TD***	Test	Kontrola	Lp.	***TD***	Test	Kontrola
1	Abemaciclib	20	31	15	Neratinib	15	31
2	Alpelisib	15	31	16	Osimertinib	20	31
3	Azd7762	20	31	17	Paclitaxel	20	31
4	Bleomycin	20	31	18	Palbociclib	17	31
5	Buparlisib	19	31	19	Panobinostat	19	31
6	Cediranib	17	31	20	Pictilisib	17	31
7	Certitinib	19	31	21	Saracatinib	17	31
8	Dasatinib	17	31	22	Taselisib	20	31
9	Doxorubicin	20	31	23	Tivantinib	19	31
10	Etoposide	19	31	24	Torin2	20	31
11	Everolimus	19	31	25	Trametinib	16	31
12	Ink128	15	31	26	Volasertib	20	31
13	Ipatasertib	18	31	27	Vorinostat	19	31
14	Luminespib	20	31				

Tab. 1 Przykładowe farmakokinetyczne zbiory danych ze zbioru HTS007

Każdy z 27 pojedynczych zbiorów danych składa się z 31 dwudziestoczteroelementowych kontrolnych szeregów czasowych oraz z od 15 do 20 (por. Tabela 1) dwudziestoczteroelementowych testowych szeregów czasowych. Tym samym każdy zbiór danych ma dwie etykiety: 1/ kontrola oraz 2/ test. Wszystkie obliczenia zostały wykonane w języku programowania R oraz w jego pakietach[25].

[25] T. Giorgino *op. cit*., U. Mori *et al.*, *op. cit*., R Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna 2022, https://www.R-project.org/ (on-line 16.02.2024), W. N. Venables, B. D. Ripley, *Modern Applied Statistics with S*, New York 2002.

6. Wyniki

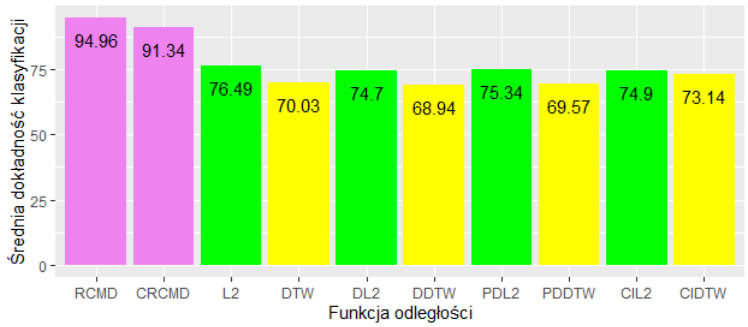
Tabela 2 zawiera wyniki analizy porównawczej wydajności algorytmów klasyfikacyjnych opartych na nowo zaproponowanych funkcjach odległości, tj. na funkcjach $RCMD$ oraz $CRCMD$, jak również na referencyjnych miarach typu $L2$ (tj. miarach $L2$, $DL2$, $PDL2$ oraz $CIL2$) oraz typu DTW (tj. miarach DTW, $DDTW$, $PDDTW$ oraz $CIDTW$). Wykres na Rycinie 1 podsumowuje wyniki z Tabeli 2. Dane te jednoznacznie wskazują, iż w 96,3% pomiarów efektywność techniki $RCMD$ jest równa lub wyższa niż 90% poprawnie sklasyfikowanych szeregów czasowych w każdym zbiorze danych. Natomiast, skuteczność techniki $CRCMD$ jest równa lub wyższa niż 90% poprawnie sklasyfikowanych szeregów w 66,67% testowanych przypadków. Z drugiej strony, na podstawie danych z Tabeli 2 można wywnioskować, iż efektywność technik typu $L2$ jest równa lub wyższa niż 90% tylko w 13,89% dokonanych pomiarów. Z kolei wydajność technik typu DTW jest równa lub wyższa niż 90% poprawnie sklasyfikowanych szeregów tylko w 5,56 % przeprowadzonych symulacji.

TD	*RCMD*	*CRCMD*	*L2*	*DTW*	*DL2*	*DDTW*	*PDL2*	*PDDTW*	*CIL2*	*CIDTW*
1	**90,2**	**92,16**	68,63	64,71	70,59	58,82	74,51	62,75	68,63	74,51
2	**97,83**	**91,3**	69,57	58,7	63,04	58,7	63,04	56,52	69,57	58,7
3	**100**	**96,08**	68,63	62,75	66,67	50,98	68,63	64,71	58,82	64,71
4	**96,08**	**96,08**	72,55	56,86	49,02	60,78	60,78	52,94	68,63	66,67
5	**92**	88	68	58	80	56	64	50	62	74
6	**91,67**	87,5	83,33	70,83	66,67	70,83	75	68,75	70,83	66,67
7	**98**	**96**	66	58	66	64	62	64	64	70
8	**100**	85,42	85,42	81,25	83,33	62,5	85,42	79,17	83,33	87,5
9	**92,16**	88,24	72,55	78,43	72,55	68,63	72,55	74,51	68,63	76,47
10	**96**	**92**	68	68	64	62	68	62	72	54
11	**98**	**100**	**100**	**100**	**98**	**100**	**100**	**100**	**100**	**100**

12	**91,3**	**93,48**	**91,3**	89,13	**91,3**	89,13	**93,48**	89,13	89,13	93,48
13	**95,92**	**93,88**	79,59	63,27	75,51	53,06	69,39	63,27	73,47	63,27
14	**92,16**	84,31	68,63	74,51	84,31	76,47	72,55	72,55	70,59	62,75
15	**100**	**93,48**	73,91	50	80,43	65,22	76,09	60,87	71,74	56,52
16	**96,08**	**92,16**	72,55	68,63	64,71	66,67	74,51	74,51	78,43	74,51
17	80,39	**96,08**	**90,2**	84,31	82,35	86,27	**92,16**	80,39	**90,2**	82,35
18	**100**	**95,83**	83,33	68,75	66,67	56,25	77,08	70,83	77,08	72,92
19	**90**	78	72	68	82	72	74	66	70	74
20	**97,92**	**95,83**	68,75	62,5	68,75	60,42	66,67	66,67	66,67	72,92
21	**95,83**	89,58	66,67	58,33	75	68,75	68,75	60,42	66,67	58,33
22	**94,12**	88,24	72,55	52,94	68,63	70,59	70,59	50,98	66,67	50,98
23	**96**	**92**	72	74	70	78	74	72	80	88
24	**90,2**	78,43	80,39	78,43	80,39	68,63	80,39	68,63	78,43	74,51
25	**100**	**97,87**	76,6	72,34	80,85	74,47	74,47	76,6	76,6	80,85
26	**92,16**	**94,12**	**92,16**	88,24	**92,16**	**96,08**	**90,2**	88,24	**90,2**	**96,08**
27	**100**	**90**	82	80	74	66	86	82	**90**	80

Tab. 2 Procentowe wyniki klasyfikacji 27 farmakokinetycznych zbiorów danych

Ponadto dane zawarte w Tabeli 2 wskazują, iż algorytmy $RCMD$, $CRCMD$, typu $L2$ oraz typu DTW osiągają najlepsze wyniki spośród wszystkich analizowanych przypadków, odpowiednio w 81,48%, 18,52%, 3,7% oraz w 5,56% wykonanych pomiarów. Tak więc można stwierdzić, iż skuteczność nowo wprowadzonych algorytmów $RCMD$ oraz $CRCMD$ jest znacznie wyższa niż skuteczność metod bazujących na klasycznych miarach odległości $L2$, DTW oraz ich udoskonaleniach.



Ryc. 1. Wykres ilustrujący średnią procentową efektywność porównywanych funkcji odległości

Ściśle rzecz biorąc, zestawiając średnie wartości efektywności porównywanych protokołów klasyfikacyjnych, można stwierdzić, iż schemat $RCMD$ średnio przewyższa najlepszy schemat typu $L2$ (tj. algorytm $L2$) oraz typu DTW (tj. algorytm $CIDTW$) o odpowiednio, 19,45% oraz o 22,98% poprawnie sklasyfikowanych szeregów temporalnych (por. Ryc. 1). Natomiast drugi z nowo zaproponowanych protokołów, tj. protokół $CRCMD$, przewyższa średnio schematy $L2$ oraz $CIDTW$ o odpowiednio, 16,26% oraz o 19,93% poprawnie sklasyfikowanych ciągów czasowych (por. Ryc. 1).

7. Dyskusja

W swojej obszernej i ważnej pracy *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Trevor Hastie, Robert Tibshirani oraz Jerome Friedman zauważają, iż „wyszczególnienie adekwatnej miary niepodobieństwa [pomiędzy analizowanymi obiektami – przyp. mój – P. W.] jest o wiele bardziej istotne w osiągnięciu sukcesu w procesie klasteryzacji niż wybór samego algorytmu analizy skupień. Ten aspekt problemu jest mniej podkreślany w literaturze poświęconej klasteryzacji, ponieważ aspekt ten zależy od specyficzności dziedziny przedmiotowej [poddawanej analizie skupień – przyp. mój – P.W.]

oraz jest mniej podatny na ogólne analizy”[26]. W naszej opinii te deklaracje, choć odnoszą się do zagadnienia „klasyfikacji bez nadzoru” (ang. *unsupervised learning*), mogą być uogólnione względem każdego rodzaju klasyfikacji, z włączeniem zagadnienia klasyfikacji szeregów czasowych. Dlatego też, aby poprawić efektywność znanych algorytmów klasyfikacji szeregów czasowych (np. algorytmu $L2$ lub DTW połączonych z klasyfikatorem $1NN$), zaproponowaliśmy nowe miary odległości pomiędzy danymi temporalnymi. Z przeprowadzonych symulacji komputerowych wynika, iż wydajność algorytmów klasyfikacji opartych na nowych bezparametrycznych funkcjach odległości jest znacznie wyższa niż wydajność protokołów opartych na funkcjach referencyjnych (np. na klasycznych miarach $L2$, DTW oraz na ich parametrycznych modyfikacjach). Rezultat ten wydaje się być bardzo istotny, ponieważ jak zauważają Tomasz Górecki oraz Maciej Łuczak „[…] prosta metoda łącząca klasyfikator jednego najbliższego sąsiada (1NN) oraz pewną formę miary odległości DTW okazała się być jedną z najwydajniejszych technik klasyfikacji szeregów czasowych. […]. Euklidesowa miara odległości ma kilka zalet. Mianowicie, złożoność wyznaczania tej metryki jest liniowa, jest ona łatwa do zaimplementowania, może być łączona z dowolnymi innymi metodami oraz jest bezparametryczna. […]. Zostało empirycznie dowiedzione, iż prosta Euklidesowa metryka odległości jest konkurencyjna lub lepsza względem wielu złożonych miar odległości oraz spełnia ważną nierówność trójkąta”[27]. Ci sami autorzy w dalszych częściach swojego tekstu twierdzą, iż „metryka odległości Euklidesowej jest najbardziej oczywistą miarą podobieństwa dla szeregów czasowych, natomiast miara DTW jest jedną z najbardziej wydajnych funkcji odległości dla danych temporalnych”[28]. Powyższe fragmenty zaczerpnięte z bieżącej literatury dotyczącej analizy i klasyfikacji szeregów czasowych mogą być bezpośrednio skonfrontowane z wynikami zamieszczonymi w Tabeli 2 oraz na Rycinie 1. Z postulowanego zestawienia można jednoznacznie wywnioskować, iż rezultaty uzyskane przez proponowany w naszej pracy rekurencyjno-kanoniczny schemat klasyfikacji oraz jego krzyżowa forma są kontrprzykładami względem stwierdzenia Góreckiego i Łuczaka, a także trudno jest znaleźć metodologię klasyfikacji danych temporalnych przewyższającą swoją skutecznością metodologię opartą na mierze DTW. Można zaznaczyć, że proponowany w tej pracy algorytm jest na tyle elastyczny, iż dzięki jego modyfikacjom można określić nowe rodziny miar odległości opartych na dyskretnych

[26] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, New York 2009, s. 506.
[27] T. Górecki, M. Łuczak, *Using*…, *op. cit.*, s. 311.
[28] *Ibidem*., s. 320.

pochodnych analizowanych szeregów czasowych. Np. dla szeregu czasowego T_g o długości n z rozpatrywanej kolekcji szeregów $\boldsymbol{T}$ można zdefiniować następujące oparte na pochodnych macierze rekurencyjne: $UR_{ij}^{(1)g} \coloneqq \left\|T_g'(i) - T_g'(j)\right\|_2$, $UR_{ij}^{(2)g} \coloneqq \left\|T_g''(i) - T_g''(j)\right\|_2$, gdzie T_g'' to druga dyskretna pochodna szeregu czasowego T_g[29] oraz następujące oparte na pochodnych krzyżowe macierze rekurencyjne: $UCR_{ij}^{(0,2)g} \coloneqq \left\|T_g(i) - T_g''(j)\right\|_2$ oraz $UCR_{ij}^{(1,2)g} \coloneqq \left\|T_g'(i) - T_g''(j)\right\|_2$[30]. Analizie tego typu 2D struktur poświęcona będzie oddzielna praca.

7. Uwagi końcowe

Podsumowując, można stwierdzić, iż cel przedkładanej pracy został osiągnięty, a nowe rekurencyjno-kanoniczne funkcje obliczania odległości pomiędzy danymi temporalnymi znajdą praktyczne zastosowanie w analizie i klasyfikacji danych farmakokinetycznych i tym samym przyczynią się do zwiększenia stopnia automatyzacji w badaniach podstawowych stosowanych.

Podziękowania: Autor powyższej pracy pragnie wyrazić podziękowania anonimowej recenzentce tekstu, której uwagi przyczyniły się do znacznego ulepszenia przedkładanej pracy.

Bibliografia

1. Batista G. E. A. P. A., Wang X., Keogh E. J., *A complexity-invariant distance measure for time series* [w:] *Proceedings of the 2011 SIAM International Conference on Data Mining*, red. B. Liu, H. Liu, C. Clifton, T. Washio, C. Kamath, 2011, s. 699–710.
2. Berndt D. J., Clifford J., *Using dynamic time warping to find patterns in time series* [w:] *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, 1994, s. 359–370.
3. Eckmann J.–P., Kamphorst S. O., Ruelle D., *Recurrence plots of dynamical systems*, „Europhysics Letters" 1987, nr 5, s. 973–977.
4. Eroglu D., T. K. DM. Peron, N. Marwan, F. A. Rodigues, L. da F. Costa, M. Sebek, I. Z. Kiss, J. Kurths, *Entropy of weighted recurrence plots*, „Physical Review E" 2014, nr 90, 042919-1–024919-7.

[29] Druga dyskretna pochodna T'' szeregu czasowego T o długości n to nowy szereg czasowy o długości $n-2$, będący pierwszą dyskretną pochodną pierwszej dyskretnej pochodnej T' wejściowego (oryginalnego) szeregu T.

[30] W powyższej notacji macierze z kroku pierwszego prezentowanego algorytmu to odpowiednio macierze $UR_{ij}^{(0)g}$ oraz $UCR_{ij}^{(0,1)g}$, gdyż wejściowy (oryginalny) szereg czasowy T_g może być traktowany jako swoja dyskretna zero-rzędowa pochodna.

5. Giorgino T., *Computing and visualizing dynamic time warping alignments in R: The dtw package*, „Journal of Statistical Software" 2009 (7), s. 1–24.
6. Górecki T., Łuczak M., *Using derivatives in time series classification*, „Data Mining and Knowledge Discovery" 2013, nr 2, s. 310–331.
7. Górecki T., Łuczak M., *Multivariate time series classification with parametric derivative dynamic time warping*, „Expert Systems with Applications" 2015, nr 42, s. 2305–2312.
8. Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, New York 2009.
9. Keogh E., Kasetty S., *On the need for time series data mining benchmarks: a survey and empirical demonstration* [w:] *Proceedings of the eight ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York 2002, s. 102–111.
10. Keogh E., Pazzani M. J., *Dynamic time warping with higher order features* [w:] *Proceedings of the 2001 SIAM International Conference on Data Mining*, red. V. Kumar, R. Grossman, 2001, s. 1–11.
11. Keogh E., Ratanamahatana C. A., *Exact indexing of dynamic time warping*, „Knowledge and Information Systems" 2005 (7), s. 358–386.
12. Lubbock A. L. R., Harris L. A., Quaranta V., Tyson D. R., Lopez C. F., *Thunor: visualization and analysis of high-throughput dose-response datasets*, „Nucleic Acids Research" 2021, nr 46, s. w633–w640.
13. Marwan N., Romano M. C., Thiel M., Kurths J., *Recurrence plots for the analysis of complex systems*, „Physics Reports" 2007, nr 438, s. 237–329.
14. Mori U., Mendiburu A., Lozano J. A., *Distance measures for time series in R: The TSdist Package*, „The R Journal" 2016, nr 2, s. 451–459.
15. Rabiner L,. Juang B.-H, *Fundamentals of Speech Recognition*, New Jersey 1993.
16. R Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna 2022, https://www.R-project.org/ (on-line 16.02.2024).
17. Sankoff D., Kruskal J., *Time Warps, String Edits, and Macromolecules. The Theory and Practice of Sequence Comparison*, 1999.
18. Todeschini R., Ballabio D., Consonni V., Manganaro A., Mauri A., *Canonical measure of correlation (CMC) and canonical measure of distance (CMD) between sets of data.*

Part 1. Theory and simple chemometric applications, „Analytica Chimica Acta” 2009, nr 648, s. 45–51.

19. Thunor, https://www.thunor.net (on-line: 16.02.2024).
20. Venables W. N., Ripley B. D., *Modern Applied Statistics with S*, New York 2002.
21. Wilczek P., *An application of the local binary pattern algorithm and its uniform variant to improve the recurrence and cross-recurrence quantification analyses of the pharmacologically important time series* [w:] *Recent Advances in Computational Oncology and Personalized Medicine. Vol. 2. The Challenges of the Future !*, red. K. Krukiewicz, M. Marczyk, M. Bugdol, S. Bajkacz, Z. Ostrowski, Gliwice 2022, s. 128–152.