Eksperymentalne badanie statystycznej zgodności selektora cech opartego o LARS/LASSO

Urszula Libal[1]

Streszczenie: *Celem artykułu jest prezentacja wyników eksperymentów sprawdzających zgodność selektora cech opartego o algorytm LARS/LASSO (Least Absolute Shrinkage and Selection Operator). W badaniach wykorzystano zbiór danych 'diabetes' zawierający cechy 442 pacjentów chorych na cukrzycę oraz odpowiedź mierzoną postępem choroby. Wyzerowano niektóre cechy i zaszumiono je, a następnie tak spreparowane dane poddano selektorowi LASSO, który wybierał cechy istotne. Sprawdzono, czy cechy uprzednio wyzerowane są cechami uznanymi przez selektor za najmniej istotne. Zwiększanie liczności próby pozwoliło określić zależność między zgodnością selektora LASSO a zakładanym rozmiarem modelu df. Na zgodność selektora nie wpływa natomiast poziom szumu σ.*

Słowa kluczowe: *LASSO, selekcja cech, selekcja modelu, statystyczna zgodność*

1. Wprowadzenie

Selekcja modelu jest kluczowym problemem, decydującym o jakości rozwiązań innych zadań opartych o wyselekcjonowany model. Selekcję cech często stosuje się w zadaniach rozpoznawania i identyfikacji, jako wstępną procedurę, która odsiewa nieistotne cechy, tzn. niezwiązane z badanym obiektem, stanowiące szum lub związane z nim w słabym stopniu. Po pierwsze, zmniejszenie rozmiaru df modelu jest ważne ze względu na złożoność obliczeniową algorytmu, który będzie wykorzystywał skurczony model do rozwiązania innego zadania. Po drugie, w niektórych problemach przy dużej liczbie cech p konieczne jest posiadanie także licznego zbioru obserwacji n, aby zapewnić odpowiednio gęstą estymację w p-wymiarowej przestrzeni. Dlatego zastosowanie selektora jest w wielu przypadkach koniecznością. Selektor oparty o algorytm LASSO oraz sposób działania samego algorytmu zostaną omówione w punkcie 2.

2. Selektor LASSO (Least Absolute Shrinkage and Selection Operator)

Dla danej macierzy obserwacji X_{nxp} (n-liczba próbek, p-liczba cech) oraz odpowiedzi y, selektor LASSO $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$ estymuje nieznany model β:

$$\hat{\beta} = argmin\, ||y - X\beta||_2 = argmin \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2, \tag{1}$$

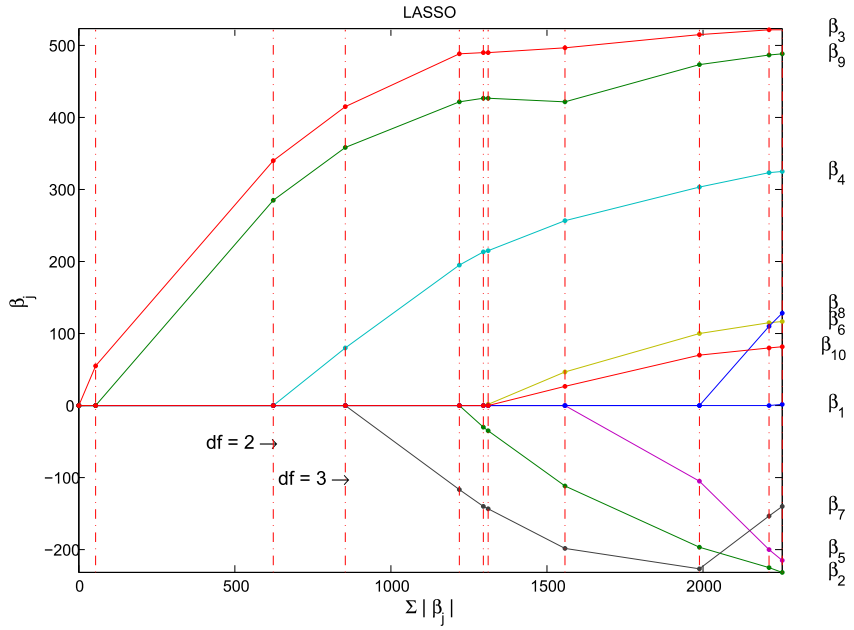
[1] Instytut Informatyki, Automatyki i Robotyki, Politechnika Wrocławska, ul. Janiszewskiego 11/17, 50-372 Wrocław, urszula.libal@pwr.wroc.pl

$$\text{przy} \quad ||\beta||_1 = \sum_{j=1}^{p} |\beta_j| \leq \lambda, \tag{2}$$

gdzie $\lambda > 0$ jest ograniczającym parametrem, który reguluje zmniejszenie modelu.

Tibshirani [3] zauważył, że regularyzacja w normie $l1$ prowadzi do wyzerowania niektórych współrzędnych wektora $\hat{\beta}$ i tym samym do zmniejszenia rozmiaru modelu. Hastie, Efron, Johnstone i Tibshirani [2] zaproponowali algorytm LASSO, jako modyfikację prostego w komputerowej implementacji algorytmu LARS. Szczegółowy opis obu algorytmów można zaleźć w [2]. Algorytm LARS przebiega w p krokach, natomiast LASSO może mieć ich więcej. W każdym kroku jest wybierana składowa $\hat{\beta}_i$ najmocniej skorelowanna z odpowiedzią y, tzn. gdy i-ta składowa $c_i = max\,(\hat{c})$ wektora korelacji $\hat{c} = X^T\,(y - \hat{y})$ jest największa, indeks i wędruje do zbioru aktywnego. Następnie uaktualniamy predyktor $\hat{y}$ o odpowiedni krok w wyliczonym kierunku i przechodzimy do następnego kroku. W modyfikacji LASSO algorytmu LARS pozwolono także na usunięcie indeksu i ze zbioru aktywnego, jeżeli w danym kroku składowa $\hat{\beta}_i$ zmieni znak na przeciwny. Zatrzymując algorytm LASSO, gdy zbiór aktywny liczy df indeksów, otrzymujemy selektor modelu $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$ o df niezerowych składowych i mówimy, że model ma rozmiar df.

Na rys. 1 przedstawiono przebiegi wartości składowych $\hat{\beta}_j$, $j = 1, 2, \ldots, p$. Przyjęcie modelu o rozmiarze df oznacza, że rozwiązanie LASSO $\hat{\beta}$ przyjmie aktualne wartości df niezerowych składowych $\hat{\beta}_j$, a pozostałe $p - df$ współrzędnych pozostanie na poziomie zero $\hat{\beta}_k = 0$.



Rys. 1. Oszacowane przez selektor LASSO składowe $\hat{\beta}_j$ modelu $\hat{\beta}$ vs. $\sum_j |\hat{\beta}_j|$. Obserwacje odpowiadające cesze $k = 6$ zastąpiono szumem Gaussowskim o $\sigma = 0.1$. Pionowymi liniami zaznaczono zmiany rozmiaru df modelu $\hat{\beta}$, tzn. wzrost lub spadek liczby df niezerowych składowych wektora $\hat{\beta}$.

W punkcie 3 zdefiniowano statystyczną zgodność selektora $\hat{\beta}$ z prawdziwym modelem β.

3. Statystyczna zgodność selektora LASSO

Statystyczną zgodność algorytmu LASSO, jako selektora cech, badali m.in. Zhao i Yu [4]. Nie udało im się jednak udowodnić zgodności lub jej braku dla selektora LASSO. Natomiast Zhao i Yu [4] podali *warunki niereprezentowalności* (*Irrepresentable Conditions*) zmiennych, których niespełnienie oznacza skorelowanie istotnych cech modelu z tymi nieistotnymi, co prowadzi w większości przypadków do niezgodności selektora. W przeprowadzonych eksperymentach obserwacje zmiennej X_k, aktualnie wybranej do skasowania, zostały zastąpione białym szumem Gaussowskim o wariancji σ^2 i nie są skorelowane z pozostałymi zmiennymi (tj. cechami X_i, gdzie $i \in \{1, 2, \ldots, p\}$ oraz $i \neq k$), ani z odpowiedzią y. Wprowadźmy ogólną definicję zgodności estymatora:

Definicja 1 *Estymator $\hat{\beta}$ jest zgodny, jeżeli dla każdego $\epsilon > 0$*

$$\lim_{n\to\infty} \mathbf{P}\left(||\hat{\beta} - \beta|| < \epsilon\right) = 1. \tag{3}$$

Ze względu na fakt, że prawdziwy model β pozostaje nieznany, a jedynie wiadomo, że model $\hat{\beta}$ nie powinien zawierać cech, które nie wnoszą istotnej informacji o odpowiedzi y, potrzebny będzie inny sposób na określenie zgodności selektora. W przeprowadzonym eksperymencie zerowaliśmy obserwacje zawierające informacje o k-tej cesze, a następnie zastępowaliśmy je niezależnymi losowymi danymi o jednakowym rozkładzie normalnym $\mathcal{N}(0, \sigma^2)$. W prawdziwym modelu składowa β_k, nieskorelowana z odpowiedzią, musi równać się zero ($\beta_k = 0$), dlatego na potrzeby eksperymentu wprowadzamy alternatywną definicję statystycznej zgodności selektora o następującej treści:

Definicja 2 *Selektor $\hat{\beta}$ jest zgodny z rzeczywistym modelem $\beta = \beta\,(df, \sigma, k)$, jeżeli*

$$\lim_{n\to\infty} \mathbf{P}\left(\hat{\beta}_k = 0\right) = 1. \tag{4}$$

Jeżeli wraz ze zwiększaniem liczby obserwacji n, prawdopodobieństwo zajścia zdarzenia loso-wego, że k-ta składowa $\hat{\beta}_k$ zostanie odrzucona z modelu $\hat{\beta}$ (tzn. wyzerowana), będzie zbiegać do jedynki, to selektor nazwiemy zgodnym z prawdziwym modelem β, dla którego $\beta_k = 0$.

Opis przeprowadzonych eksperymentów znajduje się w punkcie 4, natomiast analiza otrzymanych wyników i sformułowanie wniosków umieszczono w ostatnim, 5 punkcie.

4. Eksperymentalne badanie zgodności

W badaniu zgodności selektora LASSO użyto zbioru danych 'diabetes' [1], wykorzystanego także do zilustrowania działania algorytmów LAR, LASSO oraz Forward Stagewise w [2]. Zbiór ten zawiera $n = 442$ obserwacje $p = 10$ cech. U 442 pacjentów chorujących na cukrzycę zmierzono następujące cechy: X_1 - wiek, X_2 - płeć, X_3 - BMI, X_4 - ciśnienie krwi oraz wyniki badań osocza: X_5 - TC, X_6 - LDL, X_7 - HDL, X_8 - TCH, X_9 - LTG, X_{10} - glukoza. Jako odpowiedź y przyjęto współczynnik postępu choroby w ciągu roku. Dla ustalonego poziomu szumu $\sigma = 0.1,\ 0.01,\ 0.001$, rozmiaru modelu $df = 1, 2, \ldots, p-1$ oraz liczby obserwacji $n = 20, 40, \ldots, 420, 440$ przeprowadzono $p \cdot N = 100$ eksperymentów (po $N = 10$ dla każdej cechy X_k) polegających na poddaniu

selekcji za pomocą LASSO odpowiednio spreparowanych danych. W każdym zbiorze danych zastąpiono wygenerowanym szumem jedynie jedną zmienną X_k.

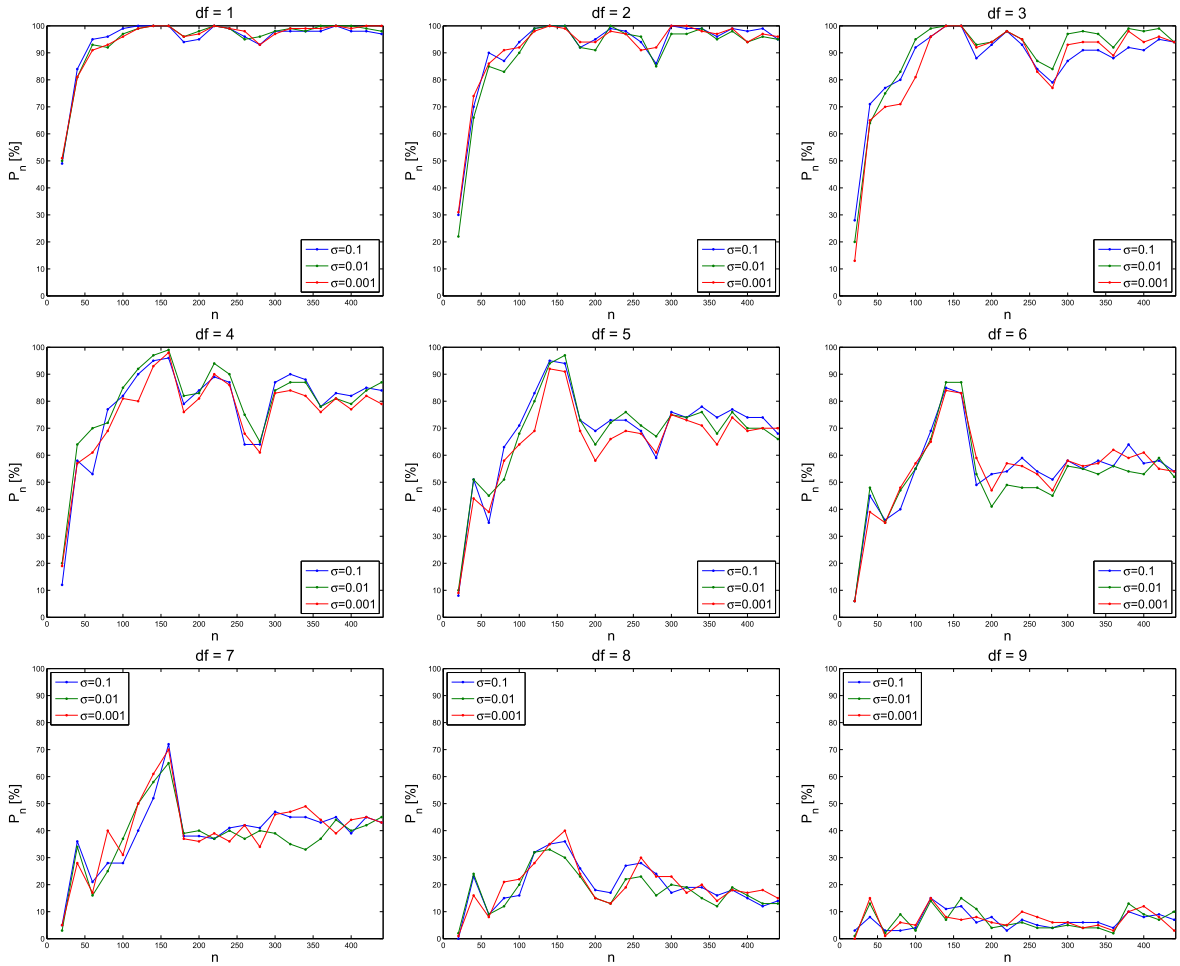
Prawdopodobieństwo $\mathbf{P}\left(\hat{\beta}_k = 0\right)$, dla ustalonego modelu $\beta = \beta\,(df, \sigma, k)$ oraz liczby obserwacji (zbadanych pacjentów) n, estymowano za pomocą częstości odrzucenia przez selektor k-tej cechy, uśrednionej dla wszystkich cech $k = 1, 2, \ldots, p$:

$$\hat{P}_n^{df,\sigma} = \frac{\sum_{k=1}^{p} \#\{\hat{\beta}_k = 0\}}{p \cdot N}. \tag{5}$$

Uwzględniając (5) przy rosnącej liczbie eksperymentów $N \to \infty$ dostajemy uśrednione po wszystkich cechach prawdopodobieństwo ich odrzucenia

$$\lim_{N\to\infty} \hat{P}_n^{df,\sigma} = \frac{1}{p} \sum_{k=1}^{p} \mathbf{P}\left(\hat{\beta}_k = 0\right). \tag{6}$$

Wyniki eksperymentów dla różnych zakładanych wielkości modelu df umieszczono na rys. 2 w postaci wykresów częstości $\hat{P}_n^{df,\sigma}$ odrzucenia nieistotnej cechy z modelu.



Rys. 2. Częstość $\hat{P}_n^{df,\sigma}$ odrzucenia przez selektor LASSO zastąpionych szumem danych odpowiadających k-tej cesze, $k \in \{1, 2, \ldots, p\}$, dla rosnącej liczby obserwacji $n = 20, 40, \ldots, 420, 440$. Każdy wykres odpowiada innemu rozmiarowi df modelu $\hat{\beta}$ ($df = 1, 2, \ldots, 9$), tzn. innej liczbie df niezerowych współrzędnych $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$.

5. Wnioski

Na podstawie przeprowadzonych eksperymentów zaobserwowano następujące zjawiska:

a) **Zgodność selektora LASSO zależy od ustalonego rozmiaru modelu df.**
Gdy rozmiar modelu jest mały $df = 1, 2, 3$, częstość odrzucenia nieistotnych cech jest bardzo wysoka i dąży do jedynki

$$\lim_{n\to\infty} \hat{P}_n^{df,\sigma} = 1,$$

czyli selektor można wtedy uznać za *zgodny*.
Natomiast dla modelu o dużym rozmiarze $df = 7, 8, 9$, dla dużych n częstość $\hat{P}_n^{df,\sigma}$ oscyluje odpowiednio na niskim poziomie 45%, 15%, czy nawet 7%, co oznacza, że selektor *nie jest zgodny*. Dla $df = 9$ zmienna zaszumiona znalazła się w modelu dla około 93% eksperymentów,

$$\lim_{n\to\infty} \hat{P}_n^{9,\sigma} \approx 0.07,$$

co może świadczyć o tym, że inne cechy też są bardzo słabo skorelowane z odpowiedzią y lub są silnie skorelowane z pozostałymi cechami włączonymi wcześniej przez selektor do modelu.

b) **Poziom szumu σ nie wpływa na zgodność selektora.** Wykresy częstości dla trzech wartości $\sigma = 0.1,\ 0.01$, czy 0.001 przeplatają się i na ich podstawie nie zauważono żadnej zależności z wartością częstości $\hat{P}_n^{df,\sigma}$.

c) Analizując wyniki eksperymentu można zauważyć, że **najczęściej występującą trójką cech X_j, dla których selektor jest zgodny, są cechy X_3 - BMI (body mass index), X_9 - LTG (lamotrygina) oraz X_4 - ciśnienie krwi.** Te trzy cechy są najmocniej skorelowane z wyjściem y. Porównaj przebiegi wartości składowych $\hat{\beta}_3$, $\hat{\beta}_9$ i $\hat{\beta}_4$ przedstawione na rys. 1. Gdy przyjmowano model o rozmiarze $df = 3$ lub mniejszym, składowe $\hat{\beta}_3$, $\hat{\beta}_9$ oraz $\hat{\beta}_4$ pozostawały najczęściej niezerowe. Składowa $\hat{\beta}_3$ była kwalifikowana najczęściej do szukanych w eksperymencie modeli o zmniejszonym rozmiarze. We wszystkich przypadkach oprócz sytuacji, gdy cechę X_3 zastąpiono szumem, cecha X_3 była wybierana jako najmocniej skorelowana z odpowiedzią y, a składowa $\hat{\beta}_3$ otrzymywała jako pierwsza wartość niezerową.

Odrzucenie cechy przez selektor nie zawsze oznacza brak jej skorelowania z odpowiedzią. Sytuacja taka może zajść, gdy odrzucona cecha jest skorelowana z innymi cechami, które już zostały wybrane przez selektor. Przy rosnącej liczbie obserwacji $n \to \infty$ zgodny selektor zbiega według prawdopodobieństwa do prawdziwego modelu. Rośnie wtedy prawdopodobieństwo, że błąd estymacji modelu $||\hat{\beta} - \beta||$ jest dowolnie mały, co jest pożądaną właściwością.

Aby uchronić się przed brakiem zgodności należy sprawdzić, czy wszystkie cechy są skorelowane z odpowiedzią, a następnie przeprowadzić analizę ich korelacji, aby wyeliminować z modelu cechy mocno skorelowane ze sobą. Algorytm LASSO w każdym kroku włącza do zbioru aktywnego cechę $\hat{\beta}_i$ o największej, aktualnej korelacji c_i, która jest składową wektora $\hat{c} = X^T(y - \hat{y})$. Aby zapewnić zgodność selektora można zakończyć działanie algorytmu i przyjąć aktualny model $\hat{\beta}$, gdy żadna z korelacji c_i nie osiągnie odpowiednio dużej wartości.

Literatura

[1] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, *The diabetes data*, [http://www-stat.stanford.edu/ hastie/Papers/LARS/diabetes.data], 2003.

[2] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, *Least Angle Regression*, Annals of Statistics, 32(2), 407–499, 2004.

[3] R. Tibshirani, *Regression Shrinkage and Selection via the Lasso*, Journal of the Royal Statistical Society, 58(1), 267–288, 1996.

[4] P. Zhao, B. Yu, *On Model Selection Consistency of Lasso*, Journal of Machine Learning Research, 7, 2541–2563, 2006.

EXPERIMENTAL STUDY ON STATISTICAL CONSISTENCY OF FEATURE SELECTOR BASED ON LARS/LASSO

The aim of this article is to present the results of experiments verifying consistency of a feature selector algorithm based on LARS / LASSO (Least Absolute Shrinkage and Selection Operator). The study used data set 'diabetes' [1] with $p = 10$ *features of 442 patients with diabetes diagnosed and a measured response* y *to disease progression. In every experiment we replace one of the covariates with white Gaussian noise from* $\mathcal{N}\left(0, \sigma^2\right)$ *distribution, and then the LASSO selector* $\hat{\beta}$ *chooses the relevant features. We repeat the experiments* $N = 10$ *times for all* $p = 10$ *covariates to average results. The simulations were done for tree noise levels* $\sigma = 0.1,\ 0.01,$ *and* 0.001*, model size* $df = 1, 2, \ldots, p-1$ *and number of observations* $n = 20, 40, \ldots, 420, 440$*. We analyze if those previously reset features* X_k *are recognized by the selector as the least important, i.e.* $\hat{\beta}_k = 0$*. We say that the LASSO selector* $\hat{\beta}$ *is consistent with the true model* $\beta = \beta\left(df, \sigma, k\right)$ *(for fixed* df*,* σ *and index of noisy covariate* k*), if*

$$\lim_{n\to\infty} \mathbf{P}\left(\hat{\beta}_k = 0\right) = 1.$$

The probability of reduction of replaced with noise covariate X_k *was estimated by frequency of setting* $\hat{\beta}_k$ *to zero:*

$$\hat{P}_n^{df,\sigma} = \frac{\sum_{k=1}^{p} \#\{\hat{\beta}_k = 0\}}{p \cdot N}.$$

Increasing the number of samples, we could observe the consistency or inconsistency of LASSO selector depending on the model size df*. The LASSO is consistent for small model size, like* $df = 1, 2$ *and* 3*, and it is not consistent for large model size* $df = 7, 8$ *and* 9*. For the model of size* $df = 3$ *the most frequently chosen features were:* X_3 *- BMI (body mass index),* X_9 *- LTG (lamotrigine) and* X_4 *- blood pressure. We noticed also that the level of noise* σ *has no influence on the selector consistency.*