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

# Zasada liniowej separowalności w selekcji modeli prognostycznych

Koncepcję liniowej separowalności wielowymiarowych zbiorów danych wiąże się z początkami metod *sieci neuropodobnych* i *rozpoznawania obrazów*. Zbiory $G^+$ i $G^-$ $n$-wymiarowych wektorów cech $x_j$ są liniowo separowalne w przestrzeni cech $F$ ($x_j \in F$), jeżeli istnieje taki wektor wag $w$ oraz próg $\theta$, że spełniony jest układ nierówności

$$\begin{aligned}(\exists w \in \mathbb{R}^n)(\exists \theta \in \mathbb{R}^1)\ (\forall x_j \in G^+)\ w^T x_j > \theta \\ \text{oraz } (\forall x_j \in G^-)\ w^T x_j < \theta\end{aligned} \tag{1}$$

W problemie liniowej separowalności zainteresowani jesteśmy zagadnieniem, czy zbiory $G^+$ i $G^-$ są liniowo separowalne, oraz wyznaczeniem takich parametrów $w^*$ oraz $\theta^*$, które w optymalny sposób spełniają (w pełni lub częściowo) układ nierówności (1). Optymalne parametry $w^*$ oraz $\theta^*$ mogą być wyznaczone na podstawie układu nierówności (1) poprzez minimalizację wypukłej i odcinkowo-liniowej (*CPL*) funkcji kryterialnej [1]. Parametry $w^*$ oraz $\theta^*$ mogą być użyte m.in. w definicji modelu prognostycznego:

$$y = (w^*)^T x - \theta^*. \tag{2}$$

Model prognostyczny (2) może być budowany na bazie przedziałowego zbioru uczącego $C_m$:

$$C_m = \{x_j, [y_j^-, y_j^+]\},\ \text{gdzie } j = 1, \ldots, m \text{ oraz } y_j^- \leq y_j^+. \tag{3}$$

Regresyjnemu zbiorowi danych $\{(x_j, y_j)\}$ można nadać postać (3), np. poprzez wprowadzenie *marginesu* $\varepsilon$ ($\varepsilon > 0$) i założenie, że $y_j^- = y_j - \varepsilon$ oraz $y_j^+ = y_j + \varepsilon$. Wyrażenie (3) pozwala uwzględniać też dane cenzurowane poprzez zastosowanie warunku $y_j^- = -\infty$ lub $y_j^+ = \infty$. W konstrukcji modelu (2) wykorzystywany jest poniższy układ postulowanych nierówności:

$$(\forall j \in \{1, \ldots, m\}) \quad y_j^- < w^T x_j - \theta < y_j^+. \tag{4}$$

Układ postulowanych nierówności (4) można sprowadzić do problemu liniowej separowalności (1) [1]. Potraktowanie konstrukcji modelu prognostycznego (2) jako problemu liniowej separowalności pozwala m.in. na zastosowanie metody *relaksowanej liniowej separowalności* (*RLS*) selekcji zestawów cech [2]. Metoda RLS umożliwiła na przykład zredukowanie zestawu genów od liczby $n = 24481$ do $n_1 = 12$. Wyselekcjonowany zestaw 12 genów pozwolił na pełną (100%) liniową separację (1) grupy pacjentów nowotworowych od nienowotworowych o liczności około 50 pacjentów w każdej z tych grup. Taka technika pozwala też na budowę modeli prognostycznych (2) na bazie wysokowymiarowych danych genetycznych z cenzurowanymi czasami przeżycia [1].

**Bibliografia**

[1] L. Bobrowski, T. Łukaszuk, *Prognostic Modeling with High Dimensional and Censored Data*, ICDM – 2012.

[2] L. Bobrowski, T. Łukaszuk, *Relaxed Linear Separability* (*RLS*) *Approach to Feature* (*Gene*) *Subset Selection*, w: Selected Works in Bioinformatics, Xuhua Xia (ed.), INTECH 2011, 103–118.