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

Selekcja podprzestrzeni cech na bazie liniowej separowalności zbiorów danych

Wiele ważnych w praktyce problemów jest reprezentowanych za pomocą zbiorów danych C_k ($k = 1, \ldots, K$) zbudowanych z „długich" wektorów cech $\boldsymbol{x}_j = [x_{j1}, \ldots, x_{jn}]^T \in \mathbb{R}^n$ ($j = 1, \ldots, m$). Termin ten oznacza sytuację, gdy liczba m wektorów $\boldsymbol{x}_j$ jest znacznie mniejsza niż liczba n ich składowych (*cech*) x_i ($m \ll n$). Taka sytuacja ma miejsce m.in. w przypadku genomicznych zbiorów danych. Mała liczba m wektorów cech w $\boldsymbol{x}_j$ w porównaniu z ich wymiarowością n ogranicza stosowanie statystycznych technik wnioskowania. W tych okolicznościach koniecznością staje się obniżanie wymiarowości przestrzeni cech $F[n]$ ($\boldsymbol{x}_j \in F[n]$). Stosuje się w tym celu *ekstrakcję cech*, m.in. poprzez liniowe transformacje typu składowych głównych [1]. Często używane w tym celu są również metody *selekcji cech* służące pomijaniu maksymalnej liczby takich cech x_i, które nie są konieczne przy rozwiązywaniu danego problemu.

Zagadnienie selekcji cech może być analizowane w oparciu o koncepcję liniowej separowalności zbiorów. Zbiory C_k są liniowo separowalne wtedy i tylko wtedy, gdy mogą być spełnione poniższe nierówności [2]:

$$\begin{aligned}(\exists k \in \{1, \ldots, K\})\ (\exists \boldsymbol{w}_k, \theta_k)(\forall \boldsymbol{x}_j \in C_k)\ \boldsymbol{w}_k^T \boldsymbol{x}_j \geq \theta_k + 1 \\ \text{i}\ (\forall \boldsymbol{x}_j \in C'_k)\ \boldsymbol{w}_k^T \boldsymbol{x}_j \leq \theta_k + 1,\end{aligned} \tag{1}$$

gdzie $\boldsymbol{w}_k$ jest *wektorem wag* ($\boldsymbol{w}_k \in \mathbb{R}^n$) oraz θ_k jest progiem ($\theta_k \in \mathbb{R}$). Zgodnie z relacją (1), każdy ze zbiorów danych C_k może być oddzielony od sumy $C'_k = \bigcup_{i \neq k} C_i$ pozostałych zbiorów C_i za pomocą hiperpłaszczyzny $H(\boldsymbol{w}_k, \theta_k) = \{\boldsymbol{x} : \boldsymbol{w}_k^T \boldsymbol{x} = \theta_k\}$ z marginesem $\delta_k = 1/||\boldsymbol{w}_k||$ w przestrzeni cech $F[n]$ [3].

Hiperpłaszczyznę $H(\boldsymbol{w}_k, \theta_k)$ rozdzielającą dwa liniowo separowalne (1) zbiory C_k i C'_k można efektywnie wyznaczać poprzez minimalizację wypukłej i odcinkowo liniowej (CPL) funkcji kryterialnej $\Phi_k(\boldsymbol{w})$. Funkcje $\Phi_k(\boldsymbol{w})$ mają związek z odległością typu L_1 (suma wartości bezwzględnych). Zostało wykazane, że jeżeli zbiory C_k i C'_k są liniowo separowalne (1) w przestrzeni cech $F[n]$, to istnieje co najmniej jedna podprzestrzeń $F_l[n']$ ($F_l[n'] \subset F[n]$) o wymiarze n' nie większym niż liczba m wektorów $\boldsymbol{x}_j$ ($n' \leq m$), która zapewnia liniową separowalność tych zbiorów. W przypadku „długich" wektorów cech $\boldsymbol{x}_j$ ($m \ll n$) istnieje zwykle wiele podprzestrzeni $F_l[n']$ zapewniających liniową separowalność zbiorów C_k i C'_k. Przy wyborze optymalnej podprzestrzeni $F_{l*}[n']$ zapewniającej liniową separowalność zbiorów C_k i C'_k można kierować się postulatem separowania tych zbiorów z maksymalnym marginesem δ_k. Postulat ten może zostać efektywnie zrealizowany poprzez minimalizację wypukłej i odcinkowo liniowej funkcji kryterialnej (CPL).

Literatura

[1] O. R. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, Wiley, New York 2001.
[2] L. Bobrowski, *Feature subset selection based on the concept of linear separability*, w: Lecture Notes of the VII-th ICB Seminar: Statistics and Clinical Practice, 2008, Warsaw.
[3] L. Bobrowski, T. Łukaszuk, *Selection of the linearly separable feature subsets*, w: Artificial Intelligence and Soft Computing — ICAISC 2004, Eds. L. Rutkowski et al., Lecture Notes in Artificial Intelligence 3070, Springer 2004, 544–549.