BIULETYN INSTYTUTU AUTOMATYKI I ROBOTYKI
NR 23, 2006

Grupowanie sekwencji czasowych

Tomasz PAŁYS

Zakład Automatyki, Instytut Teleinformatyki i Automatyki WAT,
ul. Kaliskiego 2, 00-908 Warszawa

STRESZCZENIE: W artykule przedstawiono metody grupowania sekwencji czasowych. Oryginalność tego problemu polega na tym, że grupowane elementy stanowią sekwencję, a uzyskane grupy mogą stanowić tylko segmenty sekwencji. Przedstawiono dwie metody grupowania sekwencji czasowych. Pierwsza metoda umożliwia uzyskanie grup rozłącznych. W wyniku zastosowania drugiej metody otrzymujemy grupy, które mogą się na siebie nakładać.

SŁOWA KLUCZOWE: grupowanie sekwencji, grupowanie z nakładaniem

1. Wprowadzenie

Celem grupowania jest podział zbioru obiektów na grupy (skupienia) złożone z obiektów jednorodnych bądź podobnych. Wszystkie znane metody grupowania nie uwzględniają kolejności punktów w sekwencji. W artykule zostaną przedstawione dwie metody grupowania sekwencji punktów:

- hierarchiczna metoda grupowania rozłącznego,
- hierarchiczna metoda grupowania z nakładaniem.

Do skonstruowania metody grupowania sekwencji czasowych przyjęto jako wyjściową metodę hierarchiczną [1]. Przyjmuje się, że dane wejściowe procesu grupowania stanowi zbiór punktów: $O = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_t, ..., \mathbf{o}_T\}$, gdzie: T – liczba punktów. Proces grupowania metodą hierarchiczną odbywa się przez kolejne łączenie położonych najbliżej siebie grup. Grupowanie kończy się po uzyskaniu jednej grupy złożonej ze wszystkich punktów. Taki sposób postępowania prowadzi do utworzenia drzewa grupowania, które umożliwia uzyskanie podziału na żądaną liczbę grup albo grup o zadanych właściwościach. Aby ocenić jakość grupowania, można posłużyć się współczynnikiem korelacji grupowania (ang. *cophenetic correlation coefficient*) lub współczynnikiem

niezgodności grupowania (ang. *inconsistency coefficient*) [4], [5].

2. Grupowanie sekwencji czasowych

Przedstawione niżej metody grupowania sekwencji czasowych bazują na hierarchicznej metodzie grupowania rozłącznego punktów, której opis można znaleźć w [3], [2], [1]. Dane wejściowe procesu grupowania stanowi sekwencja punktów O (a nie zbiór punktów). Dopasowanie do konkretnego zadania jest możliwe poprzez odpowiedni dobór metryki, czyli sposobu określania odległości pomiędzy punktami przestrzeni cech oraz odpowiedni dobór sposobu określenia odległości pomiędzy poszczególnymi grupami [2], [3], [1]. Opracowane metody grupowania sekwencji czasowych, w odróżnieniu od metody bazowej, polegają na łączeniu tylko grup sąsiednich. Dwie grupy nazwano sąsiednimi pod warunkiem, że w jednej z nich istnieje punkt, który w drugiej grupie ma swój poprzednik albo następnik (w sekwencji).

2.1. Hierarchiczna metoda grupowania rozłącznego sekwencji punktów

Niech $d(\mathbf{o}_n, \mathbf{o}_z)$ oznacza odległość pomiędzy punktami w przestrzeni D – wymiarowej, gdzie:

$$\mathbf{o}_n = \begin{bmatrix} \mathbf{o}_{n_1} \\ \cdots \\ \mathbf{o}_{n_D} \end{bmatrix} \in \mathbb{R}^D, \ \mathbf{o}_z = \begin{bmatrix} \mathbf{o}_{z_1} \\ \cdots \\ \mathbf{o}_{z_D} \end{bmatrix} \in \mathbb{R}^D. \qquad (1)$$

Przez $\mathbf{G}_n$ oznaczono grupę o numerze n, T_n – jej liczebność ($n = 1, ..., N$) a $\mathbf{o}_z^{(n)}, z \in \{1,...,T_n\}$ – element grupy $\mathbf{G}_n$ o indeksie z, a $dist(\mathbf{G}_n, \mathbf{G}_k)$ niech oznacza odległość pomiędzy grupą $\mathbf{G}_n$ a $\mathbf{G}_k$.

Wstępnie przyjmuje się, że każdy punkt stanowi oddzielną grupę. Punkt $\mathbf{o}_1$ grupę $\mathbf{G}_1$, punkt $\mathbf{o}_2$ grupę $\mathbf{G}_2$, itd. Na tej podstawie należy wyznaczyć odległości pomiędzy punktem sekwencji a jego następnikiem. Istotne są tylko odległości pomiędzy sąsiednimi punktami a w konsekwencji przyjętego założenia pomiędzy sąsiednimi grupami. Następnie, według jednego wybranego sposobu określania odległości pomiędzy grupami, należy wyznaczyć wektor odległości $dist(\mathbf{G}_n, \mathbf{G}_{n+1})$ pomiędzy grupami $\mathbf{G}_n$ i $\mathbf{G}_{n+1}$. W tej sytuacji każdy element wektora odpowiada odległości pomiędzy grupą a jej następnikiem: pierwszy element – odległość pomiędzy grupą $\mathbf{G}_1$ a grupą $\mathbf{G}_2$, drugi element – odległość pomiędzy grupą $\mathbf{G}_2$ a grupą $\mathbf{G}_3$ itd. Stosowne zależności, niezbędne

do wyznaczenia odległości pomiędzy punktami oraz grupami, zostały przedstawione w [3].

Dane wejściowe jednego kroku grupowania stanowi wektor odległości $dist(\mathbf{G}_r, \mathbf{G}_s)$ pomiędzy sąsiednimi grupami. Po wyszukaniu pary sąsiednich grup $(\mathbf{G}_p, \mathbf{G}_q)$, które są położone najbliżej siebie, następuje połączenie ich w jedną grupę $\mathbf{G}_p \cup \mathbf{G}_q$ i zostaje określony nowy wektor odległości. Zmianie ulegają jedynie odległości do sąsiadów grupy $\mathbf{G}_p \cup \mathbf{G}_q$. Wartości $dist(\mathbf{G}_r, \mathbf{G}_p \cup \mathbf{G}_q)$ wyznacza się na podstawie znanych wartości: $dist(\mathbf{G}_r, \mathbf{G}_p)$, $dist(\mathbf{G}_r, \mathbf{G}_q)$ oraz $dist(\mathbf{G}_p, \mathbf{G}_q)$. Grupowanie kończymy po uzyskaniu jednej grupy, złożonej ze wszystkich punktów. Algorytm grupowania rozłącznego sekwencji punktów metodą hierarchiczną przedstawiono poniżej.

Struktury danych

Stałe:

$O = (\mathbf{o}_1, ..., \mathbf{o}_T)$ – sekwencja punktów obserwacji,
T – liczba elementów sekwencji O.

Zmienne:

$\mathbf{Y}$ – kolumnowy wektor odległości składający się z $T - 1$ elementów,
$\mathbf{W}$ – macierz pomocnicza składająca się z T wierszy i 2 kolumn, kolumna numer 1 macierzy $\mathbf{W}$ będzie zawierała indeksy grup, a kolumna numer 2 liczbę elementów grupy,
k – numer kolejnego etapu grupowania,
N – liczba grup w etapie grupowania k,
$\mathbf{Z}$ – macierz grupowania, składająca się z 4 kolumn i kolejno w każdym etapie zwiększanej liczbie wierszy,
i, j – indeksy sąsiednich grup najbliżej siebie położonych,
ν – odległość pomiędzy sąsiednią parą grup najbliżej siebie położoną.

Obliczenia wstępne

Przyjmuje się, że każdy punkt stanowi oddzielną grupę. Punkt $\mathbf{o}_1$ grupę $\mathbf{G}_1$, punkt $\mathbf{o}_2$ grupę $\mathbf{G}_2$, itd. Na tej podstawie należy wyznaczyć odległości pomiędzy punktem sekwencji a jego następnikiem.

Następnie, według jednego wybranego sposobu określania odległości pomiędzy grupami, należy wyznaczyć wektor odległości $\mathbf{Y} = [dist(\mathbf{G}_n, \mathbf{G}_{n+1})]_{T-1 \times T}$, gdzie: $n = 1, ..., T - 1$. Każdy element wektora odpowiada odległości pomiędzy grupą a jej następnikiem: pierwszy element – odległość pomiędzy grupą $\mathbf{G}_1$ a grupą $\mathbf{G}_2$, drugi element – odległość pomiędzy grupą $\mathbf{G}_2$ a grupą $\mathbf{G}_3$ itd. Wiersze macierzy pomocniczej $\mathbf{W}$ będą opisywać grupy. Pierwszy element wiersza, to indeks grupy, a drugi – liczba elementów grupy. Ponieważ na początku jest $N = T$ grup, dlatego macierz $\mathbf{W}$

ma T wierszy. Wygląda ona następująco:

$$\mathbf{W} = \begin{bmatrix} 1 & 1 \\ \cdots & \cdots \\ T & 1 \end{bmatrix}.$$

Ostatnia czynność etapu wstępnego, to określenie wartości zmiennej, w której będzie przechowywany kolejny numer etapu grupowania $k := 0$.

Etap grupowania

Zwiększamy numer etapu grupowania $k := k + 1$. Znajdujemy dwie sąsiednie grupy położone najbliżej siebie. Sprowadza się do wyznaczenia najmniejszego elementu v w wektorze odległości oraz jego numeru wiersza i:

$$i = \arg\min_{n=1,\ldots,N-1} \mathbf{Y}(n), \tag{2}$$

$$v = \mathbf{Y}(i) \tag{3}$$

gdzie:

$v = dist(\mathbf{G}_{\mathbf{W}(i,1)}, \mathbf{G}_{\mathbf{W}(i+1,1)})$ – najmniejsza odległość pomiędzy sąsiednimi grupami w kroku k, jest to odległość pomiędzy grupą o indeksie $\mathbf{W}(i, 1)$ a jej następnikiem, czyli grupą o indeksie $\mathbf{W}(i + 1, 1)$.

Grupy o indeksie $\mathbf{W}(i, 1)$ i $\mathbf{W}(i + 1, 1)$ łączy się w jedną grupę. Zmniejszamy liczbę grup $N := N – 1$, a wyniki grupowania zapisujemy jako nowy wiersz macierzy $\mathbf{Z}$:

$\mathbf{Z}(k, 1) := \mathbf{W}(i, 1)$;
$\mathbf{Z}(k, 2) := \mathbf{W}(i + 1, 1)$;
$\mathbf{Z}(k, 3) := v$;
$\mathbf{Z}(k, 4) := N$.

Zgodnie z wybraną wcześniej metodą grupowania, uaktualniamy odległości pomiędzy połączoną grupą a pozostałymi grupami, tzn. $\mathbf{Y}(z) = dist(\mathbf{G}_{\mathbf{W}(z,1)}, \mathbf{G}_{\mathbf{W}(i,1)} \cup \mathbf{G}_{\mathbf{W}(i+1,1)})$, dla $z = 1, ..., i – 1, i + 1, ..., N$. Dodatkowo należy usunąć z macierzy $\mathbf{Y}$ wiersz o numerze i. Uaktualniamy macierz pomocniczą $\mathbf{W}$, indeks nowo utworzonej grupy oraz liczbę jej elementów:

$\mathbf{W}(i, 1) := T + k$,
$\mathbf{W}(i, 2) := \mathbf{W}(i, 2) + \mathbf{W}(i + 1, 2)$.

Należy jeszcze usunąć wiersz numer $i + 1$ z macierzy $\mathbf{W}$.

Etap końcowy algorytmu

Kolejne etapy grupowania powtarzamy do momentu, aż uzyskamy jedną grupę, czyli gdy $N = 1$. W wyniku otrzymujemy macierz $\mathbf{Z}$, która opisuje drzewo grupowania.

W przypadku metody grupowania sekwencji czasowych należy wyznaczyć $T-1$ liczb reprezentujących odległości pomiędzy sąsiednimi punktami sekwencji i wykonać $T-1$ kroków grupowania dla wyznaczenia drzewa grupowania. Przykład grupowania sekwencji punktów w cztery grupy przedstawiono na rys. 1. Uzyskano następujące grupy:

- 1, 2, 3, 4, 5;
- 6, 7, 8, 9, 10;
- 11, 12;
- 13, 14, 15.

Na rys. 2 przedstawiono drzewo grupowania, na podstawie którego dokonano podziału w cztery grupy.

2.2. Hierarchiczna metoda grupowania z nakładaniem sekwencji punktów

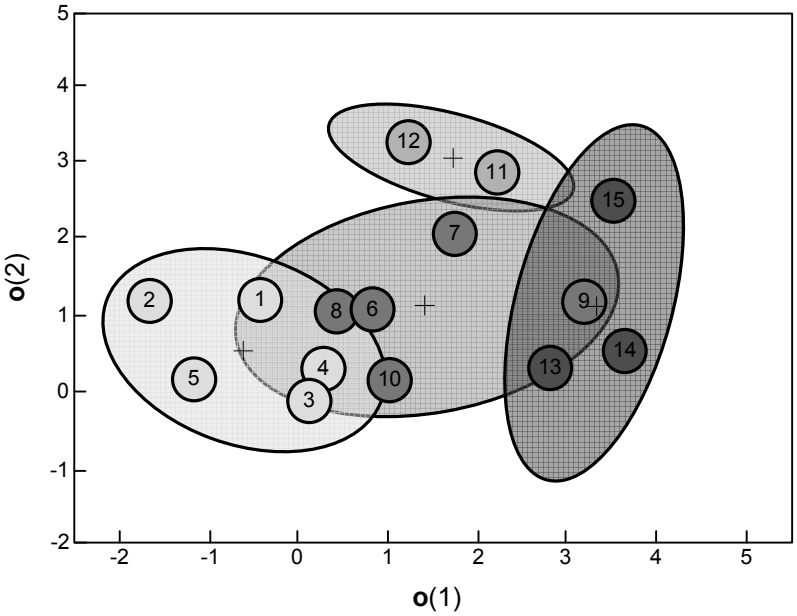
Przedstawiony poniżej algorytm umożliwia grupowanie sekwencji punktów z nakładaniem (wynikiem grupowania nie muszą być zbiory rozłączne). Grupowanie rozłączne sekwencji punktów metodą hierarchiczną polega na łączeniu w każdym kroku grupowania dwóch sąsiednich grup, które są położone najbliżej siebie. Na każdym etapie grupowania uzyskuje się grupy rozłączne. W wielu przypadkach korzystniej jest zrezygnować z tego założenia i dopuścić możliwość nakładania się grup.

Każdy etap grupowania z nakładaniem polega na znalezieniu dwóch par sąsiednich grup, które leżą najbliżej siebie. Wynik poszukiwań to 2 pary sąsiednich grup: $(\mathbf{G}_m, \mathbf{G}_n)$ i $(\mathbf{G}_p, \mathbf{G}_q)$, $m < n$, $p < q$, o odległościach l_{mn} i l_{pq}, przy czym $l_{mn} < l_{pq}$. Należy rozważyć następujące przypadki:

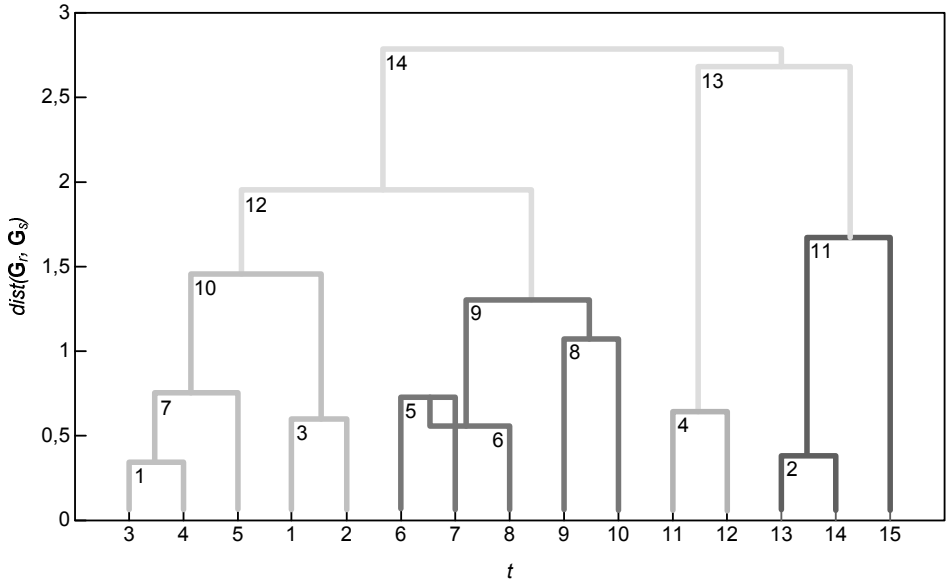
1) istnieje grupa wchodząca w skład obu par, czyli: ($m = q$ albo $n = q$) albo ($m = p$ albo $n = p$),
2) nie zachodzi pierwszy przypadek.

Wystąpienie pierwszego przypadku oznacza, że grupa $\mathbf{G}_p$ albo $\mathbf{G}_q$, jest położona względnie blisko grup $\mathbf{G}_m$ i $\mathbf{G}_n$. W tym przypadku tworzy się dwie nowe grupy: $\mathbf{G}_m \cup \mathbf{G}_n$ oraz $\mathbf{G}_p \cup \mathbf{G}_q$. Zajście drugiego przypadku oznacza, że w miejsce grupy $\mathbf{G}_m$ tworzy się tylko grupę $\mathbf{G}_m \cup \mathbf{G}_n$.

Na każdym etapie grupowania, oprócz połączenia dwóch sąsiednich grup w jedną, zapewniono dodatkowe połączenie jednej z nich do swojego sąsiada. Zasady łączenia grup są podobne jak w przypadku metody opisanej powyżej. Istotą algorytmu jest łączenie tylko grup sąsiednich, w wyniku czego uwzględniona zostaje kolejność punktów w sekwencji. Kolejne etapy algorytmu grupowania z nakładaniem sekwencji punktów przedstawiono poniżej.



Rys. 1. Grupowanie sekwencji punktów w cztery grupy rozłączne



Rys. 2. Drzewo grupowania

Struktury danych

Stałe:

$O = (\mathbf{o}_1, ..., \mathbf{o}_T)$ – sekwencja punktów obserwacji,
T – długość sekwencji O.

Zmienne:

- **Y** – kolumnowy wektor odległości składający się z $T-1$ elementów,
- **W** – macierz pomocnicza składająca się z T wierszy i 2 kolumn, kolumna numer 1 macierzy **W** będzie zawierała indeksy grup a kolumna numer 2 liczbę elementów grupy,
- k – numer kolejnych etapów grupowania,
- N – liczba grup w etapie grupowania k,
- k_d – liczba wykonanych nałożeń,
- **Z** – macierz grupowania, składająca się z 4 kolumn i kolejno w każdym etapie zwiększanej liczbie wierszy,
- i_1, j_1 – indeksy najbliżej siebie położonych sąsiednich grup,
- i_2, j_2 – indeksy drugiej w kolejności pary sąsiednich grup najbliżej siebie położonych,
- ν_1 – odległość pomiędzy sąsiednią parą grup najbliżej siebie położonych,
- ν_2 – odległość pomiędzy drugą w kolejności parą sąsiednich grup najbliżej siebie położonych.

Obliczenia wstępne

Obliczenia wstępne przebiegają tak samo, jak w przypadku hierarchicznej metody grupowania rozłącznego. Dodatkowo należy ustalić liczbę nałożeń $k_d := 0$.

Etap grupowania

Zwiększamy numer etapu grupowania $k := k + 1$. Znajdujemy najbliżej położoną siebie sąsiednią parę grup:

$$i_1 = \arg\min_{n=1,...,N-1} \mathbf{Y}(n), \tag{4}$$

$$v_1 = \mathbf{Y}(i_1), \tag{5}$$

gdzie:

$v_1 = dist\left(\mathbf{G}_{\mathbf{W}(i_1,1)}, \mathbf{G}_{\mathbf{W}(i_1+1,1)}\right)$ – najmniejsza odległość pomiędzy sąsiednią parą grup w kroku k.

Grupy o indeksie $\mathbf{W}(i_1, 1)$ i $\mathbf{W}(i_1 + 1, 1)$ łączymy w jedną grupę. Zmniejszamy liczbę grup $N := N - 1$ a wynik grupowania zapisujemy w macierzy **Z**:

$\mathbf{Z}(\mathrm{k} + \mathrm{k}_d, 1) := \mathbf{W}(\mathrm{i}_1, 1)$,
$\mathbf{Z}(k + k_d, 2) := \mathbf{W}(i_1 + 1, 1)$,
$\mathbf{Z}(k + k_d, 3) := v_1$,
$\mathbf{Z}(k + k_d, 4) := N$.

Następnie odnajdujemy drugą w kolejności, sąsiednią parę grup najbliżej siebie położonych:

$$i_2 = \arg\min_{n=1,\ldots,i_1-1,i_1+1,\ldots,N-1} \mathbf{Y}(n) \tag{6}$$

$$v_2 = \mathbf{Y}(i_2) \tag{7}$$

gdzie:

$v_2 = dist\left(\mathbf{G}_{\mathbf{W}(i_2,1)}, \mathbf{G}_{\mathbf{W}(i_2+1,1)}\right)$ – odległość pomiędzy drugą w kolejności parą sąsiednich grup w etapie k.

Jeżeli $i_1 = i_2 + 1$, to grupę o indeksie $\mathbf{W}(i_1, 1)$ łączymy z grupą o indeksie $\mathbf{W}(i_2, 1)$. Występuje tu zjawisko nakładania się grup. Elementy grupy o indeksie $\mathbf{W}(i_2, 1)$ będą występować co najwyżej w dwóch grupach, co zostaje zapisane następująco: $k_d := k_d + 1$. Uaktualniamy macierz $\mathbf{Z}$:

$\mathbf{Z}(k + k_d, 1) := \mathbf{W}(i_1, 1)$,
$\mathbf{Z}(k + k_d, 2) := \mathbf{W}(i_2, 1)$,
$\mathbf{Z}(k + k_d, 3) := v_2$,
$\mathbf{Z}(k + k_d, 4) := N$,

oraz macierz pomocniczą $\mathbf{W}$, czyli indeks nowo utworzonej grupy oraz liczbę jej elementów:

$\mathbf{W}(i_2, 1) := T + k + k_d$,
$\mathbf{W}(i_2, 2) := \mathbf{W}(i_2, 2) + \mathbf{W}(i_1, 2)$.

Jeżeli $i_1 + 1 = i_2$, to grupę o indeksie $\mathbf{W}(i_1 + 1, 1)$ łączymy z grupą o indeksie $\mathbf{W}(i_2 + 1, 1)$. Występuje tu zjawisko nakładania się grup. Elementy grupy o indeksie $\mathbf{W}(i_2 + 1, 1)$ będą występować co najwyżej w dwóch grupach, co zapisujemy: $k_d := k_d + 1$. Uaktualniamy macierz $\mathbf{Z}$:

$\mathbf{Z}(k + k_d, 1) := \mathbf{W}(i_1 + 1, 1)$,
$\mathbf{Z}(k + k_d, 2) := \mathbf{W}(i_2 + 1, 1)$,
$\mathbf{Z}(k + k_d, 3) := v_2$,
$\mathbf{Z}(k + k_d, 4) := N$.

oraz macierz pomocniczą $\mathbf{W}$, czyli indeks nowo utworzonej grupy oraz liczbę elementów grupy:

$\mathbf{W}(i_2 + 1, 1) := T + k + k_d$,
$\mathbf{W}(i_2 + 1, 2) := \mathbf{W}(i_2 + 1, 2) + \mathbf{W}(i_1 + 1, 2)$.

Zgodnie z wybranym sposobem określania odległości pomiędzy grupami [3] należy uaktualnić odległości pomiędzy nowo utworzoną grupą a pozostałymi grupami i usunąć z macierzy $\mathbf{Y}$ wiersz i_1. Na koniec etapu uaktualniamy

macierz pomocniczą **W**, czyli indeks nowo utworzonej grupy i liczbę jej elementów. Jeżeli nastąpiło nałożenie dwóch grup, to:

$$\mathbf{W}(i_1, 1) := T + k + k_d - 1,$$

w przeciwnym przypadku:

$$\mathbf{W}(i_1, 1) := T + k + k_d.$$

Liczba elementów nowo utworzonej grupy jest równa:

$$\mathbf{W}(i_1, 2) := \mathbf{W}(i_1, 2) + \mathbf{W}(i_1 + 1, 2),$$

Na koniec usuwamy wiersz numer $i_1 + 1$ z macierzy **W**.

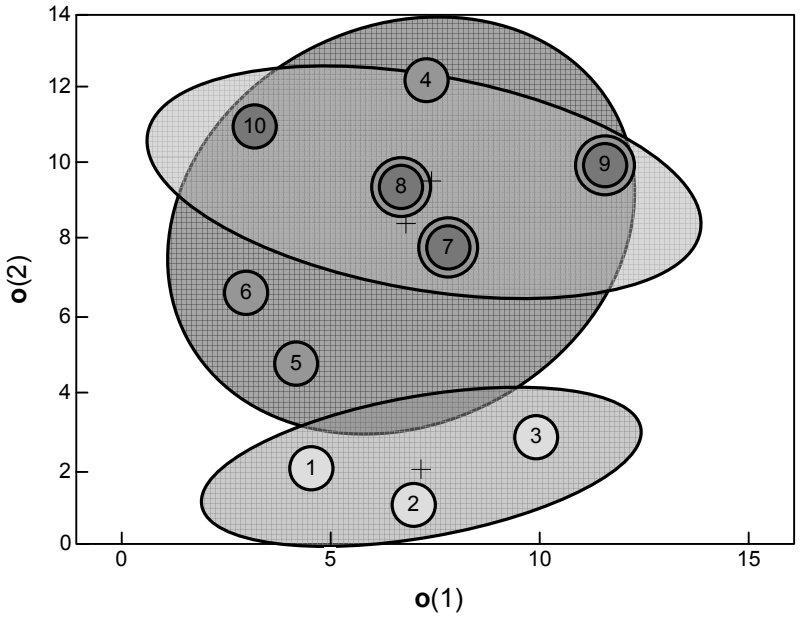
Etap końcowy algorytmu

Kolejne etapy grupowania powtarzamy do momentu, aż uzyskamy jedną grupę, czyli gdy $N = 1$. W wyniku otrzymujemy macierz **Z**, która opisuje drzewo grupowania.

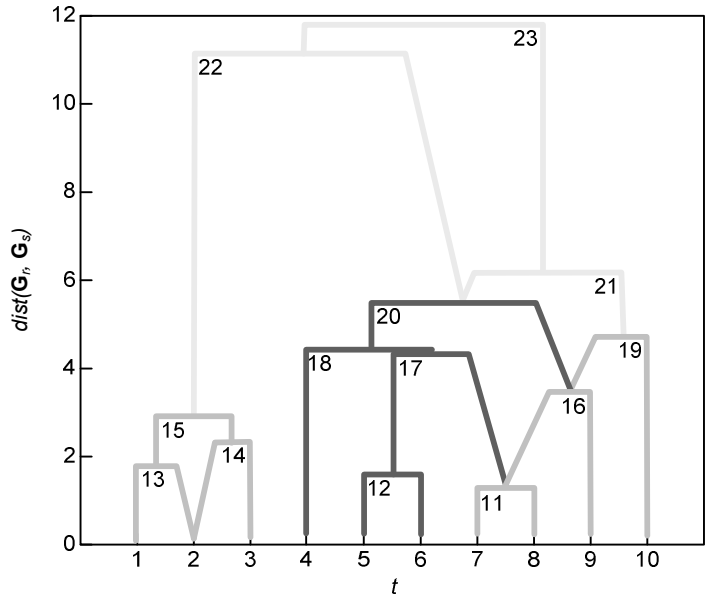
Przykład grupowania punktów w trzy grupy przedstawiono na rys. 3. Uzyskano następujące grupy (punkty: 7, 8, 9 wchodzą w skład dwóch grup):

- 1, 2, 3;
- 4, 5, 6, 7, 8, 9;
- 7, 8, 9, 10.

Grupy wydzielono na podstawie drzewa grupowania, które przedstawiono na rys. 4.



Rys. 3. Grupowanie sekwencji punktów w trzy grupy



Rys. 4. Drzewo grupowania

2.3. Wskaźniki grupowania

Do oceny jakości grupowania zaproponowano procedury środowiska MATLAB z przybornika *Statistics Toolbox*. Oceniając jakość grupowania można posłużyć się współczynnikiem korelacji grupowania (ang. *cophenetic correlation coefficient*) oraz współczynnikiem niezgodności grupowania (ang. *inconsistency coefficient*). Poniżej krótko przedstawiono zasady przeprowadzania obliczeń.

Współczynnik korelacji grupowania c wyznaczamy następująco [4]. Niech K oznacza liczbę etapów grupowania. Parę punktów, którą można utworzyć na etapie grupowania k oznaczmy przez $(\mathbf{o}_n, \mathbf{o}_s)$, a zbiór wszystkich możliwych par $\hat{\mathbf{G}}_k$. Liczbę możliwych par, które można utworzyć na etapie k, oznaczamy następująco:

$$u_k = \left\| \hat{\mathbf{G}}_k \right\|, \quad \hat{\mathbf{G}}_k = \left\{ (\mathbf{o}_n, \mathbf{o}_s) : \mathbf{o}_n \in \mathbf{G}_{\mathbf{Z}(k,1)} \ \wedge \ \mathbf{o}_s \in \mathbf{G}_{\mathbf{Z}(k,2)} \ \wedge \ \mathbf{o}_n \neq \mathbf{o}_s \right\}, \quad (8)$$

gdzie:

u_k – liczba elementów zbioru $\hat{\mathbf{G}}_k$,
k – etap grupowania, $k = 1, ..., K$,
K – liczba etapów grupowania (liczba wierszy macierzy $\mathbf{Z}$),

$\hat{\mathbf{G}}_k$ – zbiór par punktów z grup łączonych na etapie k,
$\mathbf{Z}(k, 1)$ – indeks pierwszej grupy łączonej na etapie k,
$\mathbf{Z}(k, 2)$ – indeks drugiej grupy łączonej na etapie k.

Liczba połączeń pomiędzy punktami na wszystkich etapach grupowania jest równa:

$$U = \sum_{k=1}^{K} u_k \,. \tag{9}$$

Niech S_d oznacza sumę odległości pomiędzy parami punktów na wszystkich etapach grupowania k:

$$S_d = \sum_{k=1}^{K} \left[\sum_{(\mathbf{o}_n, \mathbf{o}_s) \in \hat{\mathbf{G}}_k} d(\mathbf{o}_n, \mathbf{o}_s) \right] \tag{10}$$

natomiast R_d – sumę ich kwadratów:

$$R_d = \sum_{k=1}^{K} \left[\sum_{(\mathbf{o}_n, \mathbf{o}_s) \in \hat{\mathbf{G}}_k} d^2(\mathbf{o}_n, \mathbf{o}_s) \right]. \tag{11}$$

Przez S_z oznaczmy analogiczną do S_k sumę odległości po grupowaniu:

$$S_z = \sum_{k=1}^{K} u_k \, \mathbf{Z}(k,3) \,, \tag{12}$$

a odpowiednią sumę kwadratów tych odległości przez:

$$R_z = \sum_{k=1}^{K} u_k \left[\mathbf{Z}(k,3) \right]^2 . \tag{13}$$

Z kolei przez S_{dz} oznaczmy sumę iloczynów odległości przed grupowaniem i odległości po grupowaniu:

$$S_{dz} = \sum_{k=1}^{K} \left[\sum_{(\mathbf{o}_n, \mathbf{o}_s) \in \hat{\mathbf{G}}_k} d(\mathbf{o}_n, \mathbf{o}_s) \mathbf{Z}(k,3) \right]. \tag{14}$$

Biorąc pod uwagę wyznaczone powyżej wielkości, współczynnik korelacji grupowania c wyznaczamy następująco:

$$c = \frac{S_{dz} - \frac{1}{U} S_d S_z}{\sqrt{\left(R_d - \frac{1}{U} S_d^2 \right) \left(R_z - \frac{1}{U} S_z^2 \right)}} \,. \tag{15}$$

Współczynnik korelacji grupowania przyjmuje wartości z przedziału $\langle 0, 1 \rangle$. Wyższa wartość współczynnika c oznacza lepsze dopasowanie metryki i sposobu łączenia w grupy do sekwencji punktów. Czym mniejsza wartość c, tym gorsze grupowanie punktów.

Współczynnik niezgodności grupowania [4] na głębokość h opisuje każdy etap grupowania k. Odbywa się to przez porównanie odległości pomiędzy

dwoma połączonymi grupami na etapie k, ze średnią odległością łączenia w grupie pierwszej i drugiej.

Niech $\tilde{\mathbf{Z}}_k$ oznacza zbiór odległości pomiędzy łączonymi grupami na etapie k. Pierwszy element zbioru $\tilde{\mathbf{Z}}_k$ to $\mathbf{Z}(k, 3)$. Następnie sprawdzamy grupę o indeksie $\mathbf{Z}(k, 1)$. Jeżeli indeks tej grupy jest większy od T (długość sekwencji punktów), to obliczamy numer etapu, na którym powstała grupa: $s_1 = \mathbf{Z}(k, 1) - T$, a odległość $\mathbf{Z}(s_1, 3)$ dodajemy do zbioru $\tilde{\mathbf{Z}}_k$. W ten sam sposób sprawdzamy grupę o indeksie $\mathbf{Z}(k, 2)$. Jeżeli indeks grupy jest większy od T, to grupa powstała na etapie $s_2 = \mathbf{Z}(k, 2) - T$, a do zbioru $\tilde{\mathbf{Z}}_k$ dodajemy odległość $\mathbf{Z}(s_2, 3)$. Tak utworzony zbiór $\tilde{\mathbf{Z}}_k$ posłuży do wyznaczenia współczynnika niezgodności grupowania na głębokość $h = 2$. Jeżeli chcemy wyznaczyć współczynnik niezgodności grupowania na głębokość $h = 3$, to musimy jeszcze sprawdzić grupy o indeksie: $\mathbf{Z}(s_1, 1)$ i $\mathbf{Z}(s_1, 2)$ a także: $\mathbf{Z}(s_2, 1)$ i $\mathbf{Z}(s_2, 2)$. W przypadku, gdy ich indeksy są większe od T, to do zbioru $\tilde{\mathbf{Z}}_k$ należy dołączyć odpowiednie odległości. W przypadku zadania większej głębokości h postępujemy analogicznie do sposobu opisanego powyżej.

Przyjmijmy, że zbiór $\tilde{\mathbf{Z}}_k$ zawiera $\tilde{u}_k$ elementów:

$$\tilde{\mathbf{Z}}_k = \left\{ l_1, l_2, ..., l_{u_k} \right\}. \tag{16}$$

Wyznaczamy sumę wartości elementów zbioru $\tilde{\mathbf{Z}}$:

$$S_k = \sum_{i=1}^{\tilde{u}_k} l_i \tag{17}$$

oraz sumę ich kwadratów:

$$R_k = \sum_{i=1}^{\tilde{u}_k} l_i^2 \tag{18}$$

Wartość średnią odległości łączenia grup na etapie k wyznaczamy następująco:

$$E\left(\tilde{\mathbf{Z}}_k\right) = \frac{1}{\tilde{u}_k} S_k \tag{19}$$

a wariancję odległości łączonych grup:

$$V\left(\tilde{\mathbf{Z}}_k\right) = \frac{1}{\tilde{u}_k - 1}\left(R_k - \frac{S_k^2}{\tilde{u}_k} \right), \text{ dla } u_k > 1 \tag{20}$$

Ostatecznie wyznaczamy współczynnik niezgodności grupowania na etapie k według następującego wzoru:

$$\mathbf{Y}_k = \frac{\mathbf{Z}(k,3) - E(\tilde{\mathbf{Z}}_k)}{\sqrt{V(\tilde{\mathbf{Z}}_k)}}, \tag{21}$$

gdzie:

$k = 1, ..., T - 1$ – etap grupowania grupowania,

$\mathbf{Y}_k$ – współczynnik niezgodności grupowania na etapie k,

$\mathbf{Z}(k, 3)$ – odległość pomiędzy grupami połączonymi na etapie k,

$E(\tilde{\mathbf{Z}})$ – średnia odległość łączenia grup na etapie k,

$\sqrt{V(\tilde{\mathbf{Z}})}$ – odchylenie standardowe łączenia grup na etapie k.

Dokonanie oceny jakości grupowania polega na obliczeniu współczynnika Y_k od pierwszego do ostatniego etapu grupowania. W przypadku określenia jego maksymalnej wartości podział na grupy wyznacza ten etap grupowania, dla którego otrzymano zadaną wartość współczynnika niezgodności grupowania.

3. Podsumowanie

Podstawą do opracowania metod grupowania sekwencji czasowych była hierarchiczna metoda grupowania. W celu uwzględnienia kolejności punktów sekwencji grupowanie ograniczono tylko do grup sąsiednich. W każdym etapie grupowania łączone są dwie najbliżej położone siebie grupy sąsiednie. W wyniku tego otrzymuje się drzewo grupowania, na podstawie którego można otrzymać żądaną liczbę grup albo grupy o zadanych właściwościach.

W wyniku zastosowania hierarchicznej metody grupowania rozłącznego sekwencji punktów uzyskuje się grupy rozłączne. W hierarchicznej metodzie grupowania z nakładaniem sekwencji punktów umożliwiono łączenie jednej grupy z dwoma sąsiadami pod warunkiem, że są one położone dostatecznie blisko.

Literatura

[1] Everitt B., Landau S. Leese M.: *Cluster Analysis, 4'th edition*, Edward Arnold Publishers Ltd., London 2001.

[2] Koronacki J., Ćwik J.: *Statystyczne systemy uczące się*, Wydawnictwa Naukowo – Techniczne, Warszawa 2005.

[3] Kwiatkowski W.: *Metody automatycznego rozpoznawania wzorców*, Instytut Automatyki i Robotyki WAT, Warszawa 2001.

[4] Mathworks, Inc: *Statistics Toolbox User's Guide*, http://www.mathworks.com/access/helpdesk/help/pdf_doc/stats/stats.pdf, MathWorks, 2005.

[5] Pałys T.: *Zastosowanie metody grupowania sekwencji czasowych w rozpoznawaniu mowy na podstawie ukrytych modeli Markowa*, Rozprawa doktorska, WAT. Warszawa, 2006.

[6] Wiśniewski A. M.: *Metody oceny systemów rozpoznawania mówców*, Biuletyn Instytutu Automatyki i Robotyki WAT, 13/2000, ss. 3-35. Warszawa, 2000

Clustering of Time Sequences

ABSTRACT: Methods of time sequences grouping are presented in this paper. The originality of the problem lies in that the clustered elements determine time sequence, and received groups may determine only segments of a sequence. Two time sequences grouping methods have been elaborated. The first one gives possibility to receive separate groups. By the use of the second one it is possible to obtain groups which overlaps one another.

KEYWORDS: clustering of sequences, overlap clustering

Recenzent: prof. dr hab. inż. Włodzimierz KWIATKOWSKI

Praca wpłynęła do redakcji: 28.12.2006