Piotr Wilczek[1]
Computer Laboratory

Nowe metody komputerowej analizy sygnałów biomedycznych

Streszczenie

W niniejszej pracy zaproponowano dwie nowe funkcje odległości temporalnej, których efektywność w klasyfikowaniu medycznie istotnych szeregów czasowych została empirycznie przetestowana. Algorytmy oparte na nowo zaproponowanych funkcjach zostały porównane z algorytmami opartymi na klasycznych miarach *Lp* i *DTW*, jak również na ich ulepszeniach. Porównanie wykazało, że w prawie wszystkich przypadkach nasze techniki przewyższają lub znacznie przewyższają metody referencyjne. W związku z tym można stwierdzić, że wprowadzenie nowych miar temporalnych jest uzasadnione.

Słowa kluczowe: biosygnały, szeregi czasowe, faktoryzacja macierzy na wartości własne i wektory własne, temporalne funkcje odległości, metryki, miara *DTW*, klasyfikator jednego najbliższego sąsiada

Abstract

In the present contribution, we proposed two novel temporal distance functions whose effectiveness in classifying medically significant time series data was empirically tested. The algorithms based on the newly developed functions were juxtaposed with the algorithms based on the classical *Lp* and *DTW* measures as well as their refinements. These comparisons showcased that in almost all cases our techniques outperform or overwhelmingly outperform the reference methods. Therefore, it can be asserted that the introduction of the novel temporal distance measures is justified.

Keywords: biosignals, time series, matrix eigendecomposition, temporal distance functions, metrics, *DTW* measure, 1-nearest neighbor classifier

Wstęp

W diagnostyce medycznej, ważną grupę sygnałów biomedycznych stanowią *biopotencjały*. We współczesnej medycynie, metoda *elektrokardiografii* (EKG) jest najpowszechniej stosowaną metodą analizy medycznie ważnych biosygnałów (Kuniszyk-Jóźkowiak, 2011). Innym typem rozpatrywanego w poniższej pracy biosygnału diagnostycznego są *krzywe topnienia*, czyli *krzywe denaturacji* dwuniciowego kwasu DNA (tj. dsDNA). Proces *topnienia* (tj. *denaturacji*) makromolekuły dsDNA polega na separacji nici DNA na dwie pojedyncze nici wskutek zerwania wiązań wodorowych oraz zaniku otoczki hydratacyjnej. Jedną z technik laboratoryjnych pozwalającą na wykorzystanie denaturacji dsDNA w celu identyfikacji patogenów jest *Wysokorozdzielcza Analiza Topnienia* (ang. *High-Resolution Melting*

[1] mgr, ORCID 0000-0003-2758-343X, e-mail: piotr.wilczek.net@onet.pl

(HRM) analysis) (Lu i in., 2017). Metoda HRM polega na śledzeniu stopnia denaturacji podwójnych łańcuchów analizowanego dsDNA w funkcji temperatury za pomocą zmian w intensywności fluorescencji. Technika HRM to bardzo skuteczna metoda umożliwiająca wykrywanie mutacji, polimorfizmu oraz epigenetycznych różnic w próbkach dsDNA (Lu i in., 2017).

Prezentowana praca ma na celu zaproponowanie nowych algorytmów komputerowej analizy sygnałów biomedycznych (tj. EKG oraz HRM) w oparciu o (spektralne) -obcięte miary odległości pomiędzy szeregami czasowymi reprezentującymi owe sygnały. W drugim podrozdziale zostanie przypomniana terminologia dotycząca metod analiz szeregów czasowych oraz zaproponowane zostaną nowe techniki umożliwiająca efektywną klasyfikację rozpatrywanych biosygnałów. Podrozdział trzeci przedstawia przykładowe zbiory danych EKG oraz HRM oraz wyszczególnia narzędzia informatyczne użyte w pracy. Z kolei podrozdział czwarty zawiera wyniki symulacji komputerowych i dyskusję. Wnioski końcowe zawarte są w podrozdziale piątym.

Nowy spektralny algorytm analizy szeregów czasowych

Sygnałem nazywa się zmianę jednej (lub kilku) wielkości w zależności od zmian innej wielkości. W szerokim użyciu jest zastosowanie pojęcia sygnału do zmian pewnych wielkości fizycznych w funkcji czasu. Taki sygnał nazywa się szeregiem czasowym. Przypomnijmy, iż szereg czasowy to ciąg pomiarów (obserwacji) uporządkowany w czasie (lub w przestrzeni). W poniższej pracy przyjmujemy, iż zmienna niezależna (tj. zmienna czasowa) jest dyskretna. A więc, z formalnego punktu widzenia, szereg czasowy T to ciąg par o postaci (Shifaz i in., 2023):

$$T=[(t_1, x_1), (t_2, x_2), \ldots (t_i, x_i), \ldots (t_n, x_n) \text{ dla } t_1 < t_2 < \ldots < t_i < \ldots t_n$$

dla oraz dla zbioru indeksów $I=\{1, 2, \ldots i, \ldots n\}$ 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 poniższej pracy przyjmujemy, iż $d = 1$ dla każdego z rozpatrywanych szeregów czasowych. A więc, wszystkie dane temporalne analizowane w poniższej pracy są jednowymiarowe (ang. *univariate*). Ponadto przyjmujemy, iż punkty czasowe t_i wszystkich rozpatrywanych danych temporalnych są równomiernie rozmieszczone, a więc każdy badany szereg czasowy jest *regularny*. Liczba zarejestrowanych pomiarów (obserwacji), (n), to długość szeregu T. Dla $k \leq n$, szereg czasowy obcięty (ograniczony) do swoich k-pierwszych wyrazów oznaczony jest jako $T \upharpoonright k$. W poniższej pracy, podstawowa jednostka naszych analiz to prostokątny $m \times (n+1)$ temporalny zbiór danych (***TD***). Formalnie, ***TD*** to kolekcja $\boldsymbol{T}$ ułożonych równolegle szeregów czasowych z których każdy ma długość n, a cała kolekcja $\boldsymbol{T}$ składa się z m takich szeregów razem z określonym (ang. *predefined*) dyskretnym wektorem[2] etykiet $\boldsymbol{C}$ o długości m, a więc $\boldsymbol{T} = \{T_g\}_{g=1}^{m}$ oraz $|\boldsymbol{C}| = m$. Wektor $\boldsymbol{C}$ stanowi ostatnią (tj. $n + 1$) kolumnę temporalnego zbioru danych ***TD***.

[2] Z matematycznego punktu widzenia wektor to multizbiór. Liczba klas równoważności elementów wektora oznaczana jest przez .

Wiersze obiektu $\boldsymbol{TD}$ są indeksowane zbiorem $J = \{1, 2, \dots, g, \dots, m\}$ natomiast jego kolumny zbiorem $I^* = \{1, 2, \dots, i, \dots, n, (n+1)\}$. Podsumowując, temporalny zbiór danych $\boldsymbol{TD}$ to struktura o postaci $\boldsymbol{TD} = (\boldsymbol{T}, \boldsymbol{C}) = \left(T_g, C^{T_g}\right)$, gdzie $T_g \in \boldsymbol{T}$ to g-ty szereg czasowy, natomiast $C^{T_g} \in \boldsymbol{C}$ to jego etykieta. W poniższej 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. Temporalny zbiór danych bez ostatniej kolumny, tj. bez wektora etykiet $\boldsymbol{C}$ to kwadratowa $m \times n$ temporalna macierz danych (*TM*).

Przykład 1. Przykładowy temporalny zbiór danych $\boldsymbol{TD}$ o wymiarach 5 × (6 + 1) składa się z pięciu (m=5) ułożonych równolegle szeregów czasowych z których każdy ma długość n=6; szeregi te to: T_1 = (35,22,6,47,3,60)', T_2=(20,18,6,31,55,1), T_3=(8,79,0,2,16,36), T_4=(10,0,100,15,28,31), T_5=(3,4,12,11,8,99). Są one etykietowane multizbiorem o postaci $\boldsymbol{C}$={*A, A, B, C, C*} , gdzie $|\boldsymbol{C}| = 5$ oraz (tj. wektor $\boldsymbol{C}$ składa się z dwóch dwuelementowych klas równoważności *A* i *C* oraz z jednej jednoelementowej klasy równoważności *B*; a więc $C^{T_1} \cong C^{T_2}$ oraz $C^{T_4} \cong C^{T_5}$, ponieważ $C^{T_1} = C^{T_2} = A$ i $C^{T_4} \cong C^{T_5} = C$ oraz $C^{T_3} \cong C^{T_3}$, ponieważ $C^{T_3} = B$). Temporalna macierz danych *TM*, a więc struktura $\boldsymbol{TD}$ bez kolumny $\boldsymbol{C}$ ma wymiary 5 ×6. Jej wiersze stanowią szeregi czasowe będące elementami kolekcji $\boldsymbol{T} = \{T_g\}_{g=1}^{5}$

$$\boldsymbol{TD} = \begin{bmatrix} & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & \boldsymbol{C} \\ T_1 & 35 & 22 & 6 & 47 & 3 & 60 & A \\ T_2 & 20 & 18 & 6 & 31 & 55 & 1 & A \\ T_3 & 8 & 79 & 0 & 2 & 16 & 36 & B \\ T_4 & 10 & 0 & 100 & 15 & 28 & 31 & C \\ T_5 & 3 & 4 & 12 & 11 & 8 & 99 & C \end{bmatrix}$$

Przypomnijmy, iż jednym z głównych zadań temporalnej analizy danych jest opracowanie ilościowych miar odległości pomiędzy dwoma szeregami czasowymi, przy czym pożądaną cechą charakteryzującą owe miary powinna być ich zdolność do odzwierciedlania rzeczywistej odległości pomiędzy ciągami pomiarów (obserwacji), które te szeregi reprezentują. Celem poniższego studium jest opracowanie nowej, *spektralnej* (tj. opartej o *wektory własne* tzw. produktowej macierzy temporalnej), *kontekstowo-zależnej* (ang. *context-dependent*)[3] temporalnej miary odległości. Jej wydajność w zadaniach klasyfikacyjnych spotykanych na terenie analizy danych czasowych będzie testowana w podrozdziale czwartym.

Proponowany kontekstowo-czuły algorytm obliczania odległości pomiędzy dwoma szeregami czasowymi (tj. dwoma wierszami temporalnej macierzy danych) oparty jest na następujących krokach:

1/ **W kroku pierwszym**, temporalna macierz danych *TM* o wymiarach $m \times n$ jest mnożona (ang. *post-multiplied*) przez swoją formę transponowaną TM^T o wymiarach $n \times m$. Rezultatem tego kroku jest nowa macierz kwadratowa o wymiarach $m \times m$, tzw. temporalna macierz produktowa (*TP*), tj.

[3] Równoważnie określanej jako *kontekstowo-czułej* (ang. *context-sensitive*).

$$TP: = TM \times TM^T \qquad (1)$$

Krok ten zilustrowany jest w Przykładzie nr 2.

Przykład 2. Mnożenie macierzy *TM* z Przykładu 1 przez jej formę transponowaną TM^T. Otrzymano nową macierz produktową .

$$TM \times TM^T = TP = \begin{bmatrix} & 1 & 2 & 3 & 4 & 5 \\ 1 & 7563 & 2814 & 4320 & 3599 & 6746 \\ 2 & 2814 & 4747 & 2560 & 2836 & 1084 \\ 3 & 4320 & 2560 & 7861 & 1674 & 4054 \\ 4 & 3599 & 2836 & 1674 & 12070 & 4688 \\ 5 & 6746 & 1084 & 4054 & 4688 & 10155 \end{bmatrix}$$

2/ *W kroku drugim*, temporalna macierz produktowa *TP* jest *faktoryzowana* (ang. *factorized*) według zależności (Bloomfield, 2014):

$$TP \coloneqq Q * E * Q^{-1} \qquad (2)$$

gdzie Q to kwadratowa macierz o wymiarach $m \times m$, której g-ta kolumna to g-ty wektor własny (q_g) macierzy produktowej, natomiast E to macierz diagonalna, której wyrazy diagonalne to wartości własne macierzy (1), tj. $E_{g,g} = e_g$, gdzie e_g to g-ta wartość własna macierzy produktowej. W poniższej pracy, liniowo niezależne wektory kolumnowe $\{q_1, q_2, \dots, q_g, \dots, q_m\}$ macierzy Q są znormalizowane, chociaż w ogólnym przypadku nie muszą być. Wiersze kwadratowej $m \times m$ macierzy Q są indeksowane tym samym zbiorem J co wiersze temporalnej macierzy danych *TM* i są one *funkcjonalnymi surogatami spektralnymi* szeregów czasowych, które konstytuują wiersze prostokątnej $m \times n$ macierzy *TM*. Kolekcję funkcjonalnych surogatów spektralnych szeregów czasowych tworzących kolekcję $\boldsymbol{T} = \{T_g\}_{g=1}^{m}$ oznaczono jako $\boldsymbol{T}^* = \{T_g^*\}_{g=1}^{m}$. Długość funkcjonalnego surogatu spektralnego T_g^* szeregu czasowego T_g o długości n wynosi m i jest ona zależna od mocy temporalnego zbioru danych ***TD***, którego elementem jest szereg T_g. Przykład nr 3 przedstawia wyżej opisany etap.

Przykład 3. Temporalna macierz produktowa (1) z Przykładu 2 jest faktoryzowana według zależności (2) i otrzymano macierz :

$$Q = \begin{bmatrix} & q_1 & q_2 & q_3 & q_4 & q_5 \\ T_1^* & -0.49 & 0.26 & 0.07 & -0.48 & 0.67 \\ T_2^* & -0.23 & -0.04 & -0.61 & -0.56 & -0.5 \\ T_3^* & -0.37 & 0.46 & -0.5 & 0.63 & 0.05 \\ T_4^* & -0.5 & -0.82 & -0.08 & 0.24 & 0.13 \\ T_5^* & -0.56 & 0.21 & 0.6 & 0.03 & -0.53 \end{bmatrix}$$

Wiersze T_1^*, T_2^*, T_3^*, T_4^* oraz T_5^* macierzy Q to, odpowiednio, funkcjonalne surogaty spektralne szeregów czasowych T_1, T_2, T_3, T_4 oraz T_5. Ich długość wynosi m.

3/ **W kroku trzecim** przyjęto następującą terminologię: każda temporalna[4] funkcja odległości *DF* działająca na wiersze (ang. *in a row-wise manner*) wejściowej $m \times n$ macierzy *TM* lub jej $m \times i$ (gdzie $i \leq n$) podmacierzy to, odpowiednio, zwykła temporalna miara odległości (*F*) lub zwykła *k*-obcięta temporalna miara odległości (k – *F*). W drugim przypadku, parametr k przebiega podzbiór I_i zbioru I o postaci: $I_i = \{1, 2, \dots, i\} \subseteq I$. Z kolei, każda temporalna funkcja odległości *DF* działająca na wiersze kwadratowej $m \times m$ macierzy Q wektorów własnych struktury produktowej *TP* lub jej $m \times g$ (gdzie $g \leq m$) podmacierzy to, odpowiednio, spektralna (tj. oparta na wektorach własnych) temporalna miara odległości (evF) lub spektralna (tj. oparta na wektorach własnych) *k*-obcięta temporalna miara odległości ($k - evF$). W drugim przypadku, parametr k przebiega podzbiór J_g zbioru J o postaci: $J_g = \{1, 2, \dots, g\} \subseteq J$. Wynikiem działania każdej z zaproponowanych funkcji temporalnych $F, k - F, evF, k - evF$ jest kwadratowa $m \times m$ macierz odległości temporalnych TDM_{DF} (gdzie $DF \in \{F, k - F, evF, k - evF\}$), której wyraz $TDM_{DF}[g, h]$ jest równy odległości pomiędzy szeregami czasowymi T_g oraz T_h. Z powyższych rozważań wynikają następujące tożsamości:

$$F(T_g, T_h) := DF(T_g, T_h) \tag{3}$$

$$k - F(T_g, T_h) := DF(T_g \restriction k, T_h \restriction k) \tag{4}$$

$$evF(T_g, T_h) := DF(T_g^*, T_h^*) \tag{5}$$

oraz

$$k - evF(T_g, T_h) := DF(T_g^* \restriction k, T_h^* \restriction k) \tag{6}$$

gdzie *DF* to funkcja temporalna, $T_g, T_h \in \boldsymbol{T}$, $T_g^*, T_h^* \in \boldsymbol{T}^*$, natomiast $T_g \restriction k$ i $T_h \restriction k$ to szeregi T_g, T_h obcięte do swoich *k*-pierwszych wyrazów (dla $k \leq n$) oraz $T_g^* \restriction k$ i $T_h^* \restriction k$ to funkcjonalne surogaty spektralne obcięte do swoich *k*-pierwszych wyrazów (dla $k \leq m$). Wartości parametru *k*, dla których funkcje k – *F* oraz ($k - evF$) osiągają maksimum efektywności (tzn. wartości *k* są optymalne) oznaczone zostały przez k_{opt}, a same funkcje przez k_{opt} – *F* oraz $k_{opt} - evF$.

Rozpatrując szeregi czasowe jako realizacje pewnych procesów stochastycznych generowanych przez określone modele statystyczne można przyjąć założenie, iż dane pomiarowe uporządkowane w czasie, a więc pary (t_i, x_i) nie są niezależne. Dlatego też, na przykład, pomiar biosygnału x_i zarejestrowany w teraźniejszym punkcie czasowym t_i, czyli para (t_i, x_i) razem z pomiarami przeszłymi (t_{i-1}, x_{i-1}), (t_{i-2}, x_{i-2}), (t_{i-3}, x_{i-3})… mogą być użyte do prognozowania przyszłych pomiarów, tj. par (t_{i+1}, x_{i+1}), (t_{i+2}, x_{i+2}), (t_{i+3}, x_{i+3})… (Box, Jenkins, 1983). Można więc przypuszczać, iż analogiczne rozumowanie uprawnia do rozpatrywania funkcji *k*-obciętych o postaci danej zależnością (4), ponieważ większość informacji strukturalnych charakteryzujących dany szereg temporalny powinna być zakodowana w jego *k*-pierwszych wyrazach. Dlatego też odległość temporalna obliczona za pomocą zależności

[4] Przez pojęcie temporalnej funkcji odległości rozumie się funkcję odległości pomiędzy szeregami czasowymi, tj. wierszami macierzy .

(4) powinna odzwierciedlać rzeczywiste odległości pomiędzy danymi czasowymi w sposób porównywalny lub bardziej adekwatny niż odległość obliczona za pomocą klasycznych miar typu (3). Tak jak E. Keogh oraz M. Pazzani (2001) lub T. Górecki oraz M. Łuczak (2013, 2015) przyjęli założenie, iż rzeczywista odległość pomiędzy szeregami temporalnymi może być reprezentowana za pomocą odległości pomiędzy pierwszymi dyskretnymi pochodnymi owych szeregów, tak w niniejszej pracy przyjęto założenie, iż rzeczywista odległość pomiędzy analizowanymi szeregami czasowymi może być reprezentowana za pomocą odległości pomiędzy ich funkcjonalnymi surogatami spektralnymi. Natomiast z algebraicznych własności konstrukcji macierzy (1) oraz własności jej faktoryzacji (2) można przypuszczać, iż większa część informacji strukturalnej charakteryzującej szeregi czasowe konstytuujące wiersze temporalnej macierzy danych *TM* jest zakodowana w kilku lub kilkunastu pierwszych wyrazach funkcjonalnych surogatów spektralnych owych szeregów. Dlatego też należy przypuszczać, iż rozpatrywanie temporalnych funkcji odległości o postaci (5) oraz (6) wydaje się być celowe w ramach temporalnej analizy danych czasowych.

Można przypuszczać również, iż porównując dwa typy parametrów *k*, tj. k^I oraz k^J, gdzie symbol k^I oznacza, iż parametr przebiega zbiór indeksów *I* (a więc jest parametrem funkcji *k – F*), natomiast symbol k^J oznacza, iż parametr przebiega zbiór indeksów *J* (a więc jest parametrem funkcji $k - evF$) otrzymuje się następującą zależność: $min(k_{opt}^I) >_{\%} min(k_{opt}^J)$, gdzie symbole $min(k_{opt}^I)$ oraz $min(k_{opt}^J)$oznaczają minimalne wartości (odpowiednio) parametrów k_{opt}^I oraz k_{opt}^J, natomiast symbol $>_{\%}$ oznacza, iż nierówność > określona jest na skali procentowej.

4/ **Krok czwarty** to wybór odpowiedniej funkcji wejściowej *DF*. Zauważmy, iż wyżej zaprezentowany algorytm może być zaimplementowany z dowolną funkcją temporalną której argumenty to *równej długości* jednowymiarowe szeregi czasowe. Niemniej jednak jego walidacja przeprowadzona w podrozdziale czwartym naszego studium oparta jest na implementacji testowanego algorytmu z czterema klasycznymi metrykami L_p, czyli metrykami, wyrażonymi przez następującą ogólną formułę:

$$Lp(V_g, V_h) \coloneqq \sqrt[p]{\sum_{s=1}^{w} [V_g(s) - V_h(s)]^p} \tag{7}$$

gdzie wektor $V \in \{V_g, V_h\}$ jest indeksowany zborem $\{1, 2, \dots, s, \dots, w\}$. Symulacje komputerowe zawarte w przedkładanej pracy przeprowadzane są dla parametru *p*=1, 2, 3 oraz ∞, a więc, dla następujących funkcji metrycznych (Górecki, Piasecki, 2019):

$$L1(V_g, V_h) \coloneqq \sum_{s=1}^{w} |V_g(s) - V_h(s)| \tag{8}$$

$$L2(V_g, V_h) \coloneqq \sqrt{\sum_{s=1}^{w} [V_g(s) - V_h(s)]^2} \tag{9}$$

oraz

$$L3(V_g, V_h) \coloneqq \sqrt[3]{\sum_{s=1}^{w} \left[V_g(s) - V_h(s)\right]^3} \tag{10}$$

oraz .

$$Lmax(V_g, V_h) \coloneqq \max_s \{\left|V_g(s) - V_h(s)\right|\} \tag{11}$$

gdzie $V_g \in \{T_g, T_g \restriction k, T_g^*, T_g^* \restriction k\}$ oraz $V_h \in \{T_h, T_h \restriction k, T_h^*, T_h^* \restriction k\}$ [5].

Przykład 4. Aby zaprezentować procedurę obliczania macierzy odległości dla analizowanych szeregów czasowych, obliczona została funkcja $3 - L2$ oraz $3 - evL2$ dla szeregów czasowych z Przykładu 1. Można zauważyć, iż funkcja $3 - L2$ to zwykła odległość Euklidesowa obliczona dla 5 ×3 podmacierzy *TM** macierzy wejściowej *TM*. Wiersze podmacierzy *TM** są tak samo indeksowane jak wiersze podmacierzy *TM*, natomiast wśród kolumn podmacierzy *TM** są tylko kolumny t_1, t_2 oraz t_3 macierzy *TM*. Z kolei, funkcja $3 - evL2$ to zwykła metryka Euklidesowa obliczona dla 5 ×3 podmacierzy *Q** macierzy wektorów własnych *Q* z Przykładu 3. Wiersze podmacierzy *Q** są tak samo indeksowane jak wiersze macierzy *Q*, natomiast wśród kolumn podmacierzy *Q** są tylko kolumny q_1, q_2 oraz q_3 . W obu przypadkach, rezultatami działania funkcji $3 - L2$ oraz $3 - evL2$ są, odpowiednio, kwadratowe 5×5 macierze odległości TDM_{3-L2} oraz TDM_{3-evL2}. Są to macierze:

$$TDM_{3-L2} = \begin{bmatrix} & T_1 & T_2 & T_3 & T_4 & T_5 \\ T_1 & 0 & 15.52 & 63.36 & 99.72 & 37.2 \\ T_2 & 15.52 & 0 & 62.46 & 96.23 & 22.83 \\ T_3 & 63.36 & 62.46 & 0 & 127.46 & 76.12 \\ T_4 & 99.72 & 96.23 & 127.46 & 0 & 88.37 \\ T_5 & 37.2 & 22.83 & 76.12 & 88.37 & 0 \end{bmatrix}$$

oraz

$$TDM_{3-evL2} = \begin{bmatrix} & T_1 & T_2 & T_3 & T_4 & T_5 \\ T_1 & 0 & 0.79 & 0.62 & 1.09 & 0.54 \\ T_2 & 0.79 & 0 & 0.54 & 0.98 & 1.28 \\ T_3 & 0.62 & 0.54 & 0 & 1.36 & 1.15 \\ T_4 & 1.09 & 0.98 & 1.36 & 0 & 1.23 \\ T_5 & 0.54 & 1.28 & 1.15 & 1.23 & 0 \end{bmatrix}$$

Nowa spektralna miara odległości temporalnych jest kontekstowo-czuła, co oznacza, iż odległość temporalna pomiędzy dwoma szeregami czasowymi

[5] Oczywiście, szeregi czasowe (lub ich obcięcia) są indeksowane zbiorem *I* (lub podzbiorem zbioru *I*), a funkcjonalne surogaty spektralne (lub ich obcięcia) są indeksowane zbiorem *J* (lub jego podzbiorem). Mianowicie, *i*-ty wyraz szeregu *T* zapisujemy jako *T*(*i*), a *g*-ty wyraz funkcjonalnego surogatu spektralnego *T** jako *T**(*g*).

$T_g, T_h \in \boldsymbol{T}$ jest zależna od otoczenia w którym owe szeregi się znajdują, a więc od pozostałych m – 2 elementów kolekcji $\boldsymbol{T}$, gdzie m to moc temporalnego zbioru danych $\boldsymbol{TD}$. Z powyższego wynika, iż tylko dla dwuelementowej kolekcji $\boldsymbol{T}$ i odpowiadającej jej dwuwierszowej macierzy TM, spektralna miara temporalna nie jest kontekstowo zależna. W kolejnym kroku prześledzono powyższą zależność na Przykładzie 5.

Przykład 5. Rozpatrzone zostały te same funkcje odległości co w Przykładzie 4, tzn. $3 - L2$ oraz $3 - evL2$ lecz określone teraz na '*nowej*' macierzy temporalnej TM' powstałej z wejściowej '*starej*' macierzy TM poprzez wymazanie wierszy T_4 oraz T_5. A więc, jej wiersze stanowią szeregi czasowe będące elementami kolekcji $\boldsymbol{T}' = \{T'_g\}_{g=1}^{3}$. Przy czym, kolekcja $\boldsymbol{T}'$ powstała z kolekcji T poprzez usunięcie z niej szeregów czasowych T_4 oraz T_5. Przyjęto następującą notację: $T'_g \in \boldsymbol{T}' \leftrightarrow T_g \in \boldsymbol{T}$ dla g=1, 2 oraz 3. Z powyższego wynika, iż szeregi T_g oraz T'_g mają identyczne wyrazy, lecz różnią się tylko otoczeniem (tzn. kontekstem) w którym występują, ponieważ są elementami różnych kolekcji. '*Nowa*' macierz powstała w rezultacie faktoryzacji (2), tzn. ma teraz postać:

$$Q' = \begin{bmatrix} & q'_1 & q'_2 & q'_3 \\ T'^*_1 & -0.65 & 0.48 & 0.59 \\ T'^*_2 & -0.39 & 0.46 & -0.8 \\ T'^*_3 & -0.65 & -0.75 & -0.11 \end{bmatrix}$$

Jej wiersze T'^*_1, T'^*_2, oraz T'^*_3 to, odpowiednio, '*nowe*' funkcjonalne surogaty spektralne szeregów czasowych T_1 ,T_2 oraz T_3. Ich długość wynosi m – 2, ponieważ dwa wiersze zostały wymazane z wejściowej (tzn. '*starej*') macierzy temporalnej TM. Natomiast, obydwie '*nowe*' macierze odległości temporalnych dla funkcji $3 - L2$ oraz $3 - evL2$ mają postać:

$$TDM'_{3-L2} = \begin{bmatrix} & T_1 & T_2 & T_3 \\ T_1 & 0 & 15.52 & 63.36 \\ T_2 & 15.52 & 0 & 62.46 \\ T_3 & 63.36 & 62.46 & 0 \end{bmatrix}$$

oraz

$$TDM'_{3-evL2} = TDM'_{evL2} = \begin{bmatrix} & T'_1 & T'_2 & T'_3 \\ T'_1 & 0 & 1.41 & 1.41 \\ T'_2 & 1.41 & 0 & 1.41 \\ T'_3 & 1.41 & 1.41 & 0 \end{bmatrix}$$

5/ **Krok piąty** algorytmu, to wybór odpowiedniego klasyfikatora. Jak powyżej wspomniano, wydajność nowo zaproponowanych funkcji odległości będzie testowana w zadaniach klasyfikacyjnych w podrozdziale czwartym. Przypomnijmy, iż dla kolekcji $\boldsymbol{T} = \{T_g\}_{g=1}^{m}$ szeregów czasowych i dyskretnego wektora ich etykiet $\boldsymbol{C}$, problem klasyfikacji polega na aproksymacji funkcji o postaci $c: \boldsymbol{T} \to \boldsymbol{C}$ (gdzie $c(T_g) = C^{T_g}$ dla każdego $g \in J$) funkcją $\hat{c}: \boldsymbol{T} \to \boldsymbol{C}$ tak aby spełniony był warunek

$\hat{c}(T_g) = C^{T_g}$ dla każdego $g \in J$. Funkcja c to *klasyfikator aprioryczny* (ang. *predefined*), natomiast funkcja $\hat{c}$ to *klasyfikator aposterioryczny*[6]. A więc, jeżeli:

$$c(T_g) = \hat{c}(T_g) \text{ dla każdego } g \in J \tag{12}$$

to mówimy, że klasyfikator aposterioryczny w pełni (perfekcyjnie) aproksymuje klasyfikator aprioryczny. Natomiast jeżeli tylko dla większości $g \in J$ zachodzi warunek (12), to mówimy, iż klasyfikator aposterioryczny tylko częściowo przybliża klasyfikator aprioryczny. Zagadnienie klasyfikacji polega więc na znalezieniu jak najdokładniejszego klasyfikatora aposteriorycznego. Tym samym, wydajność klasyfikatora aposteriorycznego może być mierzona w procentach poprawnie sklasyfikowanych szeregów czasowych w analizowanym zbiorze danych[7]. Postępując zgodnie z sugestią zaproponowaną przez E. Keogha oraz E. Kasetty'ego (2003), w proponowanym algorytmie, obliczone macierze odległości temporalnych stanowią dane wejściowe dla klasyfikatora jednego najbliższego sąsiada (ang. 1-*nearest neighbor* (1NN) *classifier*). W podrozdziale czwartym, zaproponowany powyżej algorytm przetestowany będzie przez pryzmat klasyfikatora 1NN na trzech zbiorach danych empirycznych biosygnałów. W szczególności zweryfikowano hipotezę badawczą mówiącą o koncentracji informacji strukturalnej zawartej w temporalnej macierzy danych na kilku lub kilkunastu pierwszych wektorach własnych macierzy Q.

Materiały oraz metody

Efektywność nowo zaproponowanych algorytmów zostanie przetestowana na trzech przykładowych empirycznych zbiorach danych: dwóch empirycznych kolekcjach elektrokardiogramów (zbiory *ECG200* oraz *ECGFive*) oraz na jednej empirycznej kolekcji krzywych analiz HRM (zbiór *Fungi*). Podstawowe własności powyższych trzech zbiorów danych zawarte są w tabeli 1. Zbiór ECG200 składa się z zapisów EKG pochodzących od osób zdrowych oraz osób z zawałem mięśnia sercowego (Olszewski, 2001). Zbiór ECGFive zawiera zapisy EKG pochodzące od 67 letniego mężczyzny zarejestrowane w dniach 12/11/1990 i 17/11/1990 (a więc, z pięciodniowym odstępem) (Time Series Machine Learning Website, 2023). Zbiór *Fungi* składa się z krzywych topnienia (tj. krzywych HRM), będących pierwszymi pochodnymi (ze znakiem minus) intensywności fluorescencji (tj. $-\frac{dF}{dt}$, gdzie F to intensywność fluorescencji, a t to temperatura) (Lu i in., 2017). Próbki dsDNA poddane analizie HRM pochodziły od 51 szczepów wyizolowanych z 18 gatunków grzybów (por. Tabela 1 w Lu i in., 2017).

---

[6] Oczywiście, pojęcia klasyfikatora apriorycznego oraz aposteriorycznego mają sens tylko w odniesieniu do matematycznego problemu klasyfikacji. Oznacza to, iż chociaż elementy dyskretnego wektora etykiet mogły zostać przyporządkowane poszczególnym szeregom czasowym na podstawie procedur empirycznych, to podczas procesu klasyfikacji owo przyporządkowanie jest czymś z góry ustalonym (ang. *predefined*), a więc z punktu widzenia klasyfikacji jest to przyporządkowanie *aprioryczne*.

[7] W tym ujęciu, wydajność klasyfikatora apriorycznego jest zawsze równa 100 %.

Tabela 1. Podstawowe własności strukturalne przykładowych zbiorów danych

TD	m	n	$\|\|C\|\|$
ECG200	200	96	2
ECGFive	884	136	2
Fungi	204	201	18

Legenda: *m, n* oraz $\|\|C\|\|$ to, odpowiednio, moc zbioru ***TD***, długość szeregów czasowych z ***TD*** oraz liczba klas równoważności wektora etykiet C.
Źródło: opracowanie własne.

Amplikony poddane analizie stanowiły regiony ITS (ang. *internal transcribed spacer*) rDNA. Wszystkie trzy przykładowe zbiory danych są publicznie dostępne (Time Series Machine Learning Website, 2023). Zgodnie z ogólnie przyjętą metodologią walidacji nowo zaproponowanych miar odległości pomiędzy danymi temporalnymi (Górecki, Łuczak, 2013; Górecki, Piasecki, 2019; Keogh, Kasetty, 2002; Wang i in., 2013; Wilczek, 2022), porównano efektywność algorytmów opartych na zaproponowanych w tej pracy miarach odległościach z efektywnością algorytmów bazujących na klasycznych (danych ogólną zależnością (7)) oraz na niedawno wprowadzonych funkcjach odległości. Mianowicie, efektywność w klasyfikacji danych czasowych nowo zaproponowanych algorytmów będzie porównana nie tylko z wydajnością algorytmów opartych na tradycyjnych metrykach (8)-(11), ale również z wydajnością ostatnio szeroko rekomendowanego algorytmu (ang. *Dynamic Time Warping algorithm*) (Berndt, Clifford, 1994), jego trzech modyfikacji (tzn. *CIDTW*, *DDTW* oraz *PDDTW*), jak również z efektywnością algorytmów opartych na trzech modyfikacjach funkcji Euklidesowej (tzn. *CIL2*, *DL2* oraz *PDL2*). Przypomnijmy, iż algorytm *CIDF*, gdzie *DF* to *dowolna* miara temporalna (np. $DF \in \{DTW, L2\}$) to tzw. *algorytm niezmienniczy* (ze względu na złożoność strukturalną danych temporalnych) obliczania odległości pomiędzy dwoma szeregami czasowymi (por. Batista, Wang, Keogh, 2011). Natomiast algorytm *DDF*, gdzie *DF* to *dowolna* miara temporalna (np. $DF \in \{DTW, L2\}$) to tzw. algorytm *pochodnej miary* odległości pomiędzy dwoma szeregami czasowymi T_1 oraz T_2, którego obliczenie redukuje się do obliczenia odległości wejściowej *DF* pomiędzy pierwszymi dyskretnymi pochodnymi T_1^d oraz T_2^d, odpowiednio, szeregów czasowych T_1 oraz T_2 (Keogh, Pazzani, 2001). Z kolei, algorytm *PDDF*, gdzie *DF* to dowolna miara temporalna (np. $DF \in \{DTW, L2\}$) to *parametryczna pochodna miara* odległości, której obliczenie redukuje się do obliczenia *ważonej wypukłej* (ang. *weighted convex*) kombinacji wejściowej miary *DF* oraz jej miary pochodnej *DDF* (Górecki, Łuczak, 2013). W poniższej pracy, algorytm *PDDF* jest obliczony dla parametrów $a = b = \cos\alpha = \sin\alpha = 0.7853982$, gdzie $\alpha = \frac{\pi}{4}$ (por. Górecki, Łuczak, 2013, s. 317).

Wszystkie symulacje komputerowe oraz wizualizacje otrzymanych danych przeprowadzone zostały w języku programowania *R* oraz w jego pakietach (Mori, Mendiburu, Lozano, 2016; R Core Team, 2022; Venables, Ripley, 2002; Wickham, 2009).

Wyniki i dyskusja

Zgodnie z sugestią z podrozdziału drugiego, efektywność nowo zaproponowanych algorytmów klasyfikacyjnych mierzona jest w procentach poprawnie sklasyfikowanych szeregów czasowych w rozpatrywanym zbiorze danych. Z przeprowadzonych symulacji wynika, iż wydajność nowych spektralnych *k*-obciętych schematów klasyfikacyjnych mieści się w granicach od 90% (protokół $k_{opt} - evLmax$ na zbiorze *ECG200*) do 100 % (protokoły $k_{opt} - evL1$, $k_{opt} - evL3$ oraz $k_{opt} - evLmax$ na zbiorze *Fungi*), a ich wartości średnie są w zakresie od 96.4 % (protokół $k_{opt} - evLmax$) do 97.35 % (protokół $k_{opt} - evL2$) (por. tabele 2 i 3). Procentowe poprawy skuteczności w zadaniach klasyfikacyjnych osiągnięte przez nowe $k_{opt} - evLp$ techniki w stosunku do ich tradycyjnych L_p odpowiedników są w przedziale od 0.11 % (protokół $k_{opt} - evL3$ na zbiorze *ECGFive*) do 5.39 % (protokół $k_{opt} - evL3$ na zbiorze *Fungi*), zaś ich średnie wartości znajdują się w zakresie od 2.02 % (protokół $k_{opt} - evL2$) do 3.89 % (protokół $k_{opt} - evLmax$). Porównanie skuteczności najwydajniejszej (dla danego zbioru danych czasowych) nowej spektralnej *k*-obciętej metodologii z efektywnością metodologii opartych na odległości *DTW* i jej modyfikacjach jednoznacznie ukazuje, iż procentowe poprawy skuteczności nowych technik klasyfikacyjnych są w granicach od 0.45 % (poprawa nad protokołem *DDTW* na zbiorze *ECGFive*) do 18.28 % (poprawa nad protokołem *CIDTW* na zbiorze *ECG200*), zaś średnia wartości popraw skuteczności są na poziomie od 5.19 % (poprawa nad protokołem *PDDTW*) do 8.83 % (poprawa nad protokołem *CIDTW*) (por. tabela 5). Procentowa poprawa skuteczności osiągnięta przez nowo wprowadzoną $k - evL2$ technikę jest w granicy od 0 % (poprawa nad protokołem *CIL2* na zbiorze *ECGFive*) do 10.75 % (poprawa nad protokołem *DL2* na zbiorze *ECG200*), zaś średnie wartości poprawy jakości schematów klasyfikacyjnych są na poziomie od 2.28 % (poprawa nad protokołem *PDL2*) do 6.14 % (poprawa nad protokołem *DL2*) (por. tabela 6). Konkludując można stwierdzić, iż wydajność algorytmów opartych na nowych spektralnych *k*-obciętych funkcjach odległości jest wyższa lub znacznie wyższa niż wydajność algorytmów opartych na metrykach danych ogólną zależnością (7) lub na niedawno zaproponowanej funkcji *DTW* i jej udoskonaleniach, jak również na udoskonaleniach metryki *L2*. Dlatego też można stwierdzić, iż wprowadzenie nowych funkcji odległości opartych na wektorach własnych temporalnej macierzy produktowej *TP* wydaje się być w pełni uzasadnione.

Tabela 2. Wyniki klasyfikacji (w %) otrzymane przez algorytmy oparte na miarach oraz na referencyjnych metrykach

TD	$L1$	$k_{opt} - evL1$	↑	$L2$	$k_{opt} - evL2$	↑
ECG200	88	92	4.35	89.5	**93**	3.76
ECGFive	97.62	**99.77**	2.15	99.21	99.55	0.34
Fungi	99.02	**100**	0.98	97.55	99.51	1.97
Średnia z pomiarów	94.88	97.26	2.49	95.42	**97.35**	2.02

Legenda: Symbol ↑ oznacza poprawę (w %) wydajności klasyfikacji nowo zaproponowanych algorytmów względem algorytmów odniesienia. Najlepsze wyniki (spośród wyników z tabel 2 i 3) są pogrubione.
Źródło: opracowanie własne.

Tabela 3. Wyniki klasyfikacji (w %) otrzymane przez algorytmy oparte na miarach oraz na referencyjnych metrykach

TD	$L3$	$k_{opt} - evL3$	↑	$L_{\max}$	$k_{opt} - evLmax$	↑
ECG200	89.5	90.5	1.1	87	90	3.33
ECGFive	99.21	99.32	0.11	95.81	99.21	3.43
Fungi	94.61	**100**	5.39	95.1	**100**	4.9
Średnia z pomiarów	94.44	96.61	2.2	92.64	96.4	**3.89**

Legenda: Symbol ↑ oznacza poprawę (w %) wydajności klasyfikacji nowo zaproponowanych algorytmów względem algorytmów odniesienia. Najlepsze wyniki (spośród wyników z tabel 2 i 3) są pogrubione.
Źródło: opracowanie własne.

Z kolei, tabela 4 ukazuje, iż poprawa skuteczności (względem referencyjnych metod L_p) algorytmów opartych na zwykłych k-obciętych metrykach jest niższa niż algorytmów bazujących na funkcjach spektralnych i mieści się w granicy od 0 % (protokoły $k_{opt} - L2$ oraz $k_{opt} - Lmax$ oraz na zbiorze *Fungi*) do 3.87 % (protokół $k_{opt} - Lmax$ na zbiorze *ECGFive*). Ryciny 1-3 obrazują związek pomiędzy efektywnością nowych miar k – Lp oraz $k - evLp$ a parametrem k. Przypomnijmy, iż na wykresach w lewych panelach analizowanych rycin, parametr k przebiega zbiór indeksów I, a na wykresach w prawych panelach, parametr k przebiega zbiór indeksów J.

Tabela 4. Wyniki klasyfikacji (w %) otrzymane przez algorytmy oparte na miarach

TD	$k_{opt} - L1$	↑	$k_{opt} - L2$	↑	$k_{opt} - L3$	↑	$k_{opt} - Lmax$	↑
ECG200	89.5	1.68	91.5	2.19	**92.5**	3.24	90.5	**3.87**
kECGFive	98.64	**1.03**	**99.77**	0.56	99.66	0.45	95.93	0.13
Fungi	**99.51**	0.49	97.55	0	96.57	**2.03**	95.1	0
Średnia z pomiarów	95.88	1.07	**96.27**	0.92	96.24	**1.91**	93.84	1.33

Legenda: Symbol ↑ oznacza poprawę (w %) wydajności klasyfikacji nowo zaproponowanych algorytmów względem algorytmów odniesienia . Najlepsze wyniki są pogrubione.
Źródło: opracowanie własne.

Tabela 5. Wyniki klasyfikacji (w %) otrzymane przez algorytmy oparte na mierze i na jej modyfikacjach

TD	*DTW*	↑	*CIDTW*	↑	*DDTW*	↑	*PDDTW*	↑
ECG200	**82**	11.83	76	18.28	80	13.98	81.5	12.37
ECGFive	96.15	3.63	96.95	2.83	**99.32**	0.45	97.06	2.72
Fungi	**99.51**	0.49	94.61	5.39	97.06	2.94	**99.51**	0.49
Średnia z pomiarów	92.55	5.32	89.19	**8.83**	92.13	5.79	**92.69**	5.19

Legenda: Symbol ↑ oznacza poprawę (w %) wydajności klasyfikacji najlepszych nowo zaproponowanych algorytmów (por. tabele 2 i 3) względem algorytmów odniesienia typu i pokrewnych. Najlepsze wyniki są pogrubione.
Źródło: opracowanie własne.

Tabela 6. Wyniki klasyfikacji (w %) otrzymane przez algorytmy oparte na modyfikacjach miary

TD	*CIL2*	↑	*DL2*	↑	*PDL2*	↑
ECG200	86.5	6.99	83	10.75	89	4.3
ECGFive	**99.55**	0	93.89	5.69	98.98	0.57
Fungi	94.61	4.92	**97.55**	1.97	**97.55**	1.97
Średnia z pomiarów	93.55	3.97	91.48	**6.14**	95.18	2.28

Legenda: Symbol ↑ oznacza poprawę (w %) wydajności klasyfikacji najlepszych nowo zaproponowanych algorytmów k_opt-evLp (por. tabele 2 i 3) względem algorytmów odniesienia (tj. modyfikacji miary L2). Najlepsze wyniki są pogrubione.
Źródło: opracowanie własne.

Tabela 7. Minimalne wartości (w % wektorów kolumnowych macierzy) dla nowych miar

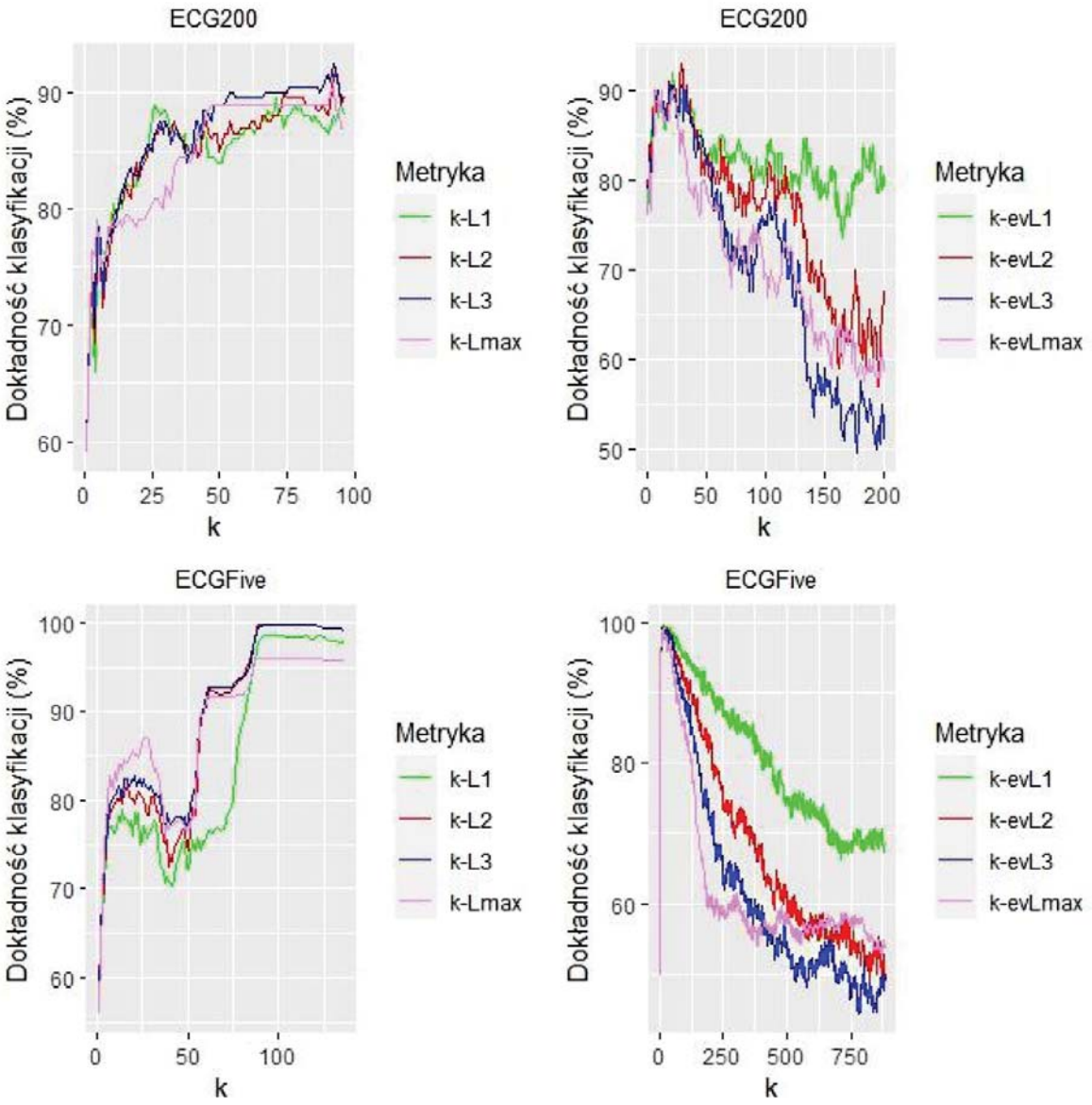
TD	$k_{opt} - L1$	$k_{opt} - L2$	$k_{opt} - L3$	$k_{opt} - Lmax$
ECG200	73.96	95.83	95.83	95.83
ECGFive	68.38	68.38	67.65	64.71
Fungi	51.74	87.56	52.74	87.06
Średnia z pomiarów	64.69	83.92	72.07	82.53

Źródło: opracowanie własne.

Tabela 8. Minimalne wartości (w % wektorów własnych macierzy produktowej) dla nowych miar

TD	$k_{opt} - evL1$	$k_{opt} - evL2$	$k_{opt} - evL3$	$k_{opt} - evLmax$
ECG200	10.5	14.5	10	3.5
ECGFive	1.58	1.36	1.36	1.36
Fungi	6.37	6.37	6.37	6.37
Średnia z pomiarów	6.15	7.41	5.91	3.74

Źródło: opracowanie własne.

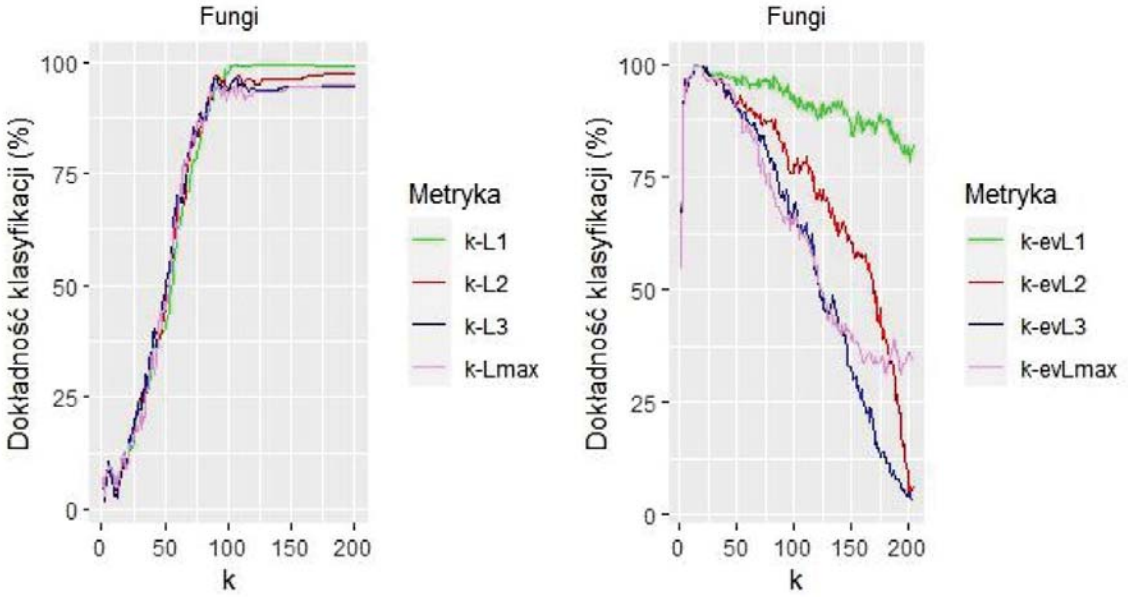


Rysunek 1. Zależność pomiędzy parametrem *k* a dokładnością algorytmów opartych na nowych funkcjach temporalnych *k – Lp* oraz *k – evLp* określonych na zbiorach danych *ECG200* oraz *ECGFive*

Źródło: opracowanie własne.

Rozpatrywane wykresy dowodzą, iż związek pomiędzy wydajnością analizowanych funkcji, a parametrem *k* nie zawsze jest monotoniczny. Jakkolwiek, analizując wydajność funkcji *k – Lp* można zauważyć, iż we wszystkich trzech przypadkach, wraz ze wzrostem wartości parametru *k*, ich wydajność również wzrasta i po przebiegnięciu około trzech czwartych wartości parametru *k* osiąga swoje maksimum. Minimalne wartości parametru k_{opt} wyrażone w procentach wektorów kolumnowych wejściowych macierzy *TM* dla których badane funkcje są najbardziej efektywne w przykładowych zadaniach klasyfikacyjnych zawarte są w tabeli 7. Dane te w sposób jednoznaczny wskazują, iż znajomość od 64.69 % do 83.92 % pierwszych wyrazów analizowanych szeregów czasowych wystarczy do ich prawidłowego sklasyfikowania. A więc można stwierdzić, iż hipoteza z podrozdziału drugiego mówiąca, iż najważniejsze informacje strukturalne charakteryzujące

dany szereg temporalny są zakodowana w jego *k*-pierwszych wyrazach wydaje się być w pełni potwierdzona i tym samym rozpatrywanie nowych -obciętych funkcji odległości wydaje się być w pełni uzasadnione.



Rysunek 2. Zależność pomiędzy parametrem *k* a dokładnością algorytmów opartych na nowych funkcjach temporalnych *k – Lp* oraz *k – evLp* określonych na zbiorze danych *Fungi*

Źródło: opracowanie własne.

Z drugiej strony, analizując efektywność algorytmów opartych na funkcjach *k – evLp* można dostrzec, iż we wszystkich trzech przypadkach, wraz ze wzrostem wartości parametru *k*, ich efektywność szybko wzrasta, po przebiegnięciu około jednej dwudziestej wartości parametru *k* osiąga maksimum i następnie stopniowo maleje. Minimalne wartości parametru k_{opt} wyrażone w procentach wektorów własnych temporalnej macierzy produktowej *TP* zawarte są w tabeli 8. Dane te w sposób jednoznaczny wskazują, iż znajomość od 3.74 % do 7.41 % pierwszych wektorów własnych macierzy (1) wystarczy do prawidłowej klasyfikacji szeregów czasowych stanowiących wiersze temporalnej macierzy danych *TM*. Porównując dane z tabel 7 i 8 widzimy, iż $k^I_{opt} >_{\%} k^J_{opt,}$ a więc informacja strukturalna charakteryzująca badane szeregi czasowe jest skondensowana na kilku pierwszych wyrazach ich funkcjonalnych surogatów spektralnych. Można zauważyć więc, iż następne dwie hipotezy z podrozdziału drugiego zostały potwierdzone i tym samym rozpatrywanie nowych spektralnych *k*-obciętych funkcji odległości wydaje się być w pełni uzasadnione.

W swojej obszernej i ważnej pracy „*The Elements of Statistical Learning. Data Mining, Inference, and Prediction*", T. Hastie, R. Tibshirani oraz J. Friedman (2009, s. 506) zauważają, iż „wyszczególnienie adekwatnej miary niepodobieństwa 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 oraz jest mniej podatny na ogólne analizy".

W naszej opinii, deklaracje T. Hastie oraz współpracowników, 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 *Lp* połączonego z klasyfikatorem 1*NN*) zaproponowano nowe miary odległości pomiędzy danymi temporalnymi. Z przeprowadzonych symulacji komputerowych wynika, iż wydajność algorytmów klasyfikacji opartych na nowych parametrycznych funkcjach odległości jest wyższa lub znacznie wyższa niż wydajność protokołów opartych na funkcjach referencyjnych (np. na klasycznych miarach *Lp*, *DTW* oraz na ich (parametrycznych) modyfikacjach). Rezultat ten wydaje się być bardzo istotny, ponieważ jak zauważa T. Górecki oraz M. Łuczak (2013, s. 311) „[…] prosta metoda łącząca klasyfikator jednego najbliższego sąsiada (1*NN*) oraz pewną formę miary odległości *DTW* okazała się być jedną z najwydajniejszych technik klasyfikacji szeregów czasowych. […]. Zostało empirycznie dowiedzione, iż prosta Euklidesowa metryka odległości jest konkurencyjna lub lepsza względem wielu złożonych miar odległości oraz spełnia ważną nierówność trójkąta". W dalszych częściach swojego tekstu T. Górecki oraz M. Łuczak (2013, s. 320) 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". Konfrontując powyższe fragmenty z wynikami z tabel 2-6 uprawnionym jest twierdzenie, iż nasze rezultaty stanowią kontrprzykład względem opinii T. Góreckiego i M. Łuczaka.

Uwagi końcowe

Podsumowując można stwierdzić, iż cel przedkładanej pracy został osiągnięty i nowe (spektralne) *k*-obcięte funkcje temporalne znajdą praktyczne zastosowanie w analizie i klasyfikacji danych diagnostycznych i tym samym przyczynią się do zwiększenia stopnia automatyzacji procedur medycznych.

Literatura

Artykuły i pozycje książkowe

1) Batista, G. E. A. P. A., Wang, X., Keogh, E. J. (2011). *A complexity-invariant distance measure for time series*. W: B. Liu, H. Liu, C. Clifton, T. Washio, C. Kamath (red.), *Proceedings of the 2011 SIAM International Conference on Data Mining,* (699-710). Arizona: Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972818.60.
2) Berndt, D. J., Clifford, J. (1994). *Using dynamic time warping to find patterns in time series*. W: *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining* (359-370).
3) Bloomfield, V. A. (2014). *Using R for Numerical Analysis in Science and Engineering. The R Series*. Boca Raton: CRC Press.

4) Box, G. E. P., Jenkins, G. M. (1983). *Analiza Szeregów Czasowych. Prognozowanie i Sterowanie*. Warszawa: Państwowe Wydawnictwo Naukowe.
5) Górecki, T., Łuczak, M. (2013). Using derivatives in time series classification. *Data Mining and Knowledge Discovery*, *26 (2)*, 310-331. DOI 10.1007/s10618-012-0251-4.
6) Górecki, T., Piasecki, P. (2019). A comprehensive comparison of distance measures for time series classification. W: A. Steland, E. Rafajłowicz, O. Okhrin (red.), *Stochastic Models, Statistics and their Applications. Springer Proceedings in Mathematics & Statistics*, Vol. 294 (409-428). Cham: Springer. https://doi.org/10.1007/978-3-030-28665-1_31.
7) Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. New York: Springer Series in Statistics. https://doi.org/10.1007/978-0-387-84858-7.
8) Keogh, E. J., Pazzani, M. J. (2001). *Dynamic time warping with higher order features*. W: V. Ku-mar, R. Grossman (red.), *Proceedings of the 2001 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics (1-11).
9) Keogh, E., Kasetty, E. (2002). *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: Association for Computing Machinary (102-111). https://doi.org/10.1023/A:1024988512476.
10) Kuniszyk-Jóżkowiak, W. (2011). *Przetwarzanie Sygnałów Biomedycznych*. Lublin: Instytut Informatyki UMCS.
11) Lu, S., Mirchevska, G., Phatak, S. S., Li, D., Luka, J., Calderone, R. A., Fonzi, W. A. (2017). Dynamic time warping assessment of high-resolution melt curves provides a robust metric for fungal identification. *PLoS ONE, 12 (3)*, e0173320. https://doi.org/10.1371/journal.pone.0173320.
12) Mori, U., Mendiburu, A., Lozano, J. A. (2016). Distance measures for time series in R: The TSdist Package. *The R Journal*, *8(2)*, 451-459. https://doi.org/10.32614/RJ-2016-058.
13) Olszewski, R. T. (2001). *Generalized Feature Extraction for Structural Pattern Recognition in Time Series Data*. Pittsburgh: Carnegie Mellon University.
14) Shifaz, A., Pelletier, C., Petitjean, F., Webb, G. I. (2023). Elastic similarity and distance measures for multivariate time series. *Knowledge and Information Systems, 65 (6)*, 2665-2698. https://doi.org/10.1007/s10115-023-01835-4.
15) Venables, W. N., Ripley, B. D. (2002). *Modern Applied Statistics with S*. Fourth Edition. New York: Springer. https://doi.org/10.1007/978-0-387-21706-2.
16) Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, *26 (2)*, 149-182. https://doi.org/10.1007/s10618-012-0250-5.
17) Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag. https://doi.org/10.1007/978-3-319-24277-4.

18) Wilczek, P. (2022). *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: K. Krukiewicz, M. Marczyk, M. Bugdol, S. Bajkacz, Z. Ostrowski Z. (red.), *Recent Advances in Computational Oncology and Personalized Medicine*. *The Challenges of the Future !* Vol. 2. (128-152). Gliwice: Wydawnictwo Politechniki Ślaskiej. DOI:10.34918/85110.

Źródła internetowe

1) R Core Team. (2022). *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Pobrane z: https://www.R-project.org/ (dostęp: 16.07.2024).
2) Time Series Machine Learning Website. (2003). Pobrane z: https://timeseriesclassification.com/index.php (dostęp: 16.07.2024).