Regresyjne metody łączenia klasyfikatorów

Tomasz Górecki, Mirosław Krzyśko

Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza

XXXV Konferencja Statystyka Matematyczna
Wisła 7-11.12.2009

Wprowadzenie (1)

Załóżmy, że dysponujemy c różnymi klasyfikatorami $\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_c$ skonstruowanymi z próby uczącej

$$\mathfrak{L}_n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\},$$

gdzie $\boldsymbol{x}_i$ jest wartością wektora $\boldsymbol{X} = (X_1, \ldots, X_p)'$ obserwowanych p cech obiektu, a $y_i \in \{1, 2, \ldots, K\}$ etykietą skalarną jednej z ustalonych K klas klasyfikowanych obiektów. Ponieważ każdy z nich może być oparty na innej koncepcji, to może się zdarzyć, że żaden z nich nie jest jednoznacznie lepszy od pozostałych oraz, że poszczególne klasyfikatory mogą błędnie klasyfikować inne obserwacje.

Wprowadzenie (2)

LeBlanc i Tibshirani (1996) zaproponowali, by ostateczną decyzję klasyfikacyjną podejmować za pomocą metaklasyfikatora (klasyfikatora łączonego), wykorzystującego informację płynącą ze wszystkich c klasyfikatorów. Klasyfikator łączony otrzymany tą metodą jest kombinacją liniową ocen prawdopodobieństw a posteriori K klas uzyskanych przez c różnych klasyfikatorów. Współczynniki tej kombinacji liniowej znajdowane są metodą najmniejszych kwadratów w modelu liniowej regresji wielokrotnej. Metoda ta została nazwana regresją stosową.

Wprowadzenie (3)

W prezentowanym referacie proponujemy w charakterze metaklasyfikatorów wykorzystać, obok regresji stosowej, następujące modele regresyjne:

- regresję grzbietową,
- regresję składowych głównych,
- regresją logistyczną,
- regresję nieparametryczną.

Klasyfikacja jako regresja (1)

Załóżmy, że każdej wartości $\boldsymbol{x}$ wektora obserwowanych cech $\boldsymbol{X}$, obok etykiety skalarnej $y \in \{1, 2, \ldots, K\}$, przypisana jest etykieta wektorowa $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_K)'$, gdzie $Y_k = 1$, jeżeli obserwacja $\boldsymbol{x}$ pochodzi z klasy k, $k = 1, 2, \ldots, K$. Niech rozkład prawdopodobieństwa pary losowej $(\boldsymbol{X}, \boldsymbol{Y})$ opisuje para (μ, r), gdzie μ jest miarą probabilistyczną wektora losowego $\boldsymbol{X}$, a r jest funkcją regresji $\boldsymbol{Y}$ względem $\boldsymbol{X}$. Dokładniej, dla dowolnego zbioru borelowskiego $\mathfrak{B} \subseteq \mathbb{R}^p$, $\mu(\mathfrak{B}) = P(\boldsymbol{X} \in \mathfrak{B})$ oraz dla każdego $\boldsymbol{x} \in \mathbb{R}^p$

$$\begin{aligned} \boldsymbol{r}(\boldsymbol{x}) &= (r_1(\boldsymbol{x}), \ldots, r_K(\boldsymbol{x}))' = E(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}) = \\ &= (E(Y_1|\boldsymbol{X} = \boldsymbol{x}), \ldots, E(Y_K|\boldsymbol{X} = \boldsymbol{x}))'. \end{aligned}$$

Klasyfikacja jako regresja (2)

Ponieważ Y_k przyjmuje tylko dwie wartości 1 lub 0, to

$$r_k(\boldsymbol{x}) = E(Y_k|\boldsymbol{X} = \boldsymbol{x}) = P(Y_k = 1|\boldsymbol{X} = \boldsymbol{x}), \quad k = 1, 2, \ldots, K.$$

Jest to zatem prawdopodobieństwo a posteriori przynależności obserwacji $\boldsymbol{x}$ do klasy o etykiecie k, gdzie $k = 1, 2, \ldots, K$.
Uzyskaliśmy zatem klasyfikator postaci

$$\hat{d}(\boldsymbol{x}) = \arg \max_{1 \leq k \leq K} \hat{r}_k(\boldsymbol{x}),$$

gdzie $\hat{r}_k(\boldsymbol{x})$ jest oceną z próby uczącej $\mathfrak{L}_n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ funkcji regresji $r_k(\boldsymbol{x})$, $k = 1, 2, \ldots, K$.
Funkcje regresji $E(Y_k|\boldsymbol{X} = \boldsymbol{x})$ możemy estymować na wiele sposobów.

Regresja stosowa (1)

Dysponujemy c różnymi klasyfikatorami oraz K klasami. Mamy zatem cK ocen prawdopodobieństw a posteriori odpowiadających obserwacji $\boldsymbol{x}_i$, $i = 1, 2, \ldots, n$. Oceny te zapisujemy w postaci wektorów wierszowych

$$\hat{\boldsymbol{p}}'(\boldsymbol{x}_i) = (\hat{p}_1(1|\boldsymbol{x}_i), \ldots, \hat{p}_1(K|\boldsymbol{x}_i), \ldots, \hat{p}_c(1|\boldsymbol{x}_i), \ldots, \hat{p}_c(K|\boldsymbol{x}_i))$$

i układamy w stos jako wiersze macierzy

$$\boldsymbol{P} = \begin{pmatrix} \hat{\boldsymbol{p}}'(\boldsymbol{x}_1) \\ \vdots \\ \hat{\boldsymbol{p}}'(\boldsymbol{x}_n) \end{pmatrix}$$

rozmiaru $n \times cK$.

Regresja stosowa (2)

Wielowymiarowy model regresji liniowej ma postać

$$\boldsymbol{Y}^* = \boldsymbol{P}\boldsymbol{B} + \boldsymbol{E},$$

gdzie

$$\boldsymbol{Y}^* = \begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1K} \\ Y_{21} & Y_{22} & \dots & Y_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{n1} & Y_{n2} & \dots & Y_{nK} \end{pmatrix}$$

jest macierzą rozmiaru $n \times K$ wartości etykiety wektorowej $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_K)'$ przypisanej obserwacjom $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$, $\boldsymbol{B}$ jest macierzą rozmiaru $cK \times K$ współczynników regresji, natomiast $\boldsymbol{E}$ jest macierzą rozmiaru $n \times cK$ błędów losowych.

Regresja stosowa (3)

Ponieważ kolumny macierzy $\boldsymbol{P}$ podlegają c ograniczeniom liniowym, to $\boldsymbol{P}$ nie jest pełnego rzędu kolumnowego i $\boldsymbol{P}'\boldsymbol{P}$ jest macierzą osobliwą.
Stąd za estymator macierzy $\boldsymbol{B}$, uzyskany metodą najmniejszych kwadratów, przyjmujemy

$$\hat{\boldsymbol{B}} = (\boldsymbol{P}'\boldsymbol{P})^{+}\boldsymbol{P}'\boldsymbol{Y}^{*},$$

gdzie $(\boldsymbol{P}'\boldsymbol{P})^{+}$ jest uogólnioną odwrotnością Moore'a–Penrose'a macierzy $\boldsymbol{P}'\boldsymbol{P}$.
Niech

$$\hat{\boldsymbol{u}}(\boldsymbol{x}) = (\hat{u}_1(\boldsymbol{x}), \hat{u}_2(\boldsymbol{x}), \ldots, \hat{u}_K(\boldsymbol{x})) = \hat{\boldsymbol{p}}'(\boldsymbol{x})\hat{\boldsymbol{B}}.$$

Wówczas

$$\hat{d}_{\text{stos}}(\boldsymbol{x}) = \arg \max_{1 \leq k \leq K} \hat{u}_k(\boldsymbol{x}).$$

Regresja grzbietowa

Ponieważ problemy w metodzie regresji stosowej pojawiają się w związku z niemożnością odwrócenia macierzy $\boldsymbol{P}'\boldsymbol{P}$, to do jej przekątnej można dodać pewną stałą $\lambda \geq 0$. Dla takiego zagadnienia otrzymuje się następujące rozwiązanie:

$$\hat{\boldsymbol{B}} = (\boldsymbol{P}'\boldsymbol{P} + \lambda \boldsymbol{I})^{-1}\boldsymbol{P}'\boldsymbol{Y}^*,$$

które nazywane jest regresją grzbietową. Optymalna wartość parametru λ wybierana jest za pomocą metody sprawdzania krzyżowego.

Regresja PCR

Próbą uniknięcia problemu zależności zmiennych objaśniających (a co za tym idzie problemów z odwracaniem macierzy) jest regresja składowych głównych (regresja PCR). Zamiast oryginalnych zmiennych objaśniających (prawdopodobieństw) używamy składowych głównych, które są nieskorelowane. W praktyce używamy jedynie kilku pierwszych składowych, które w zadowalający sposób odzwierciedlają zmienność oryginalnych danych.

Funkcję regresji

$$p_k = r_k(\boldsymbol{x}) = E(Y_k|\boldsymbol{X} = \boldsymbol{x}) = P(Y_k = 1|\boldsymbol{X} = \boldsymbol{x})$$

modelujemy za pomocą funkcji logistycznej

$$p_k = \frac{\exp(\beta_{0k} + \boldsymbol{\beta}_k' \hat{\boldsymbol{p}}(\boldsymbol{x}))}{1 + \sum_{k=1}^{K-1} \exp(\beta_{0k} + \boldsymbol{\beta}_k' \hat{\boldsymbol{p}}(\boldsymbol{x}))},$$

$$p_K = \frac{1}{1 + \sum_{k=1}^{K-1} \exp(\beta_{0k} + \boldsymbol{\beta}_k' \hat{\boldsymbol{p}}(\boldsymbol{x}))}$$

gdzie

$$\beta_k = \left(\beta_{1k}^{(1)}, \ldots, \beta_{Kk}^{(1)}, \ldots, \beta_{1k}^{(c)}, \ldots, \beta_{Kk}^{(c)}\right)', \quad k = 1, 2, \ldots K-1$$

oraz

$$p_1 + p_2 + \ldots + p_K = 1.$$

Parametry β_{0k} i β_k estymujemy metodą największej wiarogodności.

Nieparametryczny estymator Nadarayi–Watsona funkcji regresji $r_k(\boldsymbol{x})$ ma postać:

$$\hat{r}_k(\boldsymbol{x}) = \frac{\sum_{i=1}^{n} K\left(\frac{\hat{\boldsymbol{p}}(\boldsymbol{x})-\hat{\boldsymbol{p}}(\boldsymbol{x}_i)}{h}\right) I(y_i = k)}{\sum_{i=1}^{n} K\left(\frac{\hat{\boldsymbol{p}}(\boldsymbol{x})-\hat{\boldsymbol{p}}(\boldsymbol{x}_i)}{h}\right)}, \quad k = 1, 2, \ldots K,$$

gdzie $K(\cdot)$ jest funkcją jądrową zależną od parametru gładkości h. Optymalną wartość parametru gładkości h dobieramy metodą sprawdzania krzyżowego.

Metoda najbliższego sąsiada

Jedna z najpopularniejszych metod klasyfikacji. Idea jej działania jest prosta i intuicyjna. Nowy obiekt otrzymuje etykietę klasy obiektu, który jest najbliżej (w uogólnieniu tej metody, czyli metodzie k najbliższych sąsiadów, etykietę klasy, która występuje najczęściej pośród jego k sąsiadów). W tym miejscu należy podkreślić, że do oceny odległości może zostać wykorzystana bardzo szeroka klasa funkcji. Jest to metoda nieparametryczna, nie wymaga zatem żadnych założeń co do rozkładów danych w klasach.

Naiwna metoda bayesowska

Jest to prosty klasyfikator probabilistyczny oparty na założeniu, że cechy opisujące obiekty są wzajemnie niezależne. Założenie to nie ma raczej nic wspólnego z rzeczywistością i właśnie z tego powodu metoda nazywana jest naiwną. Pomimo tego klasyfikator ten często działa lepiej niż można się po nim było spodziewać (zwłaszcza jeśli jest dużo cech). W praktyce estymuje się gęstość każdej cechy w każdej klasie, a następnie bada iloczyn takich gęstości dla każdej klasy. Obserwacja klasyfikowana jest do klasy, dla której ten iloczyn jest największy. Najczęściej zakłada się, że cechy mają rozkłady normalne.

Binarne drzewa klasyfikacyjne

Jest to klasyfikator oparty na sekwencyjnym dzieleniu podzbiorów przestrzeni próby. W każdym kroku podział dokonywany jest tak, aby uzyskane części były jak najbardziej jednorodne. Cały proces przedstawiany jest za pomocą drzewa binarnego (stąd nazwa). Drzewo składa się z korzenia oraz gałęzi prowadzących z korzenia do kolejnych węzłów. W każdym węźle sprawdzany jest pewien warunek dotyczący danej obserwacji, i na jego podstawie wybierana jest jedna z gałęzi prowadząca do kolejnego węzła poniżej. Na dole znajdują się liście, z których odczytujemy do której z klas należy przypisać daną obserwację.

Zbiory danych

	Liczba przypadków	Liczba cech C	Liczba cech D	Liczba klas
Balance	625	4	0	3
Breast-W	699	0	9	2
Car	1728	0	6	4
Diabets	768	8	0	2
Echo	132	5	1	2
German	1000	7	13	2
Glass	214	9	0	6
Heart-C	303	6	7	5
Heart-H	294	6	7	5
Heart	270	7	6	2
Hepatisis	155	6	13	2
Ionosphere	351	34	0	2
Iris	150	4	0	3
Sonar	208	60	0	2
Wine	178	13	0	3

Optymalne wartości parametrów

	k	h	λ
Balance	10	0.6	0.09
Breast-W	5	0.6	0.01
Car	7	0.4	0.01
Diabets	20	0.6	0.01
Echo	9	1.2	0.01
German	14	0.8	0.01
Glass	1	0.2	0.02
Heart-C	35	0.4	0.21
Heart-H	17	0.6	0.01
Heart	24	0.8	0.01
Hepatisis	9	0.4	0.01
Ionosphere	2	0.8	0.01
Iris	10	0.4	0.01
Sonar	2	0.6	0.01
Wine	1	0.4	0.09

Porównanie regresji (% błędnej klasyfikacji)

	R	R Grzb.	R PCR	R Logist.	R NParam.
Balance	22.26	22.26	22.66	22.26	**14.93**
Breast-W	4.95	4.95	4.95	4.93	**3.97**
Car	**5.21**	**5.21**	5.46	5.39	5.43
Diabets	34.90	34.90	34.90	34.90	**33.22**
Echo	37.61	37.61	37.61	37.61	**34.99**
German	32.66	32.66	32.66	32.66	**31.94**
Glass	36.29	**36.24**	36.48	38.15	37.14
Heart-C	53.28	51.83	51.08	53.82	**41.54**
Heart-H	53.96	54.00	47.32	54.00	**35.59**
Heart	36.63	36.63	36.63	36.63	**28.59**
Hepatisis	19.78	19.78	19.78	19.78	**18.47**
Ionosphere	12.94	12.94	12.94	12.91	**12.19**
Iris	7.67	7.67	7.73	9.27	**6.13**
Sonar	30.60	30.60	30.60	29.55	**25.35**
Wine	4.94	4.82	4.88	6.46	**3.65**
Średnia	19.78	19.72	19.64	20.47	**16.80**

Podsumowanie

Jak widać najlepiej wypada metoda oparta o regresję nieparametryczną, która okazuje się najskuteczniejsza dla wszystkich zbiorów z wyjątkiem dwóch. Na tych dwóch zbiorach ustępuje jednak innym metodom minimalnie. Najgorsza wydaje się metoda oparta o regresję logistyczną. Natomiast klasyczna regresja stosowa oraz jej wersja grzbietowa oraz regresja składowych głównych wypadają bardzo podobnie.


Krzyśko M., Wołyński W., Górecki T., Skorzybut M. (2008). *Systemy uczące się.* WNT.


LeBlanc M., Tibshirani R. (1996), Combining Estimates in Regression and Classification. JASA 91:641–1650.


Merz C.J., Murphy P.M. (1998). UCI repository of machine learning databases. Machine readable data repository http://www.ics.uci.edu/∼mlearn/MLRepository.html. Irvine, CA: University of California, Department of Information and Computer Science.