УДК 519.6

СИСТЕМНАЯ АППРОКСИМАЦИЯ И АНАЛИЗ ТОЧНОСТИ ВЫБОРА МОДЕЛИ

ГРИЦЮК В.И.

Исследуются критерии качества решения в задаче выбора подмножества переменных. Рассматриваются результаты экспериментального сравнения методов оценки погрешности прогноза многомерной регрессионной модели.

1. Введение

Применение алгоритмов выбора подмножества переменных в регрессионных уравнениях приводит в ряде случаев к более устойчивым , экономным и точным моделям. В то же время эмпирические оценки качества решения, используемые для выбора подмножества переменных, дают искаженное представление о точности построенной модели. В связи с этим *актуальной* является разработка критериев качества регрессии, способных учитывать особенности применяемых алгоритмов восстановления зависимости. Имея множество псевдовыборок, в том или ином смысле похожих на множество действительных выборок, достаточно применить к каждой из них тот же алгоритм обработки данных, который использовался для получения решения на опорной выборке. В настоящей статье для оценки прогноза при использовании сложных алгоритмов восстановления зависимости применяется метод имитационных псевдовыборок, в котором генератором псевдовыборок служит восстановленная по данной выборке статистическая модель.

Цель исследования – сравнение методов оценки погрешности прогноза многомерной регрессионной модели при выборе переменных.

2. Выбор критерия для аппроксимации системы

Три главных приближений в области аппроксимации сложных моделей более простыми могут быть выделены. Одно основано на последовательном расширении. Другое предназначено для определения наиболее важных собственных значений сложной модели и основывает аппроксимацию на них. Третье приближение может быть названо ориентированным на критерий. В нем данная модель, допустим S^*, может быть аппроксимирована внутри указанного класса моделей $\{M(\Theta)|\Theta \in D_M\}$. Критерий, посредством которого член класса отбирается , может быть выбран различными путями [1,2]. Для линейных систем общее приближение состоит в минимизации разницы между импульсными откликами:

$$\min_{\Theta \in D_M} \sum_{t=1}^{\infty} \left| g_{S^*}(t) - g_{\Theta}(t) \right|^2 , \qquad (1)$$

где $g_{S^*}(\cdot)$ и $g_{\Theta}(\cdot)$ означает импульсные представления моделей S^* и $M(\Theta)$ соответственно. Другое приближение состоит в минимизации разницы между выходами S^* и $M(\Theta)$ под воздействием данного входа X :

$$\min_{\Theta \in D_M} \frac{1}{N} \sum_{1}^{N} \left| y_{S^*}(t) - y_{\Theta}(t) \right|^2 . \qquad (2)$$

Заметим, что модель, которая минимизирует (2), в общем будет зависеть от особенностей выбора входа. Критерий (1) – специальный случай (2) с импульсным входом. Для стохастических систем естественный критерий для хорошей аппроксимации – сравнение предсказания, полученного для данной модели S^*, $\bar{y}_{S^*}(t)$ с предсказаниями для модели $M(\Theta)$, $\hat{y}_M(t|\Theta)$, т.е. критерий

$$\min_{\Theta \in D_M} \lim_{N \to \infty} \frac{1}{N} \sum_{1}^{N} E\left| \hat{y}_{S^*}(t) - \hat{y}_M(t|\Theta) \right|^2 . \qquad (3)$$

Здесь мы предполагаем, что модели S^* и $M(\Theta)$ действуют под определенными условиями X . Минимизация величины Θ^* будет в общем зависеть от X и мы отметим наилучшую аппроксимацию S^* в M в смысле (3) как

$$M^*(S^*, X) = \left\{ \begin{matrix} M(\Theta) | \Theta \text{ минимизирующая} \\ \text{величину (3)} \end{matrix} \right\} . \qquad (4)$$

3. Метод имитационных псевдовыборок

Рассмотрим применение метода имитационных псевдовыборок для оценки ошибки прогноза при использовании сложных алгоритмов восстановления зависимости.

Предположим, что исследуется модель

$$Y = x_m g_m + x_r g_r + e , \qquad (5)$$

где e – $np \times 1$ -вектор ошибок измерений отклика, имеющий нормальное распределение $e \sim N(0, \sigma^2 I_{np})$; $Y_{np \times 1}$ – вектор значений отклика; $x_{m\ np \times m}$ – матрица плана; $x_r - np \times r$ матрица; g_m и g_r – регрессионные коэффициенты; n – число наблюдений. Фактическая среднеквадратичная погрешность прогноза для лучшей m членной модели:

$$J_{cp}(m) = n^{-1} E\left\| xg + \varsigma - x_m \hat{g}_m \right\|^2 , \qquad (6)$$

здесь ς – новые значения остатков, независящие от e, но с тем же распределением; $\hat{g}$ – набор выборочных регрессионных коэффициентов, соответствующих минимуму остаточной суммы квадратов RSS_m ; x_m – соответствующие этим коэффициентам m столбцов матрицы x. В модельных экспериментах использовались следующие критерии качества регрессионного уравнения: теоретическая оценка средних потерь регрессионной модели

$$\Gamma_m = RSS_m + (2m\hat{\sigma}_p^2) * n^{-1} ; \qquad (7)$$

аддитивная имитационная оценка средних потерь

$$J^{a}_{им}(m) = RSS_m + [\bar{J}_{ср}(m) - \overline{\overline{RSS_m}}]. \qquad (8)$$

Оцениваем смещение оценки $l(V)$, используя псевдовыборки $\tilde{V}$ [3], и строим более точную оценку для

$$L(V,F)\ 1^{a}(V) = l(V) + [E \sim L(\tilde{V}, F(\hat{a})) - E \sim l(\tilde{V})], \qquad (9)$$

где $l(V)$-статистика, применяемая для оценки $L(V,F)$ по данной выборке; $E^{\sim}$ – математическое ожидание по выборкам объема n из генеральной совокупности с распределением $F(\hat{a})$, средний остаточный квадрат

$$\hat{\sigma}^2 = \left\| Y - x_m \hat{g}_m \right\|^2 / (np - p), \qquad (10)$$

p – число переменных.

Генерирование псевдовыборки сводится к получению np независимых случайных величин $\tilde{e}_i$, имеющих нормальное распределение с параметрами $(0, \hat{\sigma}^2)$.

Задача состоит в том, чтобы из p экстремальных моделей, соответствующих минимуму RSS_m $(m = 1,2,\cdots,p)$, выбрать наилучшую в смысле минимума средних потерь.

Если поиск "лучшего набора переменных" основывается на выборочных данных, то традиционные оценки качества решения оказываются смещенными. Точное теоретическое решение проблемы оценки влияния выбора модели на прогностические свойства окончательного решения даже для простейших моделей оказывается невероятно сложной задачей, поэтому для изучения свойств оценок применим метод Монте-Карло. Влияние выбора экстремальных моделей на оценки качества решения очевидно зависят от таких факторов, как характер убывания регрессионных коэффициентов, дисперсия погрешности отклика, число наблюдений и число переменных. Зависимость смещения оценок средних потерь от числа наблюдений и числа переменных p тем больше, чем больше отношение p/n.

Для исследования связи с действительными значениями регрессионных коэффициентов и распределением погрешности условно нужно рассмотреть

Таблица 1

Критерий	Средний риск p=15	p=25	p=35
$J_{ср}(m)$	12,4	20,4	28,4
Γ_m	13	21,4	30
$J^{a}_{им}(m)$	12,8	21	29,4

Таблица 2

Критерий	Средний риск p=15	p=25	p=35
$J_{ср}(m)$	12,6	20,6	28,6
Γ_m	12,7	20,7	29,3
$J^{a}_{им}(m)$	12,8	20,8	29

два типа задач: задачи с небольшим числом существенных переменных, вклад которых в регрессию существенно больше σ^2/n, и задачи, в которых вклад существенных переменных в регрессию сравним с дисперсией погрешности отклика.

Рассмотрим сначала результаты применения алгоритма выбора переменных к задачам первого типа. Исследуем уравнение (5) с параметрами g=(2, 1, 43, 0,...,0), $\sigma^2/n = 0{,}2$, p=15, 25, 35.

В табл. 1 даны значения среднего риска при использовании приведенных критериев. Эти результаты получены усреднением по 400 экспериментам, для вычисления имитационных оценок использовали по 20 псевдовыборок.

Анализ результатов задачи первого типа показывает, что по мере увеличения общего числа переменных в модели выигрыш имитационных методов по сравнению с классическим критерием становится все более существенным.

Сравним критерии качества решения на моделях более общего вида. Для уравнения (5): g=(1,43, 1,29, 1,16, 1,0, 0,82, 0,58, 0,...,0), $\sigma^2/n = 0{,}1$; p=15, 25, 35.

В табл. 2 показано использование приведенных критериев для выбора окончательной модели. Если рассматривать результаты применения критериев на моделях общего вида для выбора окончательной модели, можно заметить, что для модели с медленно убывающими регрессионными коэффициентами преимущество имитационных оценок начинает проявляться при больших значениях p, так как здесь сказывается не смещение из-за выбора между одинаковыми по своему вкладу в регрессию переменными, а выбор несущественных переменных вместо существенных.

4. Заключение

Таким образом, когда число переменных сравнимо с числом наблюдений, смещение оценок средних потерь, обусловленное поиском модели, оказывается существенным. Применение имитационных псевдовыборок для исследования многомерных моделей позволяет уменьшить смещение, а для моделей с небольшим числом существенных переменных – получить практически несмещенные оценки среднего риска.

Литература: 1. *Ljung L.* Consistency of the least squares identification method // IEEE Trans. Automatic Control, 1976. V AC-21. P. 779-781. **2.** *Грицюк В. И.* Состоятельность и помехоустойчивые оценки временных рядов // АСУ и приборы автоматики. 2001. Вып. 117. С. 106-108. **3.** *Efron B.* Bootstrap methods: Another look at the jackknife. Ann. Statist., 1979. Vol. 6. P.1-26.

Поступила в редколлегию 14.01.2004

Рецензент: д-р техн. наук, проф. Бодянский Е.В.

Грицюк Вера Ильинична, канд. техн. наук, доцент кафедры системотехники ХНУРЭ. Научные интересы: стохастические системы управления. Хобби: литература, музыка. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, тел. 702-10-06.