

ВЫБОР ОПТИМАЛЬНОГО ПОДМНОЖЕСТВА РЕГРЕССОРОВ

А.В. Дружинин

Введение

Исследуется проблема выбора оптимального набора базисных функций в задаче линейной регрессии. Рассматривается некоторый набор функций $f_i(x)$, $x = (x_1, \dots, x_m)$, $i \in Q = \{1, \dots, P\}$. Для описания экспериментальных данных Y_j , $j = 1, \dots, N$, $N > P$, требуется выбрать одну из линейных регрессионных моделей вида:

$$Y_j = \sum_{i \in \tilde{Q}} b_i f_i(X_j) + \varepsilon_j, \quad j = 1, \dots, N, \quad \tilde{Q} \subseteq Q, \quad (1)$$

$$X_j = (X_{j1}, \dots, X_{jm}), \quad j = 1, \dots, N.$$

Для выбора некоторого подмножества из заданной совокупности функций имеются два противоположных по характеру критерия. С одной стороны, в модель следует включать наибольшее число функций, если, пользуясь подобранной моделью, мы хотим получить надежные прогнозы. При недостатке базисных функций оценка метода наименьших квадратов, который используется для определения коэффициентов b_i , $i \in \tilde{Q}$, является смещенной и несостоятельной. С другой стороны, с увеличением числа включенных функций возрастает дисперсия прогноза, уменьшается точность оценивания параметров. Кроме того, имея в виду затраты, связанные с получением информации, желательно иметь как можно меньше функций.

Вопросы выбора компромиссных вариантов и методы их поиска довольно широко представлены в литературе [3, 4, 7]. В данной работе предлагается способ поиска подмножества регрессоров, оптимального с точки зрения минимальной оценки дисперсии.

1. Математическая формулировка задачи

Обозначим

$$RSS(\tilde{Q}) = \min_b \sum_{j=1}^N (Y_j - \sum_{i \in \tilde{Q}} b_i f_{ij})^2, \quad \tilde{Q} \subseteq Q,$$

$$f_{ij} = f_i(X_j), j = 1, \dots, N, i \in Q.$$

Векторы $f_i = (f_{i1}, \dots, f_{iN}), i \in Q$, будем называть регрессорами.

Оценка дисперсии S^2 регрессионной модели (1) определяется по формуле [3]:

$$S^2 = \frac{RSS(\tilde{Q})}{N - |\tilde{Q}|}, \text{ для любого } \tilde{Q} \subseteq Q.$$

Введем переменные $z_i \in \{0, 1\}$, $i \in Q$. Переменная z_i принимает значение 1, если i -й регрессор включен в модель, и 0 в противном случае. Задача выбора оптимального подмножества регрессоров с использованием введенных переменных может быть записана следующим образом:

$$\min_{b \in R^P} \frac{\sum_{j=1}^N (Y_j - \sum_{i=1}^P b_i z_i f_{ij})^2}{N - \sum_{i=1}^P z_i} \rightarrow \min_z, \quad (2)$$

$$\sum_{i=1}^P z_i \leq K, \quad (3)$$

$$z_i \in \{0, 1\}, i = 1, \dots, P. \quad (4)$$

При фиксированном наборе z выражение (2) является оценкой дисперсии конкретной модели. Коэффициенты b определяются при помощи метода наименьших квадратов с кубической сложностью [3]. Ограничение (3) является обобщением задачи для поиска лучшей регрессионной модели с числом степеней свободы не менее $N - K$ [3].

Задача (2)–(4) относится к классу дискретных экстремальных задач с нелинейной целевой функцией. Задача является NP -трудной, к ней сводится задача "минимальное по весу решение системы линейных уравнений" [2]. Отметим, что при ортогональном исходном наборе регрессоров, т.е. при выполнении условия

$$\sum_{j=1}^N f_{i_1 j} f_{i_2 j} = 0 \quad \text{для всех } i_1, i_2 \in Q, i_1 \neq i_2,$$

решение находится эффективно с трудоемкостью $O(P \cdot N + P^2)$.

В общем случае для точного решения задачи (2) - (4) используется алгоритм типа ветвей и границ с односторонней схемой ветвления по переменным \bar{x}_i . Использование односторонней схемы позволяет свести к минимуму затраты оперативной памяти [1], что важно при использовании алгоритма на ПЭВМ.

2. Нижняя граница

Одним из ключевых вопросов при использовании метода ветвей и границ является вопрос о построении нижней оценки на подмножествах решений задачи. Один из распространенных приемов построения нижней границы - ослабление ограничения (4) в исходной задаче (2)-(4):

$$L(\bar{x}) = \frac{\min_{\bar{b} \in R^P} \sum_{j=1}^N (Y_j - \sum_{i=1}^P b_i \bar{x}_i f_{ij})^2}{N - \sum_{i=1}^P \bar{x}_i} \rightarrow \min_{\bar{x}}, \quad (5)$$

$$\sum_{i=1}^P \bar{x}_i \leq K, \quad 0 \leq \bar{x}_i \leq 1, \quad i = 1, \dots, P.$$

Пусть часть переменных \bar{x}_i фиксирована.

Обозначим $Q_1 = \{i \in Q \mid \bar{x}_i = 1\}$, $Q_0 = \{i \in Q \mid \bar{x}_i = 0\}$.

Л е м м а. При фиксированных множествах Q_1 , Q_0 и $|Q_1| < K$:

- 1) $\frac{RSS(Q \setminus Q_0)}{N - |Q_1|} \leq L(\bar{x})$ для любых \bar{x}_i , $i \in Q \setminus \{Q_1 \cup Q_0\}$;
- 2) для любого $\varepsilon > 0$ существует набор \bar{x}_i , $i \in Q \setminus \{Q_1 \cup Q_0\}$,

такой, что

$$\frac{RSS(Q \setminus Q_0)}{N - |Q_1|} + \varepsilon > L(\bar{x}).$$

Д о к а з а т е л ь с т в о. При фиксированных множествах Q_1 , Q_0 , если $\bar{x}_i > 0$ для всех $i \in Q \setminus \{Q_1 \cup Q_0\}$, то числитель выражения (5) равен $RSS(Q \setminus Q_0)$:

$$\begin{aligned} \min_{\bar{b}} \sum_{j=1}^N (Y_j - \sum_{i=1}^P b_i \bar{x}_i f_{ij})^2 &= \min_{\bar{b}} \sum_{j=1}^N (Y_j - \sum_{i \in Q \setminus Q_0} b_i \bar{x}_i f_{ij})^2 = \\ &= \min_d \sum_{j=1}^N (Y_j - \sum_{i \in Q \setminus Q_0} d_i f_{ij})^2 = RSS(Q \setminus Q_0), \end{aligned}$$

при этом $b_i = d_i / x_i$, $i \in Q \setminus Q_0$.

Пусть некоторые переменные x_i , $i \in Q \setminus \{Q_1 \cup Q_0\}$, принимают нулевые значения. Обозначим $\bar{Q}_0 = \{i \in Q \setminus \{Q_1 \cup Q_0\} | x_i = 0\}$, $\bar{Q}_1 = \{i \in Q \setminus \{Q_1 \cup Q_0\} | x_i > 0\}$. В этом случае числитель (5) больше либо равен $RSS(Q \setminus Q_0)$:

$$\begin{aligned} \min_b \sum_{j=1}^N (Y_j - \sum_{i=1}^P b_i x_i f_{ij})^2 &= \min_b \sum_{j=1}^N (Y_j - \sum_{i \in Q_1 \cup \bar{Q}_1} b_i x_i f_{ij})^2 \geq \\ &\geq \min_b \sum_{j=1}^N (Y_j - \sum_{i \in Q_1 \cup \bar{Q}_1} b_i x_i f_{ij} - \sum_{i \in \bar{Q}_0} b_i f_{ij})^2 = RSS(Q \setminus Q_0). \end{aligned}$$

Значит,

$$L(z) \geq \frac{RSS(Q \setminus Q_0)}{N - \sum_{i=1}^P x_i} = \frac{RSS(Q \setminus Q_0)}{N - |Q_1| - \sum_{i \in Q \setminus \{Q_1 \cup Q_0\}} x_i} \geq \frac{RSS(Q \setminus Q_0)}{N - |Q_1|}.$$

Для доказательства второго утверждения леммы достаточно задать x_i , $i \in Q \setminus \{Q_1 \cup Q_0\}$, так, чтобы $x_i > 0$ для всех $i \in Q \setminus \{Q_1 \cup Q_0\}$ и

$$\sum_{i \in Q \setminus \{Q_1 \cup Q_0\}} x_i < \delta,$$

где $\delta > 0$ легко определить:

$$\begin{aligned} \frac{RSS(Q \setminus Q_0)}{N - |Q_1|} + \varepsilon &\geq \frac{RSS(Q \setminus Q_0)}{N - |Q_1| - \delta}, \\ \frac{N - |Q_1|}{RSS(Q \setminus Q_0) + \varepsilon(N - |Q_1|)} &\leq \frac{N - |Q_1| - \delta}{RSS(Q \setminus Q_0)}, \\ \delta &\leq (N - |Q_1|) \left(1 - \frac{RSS(Q \setminus Q_0)}{RSS(Q \setminus Q_0) + \varepsilon(N - |Q_1|)} \right). \end{aligned}$$

Для выполнения ограничения (3) необходимо

$$\sum_{i \in Q \setminus \{Q_1 \cup Q_0\}} x_i \leq K - |Q_1|,$$

а $K - |Q_1| > 0$ из условия леммы. Значит,

$$\delta \leq \min \left\{ 1; (N - |Q_1|) \left(1 - \frac{RSS(Q \setminus Q_0)}{RSS(Q \setminus Q_0) + \varepsilon(N - |Q_1|)} \right) \right\},$$

что завершает доказательство леммы.

Лемма дает способ построения нижней границы на подмножестве решений задачи (2)-(4), определяемом множествами Q_1 и Q_0 :

$$LB(Q_1, Q_0) = \frac{RSS(Q \setminus Q_0)}{N - |Q_1|}.$$

При $Q_1 = \emptyset$ и $Q_0 = \emptyset$ величина $LB(Q_1, Q_0)$ является нижней границей на всем множестве решений. При $|Q_1| = K$, так как

$$\sum_{i \in Q \setminus \{Q_1 \cup Q_0\}} x_i = 0,$$

нижняя граница будет равна

$$LB_K(Q_1) = \frac{RSS(Q_1)}{N - |Q_1|}.$$

Вычисление величины $RSS(Q \setminus Q_0)$ производится при помощи метода наименьших квадратов с кубической сложностью $O(N \cdot P^2 + P^3)$. В ходе работы алгоритма величину $RSS(Q \setminus Q_0)$ требуется вычислить многократно. Отметим, что использование информации предшествующих этапов алгоритма позволяет после проведения некоторых предварительных вычислений уменьшить трудоемкость получения $RSS(Q \setminus Q_0)$ до квадратичной, т.е. $O(P^2)$. Хранение текущей информации требует $O(P^3)$ ячеек памяти, проведение предварительного этапа - $O(N \cdot P^2 + P^3)$ действий и $O(N \cdot P + P^2)$ ячеек памяти. Вычислительные методы подробно описаны в [3].

3. Поиск рекордного значения

Для получения оптимального решения задачи значения целевой функции (2) вычисляются при различных фиксированных наборах переменных x_i . Лучшее из найденных допустимых решений называется рекордом [1].

В начале работы алгоритма рекордом служит значение целевой функции на полном наборе регрессоров, т.е. при $x_i = 1$, $i \in Q$. Чтобы найти оптимальное решение задачи на каждом шаге алгоритма с включением новых членов в множество Q_1 получаем значение целевой функции

$$UB(Q_1) = \frac{RSS(Q_1)}{N - |Q_1|}$$

при $x_i = 1$, $i \in Q_1$, $x_i = 0$, $i \in Q \setminus Q_1$.

Заметим, что при $|Q_1| = K$ значение $UB(Q_1)$ совпадает со значением нижней границы $LB_K(Q_1)$. Вычисление величины $RSS(Q_1)$ на

каждом шаге алгоритма, как уже упоминалось выше, производится за $O(P^2)$ действий.

4. Выбор переменной для ветвления

В процессе ветвления [1] множество Q_1 может расширяться за счет введения в него новых номеров из множества $Q \setminus \{Q_1 \cup Q_0\}$. Выбор индекса для включения в множество Q_1 осуществляется по критерию

$$RSS(Q_1 \cup \{i\}) = \min_{k \in Q \setminus \{Q_1 \cup Q_0\}} RSS(Q_1 \cup \{k\}).$$

Основным достоинством данного критерия является использование информации об уже включенных членах. В алгоритме из [3], например, переменные выбираются в порядке возрастания величин $RSS(Q \setminus \{i\})$, $i \in Q$, которые определяются на подготовительном этапе алгоритма. При таком упорядочении слабо различающиеся регрессоры, как правило, следуют друг за другом, и их последовательное включение мешает быстрому достижению "хорошего" рекорда.

5. Результаты

Предложенный алгоритм решения задачи (2)-(4) реализован на языке Фортран-77 для ПЭВМ РС-XT. Время работы алгоритма в зависимости от числа рассматриваемых регрессоров при $K = P$ приведено в таблице.

Число регрессоров	Количество задач	Среднее время
1	2	3
10	11	3"
11	1	8"
12	8	13"
13	7	13"
14	1	7"
15	11	21"
16	4	39"
17	2	25"
18	5	45"
19	3	2' 43"
20	5	2' 42"
21	11	5' 4"
22	2	9' 38"
23	1	11' 30"

1	2	3
24	3	39'
25	1	22'
26	1	43'

6. Поиск модели среди адекватных ее вариантов

Часто исследователь может получить оценку дисперсии ошибок эксперимента. Это позволяет проверить адекватность построенной модели по критерию Фишера. Регрессионная модель считается неадекватной, если отношение оценки дисперсии модели к оценке дисперсии ошибок наблюдения превосходит статистику Фишера с соответствующим числом степеней свободы и заданным уровнем значимости [5]:

$$S_M^2 / S_0^2 > F_{\nu_M, \nu_0}^\alpha,$$

где

F_{ν_M, ν_0}^α - табличное значение статистики Фишера [6];

S_M^2 - оценка дисперсии модели;

S_0^2 - оценка дисперсии ошибки эксперимента, постоянная величина,

определяемая экспериментально;

ν_M и ν_0 - их степени свободы;

$1 - \alpha$ - выбранная вероятность неадекватности модели.

Требование адекватности модели является дополнительным ограничением для задачи (2)-(4). С учетом этого ограничения задача может быть записана в виде:

$$\min_{b \in B^P} \frac{\sum_{j=1}^N (Y_j - \sum_{i=1}^P b_i x_i f_{ij})^2}{N - \sum_{i=1}^P x_i} \rightarrow \min_z,$$

$$\frac{\min_{b \in R^P} \sum_{j=1}^N (Y_j - \sum_{i=1}^P b_i x_i f_{ij})^2}{(N - \sum_{i=1}^P x_i) \cdot S_0^2} \leq F_{N - \sum_{i=1}^P x_i, \nu_0}^\alpha, \quad (6)$$

$$\sum_{i=1}^P z_i \leq K, \quad z_i \in \{0; 1\}, \quad i = 1, \dots, P.$$

При включении в Q_1 новых членов нижняя граница $LB(Q_1, Q_0)$ растет быстрее статистики Фишера $F_{N-|Q_1|, \nu_0}^\alpha$:

$$\frac{LB(Q_1 \cup \{i\}, Q_0)}{LB(Q_1, Q_0)} = \frac{N - |Q_1|}{N - |Q_1| - 1} > \frac{F_{N-|Q_1|-1, \nu_0}^\alpha}{F_{N-|Q_1|, \nu_0}^\alpha}.$$

Значит, если

$$\frac{LB(Q_1, Q_0)}{S_0^2} > F_{N-|Q_1|, \nu_0}^\alpha,$$

то на подмножестве решений, определяемом множествами Q_1 и Q_0 , нет адекватной модели:

$$\begin{aligned} \frac{LB(Q_1 \cup \{i\}, Q_0)}{S_0^2} &= \frac{LB(Q_1 \cup \{i\}, Q_0)}{LB(Q_1, Q_0)} \frac{LB(Q_1, Q_0)}{S_0^2} > \\ &> \frac{F_{N-|Q_1|-1, \nu_0}^\alpha}{F_{N-|Q_1|, \nu_0}^\alpha} \frac{LB(Q_1, Q_0)}{S_0^2} > \frac{F_{N-|Q_1|-1, \nu_0}^\alpha}{F_{N-|Q_1|, \nu_0}^\alpha} F_{N-|Q_1|, \nu_0}^\alpha = \\ &= F_{N-|Q_1|-1, \nu_0}^\alpha. \end{aligned}$$

Таким образом, ограничение (6) позволяет уменьшить число просматриваемых вариантов в методе ветвей и границ, что сокращает время работы алгоритма.

7. Сравнение методов

Сравнение методов решения проблемы выбора базисных функций в задаче линейной регрессии можно проводить по двум направлениям: сравнение критериев оптимальности модели и сравнение по времени работы алгоритмов. Основное внимание мы уделим второму направлению, по первому сделаем лишь два замечания:

1) выбор оценки дисперсии в качестве критерия оптимальности избавляет от обсуждения вопроса о количестве базисных функций в модели;

2) регрессия с минимальной оценкой дисперсии обладает максимальной статистикой R^2 , приведенной относительно степеней свободы [3, 4, 7].

По времени работы будем сравнивать только алгоритмы, которые находят лучшую из всех возможных регрессий, не рассматривая методов, дающих прибли-

женное решение.

Наиболее быстрый, по-видимому, из предлагавшихся до сих пор методов описан в [7]. В этой работе минимизируется остаточная сумма квадратов (RSS) при заданном числе регрессоров. Для решения задачи, аналогичной (2)–(4), необходимо найти модели с числом регрессоров $1, 2, \dots, P$, обладающие минимальным значением величины RSS . Проводится одновременный поиск всех таких регрессий. Нижней границей в методе ветвей и границ при фиксированных множествах Q_1 и Q_0 служит величина $RSS(Q \setminus Q_0)$. Процесс ветвления продолжается, пока нижняя граница не достигнет рекордного значения для каждой регрессии. Преимущество описанного в настоящей работе метода состоит в том, что он избавляет от необходимости построения всех регрессий, лучших при фиксированном числе регрессоров.

Алгоритм из [7] реализован на ЭВМ IBM 370/158. Как отмечают авторы [7], время работы алгоритма – 15–50 минут, при 15–27 регрессорах и быстро уменьшается при снижении числа рассматриваемых регрессоров ниже 15. ПЭВМ РС-XT, для которой реализован алгоритм, описанный в настоящей работе, по числу операций умножения и деления чисел с плавающей запятой уступает IBM 370/158 по крайней мере вдвое. Тем не менее, время работы этого алгоритма заметно меньше, чем алгоритма из [7].

Поступила в ред.-изд. отдел

12 сентября 1989 г.

Л и т е р а т у р а

1. Береснев В.Л., Гимади Э.Х., Дементьев В.Т. Экстремальные задачи стандартизации. – Новосибирск.: Наука, 1978. – 333 с.
2. Гэри М., Джонсон Д. Вычислительные машины и труднорешаемые задачи. – М.: Мир, 1982. – 416 с.
3. Себер Дж. Линейный регрессионный анализ. – М.: Мир, 1980. – 456 с.
4. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. – М.: Финансы и статистика, 1986, Кн. 2. – 351 с.
5. Хартман К., Лецкий Э., Шефер В. Планирование эксперимента в исследовании технологических процессов. – М.: Мир, 1977. – 552 с.
6. Закс Л. Статистическое оценивание. – М.: Статистика, 1976. – 598 с.
7. Furnival G.M., Wilson R.W. Regressions by Leaps and Bounds. – Technometrics, 1974, V. 16, No 4, P. 499–511.