Transformaty w klasyfikacji szeregów czasowych

Tomasz Górecki

Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza

XXXVIII Konferencja Statystyka Matematyczna
Wisła 3-7.12.2012

Wprowadzenie

Klasyfikacja szeregów czasowych jest w ostatnim czasie intensywnie rozwijana. Stosowane są w tym celu bardzo różne podejścia. Od sieci neuronowych i bayesowskich do algorytmów genetycznych i metody SVM. Równocześnie wprowadzane są nieustannie nowe miary badające podobieństwo szeregów czasowych (doskonały przegląd takich metod znajduje się w pracy Dinga i innych (2008)). Niemniej jednak okazuje się, że w większości przypadków najlepsze wyniki klasyfikacji uzyskuje się korzystając z metody najbliższego sąsiada (1NN) jako klasyfikatora oraz odległości DTW (dynamic time warping) jako miary odległości pomiędzy dwoma szeregami.

Wprowadzenie

DTW jest klasyczną już miarą odległości doskonale sprawdzającą się w klasyfikacji szeregów czasowych. Pozwala ona na znalezienie najmniejszej odległości między dwoma szeregami czasowymi przy dopuszczeniu nieliniowej transformacji czasu dla obu szeregów. Jest bardziej odpowiednia niż odległość euklidesowa szczególnie w sytuacji gdy porównujemy szeregi o podobnej strukturze, ale przesunięte w czasie. Pozwala na znalezienie najmniejszej odległości między dwoma szeregami czasowymi przy dopuszczeniu przesunięć w czasie dla obu szeregów. Algorytm ten radzi sobie również w przypadku braku części danych lub ich niedokładności. Najważniejsza jest tutaj kolejność występowania poszczególnych faz szeregu czy sygnału.

Wprowadzenie

Oczywiście w przypadku klasyfikacji szeregów czasowych wydaje się, że jedynie ich punktowe porównanie może być niewystarczające. Zdarzają się przypadki, w których przypisanie do jednej z klas zależy nie tylko od wartości funkcji, ale również od ich kształtu. W szczególności zmienność w czasie powinna mieć duży wpływ na proces klasyfikacji.

Wprowadzenie

Wiadomo, że w matematyce za zmienność funkcji w czasie odpowiada pochodna funkcji, która określa gdzie funkcja rośnie, maleje względnie jest stała. Pochodna określa ogólny kształt funkcji, pokazuje co się dzieje w pewnym sąsiedztwie punktu. W kontekście szeregów czasowych oznacza to, że pochodna określa zachowanie się szeregu czasowego przed i po pewnym punkcie czasowym.

Wprowadzenie

Wydaje się, że takie podejście do klasyfikacji może być bardzo skuteczne. Nie oczekujmy jednak, że wystarczy porównać jedynie pochodne. W większości przypadków duże znaczenie ma również porównanie odległości pomiędzy szeregami, a nie tylko pomiędzy ich pochodnymi. W takiej sytuacji optymalne wydaje się uwzględnienie obu odległości. Wkład każdej z odległości powinien być ustalony doświadczalnie dla konkretnego zbioru danych.

Wprowadzenie

Wykorzystanie pochodnych do klasyfikacji szeregów czasowych zostało zaproponowane przez Keogha i Pazzaniego (2001). Nazwali oni swoją metodę Derivative Dynamic Time Warping (DDTW). W poprzedniej pracy (Górecki, Łuczak (2012a)) zaproponowana została poprawka ich metody uwzględniająca parametryzację wkładu składowych na miarę odległości. Jak się okazało dała ona doskonałe wyniki. Zachęciło to nas do poszukiwania kolejnych modyfikacji. Oczywistym krokiem jest dodanie kolejnej pochodnej jako trzeciej składowej odległości, oczywiście również w postaci sparametryzowanej. Podczas gdy pierwsza pochodna daje informację o monotoniczności funkcji, druga dodaje informacje o jej wypukłości. Wyniki nie są już tak spektakularne, ale wciąż uzyskuje się istotną redukcję błędu klasyfikacji (Górecki, Łuczak (2012b)).

Wprowadzenie

Dodawanie kolejnych pochodnych wydaje się niecelowe, ponieważ jak pokazują badania, druga pochodna dodaje już stosunkowo niewiele informacji o odległości pomiędzy szeregami. Trzeba zatem szukać innej klasy funkcji. Wybór padł na bardzo popularne w klasyfikacji szeregów czasowych transformaty. Były one jednak do tej pory wykorzystywane w celu redukcji wymiarowości danych bez straty informacji. My natomiast chcemy dodać je jako kolejną składową miary odległości pomiędzy szeregami czasowymi. Ostatecznie zdecydowaliśmy się na trzy transformaty: sinusową, kosinusową oraz Hilberta. Oczywiście podobnie jak poprzednio nie należy sądzić, że wystarczy porównać jedynie odległość pomiędzy transformatami. Naturalnym wydaje się dodanie transformaty jako kolejnego elementu poprawiającego precyzję klasyfikacji. Podobnie jak poprzednio zostało użyte podejście parametryczne, które pozwala dobrać wpływ każdej odległości na końcową miarę odległości pomiędzy szeregami, a w konsekwencji na jakość klasyfikacji.

Transformaty

Dla szeregu czasowego $f = \{f(i)\colon i = 1, 2, \ldots, n\}$ określamy transformatę jako $\hat{f} = \{\hat{f}(k)\colon k = 1, 2, \ldots, n\}$.

- Transformata kosinusowa:

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

- Transformata sinusowa:

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

- Transformata Hilberta:

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

Miara odległości

Jeśli dist jest miarą odległości pomiędzy dwoma szeregami czasowymi f i g, nowa miara odległości $\widehat{\text{dist}}_{abc}$ ma postać

$$\widehat{\text{dist}}_{abc}(f,g) := a\,\text{dist}(f,g) + b\,\text{dist}(f',g') + c\,\text{dist}(\hat{f},\hat{g}),$$

gdzie f', g' są dyskretnymi pochodnymi szeregów f, g; $\hat{f}$, $\hat{g}$ są transformatami oraz $a, b, c \in [0,1]$ są parametrami.
Dyskretna pochodna szeregu czasowego f o długości n ma postać

$$f'(i) = f(i) - f(i-1), \qquad i = 2,3,\ldots,n$$

gdzie f' jest szeregiem czasowym o długości $n-1$.
Parametry a, b, c są wybierane w fazie uczenia.

Przestrzeń parametrów

Nie musimy sprawdzać wszystkich możliwych wartości parametrów$a, b, c \in [0, 1]$. Jeśli $a_1 = ka_2$, $b_1 = kb_2$, $c_1 = kc_2$ gdzie $k > 0$ jest stałą (tzn. punkty (a_1, b_1, c_1), (a_2, b_2, c_2) są liniowo zależne), mamy

$$\widehat{\text{dist}}_{a_1 b_1 c_1}(f_1, g_1) \lessgtr \widehat{\text{dist}}_{a_1 b_1 c_1}(f_2, g_2) \iff \widehat{\text{dist}}_{a_2 b_2 c_2}(f_1, g_1) \lessgtr \widehat{\text{dist}}_{a_2 b_2 c_2}(f_2, g_2)$$

Zatem możemy wybrać punkty (a, b, c) z dowolnej ciągłej powierzchni pomiędzy punktami $A = (1, 0, 0)$, $B = (0, 1, 0)$, oraz $C = (0, 0, 1)$. Na ten przykład może to być powierzchnia trójkąta o wierzchołkach w tych punktach lub ósma część sfery. Dla prostoty wybierany jest trójkąt. Ten $3D$ trójkąt jest rzutowany na trójkąt $2D$ o wierzchołkach w punktach $A' = (0, 0)$, $B' = (1, 0)$ i $C' = (\frac{1}{2}, \frac{\sqrt{3}}{2})$.

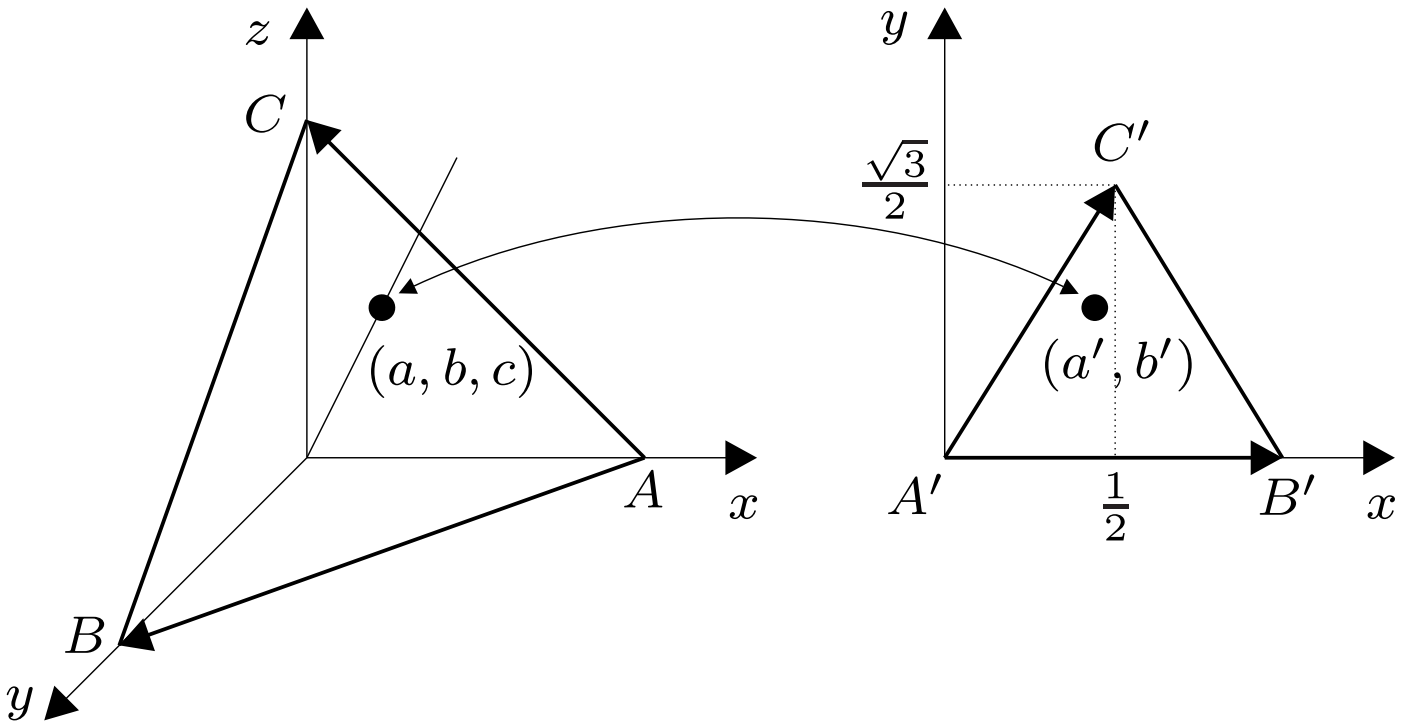
Przestrzeń parametrów

Oba trójkąty mogą być zdefiniowane w sposób parametryczny następująco:

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

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

gdzie (a, b, c) są punktami trójkąta $3D$, natomiast (a', b') są punktami trójkąta $2D$ oraz $\alpha \in [0, 1], \beta \in [0, 1 - \alpha]$ są parametrami.



Ograniczenie dolne oraz nierówność trójkąta

Jeśli low jest dolnym ograniczeniem odległości dist, wtedy

$$\widehat{\mathrm{low}}_{abc}(f,g) = a\,\mathrm{low}(f,g) + b\,\mathrm{low}(f',g') + c\,\mathrm{low}(\hat{f},\hat{g})$$

jest dolnym ograniczeniem odległości $\widehat{\mathrm{dist}}_{abc}$.
Jeśli bazowa odległość dist jest metryką, wtedy nowa odległość $\widehat{\mathrm{dist}}$ jest również metryką. Jeśli dist nie jest metryką, ale spełnia nierówność trójkąta, to odległość $\widehat{\mathrm{dist}}$ również spełnia nierówność trójkąta

$$\widehat{\mathrm{dist}}_{abc}(f,g) \leq \widehat{\mathrm{dist}}_{abc}(f,h) + \widehat{\mathrm{dist}}_{abc}(h,g).$$

Przebieg doświadczenia

Zostały wykonane obliczenia na 43 szeregach czasowych. Dane zostały zaczerpnięte z bazy UCR Time Series Classification/Clustering Homepage (Keogh i inni (2011)). Badane były trzy miary odległości: DTD^{S}_{DTW}, DTD^{C}_{DTW} oraz DTD^{H}_{DTW}. Dla każdego zbioru danych obliczony został błąd klasyfikacji na zbiorze testowym. Podczas konstrukcji modelu (uczenia) wykorzystany był zbiór uczący. W celu znalezienia optymalnych wartości parametrów wykorzystana została metoda sprawdzania krzyżowego (leave-one-out). Jeśli minimalny błąd został osiągnięty dla kilku różnych par parametrów wybierana była najmniejsza para (minimalizacja najpierw po α potem po β). Parametry poszukiwane były z przedziału od 0 do 1 z krokiem 0.01.

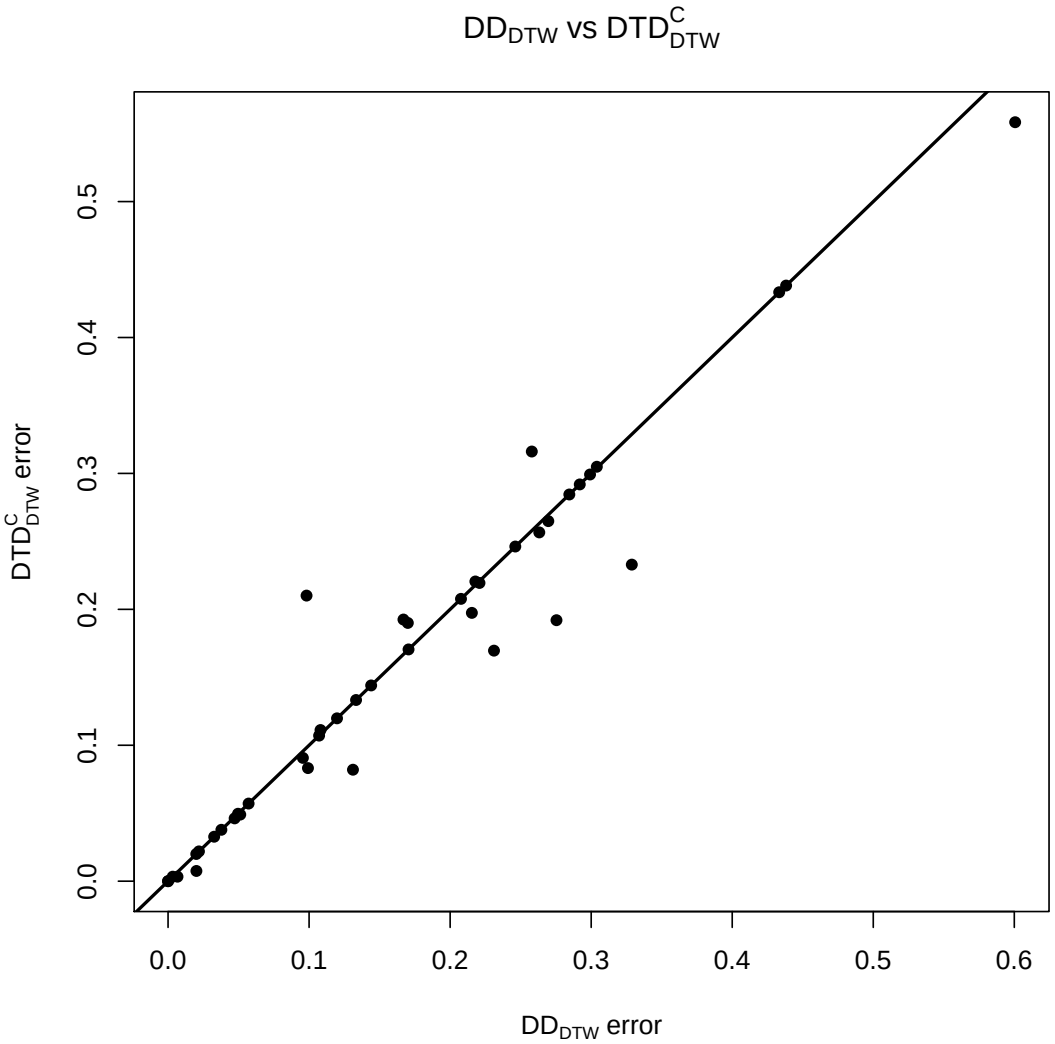
Wyniki

Tabela : Błędy względne na zbiorze testowym.

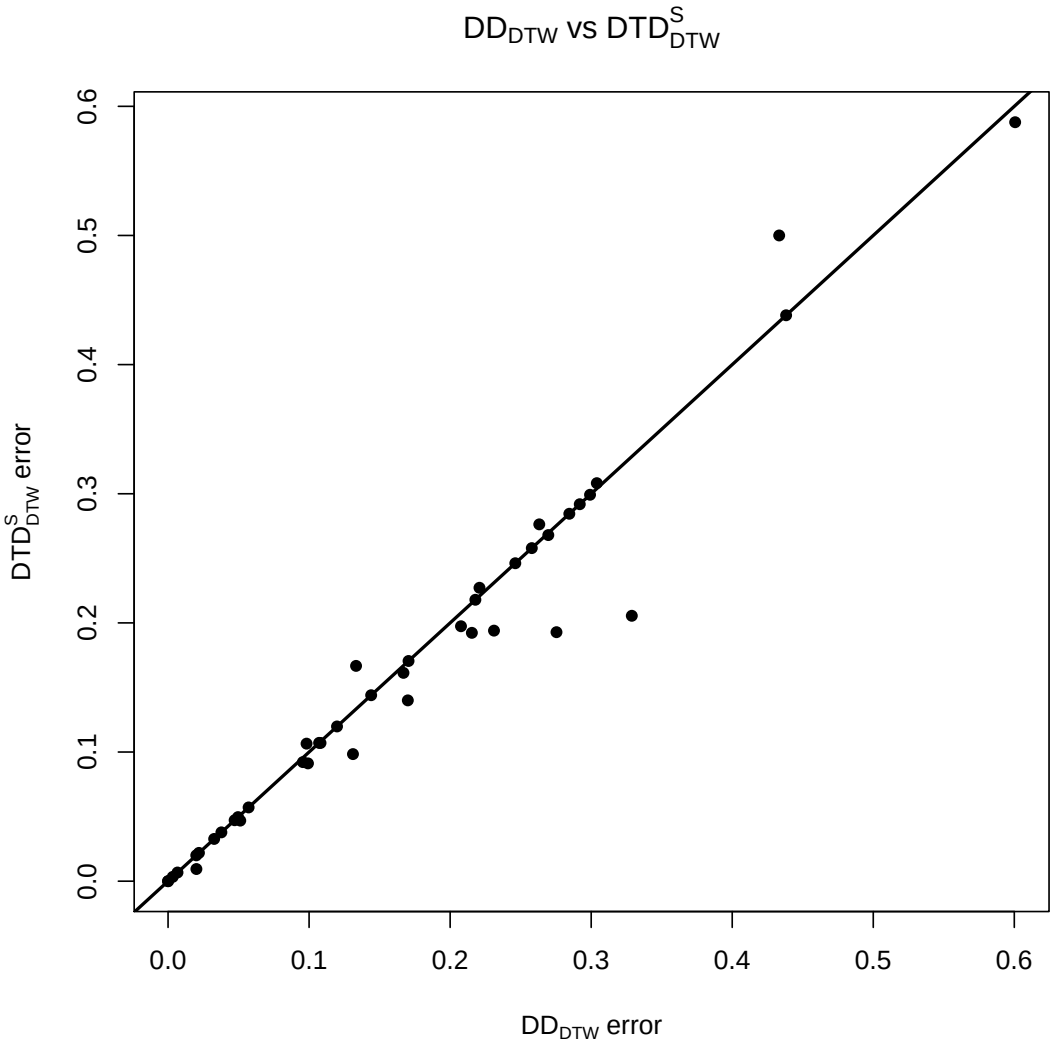
$\frac{DD_{DTW}-DTW}{DTW}$	$\frac{DTD^{C}_{DTW}-DTW}{DTW}$	$\frac{DTD^{S}_{DTW}-DTW}{DTW}$	$\frac{DTD^{H}_{DTW}-DTW}{DTW}$
-19.11	-23.18	-23.04	-19.97

Widzimy, że wszystkie proponowane metody dają redukcję względnego błędu klasyfikacji w porównaniu z klasyczną metodą DTW. Największą redukcję otrzymujemy w przypadku metody $\mathrm{DTD}^{\mathrm{C}}_{\mathrm{DTW}}$

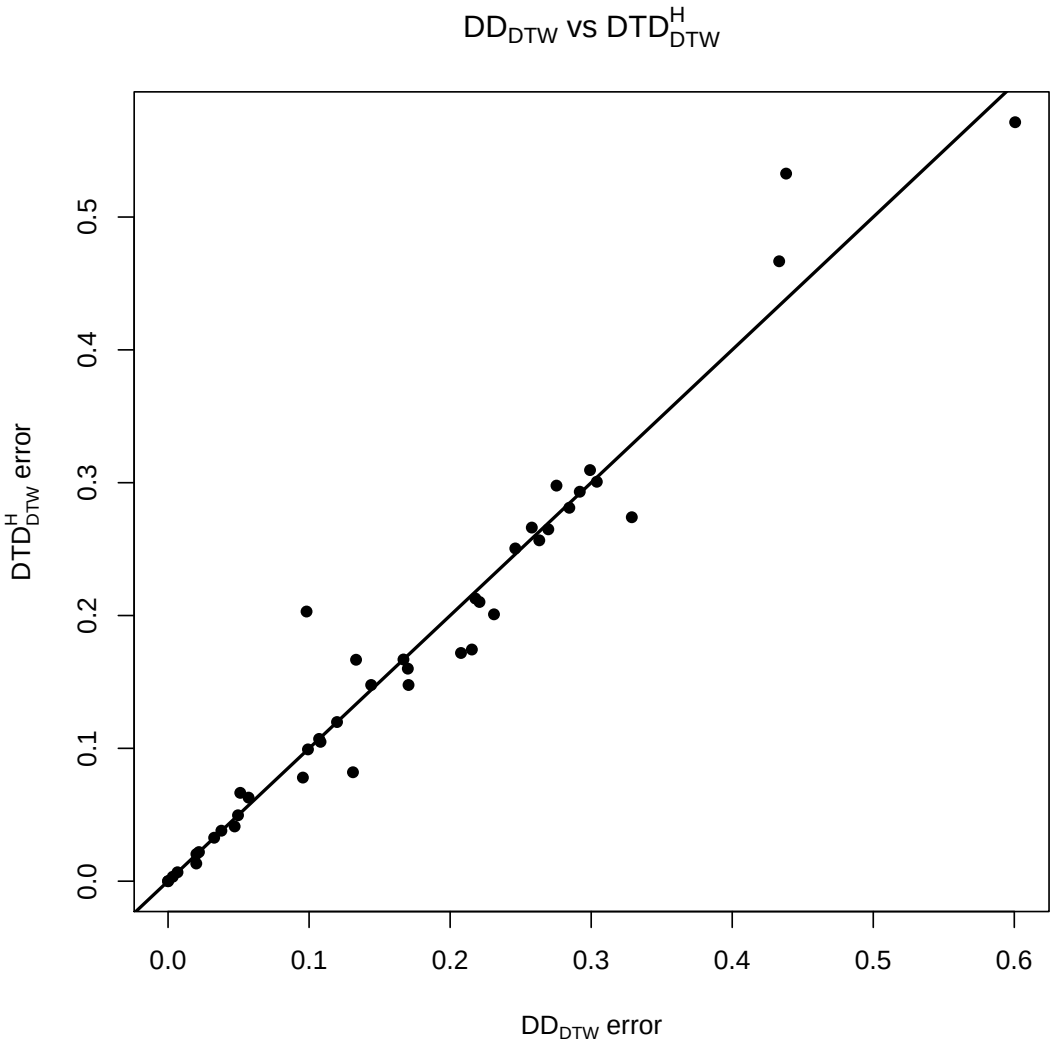
Wyniki


DD_DTW vs DTD^C_DTW
DTD^C_DTW error
DD_DTW error
0.0
0.1
0.2
0.3
0.4
0.5
0.6

Wyniki

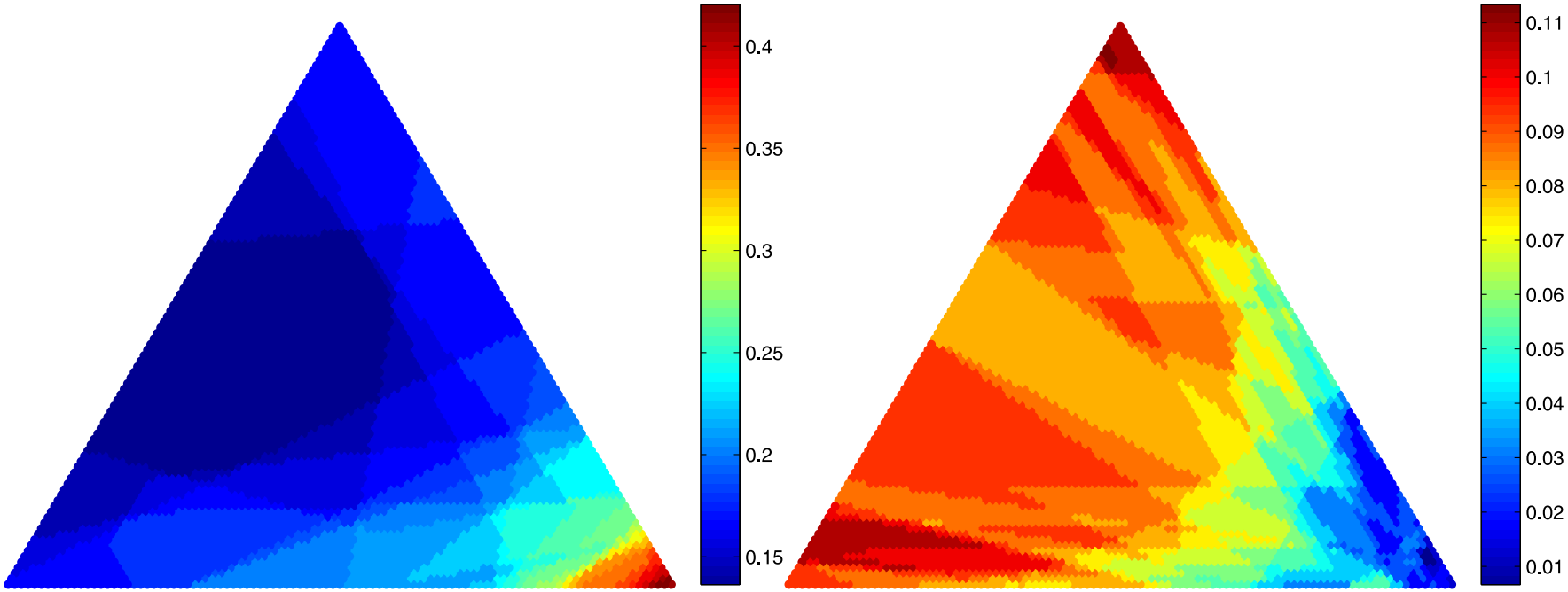

DD_DTW vs DTD^S_DTW
DTD^S_DTW error
DD_DTW error
0.0
0.1
0.2
0.3
0.4
0.5
0.6

Wyniki

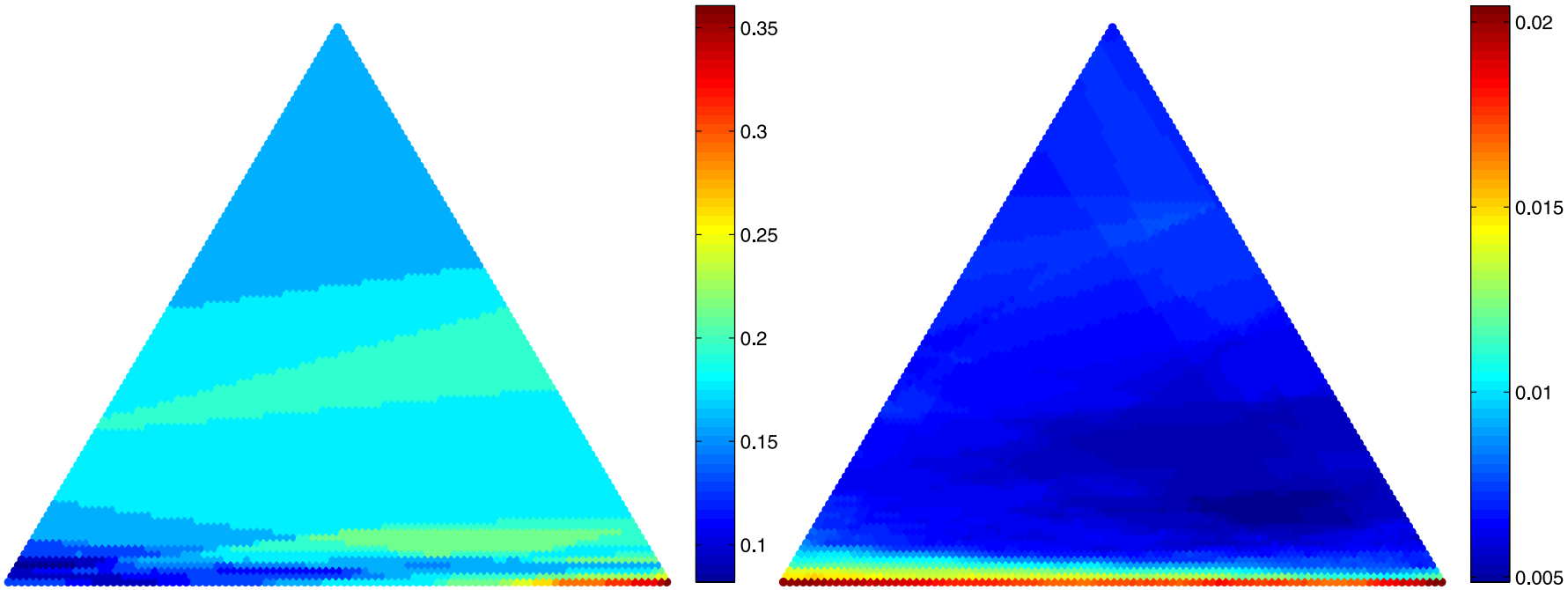

DD_DTW vs DTD^H_DTW
DTD^H_DTW error
DD_DTW error
0.0
0.1
0.2
0.3
0.4
0.5
0.6

Wyniki

Oczywiście bardzo ciekawą kwestią jest, czy istnieją pewne uniwersalne wartości współczynników określających udział poszczególnych składowych w finalnej mierze odległości. Niestety odpowiedź na to pytanie jest negatywna, co ilustrują poniższe rysunki.



Wyniki



Widzimy, że udział poszczególnych składowych jest bardzo różny dla różnych konfiguracji zbiorów i metod.

Wyniki

W celu weryfikacji statystycznej istotności chcemy przetestować hipotezę zerową, która mówi, że wszystkie metody zachowują się tak samo, a zaobserwowane różnice są jedynie dziełem przypadku. W tym celu użyta została wersja testu F opracowana przez Imana i Davenporta (1980), który jest nieparametrycznym testem ANOVA. Pod uwagę brane są wszystkie 43 zbiory danych oraz 5 metod klasyfikacji. Otrzymana p-wartość jest bliska 0. Z tego powodu odrzucamy hipotezę zerową, czyli analizowane klasyfikatory różnią się istotnie. Zatem możemy zastosować test post-hoc w celu wykrycia, które klasyfikatory istotnie różnią się jakością klasyfikacji. Kiedy używamy testów post-hoc, aby zachować poziom istotności całej procedury musimy zastosować jedną z wielu metod korekty poziomu istotności. Doskonała synteza tych metod znajduje się w pracy Garci i Herrery (2008). Pokazują oni, że pomimo dużej złożoności obliczeniowej dynamiczna procedura Bergmanna i Hommela (1988) ma największą moc. Dodatkowo szybki algorytm (Hommel, Bernhard (1994)) pozwala istotnie zredukować czas obliczeń.

Tabela : Wyniki testu post-hoc Bergmanna-Hommela.

Procedura	Średnie rangi		
DTD^{C}_{DTW}	2.49	*	
DTD^{S}_{DTW}	2.58	*	
DTD^{H}_{DTW}	2.69	*	
DD_{DTW}	2.98	*	
DTW	4.27		*

W wyniku przeprowadzenia powyższej procedury uzyskano dwie rozłączne grupy klasyfikatorów. W jednej jest metoda DTW, natomiast w drugiej, lepszej są wszystkie pozostałe metody. Daje to istotny dowód wyższości proponowanych metod nad klasyczną metodą DTW.

Bibliografia

- Bergmann, G., Hommel, G. (1988). Improvements of general multiple test procedures for redundant systems of hypotheses, in P. Bauer, G. Hommel and E. Sonnemann (Eds),Multiple Hypotheses Testing, Springer, pp. 110–115.
- Ding, H., Trajcevski, G., Scheuermann, P., Wang, X. and Keogh, E. (2008). Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures. In Proc. 34th Int. Conf. on Very Large Data Bases, 1542–1552.
- Garcia, S., Herrera, F. (2008). An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. Journal of Machine Learning Research 9: 2677–2694.
- Górecki, T., Łuczak, M. (2012a). Using derivatives in time series classification. Data Mining and Knowledge Discovery, Doi: 10.1007/s10618-012-0251-4.

Bibliografia

- Górecki, T., Łuczak, M. (2012b). First and second derivative in time series classification using DTW. Under review.
- Hommel, G., Bernhard, G. (1994). A rapid algorithm and a computer program for multiple test procedures using logical structures of hypotheses. Computer Methods and Programs in Biomedicine 43(3-4): 213–6.
- Keogh, E., Pazzani, M. (2001). Dynamic Time Warping with Higher Order Features. In First SIAM International Conference on Data Mining (SDM'2001), Chicago, USA.
- Keogh, E., Zhu, Q., Hu, B., Hao. Y., Xi, X., Wei, L. & Ratanamahatana, C. A. (2011). The UCR Time Series Classification/Clustering Homepage: `http://www.cs.ucr.edu/~eamonn/time_series_data/`