Strijov V.V. Multilevel model selection using parameters covariance matrix analysis // Proceedings of Mathematical Methods of Pattern Recognition. МАКС~Пресс, 2011 : 154-157.

Задача выбора многоуровневых моделей с анализом ковариационной матрицы параметров*

Стрижов В. В.
strijov@ccas.ru
Вычислительный центр РАН, Москва, Россия

Обсуждается метод выбора активного набора признаков и фильтрации объектов выборки при восстановлении регрессии. Предполагается, что элементы рассматриваемой выборки естественным образом были разбиты на подмножества; для каждого из которых имеется своя, отличная от других, гипотеза порождения данных. Задача заключается в том, чтобы определить это разбиение и восстановить регрессионную модель для каждой подвыборки. При этом оценивается ковариационная матрица параметров каждой модели, и на основании анализа этой матрицы определяется вероятность принадлежности некоторого объекта данной подвыборке, а и некоторого признака — данной модели.

Введение

Работа опирается на следующие результаты. Предположим, что измеряемых свободных переменных недостаточно для восстановления адекватной регрессионной модели. Для пополнения их набора используем порождающие функции и вводим при этом меры их структурной сложности, аналогичные предложенным К. Владиславлевой [1].

В работе мы исходим из того, что процедура скользящего контроля недостаточно эффективна при решении прикладных задач. В случае, когда число измеряемых или порожденных признаков многократно превосходит объем выборки, однократное разбиение выборки не исключает переобучения модели и приводит к тому, что выборку приходится разбивать на несколько подвыборок: обучающую, тестовую, контрольную и так далее, как показано С. Ватанабе [2] и С. Арло [3].

Для выбора адекватной регрессионной модели используется функция правдоподобия модели, см. Д. МакКай [4]. Эта функция является составной частью связанного байесовского вывода, см. К. Бишоп [5]. Её использование согласуется с принципом минимальной длины описания, являющимся универсальным критерием выбора модели, см. П. Грюнвальд [6, 7]. Для оценки вероятности принадлежности признаков и объектов выборки к тем или иным моделям используются методы анализа ковариационных матриц, рассмотренные Дж. Нельдером [8]. Для оценки сходства двух и более моделей используется расстояние Дженсена-Шеннона, см. [9].

Предлагаемый метод заключается в следующем. Фиксируется класс моделей; порождается множество производных признаков. Индексы элементов выборки разбиваются на подмножества. Каждое из подмножеств соответствует модели. Число моделей выбирается таким, чтобы расстояние между моделями было статистически значимым [9]. Принадлежность элемента выборки к модели определяется по результатам анализа ковариационной матрицы зависимых переменных. Структура модели определяется по результатам анализа ковариационной матрицы параметров модели.

Результатом является многоуровневая модель оптимальной сложности — набор адекватных регрессионных моделей, описывающих выборку. В качестве иллюстрации приведена задача прогнозирования периодических временных рядов.

Работа выполнена при поддержке РФФИ, грант: 10-07-00422

Постановка задачи

Задана выборка $D = \{(\mathbf{x}^i, y^i)\}$, проиндексированная $i \in \mathcal{I} = \{1, \ldots, m\}$. Элементы вектора $\mathbf{x}^i = [x^i_1, \ldots, x^i_j, \ldots, x^i_n] \in \mathbb{R}^n$ имеют индексы $j \in \mathcal{J}$. Само множество векторов представлено в виде матрицы плана $[\mathbf{x}^1, \ldots, \mathbf{x}^m]^T = X$, столбцы которой являются признаками и обозначаются нижним индексом: $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$. Соответственно, $D = (X, \mathbf{y})$, где матрица $X \in \mathbb{R}^{n\times m}$, а вектор $\mathbf{y} \in \mathbb{R}^m$.

Вектор зависимой переменной считается реализацией случайной величины; пусть ее распределение принадлежит семейству экспоненциальных распределений. Обозначим $\mathbf{f}$ — вектор восстановленных значений зависимой переменной посредством некоторой неизвестной функции регрессии, $\mathbf{f} = [f(\mathbf{w}_{\text{MP}}, \mathbf{x}^1), \ldots, f(\mathbf{w}_{\text{MP}}, \mathbf{x}^m]^T$, в которой $\mathbf{w}_{\text{MP}}$ — вектор наиболее вероятных параметров. Вектор свободных переменных $\mathbf{x}$, согласно классической постановке задачи восстановления регрессии, см. Г. Себер [10], будем считать неслучайной величиной.

Регрессионной моделью f будем называть поэлементное отображение $f : (\mathbf{w}, \mathbf{x}) \mapsto \mathrm{f}$. В терминах отображения соответствующих множеств, модель $f : \mathcal{W} \times \mathcal{X} \to \mathcal{Y}$. Будем считать, что $\mathcal{W}, \mathcal{X} \subset \mathbb{R}^n$, а $\mathcal{Y} \subset \mathbb{R}^m$.

Задача выбора модели ставится как задача нахождения такой модели f из класса допустимых моделей $\mathcal{F}$, которая имела бы максимальное правдоподобие при наиболее вероятных параметрах:

$$\hat{f}(\mathbf{w}_{\text{MP}}|\mathbf{x}) = \arg \max_{f\in\mathcal{F}, \mathbf{x}\in D} \mathcal{E}\left(f(\mathbf{w}_{\text{MP}}, \mathbf{x})\right).$$

Наиболее вероятные параметры

$$\mathbf{w}_{\text{МР}} = \arg\max_{\mathbf{w}\in\mathcal{W}} p(\mathbf{w}|D, f),$$

модели f оцениваются с помощью формулы Байеса, в которой апостериорное распределение параметров

$$p(\mathbf{w}|D, f) = \frac{p(D|\mathbf{w}, f, B)p(\mathbf{w}|f, A)}{\int p(D|\mathbf{w}', f, B)p(\mathbf{w}'|f, A)d\mathbf{w}'},$$

функция правдоподобия параметров $p(D|\mathbf{w}, f, B)$ задана распределением зависимой переменной $\mathbf{y}$. Априорное распределение параметров задано классом моделей $\mathcal{F}$ и гипотезой порождения данных.

В качестве примера приведем распределение $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, B)$ и класс линейных или линеаризованных существенно-нелинейных моделей. Параметрическая функция $\mathcal{N}$ переводится линейным отображением заданным матрицей плана X или линеаризованной матрицей плана

$$\underset{m\times n}{J} = \left[\frac{\partial f(\mathbf{w}, \mathbf{x}^i)}{\partial w_j}\right].$$

также в функцию $\mathcal{N}$ (так как линейное отображение, заданное матрицами X или J, представимо в виде произведения $U\Lambda V^T$ ортогональной, диагональной и ортогональной матриц). Параметры распределения при этом, в общем случае, изменяются. Следовательно, многомерная случайная величина $\mathbf{w}$ имеет ту же функцию распределения $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_{\text{МР}}, A)$. Прочие примеры подробно рассмотрены Нельдером [8].

Правдоподобие модели $\mathcal{E}\left(f(\mathbf{w}, \mathbf{x})\right) \stackrel{\text{def}}{=} p(D|f)$ — сомножитель правой части формулы Байеса второго уровня связного вывода, см. [11]. Согласно этому выводу, наиболее вероятная модель отыскивается исходя из сравнения апостериоых вероятностей

$$p(f|D) \propto p(D|f)p(f),$$

или же из сравнения правдоподобий моделей $p(D|f)$, считая их априорные вероятности равными.

Правдоподобие модели при этом задается выражением

$$\mathcal{E}'\left(f(\mathbf{w}, \mathbf{x})\right) = \int p(D|\mathbf{w}, f, B)p(\mathbf{w}|f, A)d\mathbf{w}.$$

Предлагается вычислять правдоподобие модели в окрестности ее наиболее правдоподобных параметров $\mathbf{w}_{\text{МР}}$, используя только подынтегральное выражение

$$\mathcal{E}(f(\mathbf{w}_{\text{МР}}, \mathbf{x})) = p(D|\mathbf{w}_{\text{МР}}, f, B)p(\mathbf{w}_{\text{МР}}|f, A). \quad (1)$$

Ковариационные матрицы A и B при этом предполагаются уже оцененными и зафиксированными на этапе нахождения наиболее правдоподобных параметров.

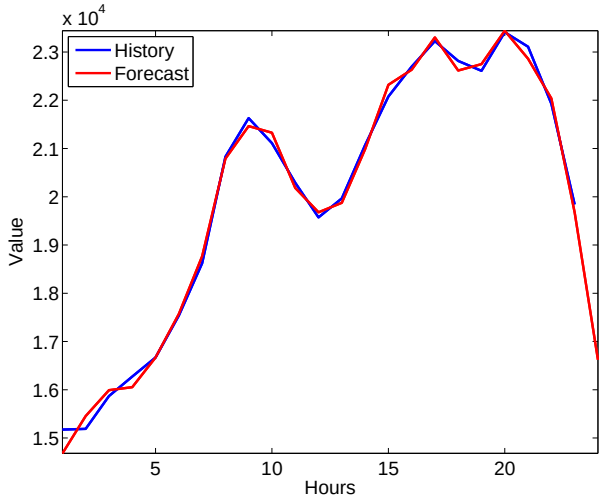

Рис. 1. Прогноз временного ряда $s(\tau)$ на 24 часа вперед

Выбор модели и фильтрация объектов

Линейная модель f однозначно задается активным множеством индексов признаков $\mathcal{A} \subseteq \mathcal{J}$. Предполагая частичную гомоскедаксичность выборки (например, среди объектов встречаются выбросы, которые должны быть исключены из рассмотрения), зададим «фильтрованную» выборку, иначе — активное множество объектов индексами $\mathcal{B} \subseteq \mathcal{I}$. Обозначим множество многомерных величин $\{\mathbf{x}^i|i \in \mathcal{B}\}$ как $\mathbf{x}^{\mathcal{B}}$. Задача выбора модели имеет вид

$$\mathcal{F} \ni \hat{f} = \arg\max_{\mathcal{A}\subseteq\mathcal{J},\mathcal{B}\subseteq\mathcal{I}} \mathcal{E}\left(f(\mathbf{w}_{\mathcal{A}}, \mathbf{x}^{\mathcal{B}})\right). \quad (2)$$

Способы решения этой задачи рассмотрены автором в [12]. Заметим, что для набору индексов признаков $\mathcal{J}$ мощности n соответствуют 2^n вершин двоичного куба. Каждая вершина задает некоторый активный набор признаков $\mathcal{A}$: считается, что j-й признак вошел в набор, если значение j-й координаты вершины единица. При решении задачи мы руководствуемся следующими предположениями.

1. Среди вершин куба существует по крайней мере одна, обозначим ее $\hat{\mathcal{A}}$, доставляющая матожидание правдоподобия модели.
2. От вершины $\mathcal{A} = \varnothing$ к вершине $\hat{\mathcal{A}}$ есть путь по ребрам куба (иначе — стратегия последовательного добавления-удаления признаков), который доставляет правдоподобию модели $\mathcal{E}\left(f(\mathbf{w}_{\mathcal{A}}, \mathbf{x})\right)$ сходимость по вероятности.

Множество индексов $\mathcal{B}$ задает выпуклую комбинацию $\{x_i|i \in \mathcal{B}\}$ — область $\mathcal{X}_{\mathcal{A}}$, «по крайней мере», в которой значения дисперсии $\{\beta_i|i \in \mathcal{B}\}$ за-

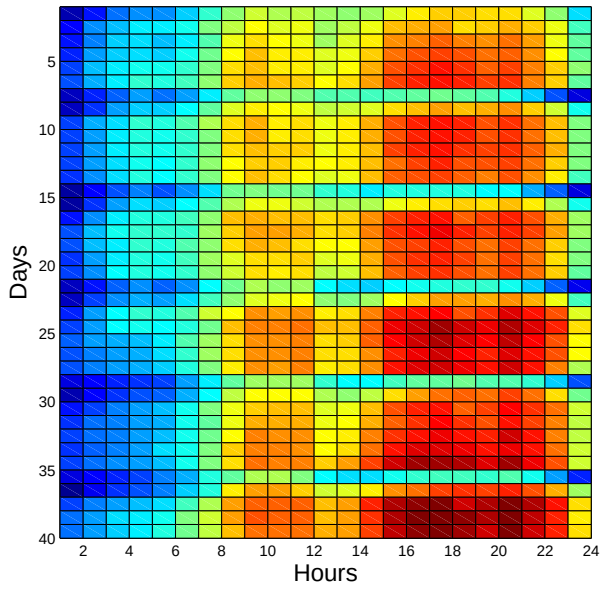

Рис. 2. Авторегрессионная матрица X, рабочие и выходные дни.

висимых переменных $\{\mathbf{y}_i | i \in \mathcal{B}\}$ меняются «незначительно». Другими словами, третий центральный момент, или коэффициент асимметрии случайной величины $\mathbf{y}$, соответствующей области $\mathcal{X}_\mathcal{A}$ равен нулю [13].

Выбор многоуровневых моделей

Многоуровневой моделью $\mathfrak{f}$ называется набор моделей $\mathfrak{f} = \{f_k | f \in \mathcal{F}\}$, $k = 1, \ldots, l$, такой, что

$$f_k : \mathcal{W}_k \times \mathcal{X}_{\mathcal{B}_k} \to \mathcal{Y}_{\mathcal{B}_k},$$

при разбиении $\mathcal{I} \supseteq \mathcal{B}^* = \sqcup \mathcal{B}_k$.

Введем функцию расстояния $\rho(f_k, f_l)$ между двумя моделями. Для этого используем дивергенцию Дженсена-Шеннона, в которой $\rho_{kl} \in [0, 1]$ является метрикой [9]:

$$\rho(p_k \| p_l) = 2^{-1} D_{\mathrm{KL}}\left(p_k \| p'\right) + 2^{-1} D_{\mathrm{KL}}\left(p' \| p_l\right),$$

где $p' = 2^{-1}(p_k + p_l)$ и здесь $p_k \stackrel{\text{def}}{=} (p(\mathbf{w}_\mathcal{A} | D, A, B, f_k)$. Несимметричная функция расстояния — дивергенция Кулльбака-Лейблера задана как

$$D_{\mathrm{KL}}\left(p \| p'\right) = \int\limits_{\mathbf{w} \in \mathcal{W}} p'(\mathbf{w}) \ln \frac{p(\mathbf{w})}{p'(\mathbf{w}) d\mathbf{w}}.$$

Отметим, что расстояние вводится только на моделях, имеющих одинаковый набор признаков $\mathcal{A}$.

Задача нахождения многоуровневых моделей ставится следующим образом:

$$\mathcal{F} \supset \hat{\mathfrak{f}} = \arg \max_{\mathcal{B}_1, \mathcal{B}_2 \subset \mathcal{B}} \rho(f_1, f_2) \tag{3}$$

при заданном множестве индексов признаков $\hat{\mathcal{A}}$, таком, что

$$\hat{\mathcal{A}} = \arg \max_{\mathcal{A} \subseteq \mathcal{J}} \mathcal{E}\left(f_1(\mathbf{w}_\mathcal{A}, \mathbf{x}^{\mathcal{B}_1})\right) \mathcal{E}\left(f_2(\mathbf{w}'_\mathcal{A}, \mathbf{x}^{\mathcal{B}_2})\right).$$

Иллюстрация: прогнозирование периодических временных рядов

Рассмотрим задачу авторегрессионного прогнозирования как одну из наиболее показательных при создании многоуровневой прогностической модели, в которых требуется одновременно выбрать объекты и признаки для каждой модели. Задан временной ряд $\{s(1), \ldots, s(\tau), \ldots, s(T-1)\}$, известен период $\varkappa$. Требуется спрогнозировать отсчет ряда в точке времени T. Для этого построим авторегрессионную матрицу X^* так, что ее строка i и столбец j отображались в номер отсчета как $(i-1)\varkappa \mapsto \tau$ при $\mathrm{mod}\frac{T}{\varkappa} = 0$. Представим X^* как матрицу, состоящую из присоединенных наборов векторов

$$X^* = \left[\begin{array}{c|c} \underset{m \times n}{X} & \underset{m \times 1}{\mathbf{y}} \\ \hline \underset{1 \times n}{\mathbf{x}^{m+1}} & \underset{1 \times 1}{s(T)} \end{array}\right].$$

Здесь X — матрица плана с числом столбцов $n = \varkappa - 1$ и $\mathbf{y}$ — последний столбец матрицы X^*. Принимая линейную модель зависимости $\mathbf{y} = X\mathbf{w}$, после оценки наиболее вероятного вектора параметров $\mathbf{w}$ получаем прогнозируемое значение $s(T) = \langle \mathbf{x}^{m+1}, \mathbf{w}_{\text{MP}} \rangle$.

Примем следующую гипотезу порождения данных: $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, B)$ из которой следует $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_{\text{MP}}, A)$. Тогда, при отсутствии гипотезы гомоскедаксичности регрессионных остатков и независимости элементов многомерной случайной величины $\mathbf{y}$, оптимизируемая функция S будет иметь вид

$$\begin{aligned} 2S(\mathbf{w} | D, f) =& (\mathbf{w} - \mathbf{w}_{\text{MP}})^T A (\mathbf{w} - \mathbf{w}_{\text{MP}}) + \\ &+ (\mathbf{f} - \mathbf{y})^T B (\mathbf{f} - \mathbf{y}). \end{aligned} \tag{4}$$

Учитываются также следующие предположения.

1. Существуют несколько типов периодов, каждый из которых должен быть спрогнозирован своей собственной моделью.
2. Не все фазы периода должны быть включены в модель.

Рисунки 1, 2 и 3 иллюстрирует результаты решения задач (2) и (3). На рис. 1 показан один период временного ряда и прогноз полученный этот период. На рис. 2 для каждого по следующего прогнозируемого значения показаны наиболее информативные признаки (имеющее наименьшие значения диагонали ковариационной матрицы A). Видно, что таковыми являются признаки, соответствующие столбцам авторегрессионной матрицы в окрестности периода данного прогнозируемого часа. На рис. 3 показана авторегрессионная матрица X. Ее строки (объекты выборки) можно условно разбить на два типа: соответствующие рабочим и выходным дням. Это дает основание для введения многоуровневой модели, состоящей из двух моделей одинаковой структуры с разными значениями параметров.

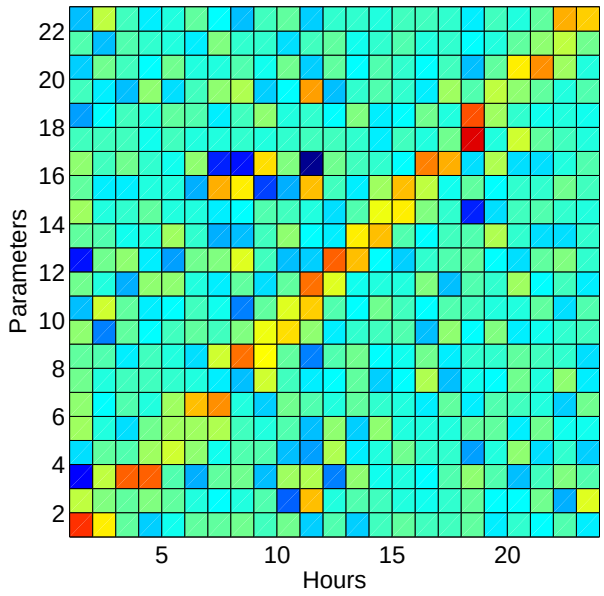

Рис. 3. Матрица информативности параметров $\mathbf{w}$ для различных значений времени τ.

Опишем алгоритм решения задачи (2).

1. Задаются единичные ковариационные матрицы A, B.
2. Для фиксированных значений матриц A, B оцениваются параметры $\mathbf{w}_{\text{MP}}$ модели f. При этом оптимизируется функция (4).
3. Оцениваются ковариационные матрицы A, B согласно гипотезе порождения данных.
4. Последние два шага повторяются до сходимости: пока изменение элементов матриц A, B не будут меньше заданных.
5. Выбираются те признаки $\mathcal{A}$ и объекты $\mathcal{B}$, которым соответствует наибольшие значения диагональных элементов матриц A, B.
6. Мощности множеств $\mathcal{A}, \mathcal{B}$ выбираются такими, чтобы они доставляли максимум функции правдоподобия (1).

Алгоритм решения задачи (3) состоит из двух основных шагов. Модели, включенные в $\mathfrak{f}$ заданы разбиением множества индексов $\mathcal{B}_1 \sqcup \mathcal{B}_2$, имеют различные ковариационные матрицы B_1, B_2 и общий набор признаков $\mathcal{A}$.

1. Решается задача максимизации правдоподобия $\mathfrak{f}$ на множестве $\mathcal{A}$ как в предыдущем алгоритме; разбиение $\mathcal{B}$ фиксировано.
2. Решается задача максимизации расстояния $\rho(f_1, f_2)$. Для этого значения диагональных элементов B_1, B_2 упорядочиваются по убыванию. Выполняется обмен b_1, b_2 индексами из разбиения $\mathcal{B}_1, \mathcal{B}_2$, соответствующими наименьшим значениям диагональных элементов. Числа b_1, b_2 выбираются такими, что расстояние $\rho(f_1, f_2)$ между двумя моделями было максимально.

Заключение

Рассмотренный метод позволяет решать задачу совместного выбора признаков и объектов как для одной регрессионной модели, так и для их набора. При этом особое внимание уделяется принятию статистических гипотез и, как следствие, корректности использования функций качества, с помощью которых отыскиваются оптимальные, в данном случае наиболее вероятные параметры моделей, а также их матрица их ковариаций.

Литература

[1] *Vladislavleva E., Smith G., Hertog D.* Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming // *EEE Transactions on Evolutionary Computation.* — 2009. — Vol. 13(2). — Pp. 333–349.

[2] *Watanabe S.* Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory // *J. Machine Learning Research.* — 2010. — Vol. 11. — Pp. 3571–3594.

[3] *Arlot S., Blanchard G., Roquain E.* Some nonasymptotic results on resampling in high dimension // *Annals of Statistics.* — 2009. — Vol. 38 — Pp. 51–82.

[4] *MacKay D.* Information Theory, Inference, and Learning Algorithms. — Cambridge University Press, 2003.

[5] *Bishop C. M.* A new framework for machine learning // Computational Intelligence.— Springer, 2008. — Pp. 1–24.

[6] *Grunwald P. D.* The Minimum Description Length Principle. — MIT Press, 2007.

[7] *Grünwald P., Myung I. J., Pitt M.* Advances in Minimum Description Length. — MIT Press, 2005.

[8] *Lee Y., Nelder J. A., Pawitan Y.* Generalized linear models with random effects: unified analysis via h-likelihood. — Chapman & Hall/CRC, 2006.

[9] *Lin J.* Divergence measures based on the shannon entropy // *IEEE Transactions on Information Theory.* — 1991. — Vol. 37, no. 1. — P. 145.

[10] *Seber G. A. F., Wild C.* Nonlinear Regression. — Wiley-IEEE, 2003.

[11] *Strijov V., Weber G. W.* Nonlinear regression model generation using hyperparameter optimization // *Computers and Mathematics with Applications.* — 2010. — Vol. 60, no. 4. — Pp. 981–988.

[12] *Strijov V. V., Krymova E. A., Weber G. W.* Evidence optimization for consequently generated models // *Mathematical and Computer Modelling.* — 2011.

[13] *Rissanen J., Roos T., Myllymäki P.* Model selection by sequentially normalized least squares // *J. Multivariate Analysis.* — 2010. — Vol. 101, no. 4. — Pp. 839–849.