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

Eksploracja wzorców wierzchołkowych w zbiorach danych

Bierzemy pod uwagę zbiór danych C, którego elementami są n-wymiarowe *wektory cech* $\boldsymbol{x}_j = [x_{j1}, \ldots, x_{jn}]^T$ ($\boldsymbol{x}_j[n] \neq \boldsymbol{0}$). Wektory te wyznaczają punkty w n-wymiarowej przestrzeni cech $F[n]$ ($\boldsymbol{x}_j \in F[n]$). Składowe x_{ji} wektora $\boldsymbol{x}_j$ mogą być 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 ($x_{ji} \in \{0, 1\}$) [1].

Każdy z m wektorów cech $\boldsymbol{x}_j$ wyznacza poniższą hiperpłaszczyznę h_j w n-wymiarowej (*dualnej*) przestrzeni parametrów R^n ($\boldsymbol{w} \in R^n$):

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

Bierzemy również pod uwagę poniższe hiperpłaszczyzny h_i w przestrzeni parametrów R^n ($\boldsymbol{w} \in R^n$) wyznaczone przez każdy z n wektorów jednostkowych $\boldsymbol{e}_i$:

$$(\forall i \in \{1, \ldots, n\})\ h_i = \{\boldsymbol{w} : \boldsymbol{e}_i^T \boldsymbol{w} = 0\}. \tag{2}$$

Dowolny zestaw S_k liniowo niezależnych wektorów złożony z l bazowych wektorów cech $\boldsymbol{x}_{j(i)}$ oraz z $n - l$ wektorów jednostkowych $\boldsymbol{e}_i$ pozwala wyznaczyć bazę $\boldsymbol{B}_k$ przestrzeni R^n oraz wierzchołek $\boldsymbol{w}_k$ rzędu l związany z tą bazą $\boldsymbol{B}_k$ za pomocą równania bazowego. Wierzchołek $\boldsymbol{w}_k$ jest punktem przecięcia l hiperpłaszczyzn $h_{j(i)}$ (1) oraz $n - l$ hiperpłaszczyzn h_i (2) wyznaczonych przez bazowe wektory $\boldsymbol{x}_{j(i)}$ oraz $\boldsymbol{e}_i$ z zestawu S_k. Wierzchołek $\boldsymbol{w}_k$ jest zdegenerowany, jeżeli przechodzi przez niego więcej niż l hiperpłaszczyzn $h_{j(i)}$ (1).

Hiperpłaszczyzna wierzchołkowa $H_k(\boldsymbol{x}_{j(1)}, \ldots, x_{j(l)})$ o wymiarze $(l-1)$ jest definiowana w przestrzeni cech $F[n]$ za pomocą l ($2 \leqslant l \leqslant n$) bazowych wektorów cech $x_{j(i)}$ z zestawu S_k:

$$H_k(\boldsymbol{x}_{j(1)}, \ldots, \boldsymbol{x}_{j(l)}) = \{\boldsymbol{x} : \boldsymbol{x} = \alpha_1 \boldsymbol{x}_{j(1)} + \ldots + \alpha_l \boldsymbol{x}_{j(l)}\} \tag{3}$$

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

Wzorzec wierzchołkowy (ang. *vertexical pattern*) P_k tworzy się z dużej liczby m_k ($m_k > l$) wektorów cech $\boldsymbol{x}_j$ ($\boldsymbol{x}_j \in C$) ulokowanych na hiperpłaszczyźnie H_k (3) lub w jej pobliżu.

Wydobycie ze zbioru danych C wzorców wierzchołkowych P_k może dać podstawy do tworzenia modeli interakcji pomiędzy wieloma cechami x_i. Szczególnie interesujące mogą tu być interakcje pomiędzy zestawami genów i czynników środowiskowych warunkujące np. powstawanie wybranych zjawisk chorobowych. Jednym z aktualnych wyzwań jest tu tworzenie wystarczająco efektywnych i precyzyjnych procedur obliczeniowych umożliwiających wykrywanie wzorców wierzchołkowych P_k w wielowymiarowych przestrzeniach cech.

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*, in: ICDM 2014, Springer, Berlin 2014, 200-213.

Praca wspierana przez projekt statutowy 4.2/st/15 IBIB PAN.