

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

Selekcja pacjentów podwyższonego ryzyka z wykorzystaniem regresji rangowej i funkcji kryterialnych typu CPL

Działania profilaktyczne w medycynie mogą mieć na celu wykrywanie pacjentów o zwiększonym ryzyku danej choroby i kierowanie ich na dodatkowe badania diagnostyczne, oraz ewentualnie na zabiegi terapeutyczne. Selekcja pacjentów podwyższonego ryzyka odbywa się rutynowo z zastosowaniem tzw. *procedur przesiewowych* (ang. *screening procedure*).

W procedurach przesiewowych stosowane są pewne modele prognostyczne, które pozwalają przewidywać przyszłe stany zdrowia danego pacjenta i selekcjonować pacjentów o zwiększonym ryzyku pojawienia się schorzenia. Przyjmujemy założenie, że poszczególni pacjenci O_j ($j = 1, \dots, m$) scharakteryzowani są w ten sam sposób za pomocą tzw. wektorów cech $\mathbf{x}_j[n] = [x_{j1}, \dots, x_{jn}]^T$. Kolejne składowe x_{ji} n -wymiarowego wektora cech $\mathbf{x}_j[n]$ są liczbowymi wynikami określonych badań diagnostycznych pacjenta O_j ($x_{ji} \in \mathbb{R}^n$ lub $x_{ji} \in \{0, 1\}$). Często używa się modeli prognostycznych opartych na liniowych równaniach regresji:

$$(\forall j \in \{1, \dots, m\}) \quad y_j = \mathbf{w}[n]^T \mathbf{x}_j[n] + w_0, \quad (1)$$

gdzie $\mathbf{w}[n] = [w_1, \dots, w_n]^T$ jest wektorem wag ($\mathbf{w}[n] \in \mathbb{R}^n$), oraz w_0 jest *progiem* ($w_0 \in \mathbb{R}$).

W konstrukcji rangowych modeli ryzyka wykorzystujemy zbiór m wektorów cech $\mathbf{x}_j[n]$ oraz informację dodatkową w postaci pewnej liczby relacji porządkowych $R_j \prec R_k$. Relacja porządkowa $R_j \prec R_k$ oznacza, że ryzyko R_j związane z pacjentem O_j jest mniejsze niż ryzyko R_k związane z pacjentem O_k . Odzworowanie (1) tworzy model rangowy, jeżeli w możliwie dużej liczbie przypadków z relacji porządkowej $R_j \prec R_k$ wynika nierówność $y_j < y_k$.

Przyjęte zostało tu założenie, że pacjent O_j obciążony był tym mniejszym ryzykiem, im dłuższy był rzeczywisty czas jego życia T_j po ustalonym zdarzeniu (np. po poważnej operacji chirurgicznej):

$$(R_j \prec R_k) \iff (T_j > T_k) \quad (2)$$

Wartość rzeczywistego czasu życia T_j często nie jest znana w praktycznych problemach rozwiązywanych w ramach *analizy przeżycia* (ang. *survival analysis*). W zbiorach danych analizy przeżycia pojawia się tzw. *prawostronne cenzorowanie* oraz *lewostronne cenzorowanie* czasu życia T_j . W przypadku *prawostronnego cenzorowania* mamy informację o minimalnym czasie życia T_j^- zamiast wartości rzeczywistego czasu życia T_j pacjenta O_j , co oznacza, że $T_j > T_j^-$. W przypadku *lewostronnego cenzorowania* mamy informację o maksymalnym czasie życia T_j^+ , co oznacza, że $T_j < T_j^+$.

Ocenzorowanie rzeczywistego czasu życia T_j nie wyklucza w pewnych przypadkach możliwości określenia zestawu relacji porządkowych $R_j \prec R_k$, który umożliwia konstrukcje rangowych modeli prognostycznych poprzez minimalizację odpowiednio dobranych funkcji kryterialnych typu CPL (ang. *convex and piecewise linear*). Z praktycznego punktu widzenia ważna jest możliwość wykorzystania minimalizacji funkcji typu CPL do konstrukcji modelu (1) wraz z doбором zestawu zmiennych x_i tego modelu.