Autoreferat

dr Maciej Łuczak

Wydział Inżynierii Lądowej, Środowiska i Geodezji
Politechnika Koszalińska

Koszalin, 2018

1. Życiorys naukowy

1.1. Imię i nazwisko: **Maciej Łuczak**

1.2. Uzyskane dyplomy i stopnie naukowe:

- **Doktor nauk matematycznych** w zakresie matematyki: 2006 r.
  Wydział Matematyki i Informatyki, Uniwersytet im. Adama Mickiewicza w Poznaniu;
  Tytuł rozprawy: Rachunki funkcyjne i twierdzenia typu Gleasona-Kahane'a-Żelazki w algebrach A-pseudowypukłych;
  Promotor: dr hab. Andrzej Sołtysiak, prof. UAM.

- **Magister matematyki**: 2001 r.
  Wydział Matematyki i Informatyki, Uniwersytet im. Adama Mickiewicza w Poznaniu;
  Tytuł rozprawy: Funkcje całkowite i algebry Banacha;
  Promotor: dr hab. Andrzej Sołtysiak, prof. UAM.

1.3. Informacje o dotychczasowym zatrudnieniu:

- 2006–: **adiunkt** na Wydziale Inżynierii Lądowej, Środowiska i Geodezji (dawniej: Wydział Budownictwa i Inżynierii Środowiska) Politechniki Koszalińskiej — Zakład Matematyki (dawniej: Katedra Matematyki);

- 2005–2006: **asystent** na Wydziale Budownictwa i Inżynierii Środowiska (obecnie: Wydział Inżynierii Lądowej, Środowiska i Geodezji) Politechniki Koszalińskiej — Katedra Matematyki (obecnie: Zakład Matematyki);

- 2001–2005: **doktorant** na Wydziale Matematyki i Informatyki Uniwersytetu im. Adama Mickiewicza w Poznaniu.

1.4. Osiągnięcie naukowe, o którym mowa w z art. 16 ust. 2 ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. 2016 r. poz. 882 ze zm. w Dz. U. z 2016 r. poz. 1311.): Jednotematyczny cykl publikacji pod wspólnym tytułem:

Kombinacje miar odległościowych na danych surowych i transformowanych w klasyfikacji szeregów czasowych.

2. Działalność naukowa

2.1. Jednotematyczny cykl publikacji:

[A1] T. Górecki, M. Łuczak (2013), Using derivatives in time series classification, Data Mining and Knowledge Discovery 26(2), 310–331. **IF: 1.743, Lista A: 40 pkt.**

[A2] T. Górecki, M. Łuczak (2014), First and second derivative in time series classification using DTW, Communications in Statistics-Simulation and Computation 43(9), 2081–2092. **IF: 0.325, Lista A: 15 pkt.**

[A3] T. Górecki, M. Łuczak (2014), Non-isometric transforms in time series classification using DTW, Knowledge-Based Systems 61, 98–108. **IF: 2.947, Lista A: 35 pkt.**

[A4] T. Górecki, M. Łuczak (2015), Multivariate time series classification with parametric derivative dynamic time warping, Expert Systems with Applications 42(5), 2305–2312. **IF: 2.981, Lista A: 35 pkt.**

[A5] M. Łuczak (2016), Hierarchical clustering of time series data with parametric derivative dynamic time warping, Expert Systems with Applications 62, 116–130. **IF: 3.928, Lista A: 35 pkt.**

[A6] M. Łuczak (2017), Univariate and multivariate time series classification with parametric integral dynamic time warping, Journal of Intelligent & Fuzzy Systems 33(4), 2403–2413. **IF: 1.261, Lista A: 20 pkt.**

[A7] M. Łuczak (2018), Combining raw and normalized data in multivariate time series classification with dynamic time warping, Journal of Intelligent & Fuzzy Systems 34(1), 373–380. **IF: 1.261, Lista A: 20 pkt.**

2.2. Wstęp

Szeregi czasowe jest to reprezentacja danych pojawiająca się w wielu zagadnieniach teoretycznych i praktycznych. Przedstawienie dyskretnego zbioru danych jako ciągu liczb rzeczywistych (dyskretny szereg czasowy) jest szeroko stosowane we wszystkich dziedzinach nauki i techniki, poczynając od zagadnień ściśle teoretycznych w matematyce, fizyce, informatyce, poprzez zastosowania inżynierskie, ze szczególnym uwzględnieniem zagadnień informatycznych, po biologię, medycynę i ekonomię. W dzisiejszych czasach pojawiają się duże zbiory danych szeregów czasowych wymagające ich przetworzenia, wyciągnięcia interesującej informacji i zastosowania jej w praktyce lub przesłania do dalszego przetworzenia. Jedne z najczęściej stosowanych metod wydobycia interesujących informacji ze zbiorów danych są klasyfikacja (klasyfikacja pod nadzorem) i analiza skupień (klasyfikacja bez nadzoru). Metody te pozwalają nam rozdzielić wyjściowy zbiór danych na odpowiednie podzbiory szeregów o podobnych własnościach mając do dyspozycji przykładowe dane w postaci zbioru uczącego (klasyfikacja — dostępność etykiet w zbiorze uczącym) lub dysponując tylko samym nieocechowanym zbiorem danych (analiza skupień — brak etykiet w zbiorze/brak zbioru uczącego). W przypadku szeregów czasowych dochodzi tutaj jeszcze szczególny charakter takich danych, gdzie każda następna wartość szeregu jest w jakiś sposób zależna od wartości poprzednich (wcześniejszych w czasie/przestrzeni).

Istnieje ogromna liczba metod klasyfikacji i analizy skupień wszelkiego typu danych. W przypadku szeregów czasowych bardzo popularne jest podejście odległościowe, to znaczy, tworzy się odpowiednią miarę odległości, która pokazuje podobieństwo/niepodobieństwo dwóch szeregów czasowych i tę wartość (odległość) wykorzystuje w procesie klasyfikacji. Tutaj bardzo popularne jest grupowanie szeregów o najmniejszej odległości między nimi, czyli tak zwana metoda najbliższego sąsiada. Wymagane jest „tylko" znalezienie odpowiedniej miary odległości pozwalającej w jak najlepszy sposób wykazać podobieństwa miedzy szeregami, zarówno w przypadku uniwersalnym (szukamy metod dających najlepsze średnie rezultaty klasyfikacji na jak największej bazie danych), jak i w przypadkach szczególnych (szukamy metod działających na pewnych podzbiorach danych o specyficznych własnościach).

W swojej pracy badawczej dotyczącej przedstawianego osiągnięcia habilitacyjnego zajmowałem się klasyfikacją szeregów czasowych wykorzystując metodę najbliższego sąsiada oraz odpowiednie miary odległości w celu uzyskania jak najlepszych wyników na jak największej bazie szeregów czasowych, czyli działających na szerokim zakresie danych czasowych o bardzo różnych charakterystykach. Jedną z metod obróbki wejściowych (surowych) danych szeregów czasowych jest ich przetransformowanie na inny szereg czasowy. Operacje tę wykonuje się najczęściej w celu zmniejszenia wymiaru surowych danych, przy jednoczesnym zachowaniu całej (lub prawie całej) informacji zawartej w danym szeregu czasowym, co pozwala na łatwiejsze/szybsze/zajmujące mniej pamięci manipulacje przetransformowanymi danymi, w szczególności ich klasyfikację czy analizę skupień. W swojej pracy skupiłem się na innym sposobie transformacji surowych danych czasowych. Mianowicie na transformacjach, które nie zachowują dokładnie informacji zawartych w szeregach wejściowych, więcej, czasami dość mocno odkształcającymi dane surowe. Jednocześnie mają to być takie transformacje, które wydobywają i uwypuklają jakiś szczególny aspekt wejściowych danych, przy jednoczesnym być może pominięciu innych cech. To podejście powoduje oczywiście, że daje ono dobre rezultaty tylko dla pewnego podzbioru ogólnej bazy danych szeregów, często na zdecydowanej mniejszości z nich. Zatem bezpośrednie

zastosowanie tak transformowanych szeregów w klasyfikacji odległościowej nie może dać dobrych rezultatów, gdy nic nie wiemy o charakterystykach poszczególnych szeregów znajdujących się w ogólnej bazie danych. Zastosowane przeze mnie podejście polegało na kombinacji informacji z szeregów surowych oraz szeregów przetransformowanych. Odpowiednia ich kombinacja pozwala przydzielić różną ważność danym surowym i transformowanym na każdym pojedynczym zbiorze danych. Jeżeli dana transformacja akurat uwypukla cechy szeregów w danym zbiorze danych, to ta informacja ma większy wpływ na dalszy proces klasyfikacji (obliczania odległości). Jeżeli natomiast transformacja nie wnosi nic nowego do istniejącej informacji w szeregu (lub wręcz psuje tę informację), to zwiększa się udział danych surowych, a minimalizuje wpływ transformacji. Co więcej w czasie badań wykazano, że nawet jeśli klasyfikacja zarówno samych danych surowych, jak i samych danych transformowanych daje złe rezultaty, to ich odpowiednia kombinacja może dawać rezultaty bardzo dobre wyciągając niejako więcej informacji z wejściowego szeregu czasowego niż można wyciągnąć osobno z danych surowych i transformowanych. Jednocześnie w pracy skupiłem się na wydajności i prostocie badanych kombinacji funkcji odległościowych, w zamyśle miała być to metoda kombinowana do bezpośredniego zastosowania w jak największej liczbie przypadków, gdzie brakuje informacji o charakterystykach zbiorów danych, a wymagane są lepsze rezultaty niż w metodach z zastosowaną pojedynczą odległością, wykorzystującą tylko dane surowe lub tylko dane przetworzone.

Za wkład przeprowadzonych przeze mnie badań w rozwój dyscypliny uważam:

- Opracowanie ogólnej kombinowanej miary odległości szeregów czasowych wykorzystującej zarówno dane surowe, jak i transformowane, z możliwością użycia dowolnych składowych miar odległości i dowolnych transformat;
- Opracowanie algorytmów wykorzystujących parametryczne wypukłe kombinacje miar odległościowych, które są równocześnie wydajne w dziedzinie klasyfikacji, jak i niewymagające pod względem wymaganych zasobów oraz łatwo interpretowalne, proste w użyciu i implementacji;
- Uogólnienie miary odległości i algorytmu klasyfikacji z przypadku jednowymiarowych szeregów czasowych na przypadek wielowymiarowy;
- Przystosowanie badanych algorytmów klasyfikacji pod nadzorem na przypadek klasyfikacji bez nadzoru (analiza skupień) z wykorzystaniem oryginalnego algorytmu korekcyjnego;
- Zastosowanie z powodzeniem niestandardowych transformacji danych (transformaty nieizometryczne, całkowanie, normalizacja) w klasyfikacji jedno i wielowymiarowych szeregów czasowych;
- Autorskie oprogramowanie wykorzystujące badane metody przystosowane do obliczeń równoległych i rozproszonych.

2.3. Wyniki prac

Praca [A1]

W pracy [A1] po raz pierwszy pojawiła się idea prostego klasyfikatora kombinowanego wykorzystującego metodę najbliższego sąsiada wraz z miarami odległości: euklidesową i DTW (Dynamic Time Warping).

(Dyskretnym jednowymiarowym) szeregiem czasowym będziemy nazywać skończony ciąg liczb rzeczywistych:

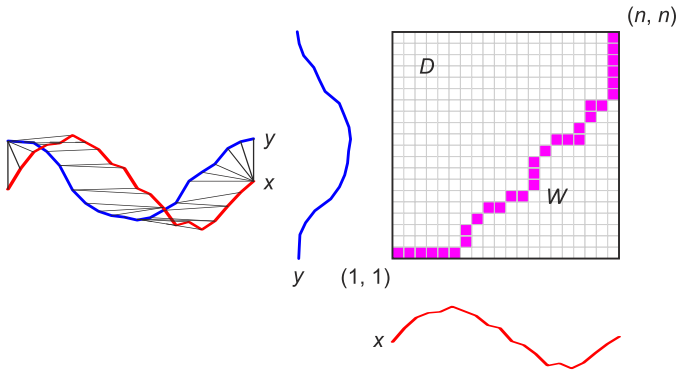
$$x = \{x(i) \in \mathbb{R} \colon i = 1, 2, \ldots, n\}, \quad n \in \mathbb{N}.$$

Euklidesową miarę odległości (ED — euclidian distance) między dwoma szeregami x i y definiujemy jako

$$\mathrm{ED}(x, y) = \sqrt{\sum_{i=1}^{n} \big(x(i) - y(i)\big)^2}.$$

Miara ED porównuje zawsze te same cechy (współrzędne) szeregu (wektora) i nie zależy od kolejności cech. Jest często stosowana przy porównywaniu obiektów nie będących szeregami i nie ma żadnych specjalnych własności związanych z sekwencyjną naturą szeregów czasowych. Jest ona jednak prosta w obliczaniu i niewymagająca wydajnościowo, więc jest często stosowana także do porównań szeregów czasowych. ED musi być obliczana na szeregach o tej samej długości $n \in \mathbb{N}$.

Inną miarą odległościową jest DTW (Dynamic Time Warping). Jest to miara przystosowana do obliczania odległości na szeregach czasowych. Nie porównuje zawsze tych samych cech. Dzięki pewnemu przesunięciu porównywanych cech wydobywa sekwencyjną naturę szeregów czasowych, porównuje raczej ogólny kształt szeregu (sygnału) nawet jeżeli występuje pewne przesunięcie lub rozszerzenie/zwężenie wykresów porównywanych szeregów czasowych (Rys. 1). DTW może być obliczane na szeregach o różnych długościach $n, m \in \mathbb{N}$.



Rysunek 1: Przesunięcie porównywanych cech oraz ścieżka skrzywienia dla dwóch szeregów o tej samej długości.

W celu obliczenia odległości DTW miedzy dwoma szeregami czasowymi x i y:

$$x = \{x(i)\colon i = 1, 2, \ldots, n\}, \quad y = \{y(i)\colon i = 1, 2, \ldots, m\}, \quad n, m \in \mathbb{N}$$

postępujemy następująco. Konstruujemy macierz D o wymiarach $n \times m$ z wartości odległości $d(x(i), y(j))$ miedzy dwoma cechami o indeksach i i j, gdzie d jest dowolną funkcja dwóch zmiennych nazywaną lokalną funkcją kosztu (local cost function). W standardowej odległości DTW funkcja d jest najczęściej odległością między dwoma liczbami rzeczywistymi $d(a, b) = (a - b)^2$. Elementy macierzy $D(i, j)$ odpowiadają przesunięciu między wartościami cech $x(i)$, $y(j)$ szeregów czasowych. Następnie konstruujemy tzw, ścieżkę skrzywienia (warping path) $W = (w_1, w_2, \ldots, w_K)$, składająca się z elementów macierzy D, $w_k = D(i, j)$. Ścieżka skrzywienia musi spełniać następujące trzy warunki:

1. warunki brzegowe (boundary conditions): $w_1 = D(1, 1)$, $w_K = D(n, m)$;
2. ciągłość (continuity): $i_{k+1} - i_k \leqslant 1$ oraz $j_{k+1} - j_k \leqslant 1$;
3. monotoniczność (monotonicity): $i_{k+1} - i_k \geqslant 0$ oraz $j_{k+1} - j_k \geqslant 0$.

Zatem aby utworzyć ścieżkę skrzywienia rozpoczynamy od elementu $w_1 = D(1, 1)$ i przesuwamy się dalej o co najwyżej jeden indeks w górę i/lub w prawo, aż do osiągnięcia ostatniego elementu $w_K = D(n, m)$ (Rys. 1). Ścieżka która minimalizuje koszt skrzywienia daje nam odległość DTW obliczaną:

$$\mathrm{DTW}(x, y) = \min_W \sqrt{\sum_{k=1}^{K} w_k}.$$

W praktyce aby obliczyć odległość DTW tworzymy tzw. macierz odległości skumulowanych (cumulative distance matrix) Γ. W obliczeniu tej macierzy korzystamy z programowania dynamicznego z następującą rekurencją

$$\Gamma(i, j) = d(x(i), y(j)) + \min\{\Gamma(i-1, j-1), \Gamma(i-1, j), \Gamma(i, j-1)\}$$

oraz warunkami początkowymi

$$\Gamma(0,0) = 0, \quad \Gamma(i,0) = \infty \ (i = 1,2,\ldots,n), \quad \Gamma(0,j) = \infty \ (j = 1,2,\ldots,m).$$

Wypełniamy macierz Γ kolumna po kolumnie (lub wiersz po wierszu) az dojdziemy do pozycji (n,m), która odpowiada szukanej odległości DTW:

$$\mathrm{DTW}(x,y) = \sqrt{\Gamma(n,m)}.$$

Miara odległości DTW nie jest metryką (nie spełnia nierówności trójkąta), jednak spełnia własności:

$$\mathrm{DTW}(x,x) = 0, \quad \mathrm{DTW}(x,y) = \mathrm{DTW}(y,x).$$

W pracy [A1] była badana transformacja szeregu czasowego x będąca dyskretną pochodną x' obliczaną:

$$x'(i) = x(i+1) - x(i), \quad i = 1,2,\ldots,n-1.$$

Na tak przetransformowanych szeregach czasowych możemy następnie obliczać odpowiednie funkcje odległości. Tworzymy parametryczną miarę odległości dist_{ab} będącą kombinacją danej odległości dist obliczanej na surowym szeregu i szeregu transformowanym, w tym przypadku na pochodnej danego szeregu.

$$\mathrm{dist}_{ab}(x,y) = a\,\mathrm{dist}(x,y) + b\,\mathrm{dist}(x',y'), \quad a,b \in [0,1].$$

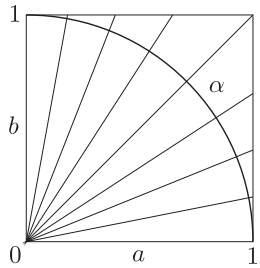
Miara odległości dist_{ab} zachowuje własności metryczne miary dist, w szczególności, jeśli dist jest metryką, to dist_{ab} także.

Odległości tej możemy następnie użyć w procesie klasyfikacji szeregów czasowych przy użyciu metody najbliższego sąsiada.

W pracy [A1] parametry a,b są zależne od pojedynczego parametru $\alpha \in [0, \frac{\pi}{2}]$.

$$a = \cos\alpha, \quad b = \sin\alpha.$$

Taki dobór parametrów pozwala zachować równe odstępy (odległości) między wartościami parametru α, co pomaga w analizie zachowania się tego parametru w procesie klasyfikacji (Rys. 2).



Rysunek 2: Zależność parametrów a, b i α.

Parametr α jest dobierany w fazie uczącej metody najbliższego sąsiada, tzn. za pomocą kroswalidacji (leave-one-out) jest wybierana wartość α, dla której błąd kroswalidacji na zbiorze uczącym jest najmniejszy. Tak dobrane α jest dalej używane w procesie klasyfikacji na zbiorze testowym.

Zatem parametr α odpowiada za udział odległości liczonej na danych surowych i na danych transformowanych, w tym przypadku odległości liczonej na pochodnej szeregu. Dla $\alpha = 0$ ($a = 1,\ b = 0$) odległość dist_{ab} jest równa odległości dist liczonej tylko na danych surowych. Wraz ze wzrostem parametru α udział danych surowych maleje, a wzrasta udział danych pochodnych, aż dla całkowitego udziału danych przetransformowanych pochodną dla $\alpha = \frac{\pi}{2}$ ($a = 0,\ b = 1$).

Konstrukcja odległości dist_{ab} jest bardzo prosta i pozwala na znaczące optymalizacje czasu obliczeń. Ponieważ parametr α (parametry a,b) znajduje się na zewnątrz odległości składowych, wystarczy raz

policzyć odległości składowe dla każdego zakresy parametru α. Ponieważ czas obliczeń samych odległości (szczególnie dla miary DTW) jest o kilka rzędów wielkości większy niż czas obliczenia kombinacji dist_α (przy obliczonych już odległościach składowych), czas obliczania dist_α praktycznie nie zależy od zakresy parametru α. Umożliwia to przyjęcie bardzo precyzyjnego (gęstego) zakresu parametru, bez wpływu na wydajność obliczeń. Gęsty zakres parametru pozwala na precyzyjniejsze dopasowanie odległości dist_α do aktualnie klasyfikowanego zbioru danych i wpływa na jakość wyników. Z drugiej strony, taka konstrukcja odległości kombinowanej pozwala w fazie uczącej (kroswalidacja) na obliczenia bez powtarzania obliczeń odległości składowych dla każdego parametru α i jednocześnie bez zapamiętywania całej macierzy odległości miedzy wszystkimi elementami zbioru danych. Możemy zatem skonstruować algorytm, który jednocześnie jest wydajny (czas obliczeń właściwie nie zależy od ilości parametrów α) i nie obciąża pamięci (nie trzeba obliczać i trzymać w pamięci całej macierzy odległości). Przykład kodu (Matlab) dla fazy uczącej (kroswalidacja) jest przedstawiony na Rys. 3. Natomiast w fazie testowej, parametr α jest juz ustalony i czas obliczeń odległości dist_α jest równy sumie czasów obliczeń odległości składowych.

```
% e - list of time series in learning data set (cell vector of vectors)
% labels - vector of labels of elements of list e
% dist - distance function
% trans - transform function

step = 0.01;
alpha = 0 : step : 1;
a = cos(alpha);
b = sin(alpha);

n = length(e);
k = length(alpha);
mistakes(1 : k) = 0; % vector of numbers of misclassified elements

for i = 1 : n
    D(1 : k) = inf;  % vector of minimal distances
    L(1 : k) = 0;    % vector of 'minimal' labels
    for j = [1 : i-1, i+1 : n] % leave-one-out
        d = a * dist(e{j}, e{i}) + b * dist(trans(e{j}), trans(e{i}));
        D(d < D) = d(d < D);
        L(d < D) = labels(j);
    end
    mistakes = mistakes + (L ~= labels(i));
end
errors = mistakes / n; % error rates for every parameter alpha
```

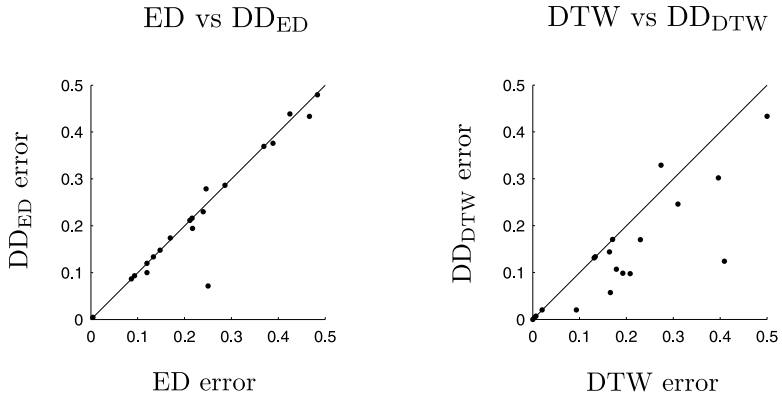
Rysunek 3: Algorytm (Matlab) — faza ucząca (kroswalidacja).

W pracy [A1] opisana kombinowana miara odległości została badana dla odległości składowych ED i DTW oraz dla pochodnej jako transformacji, czyli dla:

$$\mathrm{DD}_{\mathrm{ED}}(x,y) = \mathrm{ED}_{ab}(x,y) = a\,\mathrm{ED}(x,y) + b\,\mathrm{ED}(x',y'),$$
$$\mathrm{DD}_{\mathrm{DTW}}(x,y) = \mathrm{DTW}_{ab}(x,y) = a\,\mathrm{DTW}(x,y) + b\,\mathrm{DTW}(x',y').$$

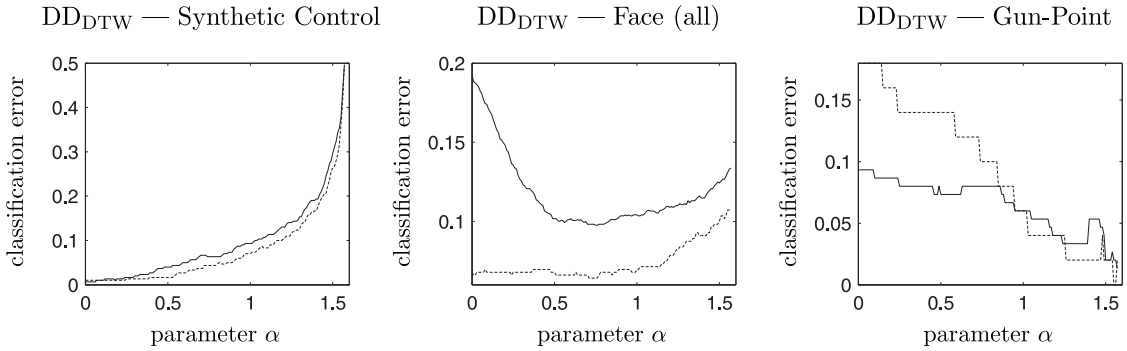
Obliczenia zostały wykonane na bazie danych UCR Time Series Classification/Clustering Homepage (`www.cs.ucr.edu/~eamonn/time_series_data/`), która jest największą wolno dostępną bazą danych jednowymiarowych rzeczywistych szeregów czasowych. Baza ta jest bardzo zróżnicowana w jej skład wchodzą zbiory danych szeregów z wielu dziedzin nauki i techniki. Każdy zbiór danych jest podzielony na cześć uczącą i testową. Baza ta jest cały czas rozwijana i powiększana, w momencie przeprowadzania badań do publikacji [A1] składała się z 20 zbiorów danych. Zostały wyliczone błędy testowe dla miar kombinowanych i miar pierwotnych. Na badanych zbiorach danych odległość kombinowana

dała jednoznacznie mniejsze błędy klasyfikacji niż odległości pierwotne. Graficzne porównanie wyników przedstawiono na Rys. 4. Każdy punkt wykresu odpowiada jednemu zbiorowi danych. Na osi X błąd klasyfikacji dla miary pierwotnej (ED, DTW), na osi Y błąd dla miary kombinowanej (ED_{ab}, DTW_{ab}). Jeżeli punkt leży poniżej przedstawionej na rysunku prostej, znaczy to, że błąd klasyfikacji miary kombinowanej jest mniejszy niż porównywanej miary pierwotnej. Im dalej od tej prostej punkt się znajduje tym większa jest różnica między błędami.



Rysunek 4: Porównanie błędów testowych.

Przebadano również zależność błędów wyliczonych na zbiorze uczącym i testowym. Przykładowe przebiegi krzywych błędu przedstawiono na Rys. 5. Dla każdego zbioru danych minimalny błąd może występować dla innej wartości parametru α. Widać dużą odpowiedniość błędów na zbiorze uczącym i testowym. Pozwala to w precyzyjny sposób dobrać parametr α, tak aby uzyskać jak najlepsze rezultaty klasyfikacji.



Rysunek 5: Zależność błędu klasyfikacji od parametru α dla miary odległości $\mathrm{DD}_{\mathrm{DTW}}$ i przykładowych trzech zbiorów danych (linia przerywana — błąd CV (zbiór uczący), linia ciągła — błąd (zbiór) testowy).

W pracy [A1] dokonano również porównań przedstawionych metod kombinowanych do innych miar odległości bazujących na pochodnej oraz dla innych definicji pochodnych. Badania wykazały, że sposób obliczania pochodnej dyskretnej ma minimalny wpływ na wyniki klasyfikacji. Pokazały też, że zaproponowane w pracy miary kombinowane dają lepsze rezultaty klasyfikacji niż inne miary bazujące na pochodnej znane z literatury.

Praca [A2]

W pracy [A2] była kontynuowana tematyka pochodnej jako transformaty w klasyfikacji szeregów czasowych. Zbadano dodatkowo wpływ drugiej pochodnej na jakość klasyfikatora kombinowanego. Zatem w pracy była badana miara odległości zależna od: danych surowych, pierwszej pochodnej, drugiej pochodnej.

Ponieważ w pracy [A1] pokazano, że sposób definicji pochodnej nie ma wpływu na jakość klasyfikacji, pierwsza i druga pochodna szeregu czasowego x została zdefiniowana w standardowy sposób:

$$x'(i) = x(i+1) - x(i), \quad i = 1, 2, \ldots, n-1;$$
$$x'' = (x')'.$$

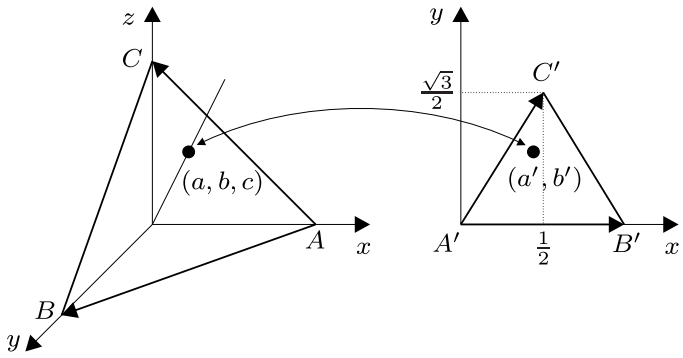
Skonstruowano ogólną kombinowaną miarę odległości:

$$\mathrm{dist}_{abc}(x, y) = a\,\mathrm{dist}(x, y) + b\,\mathrm{dist}(x', y') + c\,\mathrm{dist}(x'', y''), \quad a, b, c \in [0, 1].$$

Zatem mamy trzy parametry a, b, c, które odpowiadają za udział odpowiednich komponentów (dane surowe, pierwsza pochodna, druga pochodna) w odległości kombinowanej. Trójka parametrów (a, b, c) tworzy punkt w trójwymiarowej przestrzeni euklidesowej, a dokładnie punkt ten leży na trójkącie równobocznym o wierzchołkach $A = (1, 0, 0)$, $B = (0, 1, 0)$, $C = (0, 0, 1)$. Trójkąt ten może zostać jednoznacznie przekształcony w odpowiadający mu trójkąt na płaszczyźnie o wierzchołkach $A' = (0, 0)$, $B' = (1, 0)$, $C' = \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$ (Rys. 6). Parametrom $(a, b, c) \in \mathbb{R}^3$ odpowiadają wtedy jednoznacznie nowe parametry $(a', b') \in \mathbb{R}^2$, a zależność między nimi można wyrazić równaniami

$$(a, b, c) = A + \alpha\overrightarrow{AB} + \beta\overrightarrow{AC},$$
$$(a', b') = A' + \alpha\overrightarrow{A'B'} + \beta\overrightarrow{A'C'},$$

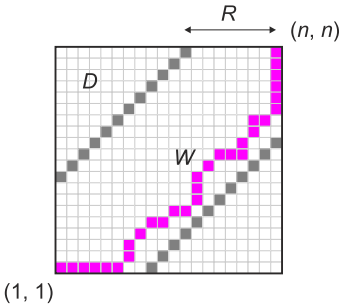
dla pewnych parametrów (współrzędnych niekartezjańskich) $\alpha, \beta \in [0, 1]$.



Rysunek 6: Zależność parametrów $a, b, c \in \mathbb{R}^3$ i parametrów $a', b' \in \mathbb{R}^2$.

Zatem możemy mówić od teraz o dwuparametrycznej odległości $\mathrm{dist}_{a'b'}$, $a', b' \in [0, 1]$.

Tak skonstruowana parametryczna kombinowana miara odległości została użyta w procesie klasyfikacji (jednowymiarowych) szeregów czasowych z wykorzystaniem metody najbliższego sąsiada i pierwotnej miary odległości DTW oraz DTWR (dist = DTW lub dist = DTWR). Miara odległości DTWR jest to zmodyfikowana miara DTW w taki sposób, że na obliczaną macierz D (Rys. 1) nakładane jest ograniczenie (okno) w postaci równoległoboku o promieniu R (Windowed Dynamic Time Warping). Takie ograniczenie powoduje zmniejszenie ilości obliczeń jednocześnie może także skutkować polepszeniem klasyfikacji (Rys. 7).



Rysunek 7: Ograniczenie macierzy D w mierze odległości DTWR.

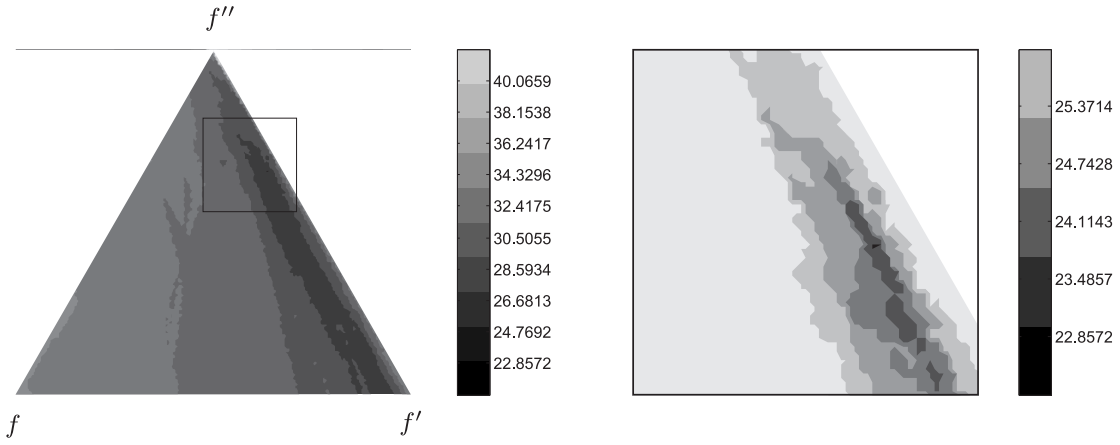
Parametry a', b' (oraz R dal DTWR) zostały dobrane na zbiorze uczącym przy pomocy kroswalidacji (leave-one-out). Podobnie jak w pracy [A1] (Rys. 3) można wykorzystać zoptymalizowany algorytm obliczania parametrów z funkcja odległości:

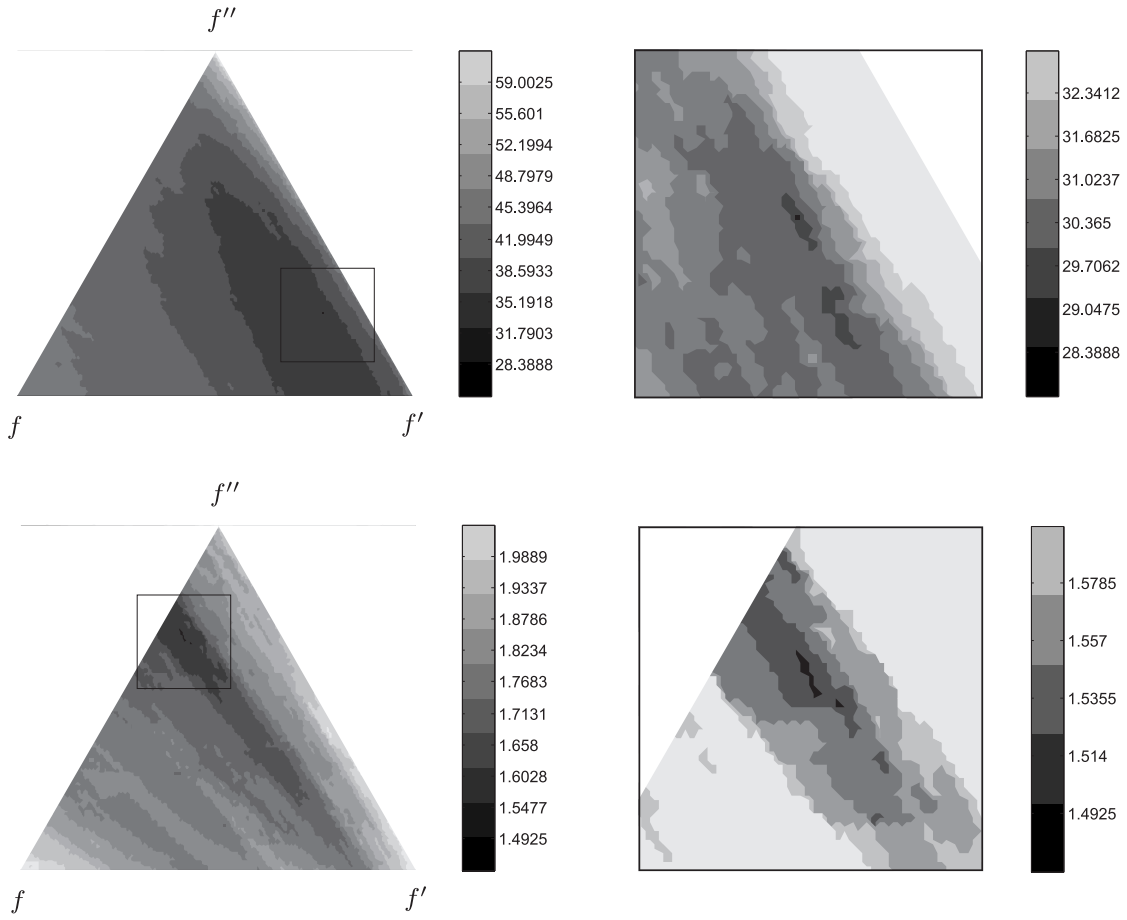
```
% e - list of time series in learning data set (cell vector of vectors)
% labels - vector of labels of elements of list e
% tr - 2d to 3d transformation
% dist - distance function (DTW, DTWR)
% trans1 - transform function (first derivative)
% trans2 - transform function (second derivative)

step = 0.01;
ap = 0 : step : 1;
bp = 0 : step : 1;
...
[a, b, c] = tr(ap, bp);
...
d = a * dist(e{j}, e{i}) +
    b * dist(trans1(e{j}), trans1(e{i})) +
    c * dist(trans2(e{j}), trans2(e{i}));
...
```

Rysunek 8: Algorytm (Matlab) — faza ucząca (kroswalidacja).

Obliczenia zostały wykonane na bazie danych UCR, która w tym czasie zawierała 47 zbiorów danych. Osobno dla odległości DTW i DTWR badane miary kombinowane zostały porównane do odległości składowych: odległość pierwotna — DTW/DTWR, odległość na pierwszej pochodnej — $\mathrm{DD_{DTW}}/\mathrm{DD_{DTWR}}$ (patrz praca [A1]). Odległość obliczana tylko na drugiej pochodnej nie była brana do porównań, gdyż jako pojedyncza odległość daje bardzo słabe wyniki klasyfikacji. Wykonane porównania wykazały, że badana odległość kombinowana daje mniejsze błędy klasyfikacji niż odległości pierwotne jak i odległości kombinowane badane w pracy [A1].

Dodatkowo przebadany został udział poszczególnych pochodnych w ostatecznym wyniku klasyfikacji. Okazuje się, że punkt parametrów $(a', b') \in \mathbb{R}^2$ dla którego otrzymujemy najniższy błąd klasyfikacji może leżeć w dowolnym miejscu trójkąta parametrów. Nawet jeżeli dla poszczególnych pochodnych (wierzchołki trójkąta) klasyfikacja daje słabe wyniki, to istnieje najczęściej taki punkt wewnątrz trójkąta parametrów, dla którego błąd klasyfikacji jest niższy (Rys. 9).





Rysunek 9: Udział poszczególnych pochodnych w błędzie klasyfikacji dla przykładowych trzech zbiorów danych i odległości DTW. Rysunek lewy — trójkąt parametrów. Rysunek prawy: powiększenie obszaru z najmniejszym błędem klasyfikacji.

Praca [A3]

Po przebadaniu wpływu pochodnej na kombinowaną miarę odległości w pracach [A1] i [A2], następny krok to wykorzystanie innych przekształceń zamiast pochodnej. Przy czym, nie chodzi tutaj o standardowe zastosowanie transformat izometrycznych (np. transformata Fouriera), które są wykorzystywane w celu zmniejszenia wymiarowości danych wejściowych, lecz wręcz przeciwnie — o wykorzystanie pewnych transformat nieizometrycznych, które nie przenoszą pełnej informacji z danych wejściowych. Pomysł jest taki, aby złożyć informację z danych surowych i transformowanych, a same transformaty mają przenosić (uwypuklać) tylko część informacji, pewien jej aspekt charakterystyczny dla danego zbioru danych szeregów czasowych.

W pracy [A3] zostały przebadane następujące transformaty szeregu czasowego x.

Transformata kosinusowa:

$$\hat{x}(k) = \sum_{i=1}^{n} x(i) \cos\left[\frac{\pi}{n}\left(i - \frac{1}{2}\right)(k-1)\right].$$

Transformata sinusowa:

$$\hat{x}(k) = \sum_{i=1}^{n} x(i) \sin\left[\frac{\pi}{n}\left(i - \frac{1}{2}\right)k\right].$$

Transformata Hilberta:

$$\hat{x}(k) = \sum_{\substack{i=1 \\ i \neq k}}^{n} \frac{x(i)}{k-i}.$$

Wszystkie te transformaty są nieizometryczne i zostały użyte w parametrycznych odległościach kombinowanych:

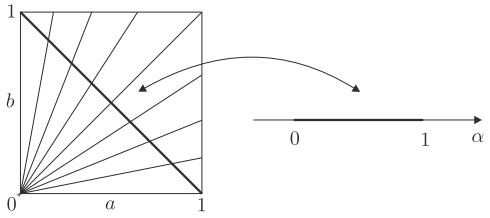
- wykorzystującej szeregi surowe i przetransformowane jedną z trzech powyższych transformat:

$$\mathrm{dist}_{ab}(x,y) = a\,\mathrm{dist}(x,y) + b\,\mathrm{dist}(\hat{x},\hat{y}). \tag{1}$$

  Wprowadzony został tutaj (podobnie jak w pracy [A1]) pojedynczy parametr $\alpha \in [0,1]$, którego zależność od parametrów a, b jest postaci:

$$a = 1-\alpha, \quad b = \alpha, \quad \alpha \in [0,1],$$

  co można przedstawić graficznie (Rys. 10).



Rysunek 10: Zależność parametrów a, b oraz α.

  Zrezygnowano tutaj z równych odstępów wartości parametrów (jak było w pracy [A1], Rys. 2), gdyż w momencie, gdy czas obliczeń właściwie nie zależy od gęstości zakresu parametru α, nie ma to znaczenia, a powyższe przekształcenie wypukłe jest prostsze i bardziej naturalne.

- wykorzystującej szeregi surowe, przetransformowane pochodną i przetransformowane jedną z trzech powyższych transformat:

$$\mathrm{dist}_{ab}(x,y) = a\,\mathrm{dist}(x,y) + b\,\mathrm{dist}(x',y') + c\,\mathrm{dist}(\hat{x},\hat{y}). \tag{2}$$

  Wprowadzone zostały tutaj (podobnie jak w pracy [A2]) niezależne parametry $a', b' \in [0,1]$ (Rys. 6).

Badania zostały wykonane na bazie UCR, która w tym czasie liczyła 47 zbiorów danych jednowymiarowych szeregów czasowych. Użyta została odległość DTW oraz przedstawione trzy transformaty: sinusowa (S), kosinusowa (C) i Hilberta (H). Zatem otrzymaliśmy następujące parametryczne kombinowane miary odległości: $\mathrm{TD}^{S}_{\mathrm{DTW}}, \mathrm{TD}^{C}_{\mathrm{DTW}}, \mathrm{TD}^{H}_{\mathrm{DTW}}$ (w przypadku złożenia szeregów surowych i transformowanych (1)) oraz $\mathrm{DTD}^{S}_{\mathrm{DTW}}, \mathrm{DTD}^{C}_{\mathrm{DTW}}, \mathrm{DTD}^{H}_{\mathrm{DTW}}$ (w przypadku złożenia danych surowych, pochodnych i transformowanych (2)).

Parametry α i a', b' zostały dobrane na podzbiorze uczącym dla każdego zbioru danych przy pomocy kroswalidacji (leave-one-out). Badania wykazały, że nowe odległości kombinowane dają mniejsze błędy klasyfikacji, zarówno porównując odległości $\mathrm{TD}^{*}_{\mathrm{DTW}}$ z odległością DTW, jak i porównując odległości $\mathrm{DTD}^{*}_{\mathrm{DTW}}$ z DTW oraz z $\mathrm{DD}_{\mathrm{DTW}}$ (praca [A1]).

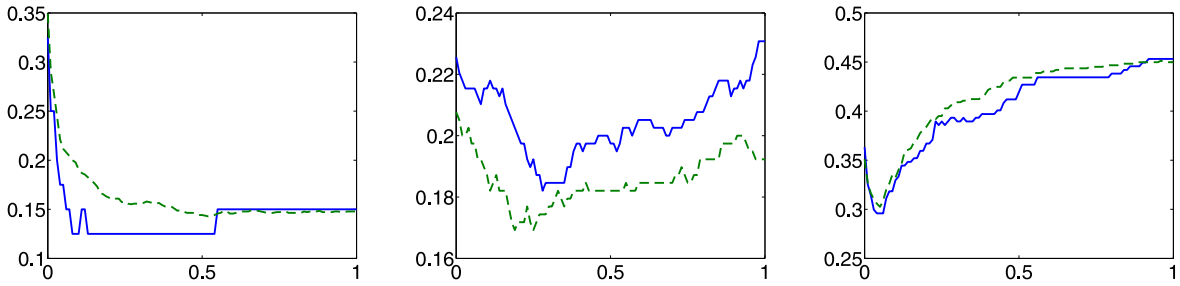
Została też przeprowadzona szczegółowa analiza statystyczna otrzymanych wyników. Do porównania błędów klasyfikacji porównywanych metod wykorzystano test post hoc Bergmanna-Hommela, który przydziel rangi poszczególnym miarom odległości i może być przedstawiony graficznie w postaci grup, do których należą poszczególne badane odległości (Tab. 1). Metody należące do tej samej grupy nie są rozróżnialne statystycznie.

Tabela 1: Wyniki testu Bergmanna-Hommela.

Procedure	Ranks mean			
$\mathrm{DTD}^{\mathrm{C}}_{\mathrm{DTW}}$	2.33	*		
$\mathrm{DTD}^{\mathrm{S}}_{\mathrm{DTW}}$	2.54	*	*	
$\mathrm{DTD}^{\mathrm{H}}_{\mathrm{DTW}}$	2.70	*	*	
$\mathrm{DD}_{\mathrm{DTW}}$	3.13		*	
DTW	4.30			*

Widać tutaj, że najlepszą odległością jest $\mathrm{DTD}^{C}_{\mathrm{DTW}}$, a najgorszą standardowe DTW.

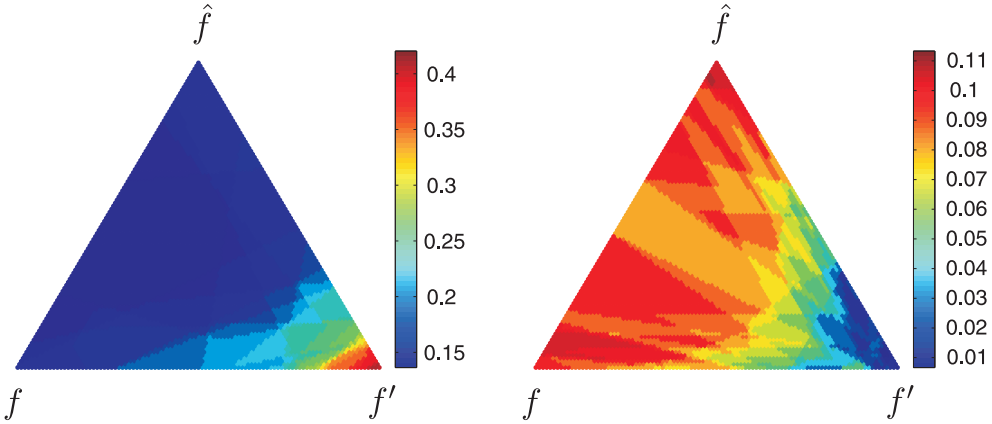
Została też przeprowadzona analiza przebiegu błędów w zależności od wartości parametrów na zbiorze uczącym i testowym. Na Rys. 11 przedstawione są przebiegi błędów dla odległości $\mathrm{TD}^{*}_{\mathrm{DTW}}$ dla przykładowych zbiorów danych.

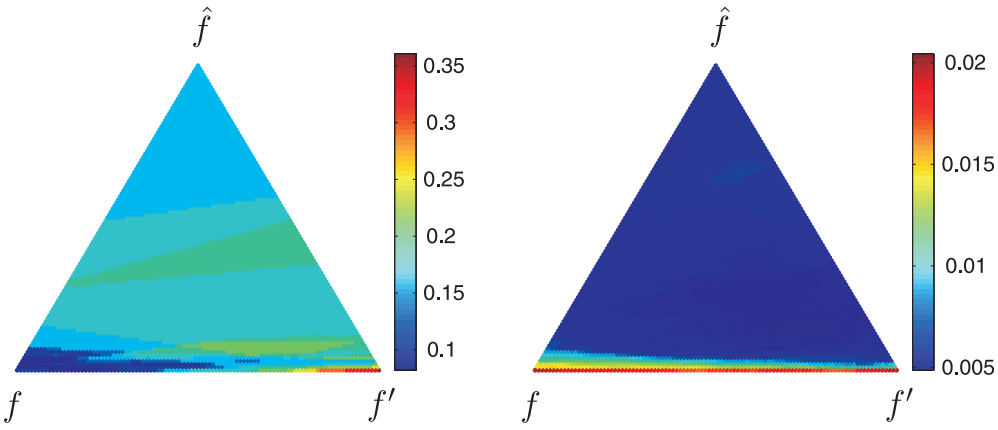


Rysunek 11: Zależność błędu klasyfikacji (parametr α) dla wybranych miar odległości $\mathrm{TD}^{*}_{\mathrm{DTW}}$ na przykładowych trzech zbiorach danych (linia przerywana — błąd (CV) na zbiorze uczącym, linia ciągła — błąd na zbiorze testowym).

Widać dużą współbieżność wykresów błędów na zbiorze uczącym i testowym oraz to, że minima błędów mogą występować w różnych punktach przedziału parametrów w zależności od przetwarzanego zbioru danych. Nawet jeśli same odległości składowe wypadają słabo (dla $\alpha = 0$, $\alpha = 1$) można znaleźć wartość parametru (wewnątrz przedziału), dla którego błąd klasyfikacji jest lepszy.

Także w przypadku odległości $\mathrm{DTD}^{*}_{\mathrm{DTW}}$ mozna zobaczyć, że najniższe wartości błędu mogą występować w bardzo różnych punktach trójkąta parametrów (Rys. 12).





Rysunek 12: Zależność błędu (testowego) klasyfikacji od udziału poszczególnych transformacji (parametry a', b') dla wybranych miar odległości $\mathrm{DTD}^*_{\mathrm{DTW}}$ na przykładowych czterech zbiorach danych.

Praca [A4]

W dotychczas opisanych pracach badane były kombinowane miary odległości w klasyfikacji jednowymiarowych szeregów czasowych. W pracy [A4] przebadano możliwość uogólnienia tego podejścia na przypadek wielowymiarowych szeregów czasowych, w szczególności chodzi tutaj o kombinację szeregów surowych i ich pochodnych.

Będziemy przyjmować, że wielowymiarowy szereg czasowy (multidimensional/multivariate time series) X jest ciągiem $m \in \mathbb{N}$ szeregów jednowymiarowych:

$$X = (x_1, x_2, \ldots, x_m),$$

gdzie $x_1, \ldots, x_m$ są jednowymiarowymi szeregami czasowymi o wspólnej długości $n \in \mathbb{N}$. Możemy zatem szereg wielowymiarowy przedstawić jako jednowymiarową trajektorię w m-wymiarowej przestrzeni euklidesowej:

$$X = \{X(i) = (x_1(i), x_2(i), \ldots, x_m(i)) \in \mathbb{R}^m \colon i = 1, 2, \ldots, n\}.$$

Przy takiej notacji możemy zdefiniować miarę odległości DTW miedzy dwoma szeregami wielowymiarowymi X i Y tak samo jak dla szeregów jednowymiarowych, tylko z lokalną funkcją kosztu postaci

$$d\big(X(i), Y(j)\big) = \sum_{k=1}^{m} \big(x_k(i) - y_k(j)\big)^2,$$

to znaczy, jako kwadrat odległości euklidesowej miedzy punktami $X(i)$ i $Y(j)$.

Natomiast pochodną wielowymiarowego szeregu czasowego definiujemy w naturalny sposób jako

$$X' = (x'_1, x'_2 \ldots, x'_m).$$

Nową parametryczną miarę odległości między wielowymiarowymi szeregami czasowymi definiujemy jako:

$$\mathrm{dist}_{ab}(X, Y) = a\,\mathrm{dist}(X, Y) + b\,\mathrm{dist}(X', Y'), \quad a, b \in [0, 1].$$

Wielowymiarowa miara odległości dist_{ab} zachowuje własności metryczne pierwotnej miary dist. W szczególności jeśli dist jest metryką, to także dist_{ab} jest metryką. W pracy została przebadana powyższa miara odległości dla dist = DTW i przedstawiona jako kombinacja wypukła (zgodnie z optymalizacjami parametrów przedstawionymi w poprzednich pracach):

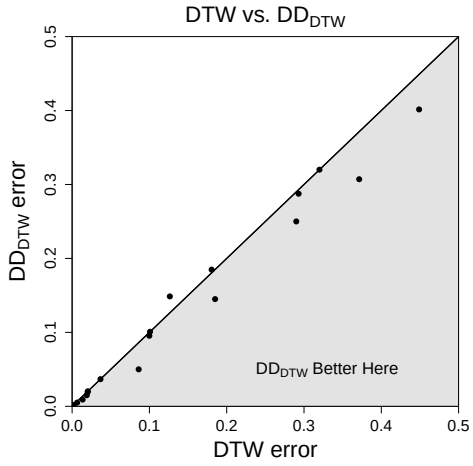
$$\mathrm{DD}_{\mathrm{DTW}}(X, Y) = (1 - \alpha)\,\mathrm{DTW}(X, Y) + \alpha\,\mathrm{DTW}(X', Y'), \quad \alpha \in [0, 1].$$

Ponieważ odległość DTW (zarówno jedno, jak i wielowymiarowa) nie jest metryką, nie jest metryką również wielowymiarowa odległość $\mathrm{DD_{DTW}}$, nie spełnia bowiem nierówności trójkąta. Natomiast zachowane są własności:

$$\mathrm{DD_{DTW}}(X, X) = 0, \quad \mathrm{DD_{DTW}}(X, Y) = \mathrm{DD_{DTW}}(Y, X).$$

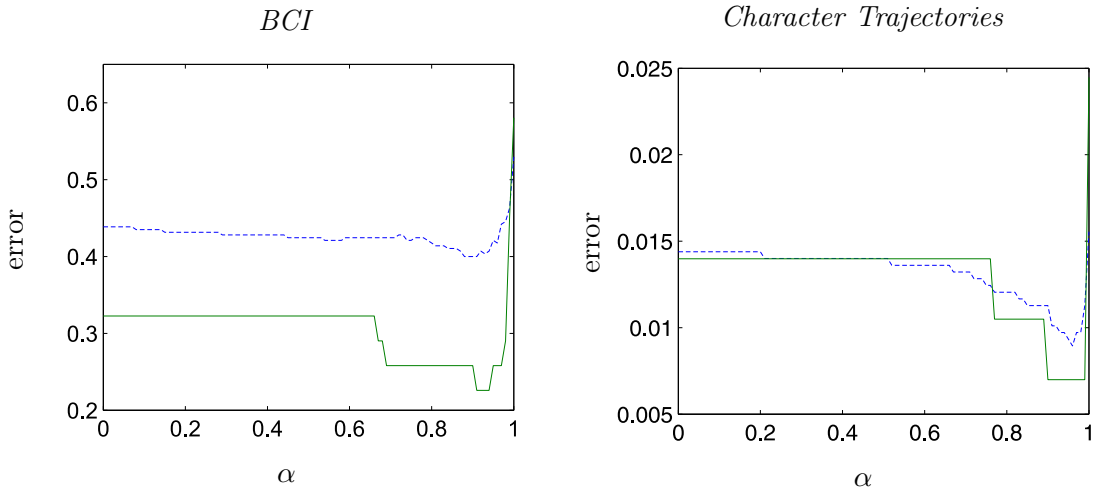
Tak zdefiniowana miara odległości została użyta w procesie klasyfikacji wielowymiarowych szeregów czasowych z użyciem metody najbliższego sąsiada. Eksperymenty obliczeniowe zostały przeprowadzone na 18 wielowymiarowych zbiorach danych znanych z literatury. Parametr $\alpha \in [0, 1]$ był dobierany na podzbiorze uczącym badanego zbioru danych w procesie kroswalidacji (leave-one-out).

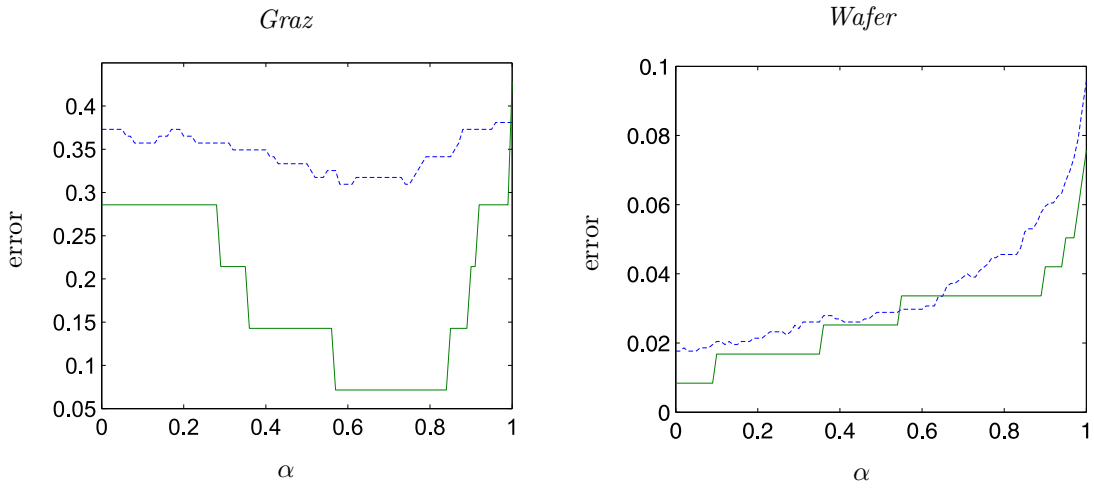
Obliczenia wykazały, że dla prawie wszystkich zbiorów danych błąd klasyfikacji metody kombinowanej $\mathrm{DD_{DTW}}$ jest mniejszy (lub równy) niż dla wyjściowej odległości DTW (Rys. 13).



Rysunek 13: Porównanie błędów klasyfikacji dla wielowymiarowych miar odległości DTW i $\mathrm{DD_{DTW}}$.

Zbadano również przebiegi błędów na zborze uczącym i testowym. Wykazują one bardzo duże podobieństwo i podobnie jak w przypadku jednowymiarowym, minimalny błąd może być osiągany dla różnych wartości parametru α w zależności od badanego zbioru danych (Rys. 14).





Rysunek 14: Zależność błędów klasyfikacji od parametru α dla przykładowych zbiorów danych. Linia przerywana — błąd (CV) na zbiorze uczącym, linia ciągła — błąd na zbiorze testowym.

Praca [A5]

Dotychczas opisane prace dotyczyły klasyfikacji (pod nadzorem) szeregów czasowych. Kolejnym krokiem w badaniach przedstawionych kombinowanych miar odległości jest sprawdzenie, czy analogiczna metoda będzie działać w przypadku analizy skupień (klasyfikacji bez nadzoru). W przypadku analizy skupień nie dysponujemy zbiorem uczącym, informacja o etykietach szeregów nie jest znana w fazie uczącej. Ponieważ w badanych odległościach występują parametry, które muszą być dobrane w czasie uczenia, musimy tutaj zastosować inne podejście niż w przypadku klasyfikacji. Standardowo wykorzystuje sie w tym przypadku tak zwane miary wewnętrzne (internal measures), które pozwalają określić w pewien sposób jakość dokonanego podziału na skupienia bez znajomości etykiet w danym zbiorze. Natomiast po dobraniu parametrów, w fazie testowej można już wykorzystać miary określające jakość analizy skupień korzystające z etykiet szeregów w danych zbiorach danych, tak zwanych miar zewnętrznych (external measures).

Okazuje się, że proste przeniesienie badanej w poprzednich pracach metody na przypadek analizy skupień nie daje dobrych rezultatów. Wymagany jest tutaj dodatkowy pośredni krok, który koryguje odpowiednie przebiegi błędów uczących (miar wewnętrznych).

W pracy [A5] badana była hierarchiczna metoda analizy skupień ze uśrednioną funkcją odległości między skupieniami (agglomerative hierarchical clustering with average linkage) ze znaną ilością skupień k:

$$\operatorname{dist}(A,B) = \frac{1}{|A||B|} \sum_{x\in A} \sum_{y\in B} \operatorname{dist}(x,y). \tag{3}$$

W tej metodzie, mając n-elementowy zbiór danych, rozpoczynamy z n skupieniami — każdy element (szereg) jest osobnym skupieniem. Następnie, w każdym kroku, najbliższe dwa skupienia (mające najmniejszą odległość (3)) są łączone (suma zbiorów) w jedno skupienie wyższego poziomu. Proces łączenie skupień kończymy, gdy dojdziemy do z góry ustalonej ilości skupień k.

W pracy [A5] jako odległość dist była użyta kombinowana parametryczna miara odległości:

$$\mathrm{DD}_{\mathrm{DTW}}(x,y) = (1-\alpha)\,\mathrm{DTW}(x,y) + \alpha\,\mathrm{DTW}(x',y'), \quad \alpha \in [0,1]$$

będąca kombinacja wypukłą miary DTW obliczanej na szeregach surowych (x, y) i przetransformowanych pochodną (x', y'). Parametr α był dobierany w fazie uczącej analizy skupień, to znaczy, z wykorzystaniem miar wewnętrznych. Przebadane zostały następujące miary wewnętrzne:

Wariancja wewnątrzgrupowa (Intra-group Variance – V):

$$\frac{1}{n-k}\sum_{i=1}^{k}\sum_{x\in C_i}\mathrm{dist}(x,c_i) \tag{4}$$

Indeks Calińskiego–Harabasza (Calinski–Harabasz Index — CH):

$$\frac{n-k}{n-1}\frac{\sum_{i=1}^{k} n_i \mathrm{dist}^2(c_i,c)}{\sum_{i=1}^{k}\sum_{x\in C_i}\mathrm{dist}^2(x,c_i)} \tag{5}$$

Zmodyfikowana statystyka Γ (Modified Hubert Γ statistic — Γ):

$$\frac{2}{n(n-1)}\sum_{x\in D}\sum_{y\in D}\mathrm{dist}(x,y)\mathrm{dist}(c_i,c_j)_{x\in C_i, y\in C_j}, \tag{6}$$

gdzie: D — zbiór danych, n — liczba elementów (szeregów) w zbiorze D, k — ustalona liczba skupień, C_i — i-te skupienie (podzbiór zbioru D, n_i — liczba elementów (szeregów) w skupieniu C_i, c_i — centrum skupienia C_i, c — centrum zbioru danych D.
Za centrum zbioru będziemy przyjmować medoid zbioru, to znaczy, element (szereg) x_j zbioru S, którego średnia odległość od pozostałych elementów zbioru jest najmniejsza:

$$x_j \in S\colon \min \mathrm{ind}_j \sum_{x\in S}\mathrm{dist}(x_j,x).$$

Natomiast w fazie testowej wykorzystane zostały następujące miary zewnętrzne:
(Purity — P):

$$\frac{1}{n}\sum_{i}\max_{j} n_{ij} \tag{7}$$

Entropia (Entropy — E):

$$-\sum_{j}\frac{n_j}{n}\sum_{i=1}^{k}\frac{n_{ij}}{n_j}\log_2\left(\frac{n_{ij}}{n_j}\right), \tag{8}$$

gdzie: n — liczba elementów (szeregów) w zbiorze danych, n_j — liczba elementów w skupieniu C_j, n_{ij} — liczba elementów z klasy i w skupieniu C_j.

Statystyka R (Rand statistic — R):

$$\frac{TP+TN}{TP+FP+FN+TN}, \tag{9}$$

gdzie następujące symbole oznaczają liczbę elementów w zbiorach par:

TP (true positive) — elementy należące do tego samego skupienia, które są w tej samej klasie;

TN (true negative) — elementy należące do różnych skupień, które są w różnych klasach;

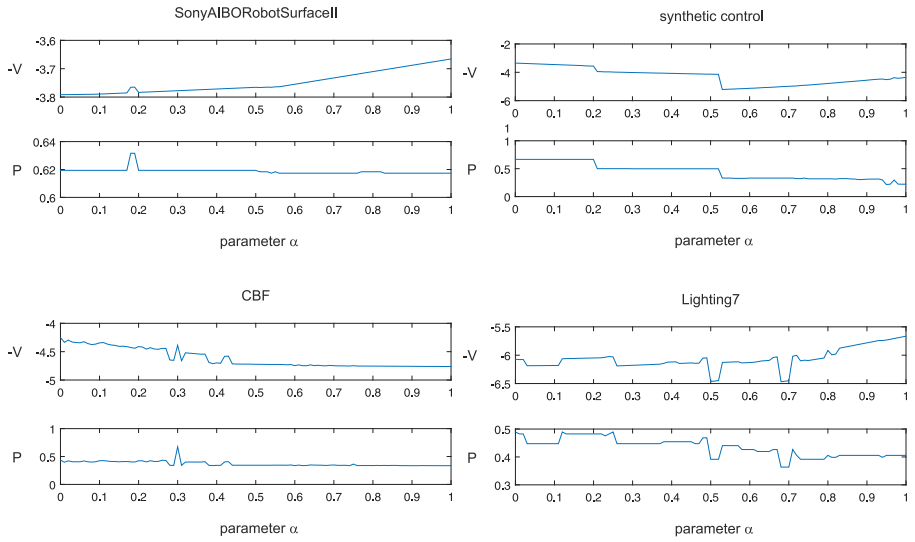
FP (false positive) — elementy należące do różnych skupień, które są w tej samej klasie;

FN (false negative) — elementy należące do tego samego skupienia, które są w różnych klasach.

Ponieważ Wariancja wewnątrzgrupowa i Entropia daje tym lepszy podział na skupienia im mniejszą ma wartość, będziemy używać ich negacji (-V, -E) dla spójności z innymi miarami, które wskazują tym lepszy podział na skupienia im większa jest ich wartość. Zatem dla wszystkich użytych miar (wewnętrznych i zewnętrznych) większa wartość miary sygnalizuje lepszy (bliższy rzeczywistemu) podział na skupienia.

Parametr α będzie dobierany prawidłowo (będzie dawać najlepsze wyniki podziału na skupienia na zbiorze testowym) jeśli wykresy miary wewnętrznej i zewnętrznej będą podobne, w szczególności chodzi o odpowiedniość maksimów tych wykresów, gdyż wybieramy ten parametr α, dla którego wartość miary wewnętrznej jest największa, a jakość podziału na skupienia jest najlepsza, gdy wartość miary

zewnętrznej jest największa. Na Rys. 15 przedstawiono przebieg przykładowej pary miar wewnętrznej i zewnętrznej względem parametru α na przykładowych zbiorach danych. Widać wyraźnie, że maksima miary wewnętrznej nie odpowiadają (nawet w przybliżeniu) maksimom miary zewnętrznej. Zatem jakość podziału na skupienia przy bezpośrednim zastosowaniu opisywanej metody jest bardzo słaba. Bezpośrednie przeniesienie metody z klasyfikacji pod nadzorem do klasyfikacji bez nadzory nie daje dobrych rezultatów.



Rysunek 15: Porównanie wykresów miary wewnętrznej (-V) i zewnętrznej (P) dla przykładowych zbiorów danych.

Jednakże, patrząc na te wykresy, ludzkie oko natychmiast wychwyci, że wykresy miary wewnętrznej i zewnętrznej są bardzo podobne tylko w pewien sposób zniekształcone. Widać, że wykres miary wewnętrznej jest niejako „wygiętą" wersją miary zewnętrznej. Widać też, że to odkształcenie jest dla każdego zbioru danych inne.

Przyczyną zniekształceń wykresu miary wewnętrznej mogą być różne zakresy wartości miary DTW na danych surowych i transformowanych, tzn. różne wartości średnie ich macierzy odległości. Jednakże wykonane eksperymenty (normalizacja macierzy odległości) nie potwierdziły tego. Po normalizacji wykresy miary wewnętrznej i zewnętrznej wciąż są odkształcone w tym samym stopniu jak bez normalizacji. Wydaje się, że przyczyną odkształceń może być nie brak globalnej normalizacji (całej macierzy odległości), lecz brak normalizacji lokalnej, czyli różnic w średnich wartościach pewnych podzbiorów macierzy odległości, które mają największy wpływ na kształt przebiegu miar wewnętrznych. Dla każdego zbioru danych ten podzbiór może być inny i nie jesteśmy w stanie go zidentyfikować w fazie (uczącej) doboru parametru α, gdy nie znamy wartości miary zewnętrznej.

Ponieważ nie jesteśmy w stanie wykryć i usunąć przyczyny zniekształcenia miar wewnętrznych możemy spróbować „leczenia objawowego". Skonstruujemy pewną procedurę ad hoc, która zlikwiduje zniekształcenia — przekształci tak wykres miary wewnętrznej by odpowiadał wykresowi miary zewnętrznej (w szczególności by odpowiadały sobie odpowiednie maksima). Przeprowadzono wiele eksperymentów by taką procedurę znaleźć. Ostatecznie najlepszy rezultat dała procedura, która wpierw usuwa trend kwadratowy z przebiegu miary wewnętrznej (QTE - quadratic trend extraction), następnie, tak przekształcony przebieg normalizuje na zakres $[0, 1]$ i na koniec eliminuje małe ruchy wykresu (skoki), przenosząc tylko duże zmiany w sygnale, większe niż pewna wartość $a \in [0, 1]$. Cały algorytm korekcyjny został przedstawiony jako Alg. 1.

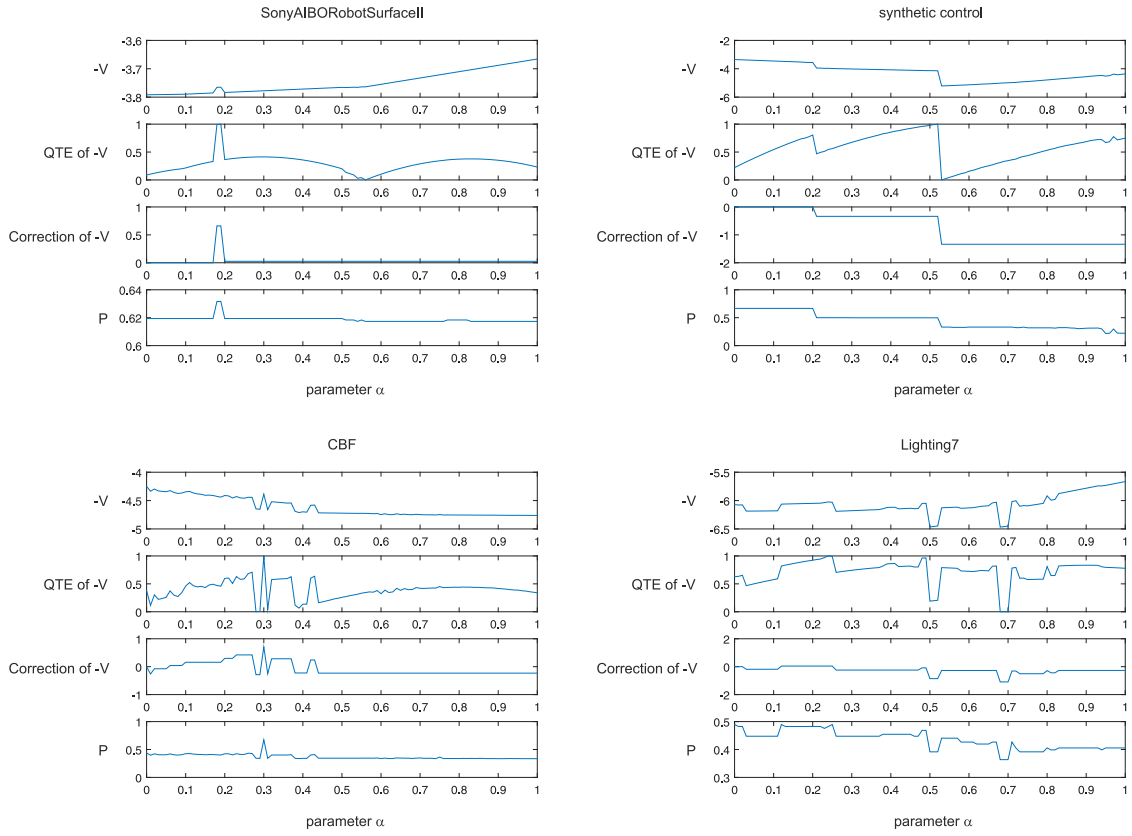
Algorithm 1 Correction algorithm

Require: $X = (X_1, X_2, \ldots, X_n)$ — internal measure, $a \in [0, 1]$ — step parameter
Ensure: $Y = (Y_1, Y_2, \ldots, Y_n)$ — corrected internal measure

```
extract quadratic trend from X (QTE)
normalize X to [0, 1] range
Y_1 ← 0
for i ← 2, 3, ..., n do
    w ← X_i − X_{i−1}
    if |w| < a then
        Y_i ← Y_{i−1}
    else
        Y_i ← Y_{i−1} + w
    end if
end for
return Y
```

Algorytm korekcyjny jest mało wrażliwy na wartość parametru kroku $a \in [0, 1]$. Uzyskano bardzo podobne rezultatu dla każdego a z zakresu $[0.025, 0.2]$. Ostatecznie ten parametr został ustalony na wartość $a = 0.1$.

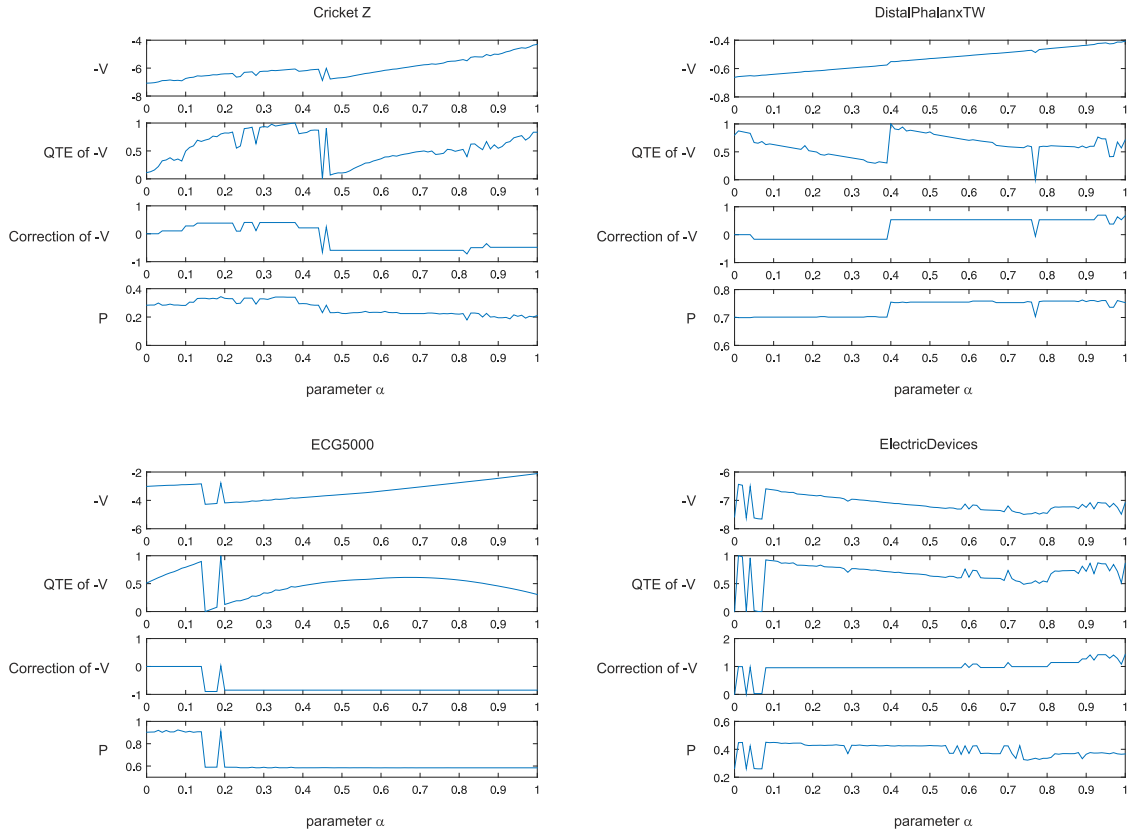
Na Rys. 16 i 17 przedstawione zostały odpowiednio przebiegi (od góry): miary wewnętrznej, miary wewnętrznej po usunięciu trendu kwadratowego, całkowicie skorygowanej miary wewnętrznej i miary zewnętrznej.



Rysunek 16: Wykres miary wewnętrznej (-V), przebieg po usunięciu trendu kwadratowego (QTE), przebieg po ostatecznym zastosowaniu algorytmu korekcyjnego i porównawczy wykres miary zewnętrznej (P).

Po zastosowaniu algorytmu korekcyjnego widać, że odpowiedniość przebiegów skorygowanej miary wewnętrznej i miary zewnętrznej jest bardzo dobra. W większości przypadków zarówno położenie maksimów i ogólny kształt przebiegów jest bardzo podobny, zachowana jest monotoniczność przebiegów i pozycje lokalnych maksimów (i minimów). Widać, że można teraz zlokalizować (ze znaczną precyzją)

maksima miary zewnętrznej znajdując maksima skorygowanej miary wewnętrznej. Ostatecznie, w prezentowanej metodzie analizy skupień, optymalna wartość parametru α jest dobierana nie na samym przebiegu miary wewnętrznej, lecz na jej przebiegu przekształconym algorytmem korekcyjnym (Alg. 1). Algorytm korekcyjny zachowuje się bardzo podobnie dla wszystkich badanych miar wewnętrznych i zewnętrznych.



Rysunek 17: Wykres miary wewnętrznej (-V), przebieg po usunięciu trendu kwadratowego (QTE), przebieg po ostatecznym zastosowaniu algorytmu korekcyjnego i porównawczy wykres miary zewnętrznej (P).

Eksperymenty zostały przeprowadzone na zbiorach danych pochodzących z bazy UCR, która w tamtym okresie (lato 2015) składała się z 84 zbiorów danych. Każdy ze zbiorów w bazie jest podzielony na podzbiór uczący i testowy. W celu badania przedstawionych algorytmów analizy skupień dla każdego zbioru danych jego podzbiór uczący i testowy zostały połączone w jeden duży zbiór. Krok parametru α został ustalony na 0.01. Parametr a algorytmu korekcyjnego został ustalony na 0.1. Procedura analizy skupień i inne procedury w tym usunięcie trendu kwadratowego i sam algorytm korekcyjny zostały zaimplementowane z wykorzystaniem odpowiednich funkcji programu Matlab.

W Tab. 2 przedstawione są średnie (po wszystkich zbiorach danych) wartości wykorzystywanych miar zewnętrznych dla optymalnej wartości parametru α (wybieranego na skorygowanych miarach wewnętrznych) dla porównywanych miar odległości DTW, DTW na pochodnej (DDTW) i parametrycznej kombinowanej mierze $\mathrm{DD}_{\mathrm{DTW}}$. Widać, że dla każdej kombinacji miary wewnętrznej i zewnętrznej miara odległości $\mathrm{DD}_{\mathrm{DTW}}$ daje lepsze wyniki niż porównywane miary DTW i DDTW (im wyższa wartość tym lepszy rezultat podziału na skupienia).

Tabela 2: Porównanie badanych miar odległości — średnie wartości miar zewnętrznych dla odpowiednich miar wewnętrznych.

		external: P			-E			R		
		DTW	DDTW	DD_{DTW}	DTW	DDTW	DD_{DTW}	DTW	DDTW	DD_{DTW}
internal:	-V			0.528			-1.442			0.606
	CH	0.512	0.447	0.527	-1.492	-1.776	-1.452	0.580	0.451	0.602
	Γ			0.524			-1.461			0.603

Przeprowadzono także porównanie statystyczne badanych metod. Porównano tylko odległości DTW i DD_{DTW}, gdyż odległość DDTW daje zawsze dużo gorsze rezultaty niż pozostałe. Do porównania użyto, rekomendowany przy porównaniu dwóch metod klasyfikacji/analizy skupień, test Wilcoxona. Ponieważ w tym przypadku nie porównujemy zewnętrznych miar każdej z każdą (interesuje nas tylko, czy dla ustalonej pary miar wewnętrznej i zewnętrznej podział na skupienia jest statystycznie różny (lepszy) dla DD_{DTW}) nie musimy stosować poprawki na testowanie wielokrotne. W Tab. 3 przedstawione są p-wartości testu Wilcoxona dla porównywanych miar odległości (wartości miar zewnętrznych). Widać, że p-wartości są bardzo małe. Wszystkie są poniżej standardowo przyjmowanej wartości 5%, gdzie większość z nich jest dużo mniejsza. Ostatecznie potwierdza to, że badana miara odległości DD_{DTW} daje statystycznie lepsze rezultaty analizy skupień niż porównywana miara DTW.

Tabela 3: Statystyczne porównanie miar odległości DTW i DD_{DTW} — p-wartości testu Wilcoxona dla wszystkich par miar wewnętrznych i zewnętrznych.

		external: P	-E	R
internal:	-V	0.00092	0.00017	0.00005
	CH	0.00341	0.00431	0.00619
	Γ	0.02246	0.03741	0.00688

Praca [A6]

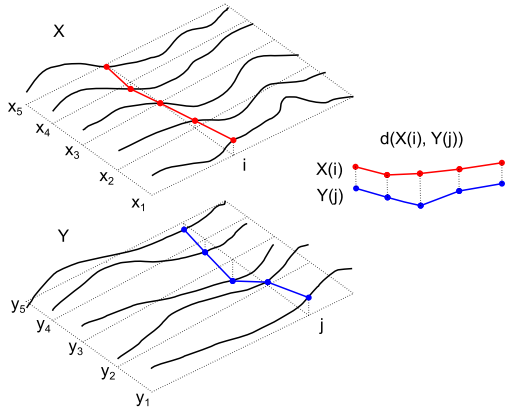
W poprzednich pracach jako transformacja danych surowych bardzo często występowała pochodna. Pokazywało to, że część informacji na temat szeregów w danym zbiorze danych uwidacznia się dopiero po zróżniczkowaniu wyjściowego szeregu czasowego. Wydaje się, że może czasami występować sytuacja odwrotna. Mianowicie, że wejściowe (surowe) szeregi czasowe występują już w pewnym stopniu w formie zróżniczkowanej i aby wydobyć z nich jakąś dodatkową informację należy odwrócić działanie pochodnej. Operacją odwrotną do różniczkowania jest całkowanie. Oczywiście, podobnie jak w przypadku pochodnej, miara odległości działająca tylko na danych scałkowanych nie może dać dobrych wyników w klasyfikacji. Należy znów skonstruować parametryczna miarę kombinowaną uwzględniającą zarówno dane surowe jak i scałkowane.

Taka kombinowana miara odległości została skonstruowana i przebadana w pracy [6]. Została ona wykorzystana w procesie klasyfikacji zarówno jedno, jak i wielowymiarowych szeregów czasowych. Pokazano tam, że transformacja w postaci całki (dyskretnej) daje pewną dodatkową informację, która pozwala znacząco zmniejszyć błąd klasyfikacji.

W pracy jako odległość bazowa została użyta miara odległościowa DTW, zarówno dla szeregów jednowymiarowych i wielowymiarowych. Podobnie jak w pracy [4], w przypadku wielowymiarowym użyto DTW ze specjalną funkcją kosztu d, zdefiniowaną dla dwóch multi-szeregów X, Y jako

$$d(X(i), Y(j)) = \sum_{k=1}^{k=m} (x_k(i) - y_k(j))^2, \tag{10}$$

to znaczy, jako kwadrat odległości euklidesowej dwóch m-wymiarowych wektorów utworzonych z wartości wzdłuż wymiarów multi-szeregu na pozycji i i j (Rys. 18).



Rysunek 18: Wyrównanie wielowymiarowych szeregów czasowych i funkcja kosztu d.

Dla jednowymiarowego szeregu czasowego x definiujemy przekształcenie całkowe I (dyskretna całka nieoznaczona), dające w wyniku nowy szereg czasowy $y = \mathrm{I}(x)$, w postaci sumy skumulowanej:

$$\begin{aligned} y(1) &= x(1), \\ y(i) &= y(i-1) + x(i), \quad i = 2, 3, \ldots, n \end{aligned}$$

lub

$$y(i) = \sum_{j=1}^{i} x(j), \quad i = 1, 2, \ldots, n. \tag{11}$$

W przypadku szeregów wielowymiarowych dokonujemy całkowania każdego z wymiarów osobno:

$$\mathrm{I}(X) = (\mathrm{I}(x_1), \mathrm{I}(x_2), \ldots, \mathrm{I}(x_m)).$$

Możemy wtedy zdefiniować miarę odległości działającą na danych scałkowanych, zarówno dla jedno i wielowymiarowych szeregów a, b postaci

$$\mathrm{IDTW}(a, b) = \mathrm{DTW}(\mathrm{I}(a), \mathrm{I}(b))$$

oraz parametryczną miarę odległościową jako kombinacje wypukłą

$$\mathrm{ID}_{\mathrm{DTW}}(a, b) = (1 - \alpha)\ \mathrm{DTW}(a, b) + \alpha\ \mathrm{IDTW}(a, b), \tag{12}$$

gdzie α jest parametrem sterującym udziałem odległości na danych surowych i scałkowanych o zakresie $\alpha \in [0, 1]$.

Tak skonstruowana miara odległości została użyta w procesie klasyfikacji jedno i wielowymiarowych szeregów czasowych z użyciem metody najbliższego sąsiada. Parametr α jest dobierany w fazie uczącej poprzez kroswalidację (leave-one-out) zbioru uczącego.

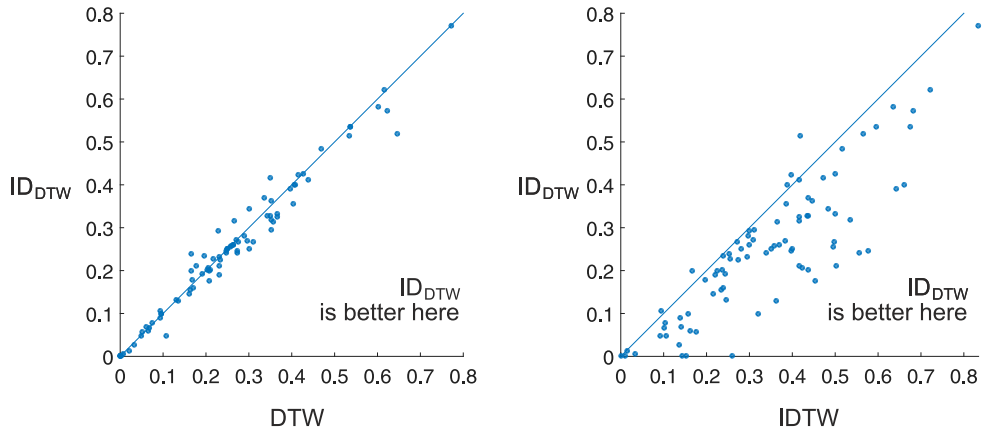
Miary IDTW i $\mathrm{ID}_{\mathrm{DTW}}$ mają takie same własności metryczne jak DTW, tzn. nie są metrykami (nie spełniają nierówności trójkata) jednak prawdziwe są własności

$$\begin{aligned} \mathrm{IDTW}(a, a) &= 0, \quad \mathrm{IDTW}(a, b) = \mathrm{IDTW}(b, a), \\ \mathrm{ID}_{\mathrm{DTW}}(a, a) &= 0, \quad \mathrm{ID}_{\mathrm{DTW}}(a, b) = \mathrm{ID}_{\mathrm{DTW}}(b, a). \end{aligned}$$

Eksperymenty obliczeniowe zostały przeprowadzone zarówno na jedno i wielowymiarowych szeregach czasowych. W przypadku jednowymiarowym użyta została baza danych UCR, która liczyła

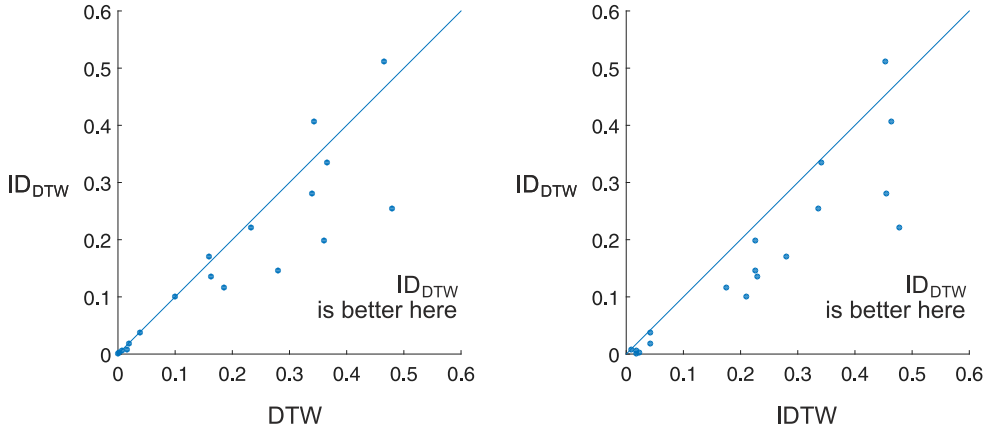
85 zbiorów danych. W przypadku wielowymiarowym wykorzystano 16 rzeczywistych zbiorów danych pochodzących z wielu dziedzin: medycyny, robotyki, rozpoznawania pisma etc.

Porównanie błędów klasyfikacji dla jednowymiarowych szeregów przedstawiono na Rys. 19.



Rysunek 19: Graficzne porównanie błędów klasyfikacji: DTW vs $\mathrm{ID}_{\mathrm{DTW}}$ i IDTW vs $\mathrm{ID}_{\mathrm{DTW}}$ dla jednowymiarowych szeregów czasowych.

Natomiast porównanie błędów klasyfikacji dla szeregów wielowymiarowych możemy obejrzeć na Rys. 20.



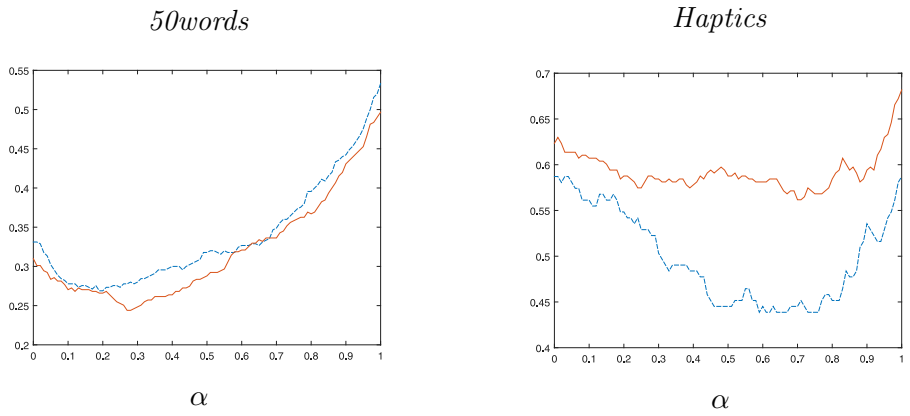
Rysunek 20: Graficzne porównanie błędów klasyfikacji: DTW vs $\mathrm{ID}_{\mathrm{DTW}}$ i IDTW vs $\mathrm{ID}_{\mathrm{DTW}}$ dla wielowymiarowych szeregów czasowych.

Widać, że w obu przypadkach (jedno i wielowymiarowym) błędy klasyfikacji metody kombinowanej $\mathrm{ID}_{\mathrm{DTW}}$ są dla znakomitej większości zbiorów danych niższe niż dla odległości składowych. W celu potwierdzenia tych wyników dokonano także analizy statystycznej za pomocą testu Wilcoxona. Odpowiednie p-wartości są przedstawione w Tab. 4.

Tabela 4: Statystyczne porównanie badanych metod — p-wartości testu Wilcoxona.

		p-value
univariate	$\mathrm{ID}_{\mathrm{DTW}}$ vs DTW	0.0272
	$\mathrm{ID}_{\mathrm{DTW}}$ vs IDTW	5.6×10^{-14}
multivariate	$\mathrm{ID}_{\mathrm{DTW}}$ vs DTW	0.0703
	$\mathrm{ID}_{\mathrm{DTW}}$ vs IDTW	0.0012

Przeprowadzono też analizę przebiegów wykresów błędów na zbiorze uczącym i testowym w zależności od wartości parametru α. Na Rys. 21 przedstawione są przebiegi błędów kroswalidacji (na zbiorze uczącym) i testowego (na zbiorze testowym) dla przykładowych zbiorów danych.



Rysunek 21: Przebiegi błędu uczącego (CV) i testowego względem parametru α dla miary $\mathrm{ID}_{\mathrm{DTW}}$ na przykładowych jednowymiarowych zbiorach danych. Linia przerywana — błąd na podzbiorze uczącym; linia ciągła — błąd na zbiorze testowym.

Widać wzajemną odpowiedniość błędów, bardzo podobną monotoniczność i miejsca występowania minimów. Jednocześnie widać, że minimum może występować dla różnych wartości parametru α w zależności od zbioru danych. Nawet jeśli błędy dla metod składowych ($\alpha = 0$, $\alpha = 1$) są wysokie, to dal pewnego α pośredniego można zredukować błędy znacząco. Pozwala to metodzie kombinowanej dostosować się do poszczególnych zbiorów danych. Jednocześnie nie widać objawów nadmiernego przystosowania się do zbioru danych (przeuczenia).

Praca [A7]

Na koniec, w pracy [A7], przebadano jeszcze jedną transformację — normalizację szeregów czasowych. Normalizacja danych jest najczęściej przeprowadzana w pierwszej fazie przetwarzania danych. Nie jest ona traktowana jako jakieś szczególne przekształcenie, raczej jako przygotowanie danych surowych do dalszych etapów wydobywania informacji z danych. Wiele procedur statystycznych wymaga wręcz, by dane wejściowe były znormalizowane. W przypadku danych wektorowych (nie szeregów) normalizację wykonuje się wzdłuż cech po wszystkich elementach zbioru danych, natomiast w przypadku szeregów czasowych normalizuje się raczej każdy szereg z osobna (wzdłuż osi czasowej). Wynika to stąd, że w przypadku danych wektorowych każda współrzędna (cecha) niesie informację innego rodzaju, natomiast w przypadku szeregów, każda następna wartość jest w jakiś sposób zależna od wartości poprzednich. W przypadku szeregów czasowych najczęściej uważa się wręcz, że wszystkie dane szeregowe powinny być normalizowane. Przykładem niech będzie baza danych UCR, która obecnie składa się tylko z danych normalizowanych i nie ma dostępu do danych nieznormalizowanych.

W pracy [A7] zastosowano techniki znane z poprzednich prac w celu pokazania, że zarówno dane nieznormalizowane (surowe), jak i znormalizowane (transformowane) niosą pewną informację dotyczącą danych i można wykorzystać zarówno jedne i drugie konstruując parametryczną miarę kombinowaną, z parametrem dobieranym w fazie uczenia na każdym ze zbiorów danych osobno.

Definiujemy normalizację (z-normalizację) szeregu czasowego x jako

$$\mathrm{norm}(x) = \frac{x - \mu(x)}{\sigma(x)},$$

gdzie $\mu(x)$ jest średnią z wartości szeregu x, a $\sigma(x)$ odchyleniem standardowym wartości szeregu x. W przypadku szeregów wielowymiarowych normalizujemy każdy wymiar szeregu X osobno

$$\mathrm{norm}(X) = (\mathrm{norm}(x_1), \mathrm{norm}(x_2), \ldots, \mathrm{norm}(x_m)).$$

Następnie tworzymy miarę odległości normDTW poprzez działanie odległością DTW na dane znormalizowane

$$\text{normDTW}(X, Y) = \text{DTW}(\text{norm}(X), \text{norm}(Y)).$$

Ostatecznie możemy utworzyć parametryczna miarę odległości jako kombinację wypukłą miar DTW i normDTW:

$$\text{combDTW}(X, Y) = (1 - \alpha)\,\text{DTW}(X, Y) + \alpha\,\text{normDTW}(X, Y), \quad \alpha \in [0, 1].$$

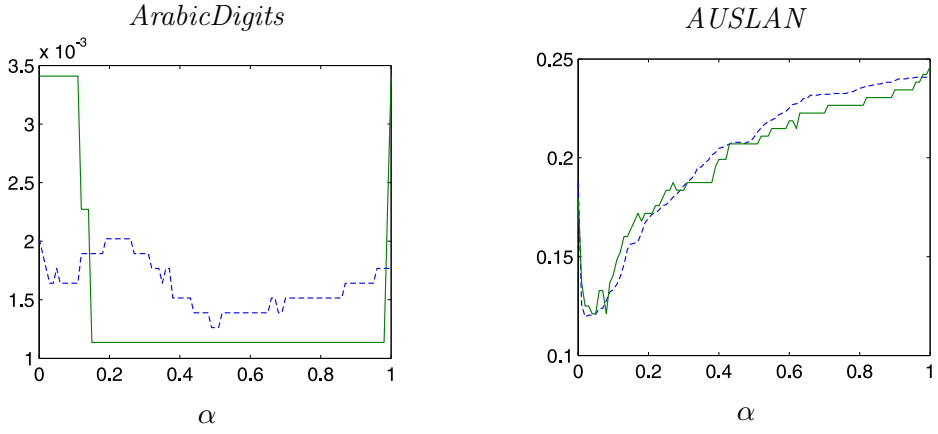
Tak skonstruowaną miarę odległości możemy wykorzystać w procesie klasyfikacji wielowymiarowych szeregów czasowych przy użyciu metody najbliższego sąsiada. Parametr α jest jak zwykle dobierany w fazie uczącej (na zbiorze uczącym w procesie kroswalidacji).

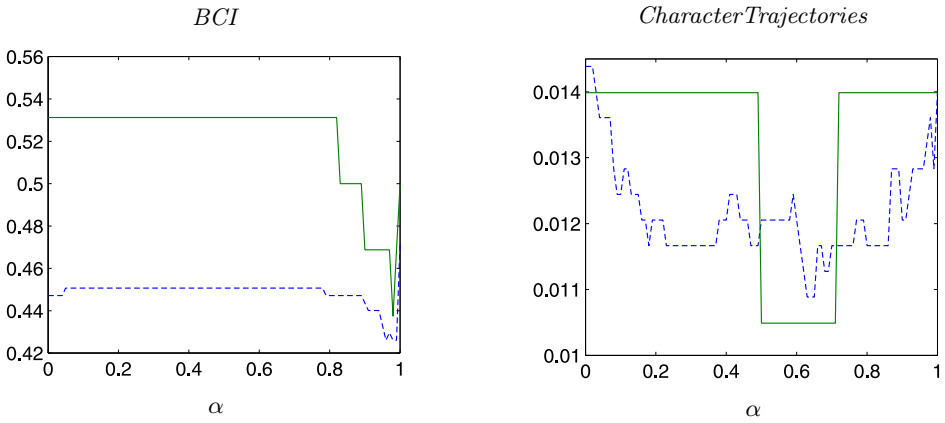
W eksperymentach obliczeniowych wykorzystano 16 zbiorów danych wielowymiarowych szeregów czasowych. Wyniki (błędy testowe) dla poszczególnych zbiorów danych przedstawione są w Tab 5.

Tabela 5: Błędy testowe dla porównywanych miar odległości (w procentach).

dataset	DTW	normDTW	combDTW
ArabicDigits	0.19	0.22	**0.14**
AUSLAN	18.05	23.20	**12.20**
BCI	44.89	46.54	**43.95**
CharacterTrajectories	1.36	1.50	**1.26**
CMUsubject16	3.67	**0.00**	**0.00**
ECG	18.50	**16.00**	**16.00**
Graz	37.14	34.29	**31.43**
JapaneseVowels	**2.03**	36.09	**2.03**
Libras	**8.61**	18.61	8.89
PenDigits	0.65	**0.63**	**0.63**
RobotFailure LP1	**12.64**	28.06	**12.64**
RobotFailure LP2	**32.00**	34.00	**32.00**
RobotFailure LP3	**29.00**	48.00	**29.00**
RobotFailure LP4	**10.08**	16.21	**10.08**
RobotFailure LP5	**29.30**	36.54	**29.30**
Wafer	**2.01**	3.85	**2.01**

Widać, że tylko dla jednego zbioru danych (Libras) metoda kombinowana jest gorsza od którejś z metod składowych. Natomiast na Rys. 22 przedstawiono przebiegi błędów uczących i testowych w zależności od parametru α.





Rysunek 22: Współzależność parametru α i wartości błędów. Linia przerywana — błąd uczący; linia ciągła — błąd testowy.

Widać, że dla poszczególnych zbiorów danych minimum błędów może występować dla różnych wartości parametru α. Równocześnie wzajemna odpowiedniość (monotoniczność, minima) błędów uczących i testowych jest bardzo duża.

Przeprowadzone badania wykazały, że zarówno w danych surowych i znormalizowanych zawiera się część informacji o danym zbiorze danych, którą mozna wykorzystać w procesie klasyfikacji wielowymiarowych szeregów czasowych. Jednocześnie widać, że optymalne wyniki daje tutaj użycie miary kombinowanej, łączącej w sobie dane surowe i znormalizowane i uzależniające udział tych informacji od konkretnego zbioru danych.

2.4. Pozostałe prace

Przed doktoratem (lub dotyczące doktoratu)

[B1] M. Łuczak (2004), On the Gleason-Kahane-Żelazko theorem, Commentationes Mathematicae 44(2), 245-253. **Lista B: 6 pkt.**

[B2] M. Łuczak (2005), Holomorphic functional calculus in A-pseudoconvex algebras, Commentationes Mathematicae 45(2), 161-169. **Lista B: 6 pkt.**

[B3] M. Łuczak (2006), A characterization of linear-multiplicative functionals in topological algebras, Commentationes Mathematicae 46(1), 79-91. **Lista B: 6 pkt.**

Po doktoracie (niedotyczące doktoratu)

[B4] I. Kierkosz, M. Łuczak (2008), Algorytm podziału i ograniczeń dla problemu rozkroju niegilotynowego, Zeszyty Naukowe Politechniki Śląskiej. Automatyka 150, 159-166.

[B5] T. Górecki, M. Łuczak (2010), Some methods of replacing the nearest neighbor method, Communications in Statistics-Simulation and Computation 39(2), 262-276. **IF: 0.343, Lista A: 13 pkt.**

[B6] T. Górecki, M. Łuczak (2010), Some methods of constructing kernels in statistical learning, Discussiones Mathematicae. Probability and Statistics 30(2), 179-201. **Lista B: 6 pkt.**

[B7] T. Górecki, M. Łuczak (2013), Linear discriminant analysis with a generalization of the Moore–Penrose pseudoinverse, International Journal of Applied Mathematics and Computer Science 23(2), 463-471. **IF: 1.39, Lista A: 25 pkt.**

[B8] T. Górecki, M. Łuczak (2014), A variant of gravitational classification, Biometrical Letters 51(1), 1-12. **Lista B: 9 pkt.**

[B9] I. Kierkosz, M. Łuczak (2014), A hybrid evolutionary algorithm for the two-dimensional packing problem, Central European Journal of Operations Research 22(4), 729-753. **IF: 0.832, Lista A: 20 pkt.**

[B10] A. Błażejewski, P. Kozioł, M. Łuczak (2014), Acoustical Analysis of Enclosure as Initial Approach to Vehicle Induced Noise Analysis Comparatevely Using STFT and Wavelets, Archives of Acoustics 39(3), 385-394. **IF: 0.661, Lista A: 15 pkt.**

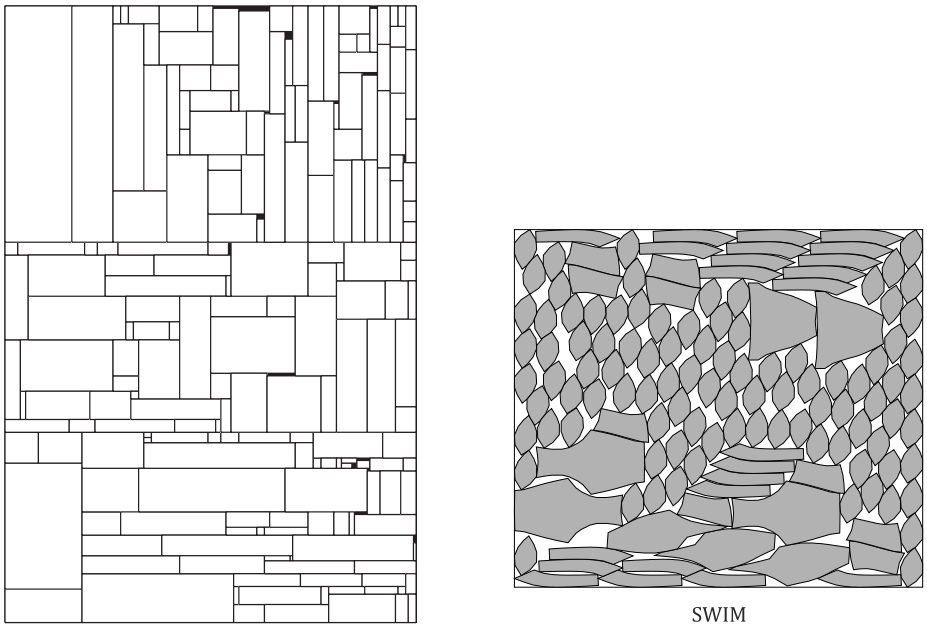
[B11] T. Górecki, M. Łuczak (2016), Evolutionarily tuned generalized pseudo-inverse in linear discriminant analysis, Computing and Informatics 35(3), 615-634. **IF: 0.488, Lista A: 15 pkt.**

[B12] T. Górecki, M. Łuczak (2017), Stacked regression with a generalization of the Moore-Penrose pseudoinverse, Statistics in Transition 18(3), 443-458. **Lista B: 15 pkt.**

Prace [B1-B3]. Prace te powstały przed doktoratem (lub dotyczą doktoratu). Są to prace z dziedziny analizy funkcjonalnej, a dokładnie pewnych zagadnień algebr topologicznych. Są w nich przedstawione i dowodzone twierdzenia dotyczące funkcjonałów liniowo-multiplikatywnych w rzeczywistych i zespolonych algebrach topologicznych: algebry Banacha, m-(pseudo)wypukłe, A-(pseudo)wypukłe.

Prace [B4, B9] i projekty [P1, P2]. Drugim dużym nurtem w moich badaniach (poza cyklem habilitacyjnym) są zagadnienia związane z dwuwymiarowym rozkrojem/pakowaniem/nestingiem elementów na danej płycie/płytach. Chodzi o uzyskanie optymalnego ułożenia elementów na płycie pod względem różnych kryteriów, np. gęstości upakowania, ilości użytych elementów, najlepszego wykorzystania materiału, ilości płyt itp. Układane elementy i płyty mogą być prostokątne lub nieregularne (Rys. 23).

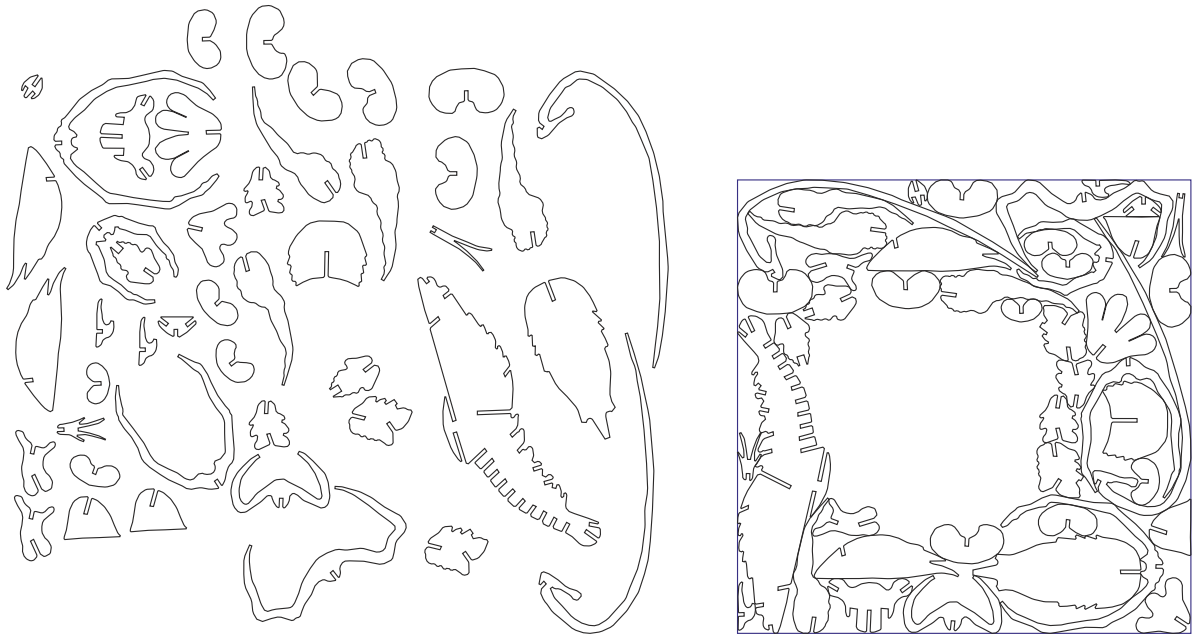
W pracy [B4] przedstawiony został porosty algorytm pakowania elementów prostokątnych z użyciem deterministycznego algorytmu z powrotami. Natomiast w pracy [B9] przedstawiono algorytm ewolucyjny rozkroju elementów prostokątnych, przeanalizowano jego wydajność i złożoność obliczeniową, porównano z innymi algorytmami rozkroju. Przykładowy rozkrój elementów z wykorzystaniem tego algorytmu przedstawiona na Rys. 23 (lewy).



Rysunek 23: Rozkrój elementów prostokątnych i nieregularnych.

Natomiast badania dotyczące nestingu elementów nieregularnych są prowadzone w ramach projektów finansowanych z otoczenia gospodarczego [P1, P2]. W ramach tych projektów (których byłem kierownikiem) powstały algorytmy rozkroju elementów nieregularnych i w postaci bibliotek użyte w oprogramowaniu ploterów/frezarek produkowanych przez firmę LYNX (`http://www.lynx-poland.com`).

W toku badań uzyskano oryginalne wyniki, które poza zastosowaniami w przemyśle zostaną także opublikowane (jeden artykuł w recenzjach, drugi przygotowywany). Przykładowy układ elementów na płycie z użyciem uzyskanych algorytmów przedstawiono na Rys. 23 (prawy). Jednym z przykładowych kryteriów rozkroju było też jak najlepsze upakowanie danych elementów z pozostawieniem jak największego spójnego niewykorzystanego obszaru płyty (Rys. 24). Charakterystyczny tutaj dookolny układ elementów okazuje się być lepszym (dającym większy niewykorzystany spójny obszar) niż inne algorytmy z tradycyjnym układem elementów (z pustym obszarem przy jednym z brzegów płyty). Wykorzystano tutaj pewien nowy algorytm jednoprzebiegowy będący bardzo efektywnym w stosunku do jakości rozkroju.



Rysunek 24: Elementy wejściowe (lewy) i układ końcowy (prawy).

Prace [B7, B11, B12]. Konstrukcja i wykorzystanie parametrycznej pseudoodwrotności w zagadnieniach klasyfikacji.

Niech A będzie macierzą liczb rzeczywistych o wymiarze $m \times n$ i rzędzie mniejszym niż $\min(m, n)$. Jeżeli M i N są macierzami dodatnio określonymi oraz istnieją faktoryzacje: $\hat{N}'\hat{N} = N$ i $\hat{M}'\hat{M} = M$, to macierz

$$A^{+}_{MN} = \hat{N}^{-1}\big(\hat{M}A\hat{N}^{-1}\big)^{+}\hat{M}, \tag{13}$$

gdzie operacja $\cdot^{+}$ jest pseudoodwrotnością Moore'a–Penrose'a (MP pseudoinverse), jest zwana ważoną pseudoodwrotnością Moore'a–Penrose'a. Ważona MP odwrotność ma wiele ważnych zastosowań np. w statystyce, teorii aproksymacji etc. Jeśli M jest nieujemnie określona oraz N jest macierzą jednostkową, to oznaczamy ją przez A^{*}_{M}.

W badanych metodach klasyfikacji zamiast MP-odwrotności A^{+} była użyta pseudoodwrotność A^{*}_{M} ze specjalną parametryczną postacią macierzy M:

$$\hat{\mathbf{N}} = \mathbf{N} = \mathbf{I}, \qquad \hat{\mathbf{M}} = \mathbf{M} = \begin{bmatrix} a_1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_m \end{bmatrix} \tag{14}$$

gdzie a_i przyjmują wartości 0 lub 1 dla $i = 1, \dots, m$. Wtedy równanie (13) przyjmuje postać

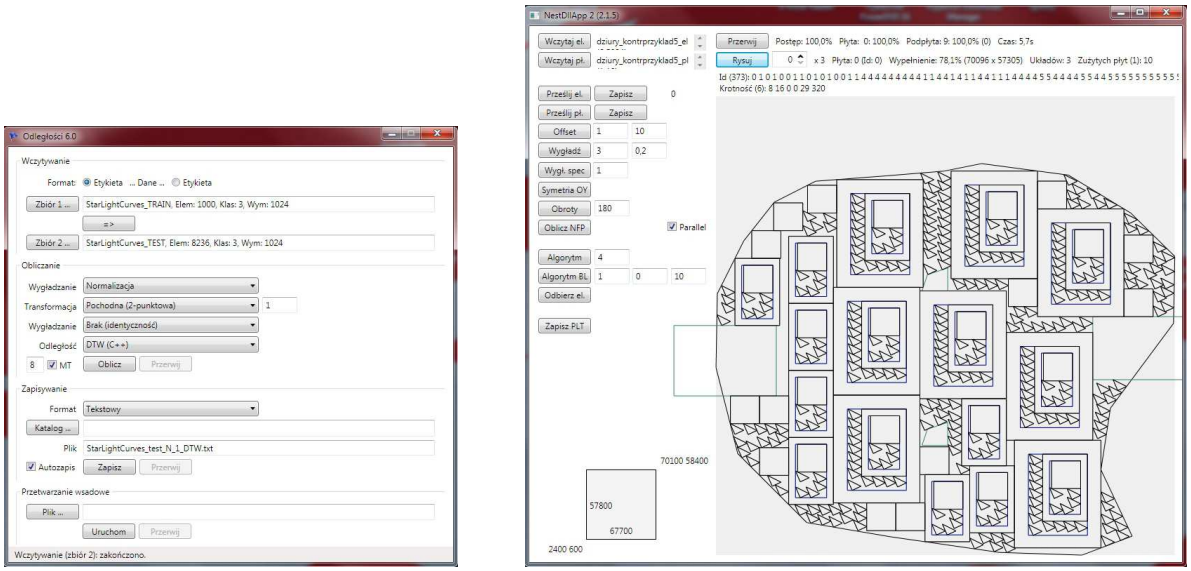
$$A^{*}_{M} = (MA)^{+}M. \tag{15}$$

Współczynniki $a_1, \ldots, a_m$ są dobierane w fazie uczącej badanych metod klasyfikacji.

Pseudoodwrotność ta była wykorzystana w klasyfikacji przy pomocy regresji stosowej [B12], dyskryminacji liniowej [B7] oraz dyskryminacji liniowej dużej skali z doborem współczynników macierzy M za pomocą algorytmu genetycznego [B11].

Eksperymenty obliczeniowe wykazały, że w przebadanych metodach, użycie pseudoodwrotności A^*_M daje lepsze rezultaty klasyfikacji niż w przypadku użycia MP-odwrotności A^+.

Autorskie oprogramowanie. Powstało autorskie oprogramowanie m.in. do obliczeń macierzy funkcji odległości, implementacje oryginalnych algorytmów klasyfikacji i analizy skupień oraz do badania i testowania algorytmów rozkroju/pakowania. Jestem autorem całego oprogramowania użytego we wszystkich przedstawionych pracach i projektach. Oprogramowanie było tworzone w językach: c, c++, c# (.net), Matlab, Mathematica. Całe oprogramowanie powstało z możliwością użycia obliczeń równoległych i rozproszonych. Może być uruchamiane zarówno na jedno i wieloprocesorowych pojedynczych komputerach, jak i na wielu takich komputerach połączonych siecią lokalną lub internetową. Część szczególnie wymagających obliczeń było wykonywanych nawet na kilkudziesięciu komputerach równocześnie. Część oprogramowania została wyposażona w wygodny graficzny interfejs (Rys. 25). Program do obliczeń macierzy odległości dla funkcji badanych w przedstawionych pracach jak i innych używanych w klasyfikacji odległościowej będzie w przyszłości (po dopracowaniu interfejsu i lokalizacji) udostępniony społeczności naukowej na otwartej licencji.



Rysunek 25: Program do obliczeń macierzy funkcji odległości (lewy) oraz program do testowania biblioteki funkcji do nestingu elementów nieregularnych (prawy).

2.5. Udział w projektach i grantach

Projekty finansowane z otoczenia gospodarczego

[P1] Temat: **Opracowanie algorytmu nestingu NFP i modułu bibliotecznego w postaci biblioteki dll**

Nr umowy: 501.01.05/2014 (Politechnika Koszalińska)

Kierownik projektu: dr Maciej Łuczak

Wykonawca projektu: dr Igor Kierkosz

Czas trwania projektu: 01.2014-09.2014

Projekt finansowany przez firmę: LYNX (Grawerki i Plotery) Piotr Wyrodow-Rakowski (Warszawa) `http://www.lynx-poland.com`.

[P2] Temat: **Opracowanie zaawansowanych algorytmów rozkroju elementów nieregularnych i ich implementacja w postaci biblioteki DLL**

Nr umowy: 501.01.08/2017 (Politechnika Koszalińska)

Kierownik projektu: dr Maciej Łuczak

Wykonawca projektu: dr Igor Kierkosz

Czas trwania projektu: 05.2017-12.2017

Projekt finansowany przez firmę: LYNX (Grawerki i Plotery) Piotr Wyrodow-Rakowski (Warszawa) `http://www.lynx-poland.com`.

2.6. Recenzowanie w czasopismach naukowych

Wykonałem 31 recenzji dla 13 czasopism.

Nazwa czasopisma	Ilość recenzji
Expert Systems with Applications	13
International Journal of Applied Mathematics and Computer Science	4
Knowledge-Based Systems	2
Econometrics and Statistics	2
Journal of Applied Statistics	2
Computational Statistics and Data Analysis	1
Earth Science Informatics	1
IEEE Access	1
IEEE Transactions on Cybernetics	1
International Journal of Data Science and Analytics	1
Knowledge and Information Systems	1
Pattern Recognition	1
Pattern Recognition Letters	1