Symulacja w Badaniach i Rozwoju
Vol. 7, No. 1-2/2016

Marcin MAZUREK
Wojskowa Akademia Techniczna, Wydział Cybernetyki
ul. Gen. S. Kaliskiego 2, 00-908 Warszawa
E-mail: marcin.mazurek@wat.edu.pl

Grupowanie trajektorii w analizie wyników badań klinicznych

1 Wstęp

Wśród gromadzonych danych klinicznych znajdują się wyniki badań pacjentów przeprowadzanych w pewnych odstępach czasu. Wynikami badań pacjentów, rozważanymi w pracy, są liczbowe wartości parametrów opisujących stan organizmu, na przykład: zawartość czerwonych krwinek, wapnia, magnezu itd. Z wynikami badań są związane daty ich przeprowadzenia, co pozwala na potraktować te dane jako wielowymiarowy szereg czasowy, opisujący trajektorię stanu zdrowia pacjenta: jej kształt wynika ze zmian chorobowych, podjętego leczenia i innych czynników.

Dostępność tego typu danych dla szerokiego zbioru pacjentów umożliwia wykorzystanie ich do wspomagania decyzji poprzez zastosowanie modeli predykcyjnych wykorzystujących modele i metody uczenia maszynowego. Duży wolumen wielowymiarowych danych, potęgowany użyciem mobilnych urządzeń monitorujących wybrane parametry stanu zdrowia, wymaga wyodrębnienia z tych danych wzorców, które mogą być przedmiotem analizy. Tak powstałe wzorce mogą zostać zwizualizowane i być przedmiotem analizy eksploracyjnej. Z drugiej strony, odzwierciedlenie przez indywidualny przypadek jednego z wyodrębnionych wzorców stanowi dodatkową zmienną objaśniającą w modelach predykcji.

Jedną z technik analizy danych jest grupowanie (klasteryzacja). Grupowanie jest metodą uczenia maszynowego bez nauczyciela. Wynikiem tej procedury jest wyodrębnienie grup obiektów w ten sposób, że obiekty należące do tej samej grupy są podobne do siebie, a powinny istotnie się różnić w przypadku przynależności do różnych grup. Grupowanie może być postrzegane jako metoda agregacji danych, ale również służy do identyfikacji anomalii oraz odkrycia typowych wzorców. Kluczowym elementem każdej procedury grupowania jest wybór miary odległości obserwacji od siebie, a ta z kolei zależy od sposobu reprezentacji i cech szeregu czasowego.

O ile zmienność pojedynczego parametru może być wizualnie oceniona i pogrupowana w typowe wzorce, w przypadku wielowymiarowych szeregów czasowych ich wizualizacja przedstawia zawsze pewien przekrój. Aby pogrupować pacjentów według charakterystyk zmian, należy zastosować metody automatycznego grupowania.

Klasteryzacja szeregów medycznych jest poprzedzona transformacją danych wejściowych do postaci regularnych bądź nieregularnych szeregów czasowych. Przygotowanie danych obejmuje: uzupełnienie brakujących danych, standaryzację zmiennych, interpolację wyników dla chwil, w których nie były wykonywane badania,

eliminacja wartości odstających. W dalszej kolejności należy dobrać liczbę skupień oraz miary odległości.

Opisana powyżej logika przetwarzania wyników danych medycznych jest dostępna w środowisku obliczeń statystycznych R, jednak jest rozproszona po wielu pakietach. W artykule opisano implementację i sposób wykorzystania autorskiego pakietu *medclust* , który udostępnia funkcjonalność grupowania szeregów czasowych.

Zaimplementowane w języku R procedury zostały wykorzystane pogrupowania trajektorii opisujących poziom PTH oraz Ca w przypadkach pooperacyjnej hipokalcemii.

2 Trajektorie wyników badań

Zakres informacyjny analizowanej bazy danych stanowią wyniki pomiarów jednego lub większej liczby parametrów opisujących stan zdrowia pacjenta, przeprowadzone wielokrotnie w przeciągu pewnego okresu.

Ze względu na regularność przeprowadzanych badań, możemy wyróżnić:

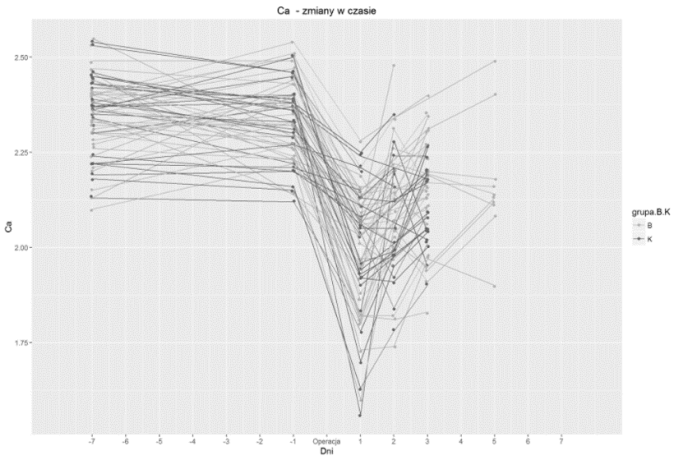
- szeregi czasowe regularne, jeżeli badanie przeprowadzane jest regularnie w tych samych odstępach czasu
- nieregularne szeregi czasowe, jeżeli odstępy pomiędzy kolejnymi badaniami są różne: w tym przypadku można dodatkowo określić, czy odstępy między konkretnymi numerami pomiarów (na przykład pomiędzy pierwszym i drugim) są takie same dla wszystkich pacjentów, czy też różnią się i są indywidualne dla każdego pacjenta. W zależności od występujących tu wzorców może zostać wykorzystana inny sposób przygotowania danych do analizy skupień.

Ze względu na umiejscowienie wyników badań w czasie możemy wyróżnić dwa podejścia:

- występowanie wyróżnionego zdarzenia, względem którego mierzymy daty, na przykład operacji, która mogła mieć miejsce dla każdego pacjenta w innym terminie. Przy tego typu analizie interesują nas przebiegi czasowe z indeksem wskazującym na położenie pomiaru względem wyróżnionego zdarzenia. Przykłady szeregów czasowych odniesionych do daty 0, która oznacza dzień operacji zostały zamieszczone na rysunku 1 oraz rysunku 2.
- wyróżnionym zdarzeniem jest data pierwszego badania, poszukujemy podobieństw pomiędzy szeregami, przy czym dopuszczamy swobodne ich przesuwanie w czasie.

W zależności od sposobu pozyskania danych, wyniki badań klinicznych mogą tworzyć

- szeroką tabelę analityczną, w której wiersze odpowiadają indywidualnemu pacjentowi, natomiast w kolumnach zapisywane są wyniki pomiarów (nazwa kolumny determinuje zarówno rodzaj pomiaru, jak również chwilę jego wykonania) . Ten sposób zapisu jest charakterystyczny dla „ręcznego" sposobu opracowywania danych \
- długą tabelę, w której każdy pomiar jest wprowadzany jako oddzielny wiersz, indeksowany przy pomocy trzech zmiennych: identyfikatora pacjenta, mierzonego parametru oraz chwili przeprowadzenia badania. Długie tabele wyników pomiarów generowane są maszynowo.



Rys. 1. Wyniki pomiaru poziomu wapnia u pacjentów przed i po operacji (ujemne indeksy oznaczają dni przed operacją)

Fig. 1. Individual patient Ca measurements before and after treatment (negative numbers are days before treatment)

W związku z powyższym, zachodzi potrzeba konwersji pomiędzy tego typu formatami zapisu danych z uwzględnieniem dodatkowej specyfiki szeregów czasowych oraz ich transformacji do uniwersalnej reprezentacji, która umożliwi wykorzystanie istniejących procedur analitycznych do przeprowadzania analizy skupień. W procesie przygotowania danych, dla każdego pacjenta należy skonstruować regularny, wielowymiarowy szereg czasowy indeksowany dyskretną wartością zmiennej czasu *t*, w którym uzupełnione zostaną brakujące wartości. Finalnie, po wstępnym przygotowaniu danych trajektoria wyników badań można zdefiniować jako:

$$X_i = \{X_i(t_l) \in R^p \; dla \; \; l = 1, \dots, L\} \tag{1}$$

gdzie

$$t_l < t_{l'}, \; dla \; \; l < l' \tag{2}$$

$$t_l \in T = \langle a, b \rangle, dla \;\; \; l = 1, \dots, L \tag{3}$$

$$X_i(t) \in R^p \; dla \; t \in T \tag{4}$$

3 Pakiet R i algorytmy grupowania szeregów czasowych

Otwarte środowisko obliczeń statystycznych R (https://www.r-project.org/) dostarcza gotowe procedury, które mogą zostać wykorzystane do grupowania trajektorii. W środowisku dostępne są pakiety, tworzone przez społeczność programistów, udostępniające przykładowe dane, klasy obiektów oraz procedury.

Podstawowe struktury do przechowywania szeregów w języku R są dostarczane w pakiecie *ts*, oraz pakiecie *zoo* który rozszerza zakres klas o takie, które umożliwiają przechowywanie nieregularny, wielowymiarowe szeregi czasowe.

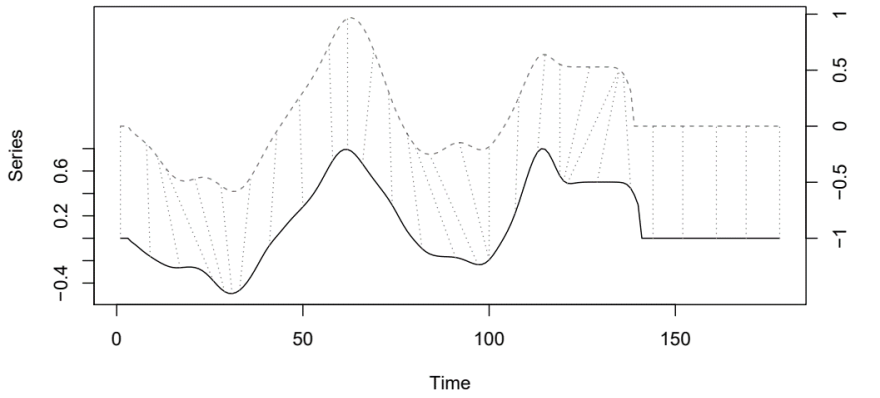
Przegląd pakietów, wraz z omówieniem algorytmów można odnaleźć w pozycji [1]. Najbardziej uniwersalnymi i kompleksowymi pakietami są:

- *dtwclust* - grupowanie wykorzystujące Dynamic Time Warping Distance.
- *pdc* – grupowanie wykorzystujące metodę Permutation Distribution Clustering

Pomocniczą rolę pełnią również pakiety:

- *TSclust* – miary odległości pomiędzy szeregami czasowymi
- *TSDist* - pakiet dostarcza algorytmy wyznaczenia miar odległości dla szeregów czasowych

Do grupowania szeregów czasowych mogą być wykorzystywane wszystkie znane algorytmy wyszukiwania skupień: k-średnich, hierarchiczne, rozmyte. Cechą rozróżniającą jest konstrukcja miary odległości pomiędzy szeregami czasowymi.



Rys. 2. Ilustracja przykładowego odwzorowania pomiędzy punktami szeregu czasowego
Fig. 2. Warping path example

Dynamic Time Warping Distance umożliwia znalezienie odległości pomiędzy nieregularnymi szeregami czasowymi o różnych długościach. Polega na znalezieniu odwzorowania (ang. *warping path*):

$$w = \left[w_1, w_2, \ldots w_K\right] \tag{5}$$

gdzie:

$$w_k = [p_k, q_k], \qquad p_k, q_k \in \{1, \ldots, L\}\, dla\ k = \{1,2, \ldots, K\} \tag{6}$$

spełniającego warunki:

- punktów krańcowych:

$$w_1 = [1,1]\,, \tag{7}$$

$$w_K = [L, L] \tag{8}$$

- monotoniczności:

$$p_k \leq p_{k+1}\ dla\ \ k = \{1,2,\dots,K-1\} \tag{9}$$

$$q_k \leq q_{k+1}\ dla\ \ k = \{1,2,\dots,K-1\} \tag{10}$$

- ciągłości:

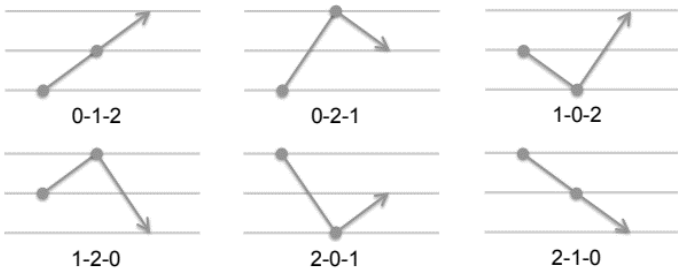
$$w_{k+1} - w_k \in \{[1,0],[0,1],[1,1]\ \} \tag{11}$$

Przykładowa ścieżka została przedstawiona na rysunku 2.

Odległość pomiędzy szeregami to minimalna wartość sumy pomiędzy punktami szeregu, uzyskana dla optymalnej ścieżki *w**:

$$c(w^*) = \sum_{k=1}^{K} D(p_k, q_k)\ .$$

Druga grupa algorytmów wykorzystuje miarę odległości pomiędzy szeregami zdefiniowana na rozkładzie permutacji (PD – *Permutation Distribution*)[4]. Szereg czasowy jest dzielony na sekwencje ustalonej długości *m*. Dla każdej sekwencji jest określany wzorzec uprządkowania: poprzez posortowanie obserwowanych wartości otrzymujemy *m*-elementowy ranking. Rozkład permutacji całego szeregu jest określony przez zliczenie relatywnej częstości występowania każdego z wzorców uporządkowania.



Rys. 3. Wzorce uporządkowania dla sekwencji 3 elementowych wyjętych z szeregu czasowego

Fig. 3. Ordinal patterns for 3-element sequences

Odległość pomiędzy dwoma szeregami czasowymi jest określana jako miara odległości pomiędzy uzyskanymi rozkładami permutacji. Ten sposób określenia miary nie zależy od skali pomiaru wartości zmiennych tworzących szereg.

4 Biblioteka *medclust*

Na potrzeby prowadzonych analiz został zbudowany pakiet R, który dostarcza funkcjonalności związane z przygotowaniem danych oraz samego grupowania danych, zorientowane na przetwarzanie wyników badań klinicznych. Biblioteka udostępnia przykładowe zbiory danych, procedury wstępnego przygotowania szeregów czasowych, obejmujące regularyzację uzupełnienie brakujących danych (imputację), normalizację oraz grupowania.

Pakiet został udostępniony publicznie w repozytorium GitHub i może zostać zaimportowany przez wywołanie jak poniżej:

```
install.packages('devtools')
devtools::install_github('mmazurek-wat/medclust')
```

Przykładowa sekwencja wywołania metod przedstawiona została na poniższym fragmencie kodu. Dla zachowania przejrzystości kodu pominięto określanie wartości parametrów wywołania procedur, pozostawiając domyślne.

```
library(medclust)
data(pthzoo)
ts<-preprocess(pthzoo)
mdc<-find_clusters_ts(ts,k=4)
showClusterStats(mdc)
```

W pierwszej linii następuje załadowanie do pamięci opisywanego pakietu. Po imporcie pakietów następuje załadowanie przykładowego zbioru danych o strukturze pokazanej na Rys. 4. Podstawową strukturą danych wykorzystywaną w pakiecie jest nazwana lista szeregów czasowych. Każdy element list reprezentuje dane jednego pacjenta. Wartość zmiennych wielowymiarowego szeregu czasowego zwizualizowane zostały w postaci kolumn, indeks czasu jest etykietą wiersza.

```
> head(pthzoo)
$`555866`
      P PTH   Mg   Ca
-7 1.92  54 0.84 2.55
-1 1.25  46 0.82 2.29
1    NA  15 0.78 1.92
2    NA  16 0.69 2.31
3    NA  11 0.66 2.16
5    NA  NA   NA   NA
6    NA  NA   NA   NA

$`574352`
      P PTH   Mg   Ca
-7 1.83  53 0.73 2.13
-1 1.19  36   NA 2.12
1    NA  20 0.71 1.83
2    NA  NA 0.55 2.28
3    NA  16 0.60 2.04
5    NA  NA   NA   NA
6    NA  NA   NA   NA

```

Rys. 4. Przykładowe dane wejściowe w formacie nazwanej listy wielowymiarowych szeregów czasowych z brakami danych

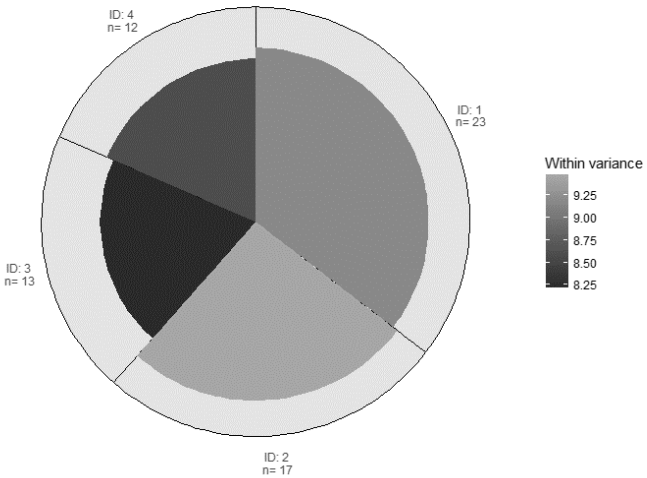
Fig. 4. Example input data stored as named list with missing values

W kolejnej linii zaprezentowanego kodu następuje transformacja danych wejściowych do postaci umożliwiającej wyznaczenie miary odległości pomiędzy szeregami. Procedura *preprocess* przykrywa logikę związaną z wszystkimi etapami wstępnego przygotowania danych. Dla zaprezentowanego przykładowego wywołania uzyskujemy dane:

		P	PTH	Mg	Ca
2	-7	0.7870968	0.7208653	0.5517241	1.0000000
3	-6	0.7150538	0.7022563	0.5459770	0.9562290
4	-5	0.6430108	0.6836474	0.5402299	0.9124579
5	-4	0.5709677	0.6650384	0.5344828	0.8686869
6	-3	0.4989247	0.6464294	0.5287356	0.8249158
7	-2	0.4268817	0.6278204	0.5229885	0.7811448
8	-1	0.3548387	0.6092114	0.5172414	0.7373737
9	0	0.3548387	0.3928821	0.4827586	0.5505051
10	1	0.3548387	0.1765527	0.4482759	0.3636364
11	2	0.3548387	0.1905094	0.2931034	0.7575758
12	3	0.3548387	0.1207258	0.2413793	0.6060606
13	4	0.3548387	0.1207258	0.2413793	0.6060606
14	5	0.3548387	0.1207258	0.2413793	0.6060606
15	6	0.3548387	0.1207258	0.2413793	0.6060606

Rys. 5. Przygotowane dane po procesie wstępnej transformacji
Fig. 5.Data after preprocessing

W kolejnej linii procedura *find_clusters_ts* przeprowadza proces analizy skupień, której zagregowane wyniki przedstawione są na rysunku 6 na wykresie uzyskanym za pomocą `showClusterStats`.



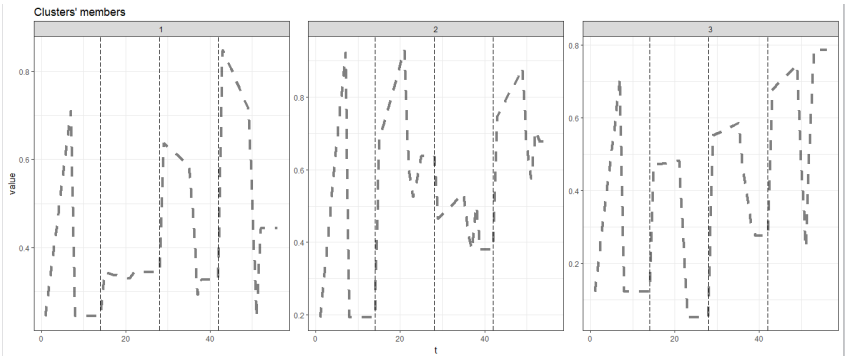
Rys. 6. Wizualizacja zagregowanych statystyk procesu klasteryzacji
Fig. 6. Clustering outcome visualization

Wizualizacja umożliwia ocenę liczebności skupień oraz ich wewnętrznego zróżnicowania.

5 Wyniki grupowania trajektorii

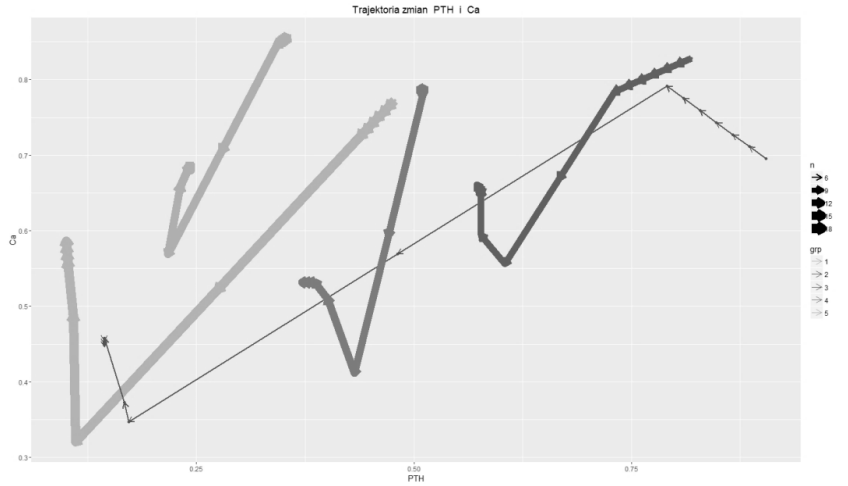
Wynikiem grupowania trajektorii są typowe przebiegi zmienności monitorowanych parametrów, za które można uważać centroidy wynikowych skupień. Zostały one przedstawione na rysunku 7.

Każdy z szeregów czasowych składał się z czterech parametrów. Przedstawione zostały trzy wybrane centroidy.



Rys. 7. Uzyskane środki skupień reprezentujące typowe trajektorie

Fig. 7. Outcome centroids representing typical trajectories



Rys. 8. Wizualizacje przebiegu dwóch parametrów uzyskane na podstawie pogrupowanych wyników badań

Fig. 8. Visualization of two-dimensional trajectories based on clustering procedure outcome

Biblioteka została wykorzystana w badaniach wpływu stosowania alfakalcydolu w okresie przedoperacyjnym na częstość występowania i przebieg pooperacyjnej hipokalcemii u pacjentów poddanych operacji wycięcia gruczołu tarczowego ze względu na raka tarczycy. Na rysunku 8 przedstawiono przebiegi dwóch z czterech badanych parametrów.

Przeprowadzając grupowanie trajektorii uzyskano grupy pacjentów o podobnych przebiegach zmienności badanych parametrów. W dalszej kolejności przeprowadzono analizę, czy sposób zmienności jest powiązany z odczuwanymi dolegliwościami.

8 Wnioski końcowe

Przedstawiona w artykule biblioteka dostarcza gotowe metody do przekształcania danych medycznych do formatu regularnych szeregów czasowych oraz grupowania wynikowych przebiegów. Ze względu na relatywnie złożone struktury danych, istniejące pakiety z algorytmami grupowania szeregów czasowych wymagają nakładu pracy na przygotowanie danych do odpowiedniego formatu, a następnie wizualizacji wybranych wyników. Zaprojektowany pakiet miał za zadanie maksymalnie uprościć ścieżkę implementacji od pobrania danych do wizualizacji skupień, przy zachowaniu możliwości parametryzacji dla zaawansowanych analityków.

Zaimplementowane procedury mogą zostać wykorzystane do budowy interaktywnego środowiska analizy skupień, wykorzystującego platformę R Shiny bądź do tworzenia statycznych opracowań wyników badań klinicznych. Implementacja w środowisku R otwiera drogę do możliwości zastosowania procedur w środowisku przetwarzania danych masowych (Big Data).

Należy zauważyć, że zastosowanie metod grupowania wielowymiarowych szeregów czasowych nie ogranicza się do wyników badań klinicznych. Przedstawiona technika analizy danych może zostać wykorzystana do analizy zachowań konsumenckich w kontekście zmian abonamentu czy stopnia utylizacji usługi.

Literatura

1. Sarda-Espinosa A.: *Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package*
2. Momtero P., Vilar J. A.: TSclust: An R Package for Time Series Clustering. *Journal of Statistical Software*, Vol. 62, 2014, www.jstatsoft.org
3. Leisch F.: *Creating R Packages: A Tutorial*. https://cran.r-project.org/doc/contrib/Leisch-CreatingPackages.pdf
4. Brandmaier A.M.: pdc: An R Package for Complexity-Based Clustering of Time-Series. *Journal of Statistical Software*, Vol. 67, 2015, www.jstatsoft.org
5. Ordóñez P, desJardins M, Feltes C, Lehmann CU, Fackler J.: Visualizing Multivariate Time Series Data to Detect Specific Medical Conditions, *AMIA Annual Symposium Proceedings*, 2008;2008:530-534

Streszczenie

Wyniki badań klinicznych mogą tworzyć wielowymiarowe szeregi czasowe, które opisują zmiany w czasie istotnych parametrów opisujących stan zdrowia i kondycję pacjenta. Analiza tego typu danych polega na wyodrębnieniu typowych przebiegów -

trajektorii w procesie analizy skupień. Klasteryzacja szeregów medycznych wiąże się z transformacją danych wejściowych: regularyzacją szeregu czasowego, uzupełnieniem brakujących danych, standaryzacją zmiennych. W dalszej kolejności należy dobrać liczbę skupień oraz wykonać grupowanie metodą k-średnich, DTW, PDC lub inną. Te algorytmy są dostępne w otwartych środowiskach obliczeń statystycznych, jednak aby ułatwić analitykom ich zastosowanie, został zbudowany pakiet *medclust*, który dostarcza wysokopoziomowych procedur, domyślnie sparametryzowanych do wyszukiwania skupień.

Słowa kluczowe:

Clustering trajectories in clinical researches

Summary

Clinical researches often involves measuring time-varying parameters of body condition, which forms multidimensional time-series. Typical, representative trajectories can be extracted with clustering algorithms. In order to apply clustering algorithms, raw data has to be preprocessed and this includes regularization of time series, imputation of missing values, values standardization. Next, one of time-series clustering can be applied: Dynamic Time Warping or Permutation Distribution Clustering. These algorithms are already available in open environments for statistical computing like R. In order to facilitate application of the clustering algorithms to the clinical reasarch data, new R package *medclust* was implemented. It provides analysts with ready-to-use high-level procedures with predefined set of parameters values to analyze clinical trajectories data.

Keywords: