Сравнение эвристических алгоритмов выбора линейных регрессионных моделей*

Крымова Е. А., Стрижов В. В.
ekkrym@mail.ru, strijov@ccas.ru
Московский физико-технический институт,
Москва, Вычислительный центр РАН

В работе описан способ построения линейных регрессионных моделей, основанный на порождении и выборе признаков. Предложены эвристические алгоритмы выбора признаков. Выполнено сравнение этих алгоритмов с общеизвестными. Особенностью данного исследования является то, что задача выбора моделей поставлена для счетного набора признаков.

Введение

Процедура построения регрессионных моделей состоит из двух шагов. На первом шаге, на основе множества свободных переменных, порождается набор признаков. Один из способов построения такого набора описан в [1]. Модель-претендент есть линейная комбинация конечного подмножества признаков. На втором шаге производится выбор признаков, при этом выполняется настройка параметров модели и оценивается ее качество. Модель с настроенными параметрами, доставляющая минимум заданному функционалу качества, называется моделью оптимальной структуры.

Целью данной работы является сравнение предложенных эвристических алгоритмов с известными алгоритмами. Мотивацией работы является тот факт, что решение практических задач восстановления регрессионной зависимости требует рассмотрения большого числа порождаемых признаков. При таком условии алгоритмы, называемые «жадными», выбирают некоторый поднабор признаков, без возможности его модификации с целью улучшения структуры модели. Переборные алгоритмы, не обладая этим недостатком, имеют высокую вычислительную сложность. В данной работе предлагается компромиссный вариант алгоритма выбора регрессионных моделей и сравниваются следующие алгоритмы:

1) LARS метод наименьших углов [2];
2) полный перебор моделей [3];
3) метод группового учета аргументов [1];
4) алгоритм Лассо [4];
5) стохастическая структурная оптимизация;
6) шаговая регрессия [5, 6, 7];
7) оптимальное прореживание в шаговой регрессии [8, 9];
8) модифицированный метод наименьших углов.

Сравнение алгоритмов показано на примере прикладной задачи, связанной с моделированием волатильности опционов по реальным историческим данным торгов опционом Brent Crude Oil.

*Работа выполнена при финансовой поддержке РФФИ, проекты № 07-07-00181, № 08-01-12022.

Постановка задачи

Задана выборка $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ — множество m пар, состоящих из вектора значений n свободных переменных $\boldsymbol{x}_i = (x_i^j)_{j=1}^n \in \mathbb{R}^n$, и значения одной зависимой переменной $y_i \in \mathbb{R}^1$. Индекс i элементов выборки и индекс j свободных переменных далее будем рассматривать как элементы множеств $i \in I = \{1, \ldots, m\}$ и $j \in J = \{1, \ldots, n\}$.

Задан класс регрессионных моделей $\mathcal{F} = \{f_s\}$ — параметрических функций, линейных относительно параметров,

$$y_i = f_s(\boldsymbol{\beta}_s, \boldsymbol{x}_i) = \sum_{j \in J_s} \beta_j x_i^j, \qquad (1)$$

в которой $s \in \{1, \ldots, 2^n\}$ является индексом модели, $\boldsymbol{\beta}_s = (\beta_j)_{j \in J_s}$ — вектор параметров, заданный индексом модели, $J_s \subseteq J$ — набор индексов свободных переменных. Введено ограничение на число элементов линейной комбинации (1). В множество $\mathcal{F}$ могут входить только модели с числом свободных переменных $|J_s| \leqslant R$.

Принята следующая гипотеза порождения данных. Пусть случайная аддитивная переменная ν регрессионной модели

$$y = f(\boldsymbol{\beta}, \boldsymbol{x}) + \nu$$

имеет нормальное распределение $\mathcal{N}(0, \sigma_\nu^2)$.

Тогда, с учетом гомоскедастичности регрессионных остатков, распределение зависимой переменной имеет вид

$$p(y|x, \boldsymbol{\beta}, \sigma_\nu^2, f) = \frac{\exp\left(-\frac{1}{\sigma_\nu^2} S(D \,|\, \boldsymbol{\beta}, f)\right)}{(2\pi\sigma_\nu^2)^{\frac{n}{2}}},$$

где S — сумма квадратов невязок $y_i - f(\boldsymbol{\beta}, \boldsymbol{x}_i)$. Это распределение задает указанный ниже критерий качества модели.

Дополнительно задано разбиение выборки $I = I^{\mathrm{T}} \sqcup I^{\mathrm{C}}$ на обучающую и контрольную. Для каждого набора данных, рассматриваемого в вычислительном эксперименте, наборы индексов $I^{\mathrm{T}}, I^{\mathrm{C}}$ определены до начала эксперимента. Алгоритм выбора модели определяет метод оптимизации, доставляющий оптимальное значение параметрам $\tilde{\boldsymbol{\beta}}$

модели f на обучающей выборке $\{(x_i, y_i)\colon i \in I^{\mathrm{T}}\}$. Принят критерий качества — сумма квадратов регрессионных остатков на контрольной выборке

$$S = \sum_{i \in I^{\mathrm{C}}} \Big(y_i - f(\tilde{\boldsymbol{\beta}}, \boldsymbol{x}_i)\Big)^2. \qquad (2)$$

Требуется найти такую модель $f_s \in \mathcal{F}$, которая доставляет наименьшее значение функционалу качества. Такая модель будет называться моделью оптимальной структуры.

Порождение свободных переменных

Предлагается следующий способ формирования выборки D, состоящий из двух шагов.

Шаг первый. Задано множество непорождаемых свободных переменных $\Xi = \{\xi^u\}_{u=1}^{U}$. Задано конечное множество функций $G = \{g_v\}_{v=1}^{V}$. Рассмотрим декартово произведение $G \times \Xi$, элементу (g_v, ξ^u) которого поставлена в соответствие суперпозиция $g_v(\xi^u)$, однозначно определяемая индексами v, u. Обозначим $a_i = g_v(\xi^u)$, где индекс $i = (v-1)U + u$.

Шаг второй. Назначается базовая модель порождения признаков. В качестве модели, описывающей отношение между зависимой переменной y и свободными переменными a_i, используется полином Колмогорова-Габора:

$$y = \beta_0 + \sum_{i=1}^{UV} \beta_i a_i + \sum_{i=1}^{UV}\sum_{j=1}^{UV} \beta_{ij} a_i a_j + \ldots,$$

где вектор коэффициентов

$$\boldsymbol{\beta} = (\beta_0, \beta_i, \beta_{ij}, \beta_{ijk}, \ldots)_{i,j,k,\ldots=1,\ldots,m}.$$

Запишем вышеприведенный ряд в виде

$$y = \sum_{j \in J} \beta_j x^j.$$

Переменные $\{x^j\}$ поставлены в однозначное соответствие мономам полинома.

Стандартизация данных

Выборка D стандартизирована таким образом, чтобы для $j \in J$ выполнялись условия нормировки

$$\sum_{i=1}^{m} x_i^j = 0, \quad \sum_{i=1}^{m} (x_i^j)^2 = 1, \quad \sum_{i=1}^{m} y_i = 0.$$

Предполагается, что векторы $\bar{\boldsymbol{x}}_j = \big(x_1^j, \ldots x_m^j\big)$ и $\bar{\boldsymbol{x}}_k = \big(x_1^k, \ldots x_m^k\big)$ для всех $j, k \in J$, $j \neq k$ линейно независимы.

LARS (метод наименьших углов)

LARS — алгоритм отбора признаков линейной модели [2]. На каждом шаге алгоритма происходит изменение вектора параметров модели так, чтобы доставить добавляемому признаку наибольшую корреляцию с вектором регрессионных остатков. Основным достоинством LARS является то, что он выполняется за число шагов, равное числу свободных переменных.

Лассо

Лассо — алгоритм оценивания коэффициентов линейной модели [4]. Введение ограничения на сумму абсолютных значений коэффициентов модели приводит к обращению в 0 некоторых коэффициентов модели. Ненулевые коэффициенты соответствуют признакам, входящим в модель.

Обозначим сумму модулей коэффициентов модели $T(\boldsymbol{\beta}) = \sum\limits_{j=1}^{n} |\beta_j|$. Вектор коэффициентов $\hat{\boldsymbol{\beta}}$ есть решения задачи минимизации $S(\boldsymbol{\beta})$ при ограничении: $T(\boldsymbol{\beta}) \leqslant t$, где t — параметр регуляризации. Для решения задачи используется метод квадратичного программирования.

Шаговая регрессия

Шаговая регрессия — алгоритм последовательного удаления/добавления признаков [5]. Алгоритм последовательного добавления признаков присоединяет к текущему набору A по одному признаку, который доставляет максимум нижеприведенному критерию,

$$\hat{\jmath} = \arg\max_{j \in J} \frac{S(A) - S(A \cup \bar{\boldsymbol{x}}_j)}{S(A \cup \bar{\boldsymbol{x}}_j)}.$$

Начальным считается пустой набор признаков.

Алгоритм последовательного удаления признаков начинает с самого большого набора, состоящего из всех признаков. На каждом шаге происходит удаление признака так, чтобы значение нижеприведенного критерия было как можно меньше:

$$\hat{\jmath} = \arg\min_{j \in J} \frac{S(A \backslash \bar{\boldsymbol{x}}_j) - S(A)}{S(A)}.$$

Останов алгоритма производится по выполнению условия C_p [10].

Алгоритм полного перебора

Этот алгоритм порождает все возможные множества мономов $\{\bar{\boldsymbol{x}}_j\}_{j \in J}$. Пусть сложность модели $|J| \leqslant R$. Под сложностью модели понимается число линейно входящих параметров. Алгоритм последовательно строит модели-претенденты неубывающей сложности. Параметры каждой модели настраиваются методом наименьших квадратов по обучающей выборке. Наилучшая модель выбирается исходя из минимума ошибки на контрольной выборке. Введем переменную выбора монома —

вектор $\boldsymbol{c} = (c_1, \ldots, c_n)$. Его элемент $c_j \in \{0, 1\}$ принимает значение 1, если $j \in J_s$, в противном случае 0. Базовая модель данного алгоритма имеет вид $y = \sum\limits_{j \in J_s} c_j \beta_j x^j$. Сложность этого алгоритма $\sum\limits_{i=1}^{R} C_n^i$.

Метод группового учета аргументов

Алгоритмы МГУА воспроизводят схему массовой селекции [1, 3]: последовательно порождаются модели возрастающей сложности. Каждая модель настраивается методом наименьших квадратов. Остановка алгоритма происходит, когда с увеличением номера шага начинается увеличение ошибки на контрольный выборке.

Стохастическая структурная оптимизация

Предложенный эвристический алгоритм состоит из итеративно выполняемых шагов. На первом шаге из множества признаков выбирается заданное число поднаборов, доставляющее соответствующей линейной модели наименьшее значение функционала качества. На втором шаге на заданном числе пар выполняется операция обмена признаками. На третьем шаге производится случайная замена произвольных признаков вновь полученных поднаборов. Шаги 2 и 3 итеративно повторяются. Алгоритм завершает работу, когда число шагов превысит заданное или когда ошибка оптимальной модели на контрольной выборке станет меньше заданной.

Оптимальное прореживание в шаговой регрессии

Оптимальное прореживание — это метод упрощения структуры регрессионной модели. Основная идея прореживания заключается в том, что те элементы модели, которые оказывают малое влияние на ошибку аппроксимации, можно исключить из модели без значительного ухудшения качества аппроксимации [8, 9].

Для построения регрессии требуется найти такие параметры $\hat{\boldsymbol{\beta}}$, которые доставляли бы наименьшее значение функции ошибки $S(\boldsymbol{\beta})$.

Локальная аппроксимация функции S в окрестности точки $\hat{\boldsymbol{\beta}}$ с помощью разложения в ряд Тейлора записывается в виде

$$S(\boldsymbol{\beta}+\Delta\boldsymbol{\beta}) = S(\boldsymbol{\beta})+\boldsymbol{g}^{\intercal}(\boldsymbol{\beta})\Delta\boldsymbol{\beta}+\tfrac{1}{2}\Delta\boldsymbol{\beta}^{\intercal}H\Delta\boldsymbol{\beta}+o(\|\boldsymbol{\beta}\|^3),$$

где $\Delta\boldsymbol{\beta}$ — возмущение вектора параметров $\boldsymbol{\beta}$, $\boldsymbol{g} = \frac{\partial S}{\partial \boldsymbol{\beta}}$ — градиент, $H = H(\boldsymbol{\beta}) = \frac{\partial^2 S}{\partial \boldsymbol{\beta}^2}$ — матрица вторых производных (матрица Гессе).

Функция $S(\boldsymbol{\beta})$ достигает своего максимума при $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, и ее поверхность квадратична. Таким образом, предыдущее выражение можно представить в виде $\Delta S = S(\boldsymbol{\beta} + \Delta\boldsymbol{\beta}) - S(\boldsymbol{\beta}) = \frac{1}{2}\Delta\boldsymbol{\beta}^{\intercal}H\Delta\boldsymbol{\beta}$.

Пусть исключение элемента модели есть исключение одного параметра модели, β_i. Исключенный параметр будем считать равным нулю. Исключение элемента эквивалентно выражению $\Delta\beta_i + \beta_i = 0$, иначе $\boldsymbol{e}_i^{\intercal}\Delta\boldsymbol{\beta} + \beta_i = 0$, где $\boldsymbol{e}_i$ — вектор, i-й элемент которого равен единице, все остальные элементы равны нулю.

Требуется минимизировать квадратичную форму $\Delta\boldsymbol{\beta}^{\intercal}H\Delta\boldsymbol{\beta}$ относительно $\Delta\boldsymbol{\beta}$ при ограничениях $\boldsymbol{e}_i^{\intercal}\Delta\boldsymbol{\beta} + \beta_i = 0$ для всех значений i. Задача условной минимизации решается с помощью введения Лагранжиана $L = \Delta\boldsymbol{\beta}^{\intercal}H\Delta\boldsymbol{\beta} - \lambda(\boldsymbol{e}_i^{\intercal} + w_i)$, в котором λ — множитель Лагранжа. Дифференцируя Лагранжиан по приращению параметров и приравнивая его к нулю получаем (для каждого индекса i параметра β_i)

$$\Delta\beta = -\frac{\beta_i}{[H^{-1}]_{ii}}H^{-1}\boldsymbol{e}_i.$$

Этому значению вектора приращений параметров соответствует минимальное значение Лагранжиана

$$L_i = \frac{\beta_i^2}{2[H^{-1}]_{ii}}.$$

Полученное выражение называется мерой выпуклости функции ошибки S при изменении параметра β_i.

Функция L_i зависит от квадрата параметра β_i. Это говорит о том, что параметр с малым значением будет удален из модели. Однако если величина $[H^{-1}]_{ii}$ достаточно мала, это означает, что данный параметр оказывает существенное влияние на качество аппроксимации модели.

EM+LARS

В данной работе предлагается алгоритм, сочетающий в себе жадную стратегию LARS и перебор моделей. Это позволяет улучшить структуру модели, несущественно увеличив вычислительные затраты. Происходит порождение подмножеств признаков с помощью EM-алгоритма. На этих подмножествах с помощью LARS находятся параметры моделей, производится перебор полученных моделей. Модель, доставляющая наименьшую среднеквадратичную ошибку на контрольной выборке, считается оптимальной.

Задано K натуральных чисел в порядке возрастания C_k, $k = 1, \ldots, K$, $C_k < n$. Пусть M — число моделей, получаемых на каждом шаге алгоритма. Рассмотрим k-й шаг алгоритма. Разобьем множество признаков X на C_k кластеров с помощью EM-алгоритма. M раз выбираем случайным образом из каждого кластера по одному элементу. Получаем M подмножеств из C_k признаков, принадлежащих разным классам. К каждому из подмножеств признаков на обучающей выборке приме-

Таблица 1. Результаты работы стандартных методов отбора признаков.

Алгоритм	$\frac{\mathrm{S}(I^{\mathrm{T}})}{\|\boldsymbol{y}_T\|^2}$	$\frac{\mathrm{S}(I^{\mathrm{C}})}{\|\boldsymbol{y}_C\|^2}$	AIC	k
LARS+EM	0.024	0.053	−863	9
Прореживв.	0.013	0.034	−1109	15
Стохаст.	0.014	0.024	−1299	20
Lasso	0.011	0.092	−491	11
Шаг. регр.	0.032	0.086	−405	30
МГУА	0.018	0.080	−584	8
LARS	0.019	0.089	−579	5
Перебор R=5	0.024	0.036	–	5

няем алгоритм LARS. В результате будут получены векторы параметров $\boldsymbol{\beta}_{kl},\ l = 1, \ldots, M$.

За K шагов алгоритм находит KM моделей, выбирается оптимальная модель.

Вычислительный эксперимент

Сравнительный анализ алгоритмов был выполнен на исторических данных торгов опционом Brent Crude Oil [11]. Срок действия опциона — полгода, с 02.01.2001 по 26.06.2001. Тип опциона — put (право на продажу базового инструмента), символ CLG01. Базовый инструмент — нефть, символ NYM. Использовались ежедневные цены закрытия опциона и базового инструмента. Сетка цен исполнения опциона $\mathcal{K} = \{19.0, 19.5, \ldots, 28.0, 28.5\}$.

Регрессионная выборка

$$\big\{(\boldsymbol{x}_i, y_i)\big\}_{i=1}^m = \big\{(\langle K_i, t_i\rangle, \sigma_i)\big\}_{i=1}^m$$

была построена по исходным данным — историческим ценам опциона $C_{K,t}$ и базового инструмента P_t, где $K \in \mathcal{K},\ t \in T$, следующим образом. Для каждого значения K_i и t_i, $i = 1, \ldots, m$ вычислялось значение предполагаемой волатильности σ_i

$$\sigma_i = \arg\min_{\sigma\in[0,1.5]}\big(C_{K_i,t_i} - C(\sigma, P_{t_i}, B, K_i, t_i)\big),$$

где справедливая цена опциона C определялась по формуле Блэка-Шоулза [12]. Длина истории составляла 112 отсчетов времени.

Множество порождающих функций было задано следующим образом: $G = \big\{1/x, \sqrt{x}, e^x\big\}$.

Максимальная степень полинома Колмогорова-Габора была выбрана равной трём. При этом число мономов составило 84. Регрессионная выборка была случайным образом разбита на контрольную и обучающую, равные по мощности. Стандартизация контрольной и обучающей выборок была проведена раздельно. Значения ошибок на обучении и контроле были усреднены по 10 запускам алгоритмов на различных разбиениях.

Результаты экспериментов представлены в Таблице 1. Для каждого алгоритма были вычислены: ошибки $\mathrm{S}(I^{\mathrm{T}})$ и $\mathrm{S}(I^{\mathrm{C}})$ на обучении и контроле (2), отнесенные к квадрату нормы соответствующего вектора ответов $||\boldsymbol{y}_T||^2 = \sum\limits_{j\in I^{\mathrm{T}}} y_j^2$ и $||\boldsymbol{y}_C||^2 = \sum\limits_{j\in I^{\mathrm{C}}} y_j^2$; значение информационного критерия Акаике $AIC = m\ln\frac{S}{m} + 2k$; сложность модели k.

Заключение

В работе выполнено сравнение предложенных алгоритмов (стохастическая структурная оптимизация, модифицированный метод наименьших углов EM+LARS) с известными алгоритмами. Вычислительный эксперимент показал, что увеличение числа признаков позволяет добиться улучшения качества модели. Результаты экспериментов подтвердили жадность алгоритма LARS и бóльшую эффективность алгоритма EM+LARS по сравнению с LARS. По результатам экспериментов наилучшими по совокупности критериев являются EM+LARS и алгоритм оптимального прореживания.

Литература

[1] *Malada H. R., Ivakhnenko A. G.* Inductive Learning Algorithms for Complex Systems Modeling. CRC Press. 1994.

[2] *Efron B., Hastie T., Johnstone I., Tbshirani R.* Least Angle Regression // The Annals of Statistics. 2004. Vol. 32, No. 2. Pp. 407–499.

[3] *Ивахненко А. Г., Степашко В. С.* Помехоустойчивость моделирования. Киев: Наукова думка. 1985.

[4] *Tibshirani R.* Regression shrinkage and Selection via the Lasso // Journal of the Royal Statistical Society. 1996. Vol. 32, No. 1. Pp. 267–288.

[5] *Draper N., Smith H.* Applied Regression Analysis. John Wiley and Sons. 1981. Pp. 307–312.

[6] *Efroymson M. A.* Multiple regression analysis. Mathematical Methods for Digital Computers. Ralston, Wiley, New York. l960.

[7] *Rawlings J. O., Pantula S. G., Dickey D. A.* Applied Regression Analysis: A Research Tool. Springer-Verlag, New York. 1998.

[8] *Стрижов В. В.* Поиск параметрической регрессионной модели в индуктивно заданном множестве // Журнал вычислительных технологий. 2007. № 1. С. 93–102.

[9] *Хайкин С.* Нейронные сети, полный курс. М: Вильямс. 2008.

[10] *Mallows C. L.* Some Comments on C_p. Technometrics. 1973. No. 15. Pp. 661–675.

[11] *Ширяев А. Н.* Основы стохастической финансовой математики. Том 1. Факты. Модели. ФАЗИС. 2004.

[12] *Hull J. C.* Options, Futures and Other Derivatives. Prentice Hall. 2000.