АЛГОРИТМЫ ПОРОЖДЕНИЯ И ВЫБОРА РЕГРЕССИОННЫХ МОДЕЛЕЙ

Стрижов В.В., Крымова Е.А[1]

Вычислительный центр РАН, e-mail: strijov@ccas.ru
[1]Московский физико-технический институт, e-mail: ekkrym@mail.ru

Рассматриваются алгоритмы порождения и выбора линейных регрессионных моделей. Модель оптимальной структуры отыскивается как линейная комбинация элементов заданной базовой модели. Критерием оптимальности служит среднеквадратичная ошибка на тестовой подвыборке.

Цель работы — анализ и сравнение эвристических алгоритмов порождения моделей: однослойного и многослойного алгоритмов МГУА [1], метода стохастической структурной оптимизации и метода оптимального прореживания [2]. При анализе используются алгоритмы выбора свободных переменных: шаговая регрессия, метод наименьших углов (LARS) и лассо Тибширани.

Решается следующая задача. Задана выборка $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$, $\boldsymbol{x} \in \boldsymbol{R}^m$, $y \in \boldsymbol{R}$, которая разбита на обучающую и тестовую подвыборки случайным образом. Разбиение определено множествами индексов ℓ и C.

Принята базовая модель — полином Колмогорова-Габора, порождающая модели-претенденты: $y = w_0 + \sum_{i=1}^m w_i x_i + \sum_{i=1}^m \sum_{j=1}^m w_{ij} x_i x_j + \ldots + \sum_{i=1}^m \cdots \sum_{z=1}^m w_{i\ldots z} \underbrace{x_i \ldots x_z}_{R}$.

В этой модели $\boldsymbol{x} = \{x_i | i = 1, \ldots, m\}$ — множество свободных переменных; $\boldsymbol{w}$ — вектор параметров $\boldsymbol{w} = \langle w_i, w_{ij}, w_{ijk}, \ldots | i, j, k, \ldots = 1, \ldots, m \rangle$ и $F_0 = F_0(R)$ — число мономов.

Базовая модель представима в виде $\boldsymbol{y} = A\boldsymbol{w}$, где столбцы матрицы — значения мономов на выборке и $\boldsymbol{y} = \{y_1, \ldots, y_N\}$. Обозначим соответствующие разбиения выборки как A_ℓ, y_ℓ и A_C, y_C.

Требуется выбрать такие столбцы матрицы A, задающие модель, которые доставляют минимум критерию оптимальности. Задача построения линейной регрессионной модели оптимальной структуры имеет вид $\boldsymbol{c} = \arg \min_{\boldsymbol{c} \in \{0,1\}^{F_0}} \|A_C(\boldsymbol{w} \times \boldsymbol{c}) - \boldsymbol{y}_C\|$, где $\times$ — знак поэлементного умножения векторов. Параметры $\boldsymbol{w}$ определены как $\boldsymbol{w} = \arg \min_{\boldsymbol{w} \in \mathbb{R}^{F_0}} \|A_\ell \boldsymbol{w} - \boldsymbol{y}_\ell\|$.

На исторических данных опционных торгов был проведен анализ предложенных методов. Работа выполнена при поддержке РФФИ, проект № 07-07-00181.

Литература.

1. *Malada H. R., Ivakhnenko A. G.* Inductive Learning Algorithms for Complex Systems Modeling. CRC Press. 1994.
2. *Стрижов В. В.* Поиск параметрической регрессионной модели в индуктивно заданном множестве. Журнал вычислительных технологий. 2007. No 1. С. 93–102.