ZESZYTY NAUKOWE POLITECHNIKI BIAŁOSTOCKIEJ 2007
Informatyka – Zeszyt 2

Tomasz Łukaszuk[1], Leon Bobrowski[1,2]

TEMPORALNOŚĆ W MODELACH RANGOWYCH

Streszczenie: W zbiorze danych określony jest pewien porządek czasowy dla wybranych obiektów. Poprzez model rangowy rozumiemy taką liniową transformację, która zachowuje w najlepszym możliwym stopniu wiedzę a priori o uporządkowaniu obiektów. W artykule przedstawiono koncepcję budowy modelu rangowego opierając się na minimalizacji wypukłej i odcinkowo-liniowej (CPL) funkcji kryterialnej. Zagadnienie zostało sprowadzone do problemu znalezienia optymalnej hiperpłaszczyzny rozdzielającej zbiory zbudowane z elementów powstałych z różnic arytmetycznych wektorów cech tworzących pary, dla których określony jest porządek czasowy.

Słowa kluczowe: model rangowy, wypukła i odcinkowo-liniowa funkcja kryterialna (CPL), liniowa separowalność zbiorów danych

1. Wprowadzenie

W związku z ogromnym wzrostem w ostatnich latach liczby gromadzonych danych, wzrosło także zainteresowanie wydobywaniem ukrytych w tych danych informacji. To wydobywanie informacji polega głównie na klasyfikowaniu, grupowaniu i odnajdywaniu zależności w danych [8], [9]. Jednym z ważniejszych problemów analizy jest postępowanie z danymi zawierającymi zależności czasowe.

Zależności czasowe mogą być określone w postaci pewnego porządku czasowego pomiędzy wybranymi obiektami ze zbioru danych. Na przykład, możemy posiadać informację, że pewne obiekty są starsze (bardziej zaawansowane w rozpatrywanym procesie) niż obiekty z jednego zbioru, natomiast te same obiekty są młodsze (mniej zaawansowane w procesie) niż obiekty z drugiego zbioru. Tego typu wiedza a priori na temat relacji wybranych obiektów może być podstawą do utworzenia modelu rangowego.

Poprzez model rangowy rozumiemy tutaj taką liniową transformację, która zachowuje w najlepszym możliwym stopniu wiedzę *a priori* o uporządkowaniu obiektów. Proces budowy modelu rangowego polega na znalezieniu parametrów

[1] Wydział Informatyki, Politechnika Białostocka, Białystok
[2] Instytut Biocybernetyki i Inżynierii Biomedycznej, PAN, Warszawa

przekształcenia liniowego na podstawie wiedzy *a priori* o porządku istniejącym w danych.

Celem autorów artykułu jest przedstawienie procedury budowy modelu rangowego, bazującej na minimalizacji wypukłej i odcinkowo-liniowej (CPL) funkcji kryterialnej. Funkcja kryterialna tego typu jest sumą dodatnich i ujemnych funkcji kary typu CPL, które są zdefiniowane na podstawie różnic arytmetycznych pomiędzy wektorami cech tworzącymi dipole (pary obiektów, co do których posiadamy wiedzę *a priori* o relacji czasowej pomiędzy nimi) [2]. W ten sposób zadanie budowy modelu rangowego może być sprowadzone do problemu zapewnienia liniowej separowalności dwóch zbiorów danych w zadanej przestrzeni cech.

2. Liniowa transformacja rangowa

Niech badane obiekty (np. pacjenci, samochody, obrazy graficzne) O_j $(j = 1,...,m)$ będą reprezentowane przez n-wymiarowe wektory cech $\mathbf{x}_j = [x_{j1},...,x_{jn}]^T$. Cecha (atrybut) x_i opisuje wartość liczbową określonego $i-$tego parametru lub wynik określonego badania obiektu O_j. Cechy mogą być binarne ($x_i \in \{0,1\}$) lub ciągłe ($x_i \in R^1$).

W zbiorze O_j obiekty są w pewien sposób uporządkowanie. Uporządkowanie to ma (może mieć) charakter jakościowy i wskazuje, że np. pewne obiekty są starsze lub lepsze pod określonym względem niż inne obiekty. Zakładamy, że wiedza *a priori* o uporządkowaniu obiektów dana jest w postaci relacji następstwa „$\prec$" wybranych par wektorów cech.

$$\mathbf{x}_j \prec \mathbf{x}_{j'} \Leftrightarrow \mathbf{x}_{j'} \quad nastepuje\_po \quad \mathbf{x}_j \tag{1}$$

Jeżeli wektory cech $\mathbf{x}_j$ i $\mathbf{x}_{j'}$ pozostają w relacji następstwa (1), oznacza to, że obiekt $O_{j'}$ reprezentowany przez wektor $\mathbf{x}_{j'}$ jest bardziej zaawansowany ze względu na rozważany czynnik niż obiekt O_j, reprezentowany przez wektor $\mathbf{x}_j$.

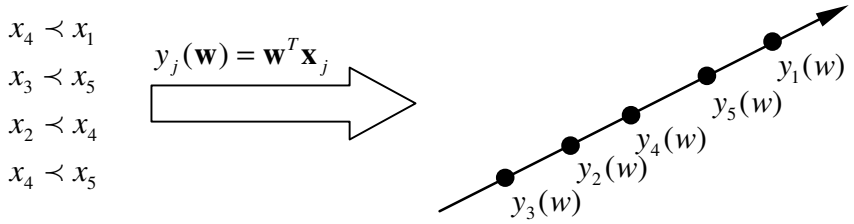
Rozważamy odwzorowanie liniowe postaci

$$y_j = \mathbf{w}^T \mathbf{x}_j \quad (j = 1,...,m) \tag{2}$$

gdzie $\mathbf{w} = [w_1,...,w_N]^T \in R^N$ jest wektorem parametrów. Odwzorowanie (2) przyporządkowuje poszczególnym wektorom cech $\mathbf{x}_j$ punkty y_j na prostej. Punkty y_j mogą być uporządkowane na prostej zgodnie z relacją większościową

$$y_{j(1)} < y_{j(2)} < ... < y_{j(m)} \tag{3}$$

Jesteśmy zainteresowani wyborem takiego wektora parametrów $\mathbf{w}$, który daje największą możliwą zgodność uporządkowania punktów y_j na prostej (2) z relacją następstwa „$\prec$" wektorów cech $\mathbf{x}_j$. Wyznaczenie wektora $\mathbf{w}$ o tej właściwości nazywamy zagadnieniem regresji rangowej.



Rys. 1. Przykład relacji następstwa oraz uporządkowania punktów na prostej zgodnie z tą relacją

3. Dodatnio i ujemnie zorientowane dipole

Relacja następstwa „$\prec$" może być użyta w określaniu orientacji dipoli $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ ($j < j'$) utworzonych z wektorów cech, dla których relacja jest dana.

Definicja 1: Para $(\mathbf{x}_j, \mathbf{x}_{j'})$ ($j < j'$) wektorów cech $\mathbf{x}_j$ i $\mathbf{x}_{j'}$ tworzy dipol z orientacją dodatnią $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ ($(j, j') \in I^+$) wtedy i tylko wtedy, gdy $\mathbf{x}_j \prec \mathbf{x}_{j'}$.

$$(\forall (j, j') \in I^+) \quad \mathbf{x}_j \prec \mathbf{x}_{j'} \tag{4}$$

gdzie I^+ jest zbiorem indeksów (j, j') dipoli z orientacją dodatnią $(\mathbf{x}_j, \mathbf{x}_{j'})$ ($j < j'$).

Definicja 2: Para $(\mathbf{x}_j, \mathbf{x}_{j'})$ $(j < j')$ wektorów cech $\mathbf{x}_j$ i $\mathbf{x}_{j'}$ tworzy dipol z orientacją ujemną $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ $((j, j') \in I^-)$ wtedy i tylko wtedy gdy $\mathbf{x}_{j'} \prec \mathbf{x}_j$.

$$(\forall (j, j') \in I^-) \quad \mathbf{x}_{j'} \prec \mathbf{x}_j \tag{5}$$

gdzie I^- jest zbiorem indeksów (j, j') dipoli z orientacją ujemną $(\mathbf{x}_j, \mathbf{x}_{j'})$ $(j < j')$.

Zgodnie z relacją (4) drugi wektor $\mathbf{x}_{j'}$ w parze $(\mathbf{x}_j, \mathbf{x}_{j'})$ następuje po $\mathbf{x}_j$. W przypadku relacji (5) pierwszy wektor $\mathbf{x}_j$ następuje po $\mathbf{x}_{j'}$.

Definicja 3: Uporządkowanie punktów y_j na prostej (2) jest zgodne z relacją „$\prec$" (1) pomiędzy wektorami cech $\mathbf{x}_j$ wtedy i tylko wtedy gdy spełnione są poniższe relacje.

$$\begin{aligned} &(\forall (j, j') \in I^+) \quad y_j < y_{j'} \\ &(\forall (j, j') \in I^-) \quad y_j > y_{j'} \end{aligned} \tag{6}$$

4. Zbiory C+ i C- i ich liniowa separowalność

Relacje (6) mogą być przedstawione w równoważnej poniższej postaci.

$$\begin{aligned} &(\forall (j, j') \in I^+) \quad \mathbf{w}^T (\mathbf{x}_{j'} - \mathbf{x}_j) > 0 \\ &(\forall (j, j') \in I^-) \quad \mathbf{w}^T (\mathbf{x}_{j'} - \mathbf{x}_j) < 0 \end{aligned} \tag{7}$$

Zdefiniujmy dwa zbiory C^+ i C^- składające się z wektorów $\mathbf{r}_{jj'}$ utworzonych z różnic arytmetycznych wektorów cech $\mathbf{x}_j$ i $\mathbf{x}_{j'}$ tworzących dipole $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$.

$$\begin{aligned} C^+ &= \{\mathbf{r}_{jj'} = (\mathbf{x}_{j'} - \mathbf{x}_j) : (j, j') \in I^+\} \\ C^- &= \{\mathbf{r}_{jj'} = (\mathbf{x}_{j'} - \mathbf{x}_j) : (j, j') \in I^-\} \end{aligned} \tag{8}$$

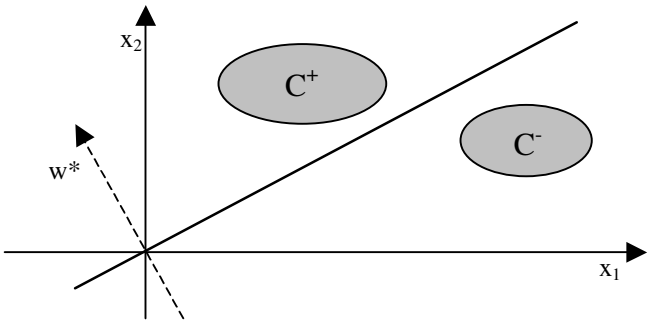
Liniowa separowalność zbiorów C^+ i C^- przez hiperpłaszczyznę $H(\mathbf{w})$ przechodzącą przez początek układu współrzędnych zapewnia spełnienie relacji (7).

$$H(\mathbf{w}) = \{\mathbf{x} : \mathbf{w}^T \mathbf{x} = 0\} \tag{9}$$

Uwzględniając definicje (8) zbiorów C^+ i C^-, relacje (7) można przedstawić w następujący sposób [5]:

$$(\exists \mathbf{w}^*) \begin{array}{ll} (\forall (j, j') \in I^+) & (\mathbf{w}^*)^T \mathbf{r}_{jj'} \geq 1 \\ (\forall (j, j') \in I^-) & (\mathbf{w}^*)^T \mathbf{r}_{jj'} < -1 \end{array} \tag{10}$$

Zagadnienie regresji rangowej sprowadza się do wyznaczenia parametrów $\mathbf{w}^*$ optymalnej hiperpłaszczyzny $H(\mathbf{w}^*)$ separującej zbiory C^+ i C^-. Optymalność hiperpłaszczyzny $H(\mathbf{w}^*)$ rozumieć należy w ten sposób, że jest to hiperpłaszczyzna zapewniająca największy możliwy margines pomiędzy nią samą a elementami zbiorów C^+ i C^- [2]. Tak wyznaczona hiperpłaszczyzna nie powinna nadmiernie dopasowywać się do danych uczących i mieć dużą zdolność generalizacji.



Rys. 2. Zbiory C^+ i C^- w przestrzeni dwuwymiarowej i hiperpłaszczyzna separująca H(w*)

5. Funkcja kryterialna CPL

Hiperpłaszczyznę $H(\mathbf{w}^*)$ optymalnie rozdzielającą zbiory C^+ i C^- można wyznaczyć minimalizując regresyjno-rangową funkcję kryterialną $\Phi_r(\mathbf{w})$ [2].

$$\Phi_r(\mathbf{w}) = \sum_{(j,j') \in I^+} \varphi^+_{jj'}(\mathbf{w}) + \sum_{(j,j') \in I^-} \varphi^-_{jj'}(\mathbf{w}) \tag{11}$$

Funkcja $\Phi_r(\mathbf{w})$ jest sumą dodatnich $\varphi^+_{jj'}(\mathbf{w})$ i ujemnych $\varphi^-_{jj'}(\mathbf{w})$ funkcji kary.

$$(\forall(j,j') \in I^+)$$
$$\varphi^+_{jj'}(\mathbf{w}) = \begin{cases} 1-\mathbf{w}^T\mathbf{r}_{jj'} & jezeli \quad \mathbf{w}^T\mathbf{r}_{jj'} < 1 \\ 0 & jezeli \quad \mathbf{w}^T\mathbf{r}_{jj'} \geq 1 \end{cases} \tag{12}$$

$$(\forall(j,j') \in I^-)$$
$$\varphi^-_{jj'}(\mathbf{w}) = \begin{cases} 1-\mathbf{w}^T\mathbf{r}_{jj'} & jezeli \quad \mathbf{w}^T\mathbf{r}_{jj'} > -1 \\ 0 & jezeli \quad \mathbf{w}^T\mathbf{r}_{jj'} \leq -1 \end{cases} \tag{13}$$

Funkcja kryterialna $\Phi_r(\mathbf{w})$ jest funkcją wypukłą i odcinkowo-liniową (CPL). Algorytm wymiany rozwiązań bazowych, technika zbliżona do programowania liniowego, pozwala znaleźć minimum tego typu funkcji w sposób efektywny, nawet przy dużych, wysokowymiarowych zbiorach danych C^+ i C^- [2].

$$\Phi^* = \Phi(\mathbf{w}^*) = \min \Phi(\mathbf{w}) \geq 0 \tag{14}$$

Wektor parametrów $\mathbf{w}^*$ definiuje prostą $y_j = (\mathbf{w}^*)^T\mathbf{x}_j$ (2), będącą najlepszą z punktu widzenia zagadnienia regresji rangowej. Jeżeli wartość funkcji kryterialnej Φ^* jest równa 0, to model prawidłowo zachowuje wszystkie relacje (1) wektorów $\mathbf{x}_j$ i $\mathbf{x}_{j'}$ ze zbioru danych [4]. Gdy wartość funkcji Φ^* jest większa od 0, model uwzględnia największą możliwą liczbę relacji (1). Taka sytuacja oznacza, że niemożliwe jest uwzględnienie wszystkich relacji w przestrzeni cech o zadanym wymiarze.

6. Wyniki eksperymentów

Przy zastosowaniu opisanych wcześniej metod wykonane były eksperymenty na dwóch zbiorach danych. W pierwszym eksperymencie użyto części zbioru Primary Biliary Cirrhosis (PBC) z repozytorium UCI [1], w drugim danych z systemu komputerowego wspierania diagnostyki chorób wątroby „Hepar" [7].

6.1. Eksperyment 1

Zbiór PBC zawiera informacje opisujące pacjentów cierpiących na pierwotną żółciową marskość wątroby. Dane były zbierane w klinice Mayo w USA w latach

1974-1984. Każdy pacjent O_j w tym zbiorze opisany jest przez 17 atrybutów ($x_1, x_2, ..., x_{17}$). Ponadto dla każdego pacjenta przypisano czas przeżycia t_j i wskaźnik niepowodzenia δ_j ($\delta_j = \{0,1\}$). Czas przeżycia jest to liczba dni od momentu stwierdzenia choroby i rozpoczęcia obserwacji do momentu śmierci pacjenta, transplantacji wątroby lub zakończenia badań w lipcu 1986 roku. Wskaźnik niepowodzenia $\delta_j = 1$ oznacza, że obserwacja pacjenta była przerwana z powodu jego śmierci, $\delta_j = 0$ oznacza, że obserwację przerwano przed śmiercią pacjenta ze względu na transplantację lub zakończenie badań. Sytuację drugiego typu nazywamy cenzorowaniem lub ucinaniem [10].

Dane użyte do eksperymentu zawierały obserwacje $(\mathbf{x}_j, t_j, \delta_j)$ 30 pacjentów O_j. Wśród nich 18 obserwacji było uciętych ($\delta_j = 0$). Zbiory C^+ i C^- wektorów różnicowych $\mathbf{r}_{jj'} = (\mathbf{x}_{j'} - \mathbf{x}_j)$ zostały utworzone na podstawie 30 wektorów cech $\mathbf{x}_j$. Bazują one na wszystkich dipolach $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ ($j < j'$) zorientowanych zgodnie z następującą zasadą

$$\mathbf{x}_j \prec \mathbf{x}_{j'} \Leftrightarrow \delta_j = 1 \quad i \quad t_j < t_{j'} \tag{15}$$

Model rangowy otrzymany w wyniku minimalizacji funkcji $\Phi_r(\mathbf{w})$ (11) ma formę

$$\begin{aligned} y_j = {} & \text{-}0{,}0016x_{j1} - 34{,}5864x_{j6} + 1{,}0555x_{j7} - 0{,}0424x_{j8} - 8{,}5786x_{j9} + \\ & + 0.0915x_{j10} - 0{,}0005x_{j11} - 0{,}1322x_{j12} - 4{,}2516x_{j16} \end{aligned} \tag{16}$$

Wartość funkcji kryterialnej $\Phi(\mathbf{w}^*) = 0$, więc model (16) zachowuje wszystkie informacje o uporządkowaniu obiektów w zbiorze danych.

Otrzymany model ma zastosowanie do prognozowania czasu przeżycia dla obserwacji uciętych. Aby móc bezpośrednio wykorzystywać generowane przez niego wyniki zastosowane zostało dodatkowe przekształcenie skalujące:

$$y_j' = \alpha y_j + \beta \tag{17}$$

gdzie α i β są parametrami skalującymi wyznaczonymi poprzez minimalizację sumy różnic $\left| t_j - \alpha y_j - \beta \right|$ dla wszystkich nieuciętych obserwacji. W wyniku tej operacji otrzymano przekształcenie o następujących parametrach:

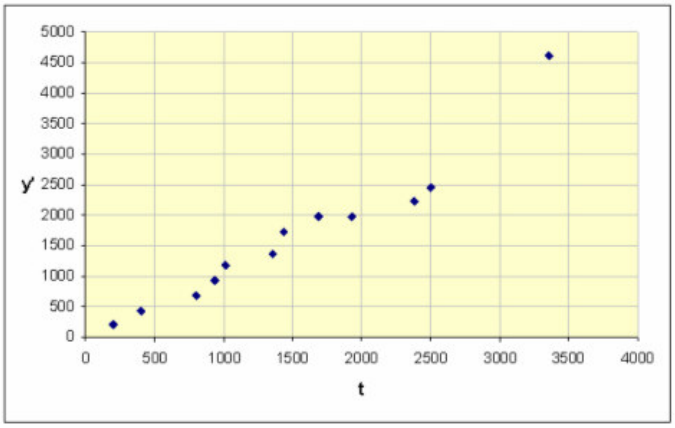
$$y_j' = 244{,}2\, y_j + 26617{,}75 \tag{18}$$

Wyznaczanie parametrów przekształcenia (17) opiera się na metryce L_1. Wybór metryki jest arbitralny i wynika z możliwości zastosowania w przypadku tej metryki minimalizacji zmodyfikowanej funkcji kryterialnej typu CPL (11), podobnie jak przy wyznaczaniu parametrów modelu (16).

Rysunki 3 i 4 obrazują wyniki eksperymentu. W górnej części tabeli na rysunku 3, nad kreską, umieszczono dane dotyczące obserwacji nieuciętych. Wartości otrzymane na podstawie modelu y_j' powinny być zbliżone do danych *a priori* rzeczywistych czasów przeżycia t_j. W dolnej części umieszczono dane dotyczące obserwacji uciętych. W tym przypadku wartości otrzymane na podstawie modelu y_j' są prognozą czasu przeżycia. W ten sposób w momencie przerwania badania z innego powodu niż śmierć pacjenta możemy określić przypuszczalną długość jego życia. Wykres na rysunku 4 przedstawia zależność rzeczywistego czasu przeżycia do czasu przeżycia wygenerowanego przez model (17), (18) dla obserwacji nieuciętych. Im bardziej zależność ta zbliżona jest do funkcji liniowej $y_j' = t_j$, tym lepszy jest model (17), (18).

tj	yj	yj'
198	-108,188	198,2751
400	-107,188	442,4751
799	-106,188	686,6751
930	-105,188	930,8751
1012	-104,188	1175,075
1360	-103,372	1374,248
1434	-101,914	1730,452
1690	-100,914	1974,652
1925	-100,914	1974,652
2386	-99,9136	2218,852
2503	-98,9136	2463,052
3358	-90,1398	4605,608
2272	-94,9121	3440,207
1615	-93,7519	3723,548
2255	-92,2662	4086,336
3099	-72,8111	8837,27
1592	-98,9656	2450,354
2318	-74,7733	8358,118
2294	-89,8265	4682,117
3069	-95,9002	3198,92
3297	-98,0101	2683,689
1701	-82,4123	6492,676
3255	-90,4258	4535,769
2944	-96,5152	3048,734
2468	-98,6327	2531,644
1614	-92,9708	3914,269
1702	-85,9536	5627,88
2033	-52,5628	13781,9
737	-77,7601	7628,724
1735	-76,7314	7879,937

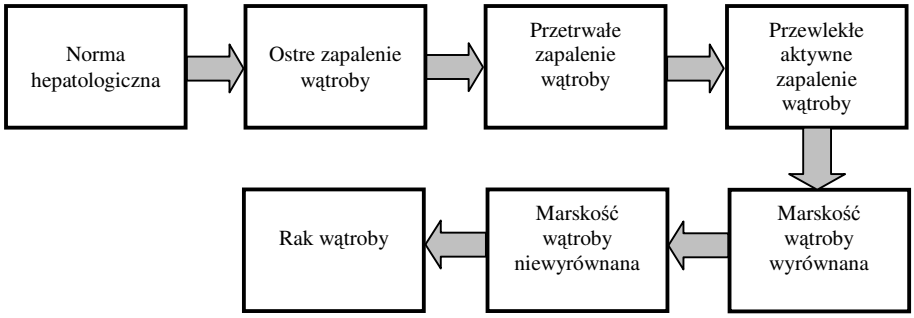
Rys. 3. Wyniki eksperymentu na zbiorze danych PBC



Rys. 4. Wyniki eksperymentu na zbiorze danych PBC, wykres zależności rzeczywistego czasu przeżycia do czasu przeżycia wygenerowanego przez model

6.2. Eksperyment 2

W drugim eksperymencie użyto danych zgromadzonych w systemie „Hepar". System ten zbudowano w Instytucie Biocybernetyki i Inżynierii Biomedycznej PAN. Baza danych systemu zawiera opisy pacjentów z przewlekłymi chorobami wątroby, leczonych w Klinice Gastroenterologii Instytutu Żywności i Żywienia w Warszawie. Opis każdego pacjenta składa się z około 200 atrybutów. Są to odpowiedzi uzyskane od pacjenta podczas wywiadu lekarskiego, symptomy z badania przedmiotowego oraz rezultaty diagnostycznych testów laboratoryjnych. Z powodu braków w danych do eksperymentu wybrano 62 atrybuty, których wartości są ustalone dla wszystkich pacjentów. Czynnikiem czasowym jest stopień zaawansowania choroby η. Uporządkowanie stopni zaawansowania choroby przedstawia rysunek 5 (Norma hepatologiczna $\eta = 1$, Ostre zapalenie wątroby $\eta = 2$, Przetrwałe zapalenie wątroby $\eta = 3$, Przewlekłe aktywne zapalenie wątroby $\eta = 4$, Marskość wątroby wyrównana $\eta = 5$, Marskość wątroby niewyrównana $\eta = 6$, Rak wątroby $\eta = 7$) [6].



Rys. 5. Kolejne etapy chorób wątroby według systemu „Hepar"

Dane użyte w eksperymencie zawierały opisy 272 pacjentów O_j. Zbiory C^+ i C^- wektorów różnicowych $\mathbf{r}_{jj'} = (\mathbf{x}_{j'} - \mathbf{x}_j)$ zbudowano na podstawie wszystkich możliwych 28769 dipoli $\{\mathbf{x}_j, \mathbf{x}_{j'}\}$ ($j < j'$). Dipol tworzą dwa wektory $\mathbf{x}_j$ i $\mathbf{x}_{j'}$ opisujące pacjentów z różnym stopniem zaawansowania choroby. Otrzymany w wyniku minimalizacji funkcji kryterialnej $\Phi_r(\mathbf{w})$ (11) model rangowy ma następującą postać.

$$\begin{aligned} y_j = [&0{,}721\ -0{,}186\ -0{,}16\ 0{,}89\ 0{,}249\ -0{,}081\ 0{,}051\ -0{,}274\ -0{,}511 \\ &-0{,}3\ 0{,}358\ 0{,}627\ 0{,}038\ -0{,}406\ 0{,}393\ 0{,}55\ -0{,}09\ 0{,}297\ -0{,}367\ 0{,}064 \\ &-0{,}338\ -0{,}156\ 0{,}048\ 0{,}147\ 0{,}566\ 0{,}602\ -0{,}354\ -2{,}553\ -0{,}1\ 0{,}42 \\ &-0{,}973\ 0{,}968\ -0{,}727\ -0{,}412\ 0{,}115\ 0{,}09\ -0{,}697\ 0{,}196\ 0{,}638\ -2{,}468 \\ &0{,}895\ -0{,}275\ -0{,}58\ 0{,}11\ -0{,}041\ ,125\ -0{,}064\ 0{,}082\ -0{,}196\ 0{,}259 \\ &-0{,}504\ 0{,}292\ 0{,}68\ -0{,}109\ 0{,}767\ 0{,}333\ -0{,}042\ 1{,}048\ 0{,}374\ -0{,}448]^T \cdot \\ &\cdot [x_{j1},...,x_{j51},x_{j54},...,x_{j62}] \end{aligned} \quad (19)$$

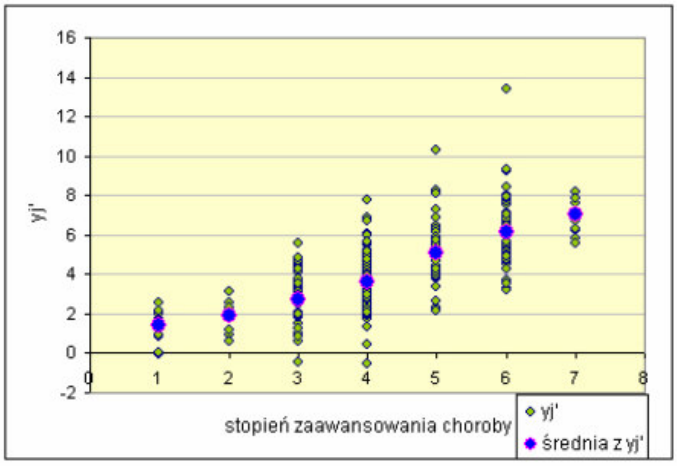
Podobnie jak w opisanym wcześniej eksperymencie 1 zastosowano przekształcenie skalujące. Parametry α i β wyznaczono poprzez minimalizację sumy różnic $|\eta - \alpha y_j - \beta|$ dla wszystkich obiektów ze zbioru danych.

$$y_j' = 0{,}58539\, y_j + 3{,}50484 \quad (20)$$

Wartość funkcji kryterialnej $\Phi(\mathbf{w}^*) > 0$. Model (19) nie zachowuje wszystkich informacji o uporządkowaniu obiektów O_j. Niedoskonałość modelu wynika z dużej liczby obiektów w stosunku do liczby opisujących je atrybutów. W zadanej

przestrzeni atrybutów nie ma możliwości znalezienia hiperpłaszczyzny liniowo separującej zbiory C^+ i C^-. Jednak, jak widać na rysunku 6, w ujęciu średnich wartości y_j' tendencja jest zachowana.

η	Liczność klasy	$\overline{y_j'}$
1	16	1,413581
2	8	1,928993
3	44	2,751939
4	95	3,649423
5	38	5,081494
6	60	6,143924
7	11	7,030914



Rys. 6. Wyniki eksperymentu na zbiorze danych „Hepar"

Na podstawie otrzymanego modelu (19), (20) możliwe jest określenie z pewnym prawdopodobieństwem stopnia zaawansowania choroby w przypadku nowego pacjenta.

7. Podsumowanie

W artykule przedstawiono metodę budowy modelu rangowego, opierając się na minimalizacji wypukłej i odcinkowo linowej (CPL) funkcji kryterialnej (11). Zagadnienie regresji rangowej zostało sprowadzone do problemu znalezienia hiperpłaszczyzny (9) optymalnie rozdzielającej zbiory C^+ i C^-, zbudowane na podstawie dipoli utworzonych z wektorów cech, dla których określony jest *a priori* porządek czasowy. Minimalizacja funkcji kryterialnej (11) może być efektywnie przeprowadzona poprzez zastosowanie algorytmu wymiany rozwiązań bazowych [2], techniki zbliżonej do programowania liniowego.

Transformację rangową stosuje się między innymi w celu uzyskania poprawy procesu wspomagania podejmowania decyzji klasyfikacyjnych. Model rangowy zbudowany na podstawie zbioru uczącego pozwala na późniejsze klasyfikowanie obiektów niebiorących udziału w procesie uczenia. Takie zastosowanie modelu rangowego może być rozszerzeniem opisanego eksperymentu 2.

Model rangowy może być także używany przy prognozowaniu nieznanych wartości określonych parametrów, np. czasu życia dla obserwacji uciętych w analizie przeżyć. To zastosowanie jest przedmiotem zaprezentowanego eksperymentu 1.

Literatura:

[1] Blake, C., Merz, C.: *UCI Repository of machine learning databases* [http://www.ics.uci.edu/~mlearn/MLRepository.html] Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[2] Bobrowski, L.: *Eksploracja danych oparta na wypukłych i odcinkowo-liniowych funkcjach kryterialnych*, Białystok: Wydaw. Politechniki Białostockiej, 2005.

[3] Bobrowski L.: *Ranked modeling with feature selection based on the CPL criterion functions*, s.218-227 w: Machine learning and data mining in pattern recognition, eds. Petra Perner, Atsushi Imiya, Lecture Notes in Computer Science, vol.3587, 2005.

[4] Bobrowski L., Łukaszuk T.: *Ranked linear modeling in survival analysis*, s.61-67 w: [Sixth International Seminar] Statistics and Clinical Practice, Warsaw, June, 2005, ed by L. Bobrowski, J. Doroszewski, C. Kulikowski, N.Victor, Lecture Notes of the ICB Seminars 70, Warsaw, 2005.

[5] Bobrowski L, Łukaszuk T.: *Selection of the linearly separable feature subsets*, s.544-549 w: Artificial intelligence and soft computing : ICAISC'2004, eds. Leszek Rutkowski, Jörg Siekmann, Ryszard Tadusiewicz, Lotfi A. Zadeh, Lecture Notes in Computer Science, vol.3070, 2004.

[6] Bobrowski L., Łukaszuk T, Wasyluk H.: *Ranked modeling of liver diseases sequence*, wysłane do European Journal of Biomedical Informatics.

[7] Bobrowski L., Wasyluk H.: *Diagnosis supporting rules of the HEPAR system*, s.1309-1313 w: MEDINFO 2001: Proceedings of the 10th World Congress on Medical Informatics. P.2, London, September 2-5, 2001, ed. by V. L. Patel, R. Rogers, R. Haux, Amsterdam: IOS Press, 2001.

[8] Duda, O.R., Heart, P.E., Stork D.G.: *Pattern Classification*, Wydanie drugie, zmienione, John Wiley & Sons, 2001.

[9] K. Fukunaga: *Statistical Pattern Recognition*, Academic Press, Inc., San Diego, 1990.

[10] Klein J. P., Moeschberger M. L.: *Survival Analysis, Techniques for Censored and Truncated Data*, Springer, NY 1997.

TEMPORALITY IN RANKED MODELS

Abstract: A known temporal order between selected objects in a data set is given. We assume the ranked model is such a linear transformation, which preserve in the most possible manner the a priori knowledge of the order between objects. The procedure of the ranked models design which is based on the minimisation of the convex and piecewise linear (CPL) criterion functions is presented in the paper. The task of the ranked model design is boiled down to the problem of searching an optimal hyperplane separated the sets constructed on the basis of the elements created from the arithmetic substractions of the vectors – the pairs with the given temporal order.

Keywords: ranked model, convex and piecewise linear (CPL) criterion functions, linear separability of data sets

Artykuł zrealizowano w ramach projektu badawczego „Temporalna reprezentacja wiedzy i jej implementacja w informatycznych systemach wspomagania postępowania medycznego”, nr 3 T11F 011 30.