prof. Leon Bobrowski
Wydział Informatyki Politechniki Białostockiej,
Instytut Biocybernetyki i Inżynierii Biomedycznej PAN, Warszawa

Płaszczyzny wierzchołkowe i płaskie wzorce w wielowymiarowej przestrzeni cech

Celem analizy eksploracyjnej danych (ang. *data mining*) może być odkrywanie użytecznych wzorców (ang. *patterns*) w wielowymiarowych zbiorach danych [1]. Wzorcami są różnego rodzaju prawidłowości w zbiorach danych, takie jak skupiska lub zależności pomiędzy elementami tych zbiorów. Przyjmujemy, że elementami zbiorów danych są n-wymiarowe wektory cech $\mathbf{x}_j = [x_{j1}, \ldots, x_{jn}]^T$ ($\mathbf{x}_j[n] \neq \mathbf{0}$). Wektory te wyznaczają punkty w n-wymiarowej przestrzeni cech $F[n]$ ($\mathbf{x}_j \in F[n]$). Składowe x_{ji} wektora $\mathbf{x}_j$ są liczbowymi wynikami pomiarów (*cech*) x_i danego obiektu O_j, gdzie $j = 1, \ldots, m$. Zakładamy, że wartościami poszczególnych cech x_i mogą być liczby rzeczywiste ($x_{ji} \in R^1$) lub binarne ($x_{ji} \in \{0, 1\}$).

Płaskie wzorce w zbiorze danych wynikają z usytuowania się dużej liczby wektorów cech $\mathbf{x}_j$ na pewnych płaszczyznach (lub w ich pobliżu) w przestrzeni cech $F[n]$. *Płaszczyzna wierzchołkowa* $P_k(\mathbf{x}_{j(1)}, \ldots, \mathbf{x}_{j(l)})$ o wymiarze $(l-1)$ w n-wymiarowej przestrzeni cech $F[n]$ może być zdefiniowana za pomocą l liniowo niezależnych wektorów cech $\mathbf{x}_{j(i)}$:

$$P_k(\mathbf{x}_{j(1)}, \ldots, \mathbf{x}_{j(l)}) = \{\mathbf{x} : \mathbf{x} = \alpha_1 \mathbf{x}_{j(1)} + \ldots + \alpha_l \mathbf{x}_{j(l)}\}, \tag{1}$$

gdzie parametry α_i ($\alpha_i \in R^1$) spełniają warunek normalizacyjny $\alpha_1 + \ldots + \alpha_l = 1$ [2].

Płaszczyzna wierzchołkowa (1) jest związana z bazą

$$\mathbf{B}_k = [\mathbf{x}_{j(1)}, \ldots, x_{j(l)}, \mathbf{e}_{i(1)}, \ldots, \mathbf{e}_{i(n-l)}]^T$$

oraz z wierzchołkiem l-tego rzędu $\mathbf{w}_k = [w_{k1}, \ldots, w_{kn}]^T \in R^n$ za pomocą równania bazowego:

$$\mathbf{B}_k \mathbf{w}_k = \mathbf{1}', \tag{2}$$

gdzie $\mathbf{e_i}$ jest i-tym wektorem jednostkowym, natomiast wektor $\mathbf{1}'$ ma składowe równe 1 lub 0 odpowiednio do wektorów $\mathbf{x}_j$ lub $\mathbf{e}_i$. Wierzchołek l-tego rzędu $\mathbf{w}_k$ jest punktem przecięcia co najmniej l hiperpłaszczyzn h_j w przestrzeni parametrów R^n ($\mathbf{w} = [w_1, \ldots, w_n]^T \in R^n$):

$$(\forall j \in \{1, \ldots, m\})\ h_j = \{\mathbf{w} : (\mathbf{x}_j)^T \mathbf{w} = 1\}. \tag{3}$$

Wierzchołek l-tego rzędu $\mathbf{w}_k$ (2) jest wierzchołkiem zdegenerowanym, gdy jest punktem przecięcia więcej niż l hiperpłaszczyzn h_j (3). Płaskie wzorce w zbiorze danych powiązane są z wierzchołkami zdegenerowanymi $\mathbf{w}_k$. Odkrywanie płaskich wzorców można oprzeć na znajdowaniu wierzchołków zdegenerowanych $\mathbf{w}_k$ poprzez minimalizację wypukłych i odcinkowo-liniowych (typu CPL) funkcji kryterialnych [3]. Odkrywanie płaskich wzorców może znaleźć zastosowanie m.in. w modelowaniu interakcji pomiędzy genami.

Bibliografia

[1] D. Hand, P. Smyth, H. Mannila, *Principles of data mining*, MIT Press, Cambridge 2001.
[2] L. Bobrowski, *Discovering main vertexical planes in a multivariate data space by using CPL functions*, w: ICDM 2014, Ed. P. Perner, Springer, Berlin 2014, 200–213.
[3] L. Bobrowski, *Eksploracja danych oparta na wypukłych i odcinkowo-liniowych funkcjach kryterialnych*, Wydawnictwa Politechniki Białostockiej, Białystok, 2005.

Praca wspierana przez projekt statutowy 4.2/st/14 IBIB PAN.