Проблема переобучения при отборе признаков в линейной регрессии с фиксированными коэффициентами

Венжега А. В., Ументаев С. А., Орлов А. А., Воронцов К. В.
voron@ccas.ru
Москва, Вычислительный центр РАН

Проблема отбора информативных признаков часто возникает при решении задач классификации и прогнозирования. В условиях, когда обучающая выборка мала, а признаков много, увеличивается риск переобучения — найденная функция регрессии может допускать на новых данных существенно больше ошибок, чем на обучении. Это может быть связано как с неудачным отбором признаков, так и с переоптимизацией параметров модели [1, 2]. В данной работе рассматривается задача линейной регрессии, в которой коэффициенты регрессии фиксированы, т. е. не настраиваются по обучающей выборке. Это позволяет исследовать переобучение, обусловленное исключительно отбором признаков.

Задача линейной регрессии с фиксированными коэффициентами

Рассмотрим задачу восстановления зависимости $y^*\colon X \to \mathbb{R}$ по выборке объектов $X^\ell = \{x_i\}_{i=1}^\ell \subset X$ с известными значениями $y_i = y^*(x_i)$. Объекты описываются p признаками $f_j\colon X \to \mathbb{R},\ j = 1, \ldots, p$. Зависимость восстанавливается в классе линейных функций $a_J(x) = \sum_{j\in J} w_j f_j(x)$, где $J \subseteq \{1, \ldots, n\}$ — подмножество признаков, веса w_j фиксированы, например, $w_j \equiv 1$. Требуется по обучающей выборке X^ℓ сформировать набор признаков $J = J(X^\ell)$ как можно меньшей мощности, при котором функция $a_J(x)$ аппроксимирует $y^*(x)$ на всём X как можно точнее. Качество аппроксимации на выборке $U \subset X$ будем характеризовать либо средним отклонением $E(J,U) = \frac{1}{|U|}\sum_{u\in U} |a_J(u) - y^*(u)|$, либо частотой ошибок $\nu(J,U) = \frac{1}{|U|}\sum_{u\in U}\big[|a_J(u) - y^*(u)| > \theta\big]$, где θ — порог ошибки.

Несмотря на упрощённость постановки, данная регрессионная задача имеет ряд приложений в социологии и экономике. Приведём примеры.

— Требуется выбрать представительный набор магазинов для оценивания суммарного объёма потребительского спроса. Здесь объекты — это промежутки времени, признаки соответствуют магазинам, значения признаков — это объёмы продаж некоторого товара или группы товаров в данном магазине за данный промежуток времени.
— Требуется выбрать представительный набор территориальных округов для прогнозирования результатов политических выборов по всей стране. Здесь объектами являются партии, либо пары ⟨партия,

год выборов⟩. Признаки соответствуют регионам; значения признаков — это число голосов, отданных за партию в регионе.

— Требуется выявить крупных участников биржевых торгов, совершающих покупки и продажи крупных пакетов акций синхронно с движением цены. Роль объектов играют промежутки времени, признаки соответствуют участникам, значениями признаков являются разности объёмов покупки и продажи.

Отметим, что в первых двух задачах целевая зависимость по определению есть сумма всех признаков, $y^*(x) = \sum_{j=1}^{p} f_j(x)$.

Задачи такого типа возникают, в частности, когда регулярный сбор данных по всем признакам (регионам, магазинам, участникам торгов, и т. п.) слишком дорог, либо вообще невозможен; но имеются результаты однократного сбора данных по всем признакам и возможность организовать регулярный сбор данных по части признаков. Необходимо найти набор признаков, который будет давать наиболее точные прогнозы.

Величину $\delta(X_n^\ell, X_n^k) = \nu(J_n, X_n^k) - \nu(J_n, X_n^\ell)$ будем называть *переобученностью* набора $J_n = J(X_n^\ell)$ при n-м разбиении выборки X^L на обучающую подвыборку X_n^ℓ и контрольную X_n^k, где $n \in N \subseteq \{1, \ldots, C_L^\ell\}$ — множество разбиений. В данной работе исследуется зависимость переобученности от параметров метода отбора признаков.

Методы отбора признаков

Рассматриваются три эвристических метода отбора признаков.

М1. Выбор t лучших признаков. Для данной обучающей выборки X^ℓ признаки упорядочиваются по возрастанию среднего отклонения $E(\{j\}, X^\ell)$, $j = 1, \ldots, p$, и в набор $J(X^\ell)$ включаются t первых признаков. Фактически, перебор подмножеств в этом методе отсутствует.

М2. Перебор подмножеств t из T лучших признаков является обобщением предыдущего. Признаки также упорядочиваются по возрастанию среднего отклонения, и из первых T признаков выбирается набор J, $|J| = t$, для которого значение $E(J, X^\ell)$ минимально. Если число вариантов C_T^t превышает R, перебираются R случайных наборов.

М3. Жадное добавление признаков. К набору J последовательно добавляется по одному признаку. Каждый раз добавляется такой признак j, для которого среднее отклонение $E(J \cup \{j\}, X^\ell)$ минимально.

Эксперименты и выводы

Эксперименты проводились на данных по результатам выборов в Государственную думу РФ. Число признаков (субъектов федерации) $p = 89$, число объектов (политических партий) $L = 35$, значения признаков $f_j(x_i)$ — это число голосов, отданных за i-ю партию в j-м регионе.

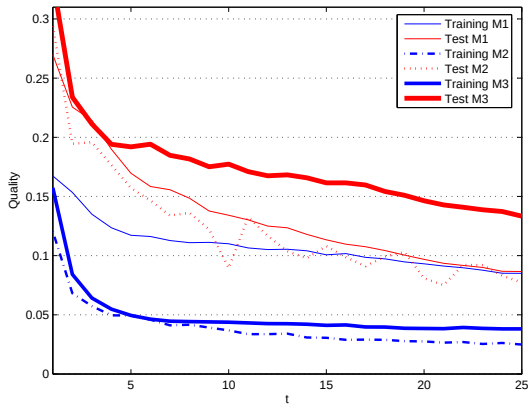

Рис. 1. Зависимость средней частоты ошибок на обучении и на контроле от числа выбранных признаков t для методов M1, M2, M3. Параметры эксперимента: $\ell = 17$, $k = 18$, $T = 35$, $R = 1300$; усреднение проводилось по $|N| = 30$ случайным разбиениям.

Оказалось, что в данной задаче переобучение, связанное с отбором признаков, возникает практически всегда.

Чем больше возможных вариантов порождает процедура отбора признаков, тем сильнее переобучение. Метод M1 наименее подвержен переобучению. Однако по критерию средней частоты ошибок на контрольной выборке он уступает методу M2, который является лидером соревнования. Наилучшие результаты M2 показывает при $T - t = 2$ или 3, то есть когда перебор делается с целью выкинуть 2–3 наименее удачных признака из T лучших признаков. Метод M3 наиболее переобучен и показывает наихудшую точность прогнозов на контроле.

Работа выполнена при поддержке РФФИ, проект № 05-01-00877 и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики».

Литература

[1] *Miller A.* Subset selection in regression. — Chapman & Hall/CRC, 2002.

[2] *Zhang P.* Inference after variable selection in linear regression models // Biometrika. — 1992. — No. 79. — Pp. 741–746.