Odległość geodezyjna w klasyfikacji wielowymiarowych szeregów czasowych

Tomasz Górecki, Sławomir Pioroński

Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza w Poznaniu

Metodologia Badań Statystycznych (MET)
Warszawa 03-05.07.2023

Plan prezentacji

Wprowadzenie

W obliczu nieustannej ekspansji cyfrowego świata, nasze społeczeństwo staje się coraz bardziej zależne od danych. Wszystko, od decyzji biznesowych po politykę publiczną, jest kształtowane przez naszą zdolność do gromadzenia, analizowania i interpretowania tych danych. W tym kontekście, wielowymiarowe szeregi czasowe stają się niezwykle ważne, jako fundamentalne narzędzie do analizy i modelowania danych.

Wprowadzenie

Wielowymiarowe szeregi czasowe są szczególnie użyteczne w obszarach, które wykraczają poza prostą analizę jednowymiarową. Przykładowo, mogą one być stosowane do analizy wzorców zachowań konsumentów w różnych regionach lub do prognozowania trendów w wielu sektorach gospodarki jednocześnie. Co więcej, są one nieodzownym narzędziem w nowoczesnych technologiach, takich jak sztuczna inteligencja czy uczenie maszynowe, które wymagają analizy wielowymiarowych danych w czasie.

Wprowadzenie

Jednak pomimo ich znaczenia, wielowymiarowe szeregi czasowe nie są jeszcze w pełni zrozumiane i zbadane. Wciąż wiele jest do zrozumienia na temat ich właściwości, metod analizy i interpretacji. W konsekwencji w ostatnim czasie nastąpił znaczy wzrost zainteresowania w badaniu tego typu danych, co z kolei spowodowało wysyp prac wprowadzających nowe metody indeksowania, klasyfikacji, grupowania oraz aproksymacji szeregów czasowych. W szczególności, wprowadzono wiele nowych miar odległości pomiędzy szeregami.

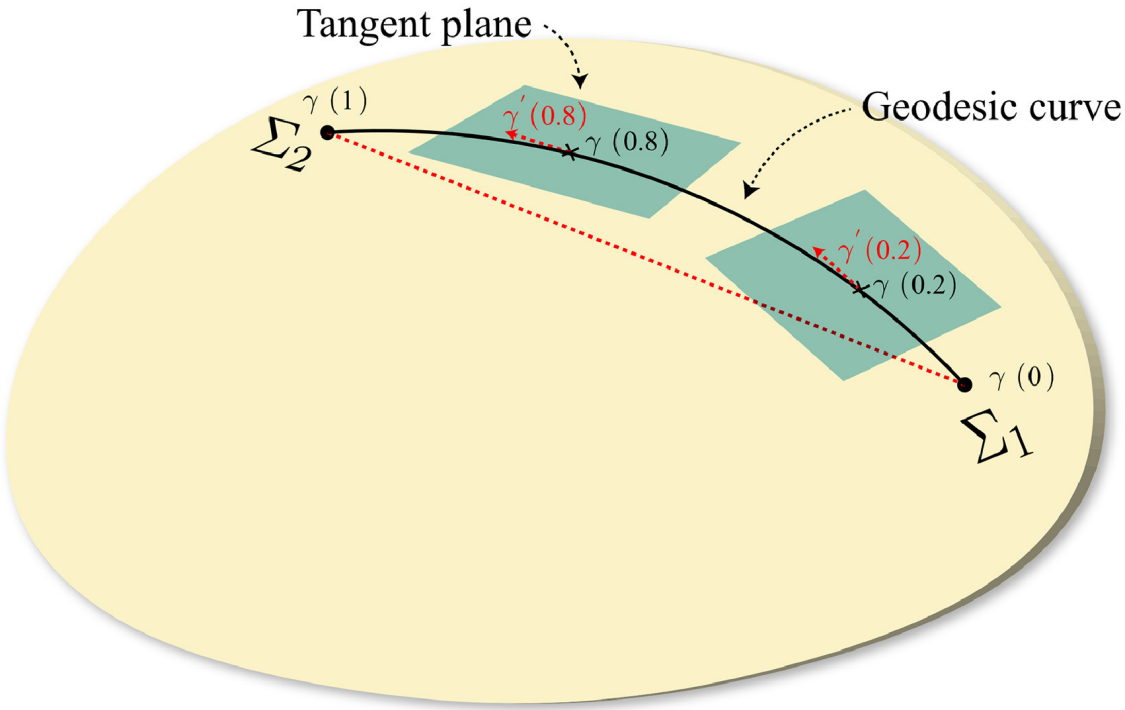
Odległość geodezyjna – definicja

Niech każdy wielowymiarowy szereg czasowy będzie reprezentowany przez macierz $\boldsymbol{X}_i \in \mathbb{R}^{c \times n}$, gdzie c jest wymiarem szeregu (liczbą zmiennych), a n jego długością. Możemy dla niego wyznaczyć oszacowanie macierzy kowariancji $\boldsymbol{\Sigma}_i$. Macierze kowariancji leżą na rozmaitości riemannowskiej, ponieważ zawsze są symetrycznymi macierzami dodatnio określonymi. Rozmaitość to przestrzeń topologiczna, która jest lokalnie podobna do przestrzeni euklidesowej. Krzywa o minimalnej długości łącząca dwa punkty na rozmaitości nazywana jest geodezyjną, a odległość między punktami jest określona przez długość tej krzywej.

Szacowanie macierzy kowariancji

W całej pracy do szacowania macierzy kowariancji została wykorzystana metoda największej wiarogodości.

Odległość geodezyjna – definicja



Rysunek jest koncepcyjną ilustracją krzywej geodezyjnej i płaszczyzny stycznej w punkcie na rozmaitości. Pokazuje również, że prosta łącząca $\mathbf{\Sigma}_1$ i $\mathbf{\Sigma}_2$ (przerywana czerwona linia) i krzywa geodezyjna mogą mieć różne długości. Punktem krytycznym, jest to, że długość krzywej geodezyjnej jest równa długości linii łączącej $\mathbf{\Sigma}_1$ z obrazem $\mathbf{\Sigma}_2$ na płaszczyźnie stycznej do $\mathbf{\Sigma}_1$.

Odległość geodezyjna – definicja

Odległość geodezyjna

Odległość geodezyjna pomiędzy dwoma macierzami kowariancji ma postać

$$d(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) = \text{tr}\left(\log^2\left(\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2}\right)\right) = \sqrt{\sum_{k=1}^{c}\log^2\lambda_k},$$

gdzie $\lambda_k,\ k = 1, \ldots, c$ są rzeczywistymi wartościami własnymi macierzy $\mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2}$.

Odległość geodezyjna – własności

1. Niezmienniczość na odwracanie:

$$d(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = d(\boldsymbol{\Sigma}_1^{-1}, \boldsymbol{\Sigma}_2^{-1}).$$

2. Niezmienniczość na przekształcenia kongruentne:

$$d(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = d(\boldsymbol{P}^\top \boldsymbol{\Sigma}_1 \boldsymbol{P}, \boldsymbol{P}^\top \boldsymbol{\Sigma}_2 \boldsymbol{P}),$$

gdzie $\boldsymbol{P}$ jest odwracalną macierzą kwadratową wymiaru $c \times c$. Własność ta jest bardzo ważna w przypadku szeregów czasowych, ponieważ zapewnia, że wszelkie operacje liniowe, które można modelować za pomocą macierzy odwracalnej $\boldsymbol{P}$, nie mają wpływu na odległość d. Ten rodzaj transformacji obejmuje przeskalowanie i normalizację szeregów czasowych, filtrowanie itp.

Zbiory danych

Wykonane zostały eksperymenty na 24 zbiorach danych. Zbiory pochodzą z Time Series Machine Learning Website[1] oraz z pakietu R mdfs[2].

[1] http://www.timeseriesclassification.com/

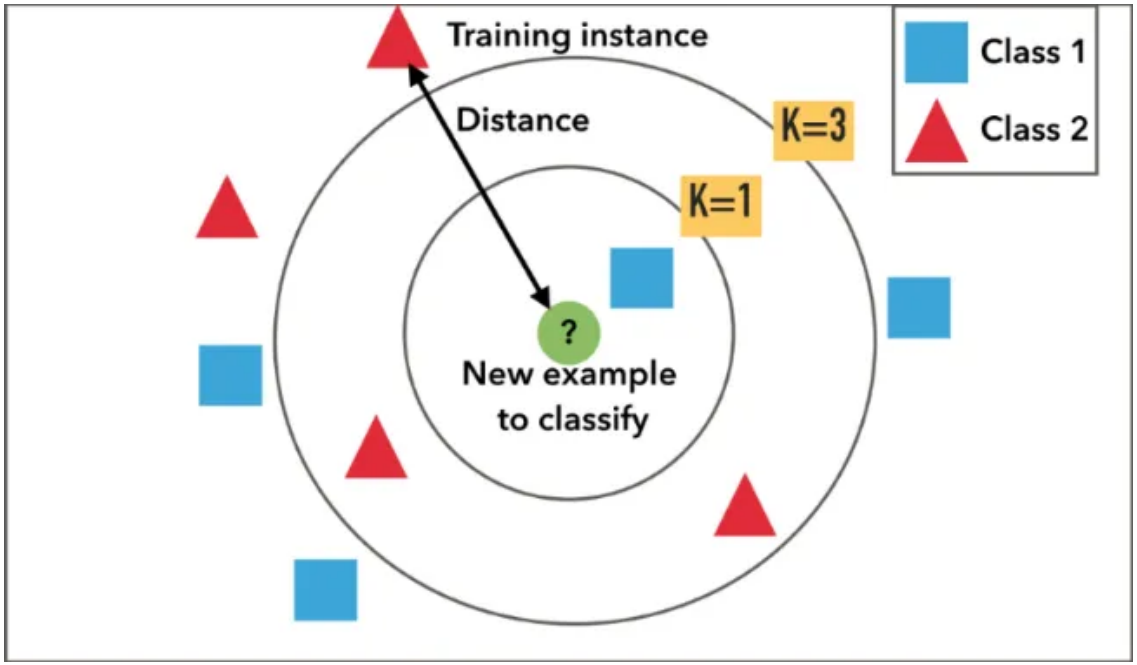
[2] https://github.com/Halmaris/mfds

Zbiory danych

Zbiór danych	Liczba obserwacji	Wymiar	Długość	Liczba klas
ArticularyWordRecognition	575	9	144	25
AtrialFibrillation	30	2	640	3
BasicMotions	80	6	100	4
CharacterTrajectories	2 858	3	182	20
Cricket	180	6	1 197	12
ECG	200	2	152	2
EigenWorms	259	6	17 984	5
ERing	300	4	65	6
EthanolConcentration	524	3	1751	4
Graz	140	3	1152	3
HandMovementDirection	234	10	400	4
JapaneseVowels	640	12	29	9
Libras	360	2	45	15
NATOPS	360	24	51	6
RacketSports	303	6	30	4
RobotFailure_LP1	88	6	15	4
RobotFailure_LP2	47	6	15	5
RobotFailure_LP3	47	6	15	4
RobotFailure_LP4	117	6	15	3
RobotFailure_LP5	164	6	15	5
SelfRegulationSCP1	561	6	896	2
SelfRegulationSCP2	380	7	1 152	2
StandWalkJump	27	4	2 500	3
Wafer	1 194	6	198	2

Porównywane metody

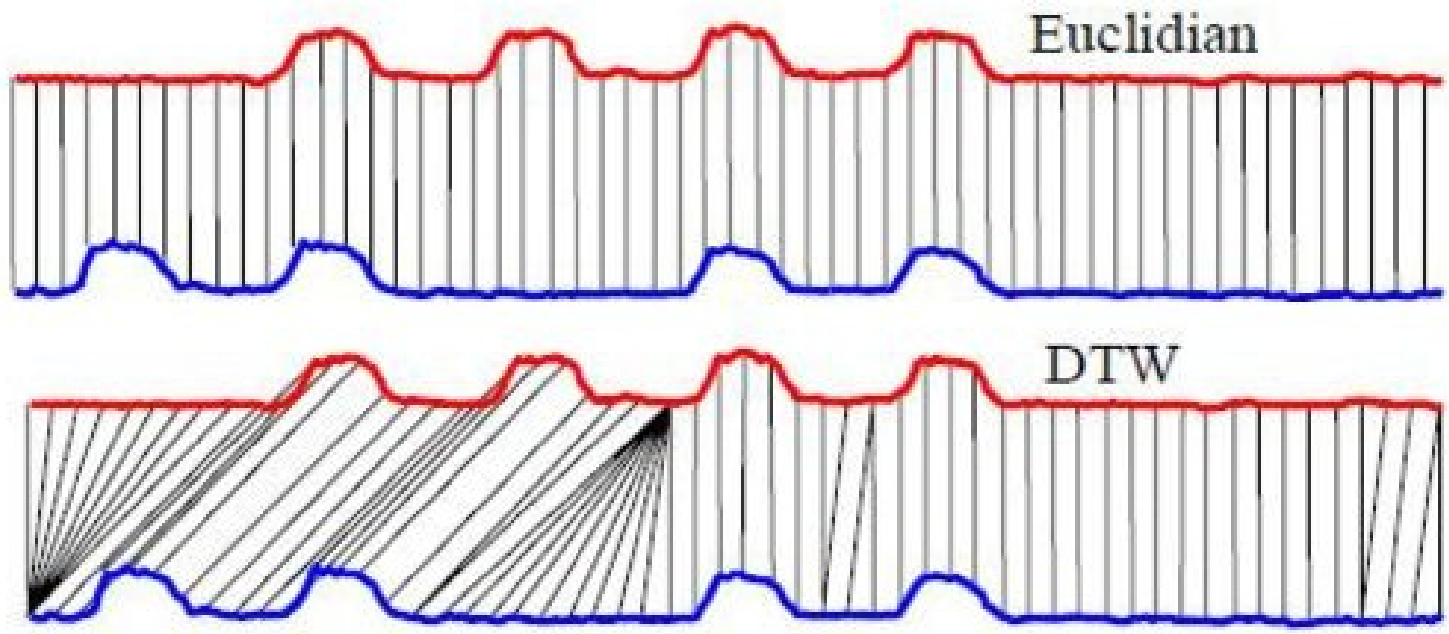
Jako algorytm klasyfikacyjny wykorzystana została metoda 1NN (metoda najbliższego sąsiada). W tej metodzie, klasa obserwacji testowej jest przewidywana na podstawie klasy najbliższej obserwacji ze zbioru uczącego. „Najbliższy” jest tutaj definiowany za pomocą pewnej miary odległości, takiej jak odległość euklidesowa. Główną zaletą 1NN jest jego prostota: nie wymaga żadnego treningu, a jedynie przechowywania całego zestawu danych uczących.



Porównywane metody

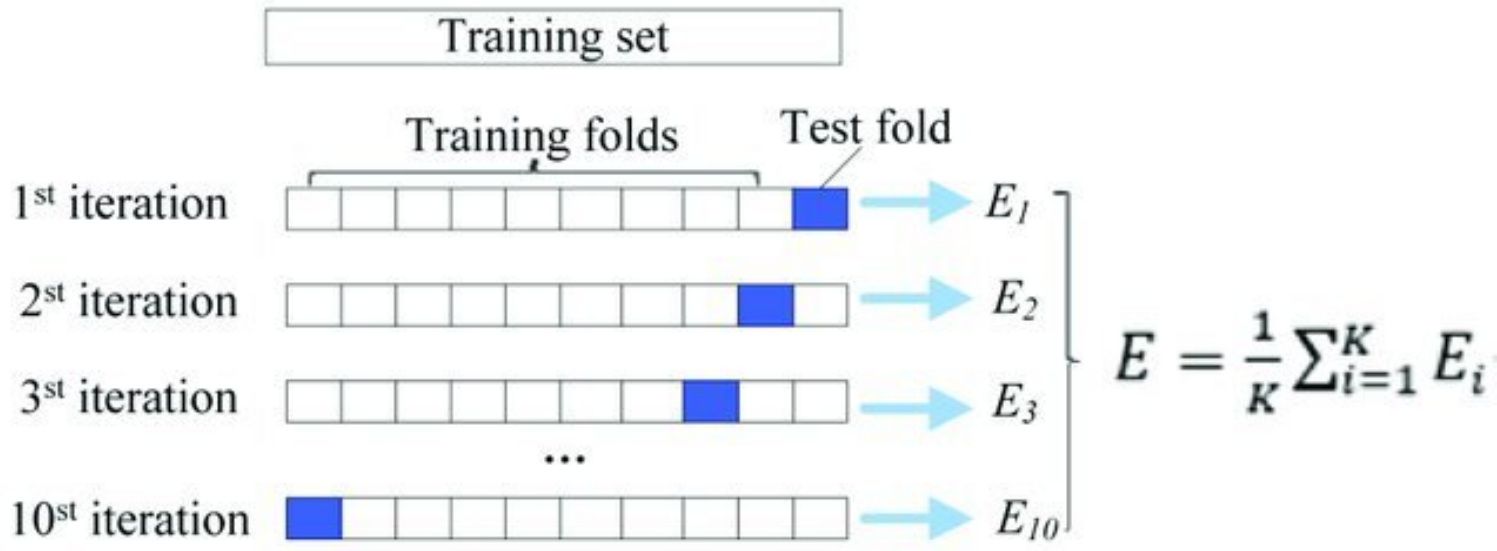
Jako odległość do porównań została wybrana odległość DTW. Odległość DTW (ang. *Dynamic Time Warping*) to technika służąca do mierzenia podobieństwa pomiędzy dwoma sekwencjami czasowymi, które mogą mieć różne długości. Jest to szczególnie przydatne w przypadku analizy szeregów czasowych, gdzie klasyczne metryki, takie jak odległość euklidesowa, mogą nie być odpowiednie. DTW działa poprzez „wykrzywianie" czasu w obu sekwencjach tak, aby optymalnie dopasować je do siebie. To „wykrzywianie" czasu jest realizowane poprzez znalezienie ścieżki z minimalnym kosztem pomiędzy początkiem, a końcem sekwencji, gdzie koszt jest mierzony jako suma różnic między dopasowanymi punktami w obu sekwencjach.

Porównywane metody



Porównywane metody

Dla każdego zbioru danych policzony został błąd klasyfikacji typu 10CV (ang. *10-fold cross-validation*) Błąd typu 10CV, znany też jako błąd walidacji krzyżowej 10-krotnej, jest to metoda oceny jakości modelu w uczeniu maszynowym. Polega na podziale danych na 10 równych części, a następnie na przeprowadzeniu 10 oddzielnych rund uczenia i testowania, za każdym razem używając innej części jako zestawu testowego, a pozostałych 9 jako zestawu treningowego.



Wyniki

W tabeli pokazano ocenę błędu klasyfikacji uzyskaną za pomocą metody 10CV. Wszystkie obliczenia wykonano w języku `Python`.

Wyniki

Zbiór danych	DTW	GEO	$\frac{\text{GEO-DTW}}{\text{DTW}}$
ArticularyWordRecognition		2,26	
AtrialFibrillation		70,00	
BasicMotions		1,25	
CharacterTrajectories	1,36	16,30	10,99
Cricket		1,11	
ECG	18,50	22,00	0,19
EigenWorms		13,11	
ERing		14,00	
EthanolConcentration		71,18	
Graz	37,14	35,00	-0,06
HandMovementDirection		67,08	
JapaneseVowels	2,03	41,87	19,62
Libras	8,61	31,94	2,71
NATOPS		34,44	
RacketSports		29,40	
RobotFailure_LP1	12,64	8,33	-0,34
RobotFailure_LP2	32,00	32,00	0,00
RobotFailure_LP3	29,00	26,50	-0,09
RobotFailure_LP4	10,08	18,03	0,79
RobotFailure_LP5	29,30	42,76	0,46
SelfRegulationSCP1		43,13	
SelfRegulationSCP2		53,95	
StandWalkJump		68,00	
Wafer	2,01	3,10	0,54

Literatura

- Górecki, T. Łuczak, M. (2015). Multivariate time series classification with parametric derivative dynamic time warping. *Expert Systems with Applications* 42:2305–2312.
- Pennec, X., Fillard, P., Ayache, N. (2006). A Riemannian framework for tensor computing. *International Journal of Computer Vision* 66(1):41–66.
- Shahbazi, M., Shirali, A., Aghajan, H., Nili, H. (2021). Using distance on the Riemannian manifold to compare representations in brain and in models. *Neuroimage* 239:118271.
- Sun, J., Yang, Y., Xiong, N.N., Dai, L., Peng, X. Luo, J. (2019). Complex network construction of multivariate time series using information geometry. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 49(1):107–122.

Dziękuję za uwagę!