*Symulacja w Badaniach i Rozwoju*
Vol. 1, No. 2/2010

**Leon BOBROWSKI**[1,2]
[1]Wydział Informatyki Politechniki Białostockiej,
[2]Instytut Biocybernetyki i Inżynierii Biomedycznej PAN, Warszawa
E-mail: leon@ibib.waw.pl

# Liniowe modele prognostyczne oparte na regresji przedziałowej z funkcjami typu *CPL*

## 1 Wstęp

Modele prognostyczne typu regresyjnego konstruujemy często w postaci liniowej (afinicznej) zależności zmiennej prognozowanej (zależnej) $Y$ od ustalonej liczby $n$ zmiennych niezależnych $X_i$ ($i = 1, \ldots, n$) [1]. Klasycznym przykładem może być tu prognozowanie wartości rynkowej mieszkania na podstawie równania regresji opartego na odpowiednio dobranej kombinacji liniowej takich parametrów charakteryzujących wybrane mieszkanie jak jego powierzchnia, wyposażenie, atrakcyjność lokalizacji czy też infrastruktura komunikacyjna.

Opracowanych zostało wiele metod budowy regresyjnych modeli prognostycznych na podstawie uczących zbiorów danych [1]. Zbiór uczący zawiera na ogół przykładowe wartości $x_{ji}$ zmiennych niezależnych $X_i$, zebrane w postaci tzw. *wektora cech* $\mathbf{x}_j=[x_{j1}, \ldots, x_{jn}]^T$, któremu towarzyszy odpowiadająca mu wartość $y_j$ zmiennej zależnej $Y$ ($j$=1, … ,$m$). Klasyczny model liniowej regresji wielowymiarowej konstruowany jest na podstawie zbioru uczącego w postaci $m$ par $(\mathbf{x}_j, y_j)$ poprzez minimalizację sumy kwadratów odchyleń $(\hat{y_j} - y_j)^2$ wartości prognozowanych (wynikających z modelu) $\hat{y_j}$ od wartości zaobserwowanych $y_j$ zmiennej zależnej $Y$. W przypadku budowania zależności zmiennej $Y$ od jednej zmiennej $X$ taka metoda konstrukcji modelu regresyjnego nazywana jest metodą najmniejszych kwadratów. Zaletą tego podejścia jest możliwość analitycznego wyznaczenia parametrów liniowego modelu regresyjnego na podstawie zbioru uczącego. Jedna z ważnych modyfikacji w projektowaniu liniowych modeli regresyjnych polega na tym, że kwadraty odchyleń $(\hat{y_j} - y_j)^2$ zastępujemy poprzez wartości bezwzględne $|\hat{y_j} - y_j|$. Zamiana taka uniemożliwia co prawda analityczne wyznaczanie parametrów modelu regresyjnego, ale znane są bardzo efektywne procedury iteracyjne służące temu celowi, bliskie programowaniu liniowemu. Konstrukcja modeli regresyjnych na podstawie wartości bezwzględnych $|\hat{y_j} - y_j|$ zamiast kwadratów odchyleń $(\hat{y_j} - y_j)^2$ ma zalety np. w postaci zmniejszenia wpływu tzw. *obserwacji odstających* (ang. *outliers*) na budowany model [2].

W zbiorze uczącym $\{(\mathbf{x}_j, y_j)\}$ o postaci $m$ par $(\mathbf{x}_j, y_j)$, wartości $y_j$ mogą być traktowane jako pewnego rodzaju wiedza dodatkowa o wektorach cech $\mathbf{x}_j$. W wielu zastosowaniach praktycznych nie dysponujemy jednak zbiorami uczących z dokładnymi wartościami $y_j$ zmiennej zależnej $Y$. Zamiast takich informacji możemy dysponować wiedzą dodatkową w postaci pewnej liczby relacji porządkowych $''\mathbf{x}_j \prec \mathbf{x}_k''$ ($''\mathbf{x}_j$ *poprzedza* $\mathbf{x}_k''$) pomiędzy wybranymi wektorami cech $\mathbf{x}_j$ i $\mathbf{x}_k$ [3]. Tego typu relacje $''\mathbf{x}_j \prec \mathbf{x}_k''$ można tworzyć np. na podstawie faktów o postaci: ″Pacjent $O_j$ reprezentowany za pomocą

wektora cech $\mathbf{x}_j$ *żył krócej*, niż pacjent $O_k$ reprezentowany za pomocą wektora $\mathbf{x}_k''$. Problem budowy modelu *regresji rangowej* został sformułowany jako wyznaczanie takiej kombinacji liniowej cech $X_i$, która w maksymalnym stopniu zachowuje relacje porządkowe $''\mathbf{x}_j \prec \mathbf{x}_k''$. Konstrukcja liniowego modelu *regresji rangowej* została oparte na sprawdzaniu warunku liniowej separowalności dwu zbiorów różnicowych $C_1$ i $C_0$, które zostały zbudowane z różnic $\mathbf{r}_{jk} = \mathbf{x}_j - \mathbf{x}_k$ wektorów tworzących relacje porządkowe $''\mathbf{x}_j \prec \mathbf{x}_k''$. Sprawdzanie liniowej separowalności zbiorów różnicowych $C_1$ i $C_0$, przeprowadza się poprzez minimalizację wypukłej i odcinkowo liniowej (ang.: *convex and piecewise linear - CPL*) funkcji kryterialnej. Funkcja kryterialna typu *CPL* została zbudowana w tym przypadku na wektorach różnicowych $\mathbf{r}_{jk} = \mathbf{x}_j - \mathbf{x}_k$ [4].

W niniejszej pracy analizowany jest problem konstrukcji modelu *regresji przedziałowej* (ang. *interval regression*) [5]. W tym przypadku wiedza dodatkowa o wektorach cech $\mathbf{x}_j$ reprezentowana jest w postaci odcinków $[y_j^-, y_j^+]$. Przyjmujemy tu, że dokładna wartość $y_j$ zmiennej zależnej $Y$ nie jest znana. Wiemy natomiast, że wartość $y_j$ zawarta jest w przedziale $[y_j^-, y_j^+]$, tj.: $y_j^- < y_j < y_j^+$. Modele regresji przedziałowej pojawiają się miedzy innymi w kontekście *analizy przeżycia* (ang. *survival analysis*) [6], gdy dysponujemy tylko taką wiedzą dodatkową o wektorze cech $\mathbf{x}_j$, że pacjent $O_j$ reprezentowany za pomocą wektora $\mathbf{x}_j$ żył po zabiegu dłużej niż $y_j^-$ lat a krócej niż $y_j^+$ lat. Z literatury znana jest konstrukcja modeli regresji przedziałowej oparta na metodzie *EM* (ang. *Expectation Maximization*) [5] W przedstawionej pracy analizowana jest możliwość wykorzystywania funkcji kryterialnych typu *CPL* w celu konstrukcji modeli regresji przedziałowej [7], [8].

## 2 Model liniowej regresji przedziałowej

W artykule używana jest terminologia z zakresu rozpoznawania obrazów (ang. *pattern recognition*) [2]. Obiekty (zdarzenia, pacjenci) $O_j$ reprezentowane są tu za pomocą $n$-wymiarowych wektorów cech $\mathbf{x}_j[n] = [x_{j1},...,x_{jn}]^T$. Symbol $\mathbf{x}_j[n]$ może oznaczać też punkt w $n$-wymiarowej przestrzeni cech $F[n]$ ($\mathbf{x}_j[n] \in F[n]$). Poszczególne składowe $x_{ji}$ wektora $\mathbf{x}_j[n]$ są liczbowymi wynikami ($x_{ji} \in R^1$ lub $x_{ji} \in \{0,1\}$) ustalonych wcześniej $n$ pomiarów dokonanych na obiekcie $O_j$.

Bierzemy pod uwagę liniowe (afiniczne) transformacje $n$-wymiarowych wektorów cech $\mathbf{x}[n]$ na punkty $y$ linii prostej:

$$y = \theta + \mathbf{w}[n]^T\mathbf{x}[n] = \mathbf{w}'[n+1]^T\mathbf{x}'[n+1]] \tag{1}$$

gdzie $\theta$ jest *progiem* ($\theta \in R^1$), $\mathbf{w}[n] = [w_1,...,w_n]^T$ jest wektorem parametrów (*wag*) $w_i$ ($w_i \in R^1$), $\mathbf{w}'[n+1]$ jest poszerzonym (ang. *augmented*) wektorem parametrów ($\mathbf{w}'[n+1]=[\theta, \mathbf{w}[n]^T]^T=[\theta, w_1, \ldots, w_n]^T=[w_0\ w_1, \ldots, w_n]^T \in R^{n+1}$), $\mathbf{x}'[n+1]$ jest poszerzonym wektorem cech ($\mathbf{x}'[n+1]=[1,\mathbf{x}[n]^T]^T=[1, x_1, \ldots, x_n]^T$).

Parametry $\mathbf{w}'[n+1]$ modelu (1) ustalane są na podstawie zbioru uczącego $C_m$. W przypadku regresji przedziałowej zbiór uczący $C_m$ ma poniższą strukturę:

$$C_m = \{\mathbf{x}_j[n], [y_j^-, y_j^+]\},\ \textit{gdzie}\ j = 1,\ldots..,m\ \textit{oraz}\ y_j \le y_j^+ \tag{2}$$

Transformacja (1) tworzy model liniowej regresji przedziałowej, jeżeli w możliwie największym stopniu spełniony jest poniższy układ nierówności:

$$(\forall j \in \{1,\ldots, m\}) \qquad y_j^- \le \theta + \mathbf{w}[n]^T\mathbf{x}_j[n] \le y_j^+ \tag{3}$$

W literaturze została opisana konstrukcja modeli regresji przedziałowej oparta na metodzie *EM* (ang. *Expectation Maximization*) [1], [5]. Jest to procedura raczej mało efektywna, co ma szczególne znaczenie w przypadku wielowymiarowych wektorów cech $\mathbf{x}_j[n]$. W przedstawionej pracy analizowana jest możliwość wykorzystywania wypukłych i odcinkowo liniowych funkcji kryterialnych (typu *CPL*) w celu konstrukcji liniowych modeli regresji przedziałowej [7].

## 3 Perceptronowa funkcja kryterialna

Pierwowzorem funkcji kryterialnych typu *CPL* jest perceptronowa funkcja kryterialna wiązana z początkami teorii sieci neuropodobnych (sieci neuronów formalnych) [2]. Funkcje kryterialne typu *CPL* definiowane są na podstawie koncepcji liniowej separowalności dwu zbiorów danych $C_1$ i $C_0$, które zostały zbudowane z wektorów cech $x_j[n]$ [8].

*Definicja* 1: Zbiory uczące $C_1$ i $C_0$ są liniowo separowalne wtedy i tylko wtedy, gdy istnieje taki poszerzony wektor parametrów $w^*[n+1]$ (1), że spełniony jest poniższy układ nierówności liniowych:

$$\begin{aligned}(\exists w^*[n+1])\ (\forall x_i[n] \in C_1)\ & w^*[n+1]^T x'_i[n+1] > 0 \\ (\forall x_j[n] \in C_0)\ & w^*[n+1]^T x'_j[n+1] < 0\end{aligned} \tag{4}$$

gdzie $x'_j[n+1]$ jest poszerzonym wektorem cech (1).

Dla celów konstrukcji funkcji kary typu *CPL* układ nierówności (4) modyfikuje się w poniższy sposób [8]:

$$\begin{aligned}(\exists \mathbf{w}^*[n+1])\ (\forall \mathbf{x}_i[n] \in C_1)\ & \mathbf{w}^*[n+1]^T \mathbf{x}'_i[n+1] \ge 1 \\ (\forall \mathbf{x}_j[n] \in C_0)\ & \mathbf{w}^*[n+1]^T \mathbf{x}'_j[n+1] \le -1\end{aligned} \tag{5}$$

Układy nierówności (4) i (5) są sobie równoważne ze względu na definicję liniowej separowalności (*Def.* 1).

W oparciu o układy nierówności (5) definiowane są pozytywne funkcje kary $\varphi_i^+(w[n+1])$ oraz negatywne funkcje kary $\varphi_j^-(w[n+1])$. Dla każdego elementu $x_j[n]$ zbioru $C_1$ definiowana jest pozytywna funkcja kary $\varphi_j^+(w[n+1])$ [8]:

$$(\forall x_j[n] \in C_1)$$

$$\varphi_j^+(\mathbf{w}[n+1]) = \begin{cases} 1 - \mathbf{w}[n+1]^T\mathbf{x}'_j[n+1] & \textit{jeżeli} \quad \mathbf{w}[n+1]^T\mathbf{x}'_j[n+1] \le 1 \\ 0 & \textit{jeżeli} \quad \mathbf{w}[n+1]^T\mathbf{x}'_j[n+1] > 1 \end{cases} \tag{6}$$

Podobnie, dla każdego elementu $x_j[n]$ zbioru $C_0$ definiowana jest negatywna funkcja kary $\varphi_i^-(w[n+1])$:

$$(\forall x_j[n] \in C_0)$$

$$\varphi_j^-(\mathbf{w}[n+1]) = \begin{cases} 1 + \mathbf{w}[n+1]^T\mathbf{x}'_j[n+1] & \textit{jeżeli} \quad \mathbf{w}[n+1]^T\mathbf{x}'_j[n+1] \ge -1 \\ 0 & \textit{jeżeli} \quad \mathbf{w}[n+1]^T\mathbf{x}'_j[n+1] < -1 \end{cases} \tag{7}$$

Obie powyższe funkcje kary $\varphi_j^+(\mathbf{w}[n+1])$ i $\varphi_j^-(\mathbf{w}[n+1])$ są wypukłe i odcinkowo liniowe.

Perceptronowa funkcja kryterialna $\Phi(\mathbf{w}[n+1])$ jest dodatnio ważoną sumą funkcji kary $\varphi_j^+(\mathbf{w}[n+1])$ i $\varphi_j^-(\mathbf{w}[n+1])$:

$$\Phi(\mathbf{w}[n+1]) = \sum_{j \in J_1} \alpha_j \varphi_j^+(\mathbf{w}[n+1]) + \sum_{j \in J_0} \alpha_j \varphi_j^-(\mathbf{w}[n+1]) \tag{8}$$

gdzie $\alpha_j$ ($\alpha_j > 0$) jest dodatnim parametrem (*ceną*) związaną z wektorem cech $\mathbf{x}_j[n]$, $J_1$ jest zbiorem indeksów $j$ wektorów cech $\mathbf{x}_j[n]$ ze zbioru $C_1$, oraz $J_0$ jest zbiorem indeksów $j$ wektorów cech $\mathbf{x}_j[n]$ ze zbioru $C_0$.

Perceptronowa funkcja kryterialna używana w teorii sieci neuropodobnych i rozpoznawania obrazów ma postać podobną do $\Phi(\mathbf{w}[n+1])$ (8) [2]. Funkcja kryterialna $\Phi(\mathbf{w}[n+1])$ (8) jest funkcją wypukłą i odcinkowo-liniową jako suma tego typu funkcji kary $\varphi_j^+(\mathbf{w}[n+1])$ (6) i $\varphi_j^-\mathbf{w}[n+1])$ (7). Algorytmy wymiany rozwiązań bazowych, zbliżone do programowania liniowego, pozwalają znaleźć minimum $\Phi(\mathbf{w}^*)$ funkcji $\Phi(\mathbf{w}[n+1])$ (8) w sposób efektywny nawet w przypadku dużych, wielowymiarowych zbiorów $C_1$ i $C_0$ [9]:

$$\Phi^* = \Phi(\mathbf{w}^*[n+1]) = \min_{\mathbf{w}[n+1]} \Phi(\mathbf{w}[n+1]) \geq 0 \tag{9}$$

Optymalny wektor parametrów $\mathbf{w}^*[n+1]$ oraz wartość minimalna $\Phi^*$ funkcji kryterialnej $\Phi(\mathbf{w})$ (8) mogą być stosowane w rozwiązywaniu wielu problemów eksploracyjnej analizy danych. W szczególności, wektor $\mathbf{w}^*[n+1]$ pozwala wyznaczyć hiperpłaszczyznę $H(\mathbf{w}^*[n+1])$ rozdzielającą zbiory $C_1$ i $C_0$ w sposób optymalny:

$$H(\mathbf{w}^*[n+1]) = \{\mathbf{x}[n+1] \in F[n+2]: \mathbf{w}^*[n+1]^T \mathbf{x}[n+1] = 0\} \tag{10}$$

*Lemat* 1: Wartość minimalna $\Phi^*$ (9) funkcji kryterialnej $\Phi(\mathbf{w}[n+1])$ (8) jest równa zeru ($\Phi^* = 0$) wtedy i tylko wtedy, gdy istnieje taki wektor parametrów $\mathbf{w}'[n+1]$, że hiperpłaszczyzna $H(\mathbf{w}'[n+1])$ dokładnie rozdziela wszystkie wektory cech $\mathbf{x}_j[n]$ ze zbioru $C_1$ oraz $C_0$.

Dokładne rozdzielenie zbiorów $C_1$ i $C_0$ oznacza, że każdy punkt $\mathbf{x}_j[n]$ ze zbioru $C_1$ leży po dodatniej stronie hiperpłaszczyzny $H(\mathbf{w}^*[n+1])$ (10) (tj. $\mathbf{w}^*[n+1]^T \mathbf{x}_i[n+1] > 0$) a każdy punkt $\mathbf{x}_i[n]$ ze zbioru $C_0$ leży po ujemnej stronie $H(\mathbf{w}^*[n+1])$ (10) (tj. $\mathbf{w}^*[n+1]^T \mathbf{x}_j[n+1] < 0$).

## 4 Wypukłe i odcinkowo liniowe funkcje kryterialne (typu *CPL*) w regresji przedziałowej

Funkcję kryterialnej $\Phi(\mathbf{w}[n+1])$ (8) można przystosować do układu nierówności przedziałowych (3). W tym celu układ ten zmodyfikujemy do postaci podobnej do (4) i (5):

$$\begin{array}{ll} (\forall j \in \{1, \ldots, m\}) & \mathbf{w}[n]^T \mathbf{x}_j[n] + \theta \geq y_i^- \\ \textit{oraz} & \mathbf{w}[n]^T \mathbf{x}_j[n] + \theta \leq y_j^+ \end{array} \tag{11}$$

lub używając poszerzonych wektorów cech x′[*n*+1] i wag w[*n*+1] (1):

$$(\forall j \in \{1,\ldots, m\}) \quad \mathbf{w}[n+1]^T\mathbf{x}'_j[n+1] \geq y_i^- \qquad \textit{oraz} \quad \mathbf{w}[n+1]^T\mathbf{x}'_j[n+1] \leq y_j^+ \tag{12}$$

Na bazie układu nierówności (12) można sformułować problem liniowej separowalności dwu zbiorów i zdefiniować w poniższy sposób pozytywne funkcje kary $\phi_j^+(w[n+1])$ podobne do $\varphi_j^+(w[n+1])$ (6) oraz negatywne funkcje kary $\phi_j^-(w[n+1])$ podobne do $\varphi_j^-(w[n+1])$ (7):

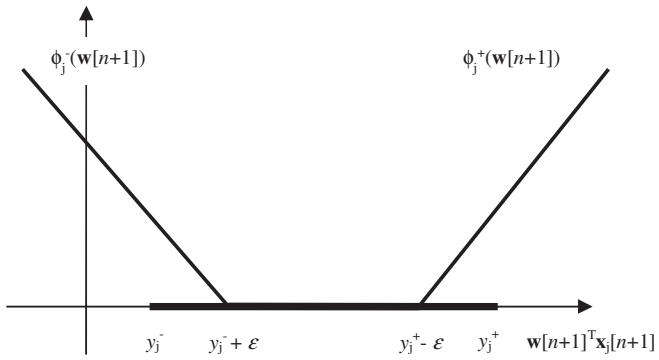
$$(\forall j \in \{1,\ldots, m\}) \qquad \phi_j^+(\mathbf{w}[n+1]) = \begin{cases} y_j^- + \varepsilon - \mathbf{w}[n+1]^T\mathbf{x}'_j[n+1] & \textit{jeżeli } \mathbf{w}[n+1]^T\mathbf{x}'_j[n+1] \leq y_j^- + \varepsilon \\ 0 & \textit{jeżeli } \mathbf{w}[n+1]^T\mathbf{x}'_j[n+1] > y_j^- + \varepsilon \end{cases} \tag{13}$$

$$\phi_j^-(\mathbf{w}[n+1]) = \begin{cases} -y_j^+ + \varepsilon + \mathbf{w}[n+1]^T\mathbf{x}'_j[n+1] & \textit{jeżeli } \mathbf{w}[n+1]^T\mathbf{x}'_j[n+1] \geq y_j^+ - \varepsilon \\ 0 & \textit{jeżeli } \mathbf{w}[n+1]^T\mathbf{x}'_j[n+1] < y_j^+ - \varepsilon \end{cases} \tag{14}$$

gdzie nieujemny margines $\varepsilon$ ($\varepsilon \geq 0$) jest zdefiniowany jako równy minimalnej różnicy $y_j^+ - y_j^-$:

$$\varepsilon = \min_j \{y_j^+ - y_j^-\} \tag{15}$$

Rola marginesu $\varepsilon$ jest zilustrowana na poniższym rysunku:




*Rys. 1: Funkcje kary $\phi_j^+(\mathbf{w}[n+1])$ (13) i $\phi_j^-(\mathbf{w}[n+1])$ (14) z marginesem $\varepsilon$ spełniającym warunek (15)*

*Fig. 1: Penalty function $\phi_j^+(\mathbf{w}[n+1])$ (13) and $\phi_j^-(\mathbf{w}[n+1])$ (14) with margin $\varepsilon$ satisfying (15)*

Obie funkcje kary $\phi_j^+(\mathbf{w}[n+1])$ (13) i $\phi_j^-(\mathbf{w}[n+1])$ (14) są wypukłe i odcinkowo liniowe.

Interwałowa funkcja kryterialna $\Psi(\mathbf{w}[n+1])$ jest zdefiniowana jako dodatnio ważona suma (8) funkcji kary $\phi_j^+(\mathbf{w}[n+1])$ (13) i $\phi_j^-(\mathbf{w}[n+1])$ (14):

$$\Psi(\mathbf{w}[n+1]) = \sum_{j\in\{1,\ldots,m\}} \alpha_j (\phi_j^+(\mathbf{w}[n+1]) + \phi_j^-(\mathbf{w}[n+1])) \tag{16}$$

gdzie $\alpha_j$ ($\alpha_j > 0$) jest dodatnim parametrem (*ceną*) związaną z wektorem cech $\mathbf{x}_j[n]$.

Można zauważyć, że funkcja kryterialna $\Psi(\mathbf{w}[n+1])$ (16) jest funkcją wypukłą i odcinkowo liniową (typu *CPL*).

Algorytm wymiany rozwiązań bazowych, który jest podobny do programowania liniowego, pozwala wyznaczyć minimum funkcji kryterialnej $\Psi(\mathbf{w}[n+1])$ (16) w sposób efektywny, nawet w przypadku dużej liczby $m$ wielowymiarowych wektorów cech $\mathbf{x}_j[n]$ [9].

$$\Psi^* = \Psi(\mathbf{w}^*[n+1]) = \min_{\mathbf{w}[n+1]} \Psi(\mathbf{w}[n+1]) \geq 0 \tag{17}$$

*Twierdzenie* 1: Wartość minimalna $\Psi(\mathbf{w}^*[n+1])$ (17) funkcji kryterialnej $\Psi(\mathbf{w}[n+1])$ (16) jest równa zeru wtedy i tylko wtedy, gdy istnieje taki wektor parametrów $\mathbf{w}\hat{}[n+1]$, który spełnia wszystkie nierówności (3):

$$(\forall j \in \{1,\ldots, m\}) \qquad y_j^- \leq \mathbf{w}\hat{}[n+1]^T \mathbf{x}'[n+1] \leq y_j^+ \tag{18}$$

Wektor $\mathbf{w}^*[n+1]$ tworzący minimum $\Psi(\mathbf{w}^*[n+1])$ funkcji kryterialnej $\Psi(\mathbf{w}[n+1])$ (16) wyznacza optymalny model regresji przedziałowej (1):

$$y = \mathbf{w}^*[n+1]^T \mathbf{x}'[n+1]] \tag{19}$$

*Twierdzenie* 2: Jeżeli wartość minimalna $\Psi(\mathbf{w}^*[n+1])$ (17) funkcji kryterialnej $\Psi(\mathbf{w}[n+1])$ (16) jest równa zeru, to wektor $\mathbf{w}^*[n+1]$ tworzący minimum tej funkcji (16) spełnia wszystkie nierówności (18).

Dowód powyższych dwu twierdzeń może być przeprowadzony w sposób podobny do dowodów twierdzeń zawartych w książce [7].

Minimalizacja funkcji kryterialnej $\Psi(\mathbf{w}[n+1])$ (16) pozwala wyznaczyć wektor optymalny $\mathbf{w}^*[n+1]$ nie tylko w tym przypadku gdy $\Psi(\mathbf{w}^*[n+1]) = 0$, lecz również w tym przypadku gdy $\Psi(\mathbf{w}^*[n+1]) > 0$. W tym ostatnim przypadku, optymalny model regresji przedziałowej (19) wyznaczony przez wektor $\mathbf{w}^*[n+1]$ (17) nie spełnia wszystkich nierówności (18).

## 5 Przykłady zastosowań modeli regresji przedziałowej

W przedstawionej pracy opisana została konstrukcja modeli regresji przedziałowej (19) oparta na minimalizacji interwałowej funkcji kryterialnej $\Psi(\mathbf{w}[n+1])$ (16) zdefiniowanej na podstawie zbioru uczącego $C_m$ o strukturze (2). Struktura interwałowego zbioru uczącego $C_m$ (2) może być traktowana jako uogólnienie klasycznego regresyjnego zbioru uczącego, gdy znane są dokładne wartości $y_j$ zmiennej zależnej. W takim przypadku mamy równości:

$$(\forall j \in \{1,\ldots, m\}) \qquad y_j^- = y_j^+ \tag{20}$$

Tak więc opisana konstrukcja może być stosowana również do budowy liniowych modeli regresyjnych na podstawie klasycznych regresyjnych zbiorów uczących.

Może to mieć szczególne znaczenie zwłaszcza wtedy, gdy pojawiają się trudności z klasyczną konstrukcją modelu liniowej regresji wielowymiarowej na podstawie minimalizacji sumy kwadratów odchyleń $(\hat{y_j} - y_j)^2$ z powodu pojawienia się osobliwości (braku odwracalności) odpowiednich macierzy typu kowariancji. Trudności tego typu pojawiają się m.in. wtedy, liczba $m$ wektorów cech $\mathbf{x}_i[n]$ jest mała w porównaniu z ich wymiarem $n$ ($m << n$). Przykładowo, zbiory dane genetycznych charakteryzują się często taką właściwością, że $m << n$. Minimalizacja interwałowej funkcji kryterialnej $\Psi(\mathbf{w}[n+1])$ pozwala wyznaczać modele regresji przedziałowej (19) zarówno dla przypadku gdy liczba $m$ wektorów cech $\mathbf{x}_j[n]$ jest duża jak również gdy jest mała w porównaniu z wymiarem $n$ .

Modele regresji przedziałowej (19) mogą mieć szczególnie duże znaczenie w przypadku występowania danych cenzorowanych (danych z brakami) pojawiających się w kontekście *analizy przeżycia* (ang. *survival analysis*) [6]. Dane cenzorowane mają strukturę przedziałowego zbioru uczącego $C_m$ (2) z dodatkowym warunkiem, że przedział jest nieograniczony z jednej strony ($y_j^- = -\infty$, lub $y_j^+ = \infty$):

$$(\exists j \in \{1,\ldots, m\}) \qquad \{\mathbf{x}_j[n], [-\infty, y_j^+]\} \tag{21}$$

lub

$$(\exists j \in \{1,\ldots, m\}) \qquad \{\mathbf{x}_j[n], [y_j^-, \infty]\} \tag{22}$$

W przypadku (21) mówimy o *lewostronnym cenzorowaniu* danych (np. pacjent $O_i$ reprezentowany za pomocą wektora $x_i[n]$ żył po zabiegu <u>*nie dłużej*</u> niż $y_i^+$ lat) [6]. W przypadku (22) mówimy o *prawostronnym cenzorowaniu* danych (np. pacjent $O_i$ reprezentowany za pomocą wektora $x_i[n]$ żył po zabiegu <u>*nie krócej*</u> niż $y_i^-$ lat). Model liniowej regresji przedziałowej (19) może być zbudowany wyłącznie na danych cenzorowowanych (21) lub (22). Model (19) może być używany w prognozowaniu, w tym przypadku pozawala on przewidywać czas życia nowego pacjenta $O_0$ reprezentowanego za pomocą wektora cech $x_0[n]$.

## 6 Uwagi końcowe

Zbiory uczące $C_m$ typu interwałowego (2) pojawiają się w wielu zagadnieniach praktycznych. Zbiory tego typu pojawiają się między innymi w warunkach nieprecyzyjności pomiarów, co na ogół występuje w praktyce. Jeżeli zmienna zależna $Y$ mierzona jest na obiekcie $O_i$ nieprecyzyjnie, to naturalnym przedstawieniem wyniku takiego pomiaru może być przedział liczbowy $[y_j^-, y_i^+]$, gdzie liczba $y_j^-$ ogranicza nieznaną wartość zmiennej $Y$ od dołu, natomiast liczba $y_j^+$ ogranicza tę wartość od góry. W tym aspekcie, konstrukcja liniowych modeli regresyjnych na bazie interwałowych zbiorów uczących $C_m$ (2) może mieć znaczące praktyczne zastosowania.

Szczególnie interesującym obszarem zastosowań interwałowego projektowania modeli regresyjnych (19) jest *analiza przeżycia* (ang. *survival analysis*) [6]. W tym przypadku, projektowanie prognostycznych modeli regresyjnych (19) odbywa się poprzez

minimalizację wypukłych i odcinkowo liniowych (typu *CPL*) funkcji kryterialnych $\Psi(w[n+1])$ (16) definiowanych na zbiorach uczących $C_m$ (2) z przedziałami $[y_j^-, y_j^+]$ ograniczonymi tylko z jednej strony ($[-\infty, y_j^+]$ (21) ) lub $[y_j^-, \infty]$ (22)). Analiza przeżycia jest standardowo stosowana przy ocenie ryzyka związanego ze stosowaniem nowych leków lub nowych sposobów terapii w medycynie. Metody analizy przeżycia stosuje się jednak również w innych obszarach, np. w zagadnieniach ekonometrycznych przy prognozowaniu działalności, rozwoju i upadku poszczególnych firm [10].

Model Cox'a gra obecnie fundamentalną rolę w konstrukcji modeli prognostycznych na bazie cenzorowanych danych i przy rozwiązywaniu wielu praktycznych problemów metodami analizy przeżycia [9]. Proponowana w tej pracy konstrukcja modeli prognostycznych bazująca na minimalizacji wypukłych i odcinkowo liniowych (typu *CPL*) funkcji kryterialnych $\Psi(w[n+1])$ (16) może być traktowana jako rozwiązanie w pewnych aspektach konkurencyjne względem rozwiązań opartych na modelu Cox'a. Do najważniejszych zalet proponowanej metody jest efektywność obliczeniowa algorytmów wymiany rozwiązań bazowych stosowanych minimalizacji funkcji $\Psi(w[n+1])$ (16) [8]. Innym ważnym aspektem proponowanej metody jest możliwość wkomponowania procesu selekcji cech w proces minimalizacji funkcji $\Psi(w[n+1])$ (16). Zaproponowana ostatnio metoda selekcji cech o nazwie *relaksowana separowalność liniowa* (ang. *relaxed linear separabilty*) [11] pozwala no wydobywanie zestawów cech o największym wpływie na prognozowany proces. Posłużyło to m.in. przy próbach identyfikacji zestawów genetycznych czynników zwiększających ryzyko pojawienia się chorób nowotworowych.

## Literatura


1. Johnson R. A., Wichern D. W.: *Applied Multivariate Statistical Analysis*, Prentice-Hall, Inc., Englewood Cliffs, New York, 1991
2. Duda O. R, Hart P. E., Stork D. G.: *Pattern Classification,* J. Wiley, New York, 2001.
3. Bobrowski L.: Ranked linear models and sequential patterns recognition, pp. 1-7 in: *Pattern Analysis & Applications*, Volume 12, Issue1 (2009)
4. Bobrowski L., Łukaszuk T., Wasyluk H.: Ranked modeling of causal sequences of diseases for the purpose of early diagnosis, pp. 23-31 in: *Computers in Medical Activity*, E. Kącki, M. Rudnicki, J. Stempczyńska (Eds.), *Advances in Intelligence and Soft Computing* 65, Springer Verlag 2009.
5. Li G., Zhang C.: Linear regression with interval censored data, *The Annals of Statistics*, 1998, Vol. 26, No. 4, 1306-1327
6. Klein J. P. Moeschberger M. L.: *Survival Analysis*, Techniques for Censored and Truncated Data, Springer, NY 1997
7. Bobrowski L.: Regresja przedziałowa oparta na funkcjach kryterialnych typu *CPL*, w streszczeniach referatów *XXXVIII Ogólnopolskiej Konferencji Zastosowań Matematyki*, Zakopane 2009
8. Bobrowski L.: *Eksploracja danych oparta na wypukłych i odcinkowo-liniowych funkcjach kryterialnych* (*Data mining based on convex and piecewise linear (CPL) criterion functions*) (in Polish), Technical University Białystok, 2005
9. Bobrowski L.: Design of piecewise linear classifiers from formal neurons by some basis exchange technique *Pattern Recognition*, 24(9), pp. 863-870, 1991

10. Frątczak E., Gach-Ciepela U., Babiker H.: *Analiza Historii Zdarzeń*, SGH, Warszawa 2005
11. Bobrowski L., Łukaszuk T.: Feature selection based on relaxed linear separability, *Biocybernetics and Biomedical Engineering* 2009, Volume 29, Number 2, pp. 43-59

## Streszczenie

Wielowymiarowe modele regresyjne są używane dla celów prognozowania. Parametry takich modeli estymowane są na podstawie zbioru wektorów cech grupujących wartości zmiennych niezależnych oraz wartości zmiennej zależnej. W wielu ważnych zastosowaniach dokładne wartości zmiennej zależnej nie mogą być określone a znane są tylko przedziały, które zawierają te wartości. W takich przypadkach stosuje się metody regresji przedziałowej. W pracy opisana jest konstrukcja liniowych modeli regresyjnych oparta na minimalizacji wypukłych i odcinkowo liniowych funkcji kryterialnych (typu *CPL*), które są zdefiniowane na przedziałowych zbiorach uczących.

# Linear prognostic models based on interval regression with the *CPL* criterion functions

## Summary

Multivariate regression models are used for the prognosis purposes. Parameters of such models are estimated on the basis of feature vectors (independent variables) combined with values of response (target) variable. The exact values of response variable can be not determined exactly in some important applications. For example, the values of response variable can be censored and given as intervals. The interval regression approach has been proposed for designing prognostic tools in such circumstances. The possibility of using the convex and piecewise linear (*CPL*) functions for designing interval regression models is examined in the paper.

Praca wspierana przez projekt S/WI/2/2010 Politechniki Białostockiej