Адаптивное планирование регрессионных экспериментов на основе гауссовских процессов*

Максим Панов, Евгений Бурнаев
{maxim.panov, evgeny.burnaev}@datadvance.net
Москва, Институт Проблем Передачи Информации РАН,
Москва, DATADVANCE,
Долгопрудный, PreMoLab

В работе рассматривается задача адаптивного планирования эксперимента для задачи регрессии. В качестве регрессионной модели используется стохастическая модель, основанная на понятии гауссовского процесса. В работе рассмотрен как ряд классических эвристических критериев адаптивного планирования эксперимента, так и ряд новых критериев, основанных на более строгой теоретической постановке. Проведено сравнение рассматриваемых методов на большом количестве тестовых функций различных размерностей.

Введение

Одной из основных задач, которые приходится решать при построении метамоделей (моделей на основе данных), является задача аппроксимации неизвестной зависимости по данным [1, 2]. Наиболее популярная модель для построения аппроксиматоров, основанная на гауссовских процессах [3, 4, 5], используется в большом количестве разнообразных прикладных задач, включая концептуальное проектирование, структурную оптимизацию, многокритериальную оптимизацию при проектировании, конструирование в аэрокосмической и автомобильной отраслях.

Во многих инженерных задачах количество вычислений целевой функции $f(\mathbf{x})$ существенно ограничено по причине очень большого времени, требуемого для одного вычисления, и/или его высокой стоимости. В связи с этим крайне важной является задача построения методов выбора обучающей выборки D_N ограниченного объема N таким образом, чтобы максимизировать качество аппроксимации. Задача, в которой план эксперимента (обучающая выборка) выбирается оптимальным образом в смысле некоторого статистического критерия, называется задачей оптимального планирования эксперимента. Для параметрических моделей задача адаптивного планирования эксперимента, в которой критерием оптимальности выступает качество аппроксимации, совпадает с задачей оценки параметров модели оптимальным образом. Теория оптимального планирования эксперимента для параметрических моделей широко развита в работах Федорова [15], Пукельшейма [17] и мн. др. Однако, в случае, когда регрессионная модель является непараметрической (например, модель гауссовского процесса), возникает другая постановка задачи оптимального планирования эксперимента [16, 9], в которой план эксперимента строится таким образом, чтобы обеспечить наилучшее предсказание значений функции.

В данной работе мы сосредоточимся на этой постановке, причем будем рассматривать её адаптивный вариант, который в последнее время получил большое развитие [2, 6]. При адаптивном планировании эксперимента обучающая выборка строится итеративно, причем на каждой новой итерации для выбора новых точек используется аппроксимация, построенная на предыдущей итерации.

Постановка задачи аппроксимации

В наиболее общем виде задача аппроксимации может быть сформулирована следующим образом. Пусть $y = f(\mathbf{x})$ некоторая неизвестная функция со входом $\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^n$ и выходом $y \in \mathbb{R}$. Пусть $D_N = \big(X_N, \mathbf{y}_N\big) = \big\{\big(\mathbf{x}_i, y_i = f(\mathbf{x}_i)\big), i = 1, \ldots, N\big\}$ - обучающая выборка. Задача состоит в построении аппроксимации $\hat{y} = \hat{f}(\mathbf{x}) = \hat{f}(\mathbf{x}|D_N)$ для исходной зависимости $y = f(\mathbf{x})$ по обучающей выборке D_N.

Если для всех $\mathbf{x} \in \mathbb{X}$ (не только для $\mathbf{x} \in D_N$) имеет место примерное равенство $\hat{f}(\mathbf{x}) \approx f(\mathbf{x})$, то считается, что аппроксиматор хорошо воспроизводит исходную зависимость. Это факт проверяется на независимой тестовой выборке $D_* = \big(X^*, \mathbf{y}^*\big) = \big\{\big(\mathbf{x}_j^*, y_j^* = f(\mathbf{x}_j^*)\big), j = 1, \ldots, N_*\big\}$. Мерой качества аппроксимации является среднеквадратичная ошибка на тестовой выборке:

$$Q(\hat{f}|D_*) = \sqrt{\frac{1}{N_*}\sum_{j=1}^{N_*}\big(y_j^* - \hat{f}(\mathbf{x}_j^*)\big)^2}. \qquad (1)$$

В следующем разделе рассматривается построение аппроксиматора на основе гауссовских процессов [7], который в дальнейшем используется для адаптивного планирования эксперимента.

Аппроксиматор на основе гауссовского процесса

Предположим, что целевая функция $f(\mathbf{x})$ является реализацией гауссовского процесса. Гауссовский процесс является одним из возможных способов задания распределения на пространстве

Работа выполнена при поддержке Лаборатории структурных методов анализа данных в предсказательном моделировании, МФТИ, грант правительства РФ дог. 11.G34.31.0073.

функций, которое полностью определяется функцией среднего $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ и ковариационной функцией $cov(f(\mathbf{x}), f(\mathbf{x}')) = k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$. В типичных, реалистичных ситуациях при моделировании мы не имеем доступа непосредственно к значениям функции, а наблюдаем их только в зашумленном виде:

$$y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon(\mathbf{x}), \qquad (2)$$

где шум $\varepsilon(\mathbf{x})$ моделируется независимыми одинаково распределенными нормальными случайными величинами с нулевым средним и дисперсией $\tilde{\sigma}^2$. В таком случае наблюдения $y(\mathbf{x})$ будут гауссовским процессом с нулевым средним и ковариационной функцией $cov(y(\mathbf{x}), y(\mathbf{x}')) = k(\mathbf{x}, \mathbf{x}') + \tilde{\sigma}^2\delta_{\mathbf{x},\mathbf{x}'}$, где $\delta_{\mathbf{x},\mathbf{x}'}$ - символ Кронекера.

Если положить функцию среднего процесса $f(\mathbf{x})$ нулевой, т.е. $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] = 0$, а ковариационную функцию $k(\mathbf{x}, \mathbf{x}')$ считать известной, то функция апостериорного (для заданной обучающей выборки) среднего значения гауссовского процесса $f(\mathbf{x})$ в точках контрольной выборки X^* выглядит следующим образом [7]:

$$\hat{f}_N(X^*) = \hat{f}(X^*|D_N) = K_*\big(K + \tilde{\sigma}^2 I\big)^{-1}\mathbf{y}, \qquad (3)$$

где $K_* = \big[k(\mathbf{x}_i^*, \mathbf{x}_j), i = 1, \ldots, N_*; j = 1, \ldots, N\big]$, $K = \big[k(\mathbf{x}_i, \mathbf{x}_j), i, j = 1, \ldots, N\big]$, I - единичная матрица размера $N \times N$.

При этом апостериорная дисперсия гауссовского процесса в точках контрольной выборки имеет вид:

$$\hat{\sigma}^2_N(X^*) = \hat{\sigma}^2(X^*|X_N) =$$
$$= \mathbf{k}_* + \tilde{\sigma}^2\mathbf{e}_* - diag\big(K_*\big(K + \tilde{\sigma}^2 I\big)^{-1}K_*^T\big), \qquad (4)$$

где $\mathbf{k}_* = \big[k(\mathbf{x}_i^*, \mathbf{x}_i^*), i = 1, \ldots, N_*\big]$, $\mathbf{e}_*$ - единичный вектор длины N_*.

Апостериорное среднее в точках контрольной выборки используется для прогноза значений функции, а соответствующие дисперсии могут быть использованы как оценки ожидаемой ошибки аппроксимации в этих точках.

Заметим, что в данном разделе нами описан только базовый алгоритм регрессии на основе гауссовских процессов. Для более подробного ознакомления см. [8].

Адаптивное планирование регрессионного эксперимента

Для постановки задачи адаптивного планирования эксперимента нам понадобятся следующие определения.

Определение 1. *Пусть*

$$\rho = \rho(f, \hat{f})$$

- некоторый измеримый функционал. Назовем его функцией ошибки.

Определение 2. *Пусть*

$$q_\rho\big(D_l\big) = \mathbb{E}\big[\rho(f, \hat{f}_l)\big|D_l\big]$$

где математическое ожидание понимается как условное при фиксированной выборке D_l. Назовем $q_\rho\big(D_l\big)$ средним апостериорным риском.

С целью нахождения оптимального плана эксперимента будем решать следующую задачу:

$$J^* = \min_{X_N \in \mathbb{X}^N} \mathbb{E}_{\mathbf{y}_N}\left[\sum_{i=1}^{N} Q_\rho\big(\mathbf{x}_i, y_i, D_{i-1}\big)\right], \qquad (5)$$

где $Q_\rho\big(\mathbf{x}_i, y_i, D_{i-1}\big) = q_\rho\big(D_{i-1} \cup (\mathbf{x}_i, y_i)\big) - q_\rho\big(D_{i-1}\big)$.

Заметим, что данная задача эквивалента задаче минимизации итоговой ошибки на последней итерации:

$$J^* = \min_{X_N \in \mathbb{X}^N} \mathbb{E}_{\mathbf{y}_N}\big[q_\rho\big(D_N\big)\big] - q_\rho\big(D_0\big),$$

где введено обозначение $q_\rho\big(D_0\big) = \mathbb{E}\rho(f, \hat{f}_0)$, $\hat{f}_0 = 0$.

Заметим также, что рассматриваемая задача является задачей стохастического динамического программирования. Выпишем для нее уравнение Беллмана назад:

$$J_r\big(D_{r-1}\big) = \min_{\mathbf{x}_r \in \mathbb{X}} \mathbb{E}_{y_r}\Big[Q_\rho\big(\mathbf{x}_r, y_r, D_{r-1}\big) +$$

$$+ J_{r+1}\big(D_{r-1} \cup (\mathbf{x}_r, y_r)\big)\Big], r = 1, \ldots, N, J_{N+1} \equiv 0,$$

где $J_r\big(D_{r-1}\big)$ - функция цены на r-ом шаге решения задачи.

Планирование на один шаг вперед

Функция цены в случае, когда осталось сделать один последний шаг алгоритма, дается следующим выражением:

$$J_N\big(D_{N-1}\big) = \min_{\mathbf{x}_N \in \mathbb{X}} \mathbb{E}_{y_N}\Big[Q_\rho\big(\mathbf{x}_N, y_N, D_{N-1}\big)\Big].$$

Таким образом критерий выбора оптимальной точки:

$$\mathbf{x}_N = \arg\min_{\mathbf{x} \in \mathbb{X}} \mathbb{E}_{y_N}\Big[Q_\rho\big(\mathbf{x}, y_N, D_{N-1}\big)\Big].$$

В зависимости от выбора вида функции ошибки $\rho(f, \hat{f})$ будут получаться различные виды критериев. Рассмотрим следующие варианты выбора функции ошибки:

1. L_2-норма разности f и $\hat{f}_N$:

$$\rho_2(f, \hat{f}_N) = \frac{1}{|\mathbb{X}|}\int_{\mathbb{X}} (f(\mathbf{u}) - \hat{f}_N(\mathbf{u}))^2 \mathrm{d}\mathbf{u}.$$

2. L_1-норма разности f и $\hat{f}_N$:

$$\rho_1(f,\hat{f}_N)=\frac{1}{|\mathbb{X}|}\int_{\mathbb{X}}|f(\mathbf{u})-\hat{f}_N(\mathbf{u})|\mathrm{d}\mathbf{u}.$$

3. L_∞-норма разности f и $\hat{f}_N$:

$$\rho_\infty(f,\hat{f}_N)=\max_{\mathbf{u}\in\mathbb{X}}|f(\mathbf{u})-\hat{f}_N(\mathbf{u})|.$$

L_2 функция ошибки

Утверждение 1. *Если функция ошибки* $\rho(f,\hat{f}_N)=\rho_2(f,\hat{f}_N)$, *то оптимальное решение для последнего шага алгоритма имеет вид:*

$$\mathbf{x}_N=\arg\min_{\mathbf{x}\in\mathbb{X}}\frac{1}{|\mathbb{X}|}\int_{\mathbb{X}}\big(\hat{\sigma}^2(\mathbf{u}|X_{N-1}\cup\mathbf{x})-\hat{\sigma}^2(\mathbf{u}|X_{N-1})\big)\mathrm{d}\mathbf{u}.$$

Заметим, что данный критерий известен под названием ImseGain, а в силу независимости $\hat{\sigma}^2(\mathbf{u}|X_{N-1})$ от $\mathbf{x}_N$ данный критерий совпадает с известным в литературе критерием Imse:

$$\mathbf{x}_N=\arg\min_{\mathbf{x}\in\mathbb{X}}\frac{1}{|\mathbb{X}|}\int_{\mathbb{X}}\hat{\sigma}^2(\mathbf{u}|X_{N-1}\cup\mathbf{x})\mathrm{d}\mathbf{u}.$$

В частных случаях критерий может быть вычислен аналитически.

L_1 функция ошибки

Утверждение 2. *Если функция ошибки* $\rho(f,\hat{f}_N)=\rho_1(f,\hat{f}_N)$, *то оптимальное решение для последнего шага алгоритма имеет вид:*

$$\mathbf{x}_N=\arg\min_{\mathbf{x}\in\mathbb{X}}\frac{1}{|\mathbb{X}|}\int_{\mathbb{X}}\sqrt{\frac{2}{\pi}}\big(\hat{\sigma}(\mathbf{u}|X_{N-1}\cup\mathbf{x})-\hat{\sigma}(\mathbf{u}|X_{N-1})\big)\mathrm{d}\mathbf{u}.$$

Заметим, что аналитическое вычисление данного критерия представляется затруднительным в силу особенностей подынтегрального выражения.

L_∞ функция ошибки

Утверждение 3. *Если функция ошибки* $\rho(f,\hat{f}_N)=\rho_\infty(f,\hat{f}_N)$, *то оптимальное решение для последнего шага алгоритма имеет вид:*

$$\mathbf{x}_N=\arg\min_{\mathbf{x}\in\mathbb{X}}\mathbb{E}\left[\max_{\mathbf{u}\in\mathbb{X}}\big|f(\mathbf{u})-\hat{f}(\mathbf{u}|D_{N-1}\cup(\mathbf{x},y_N))\big|-\max_{\mathbf{u}\in\mathbb{X}}\big|f(\mathbf{u})-\hat{f}(\mathbf{u}|D_{N-1})\big|\,\middle|\,D_{N-1}\right].$$

Заметим, что аналитическое вычисление данного критерия представляется невозможным. Для его вычисления требуется применение аппроксимационных техник для асимптотик распределения максимума гауссовского процесса. Обычно, распределение максимума пропорционально максимальному значению апостериорной дисперсии процесса, что обосновывает применения стандартного критерия Maximum Variance:

$$\mathbf{x}_N=\arg\max_{\mathbf{x}\in\mathbb{X}}\hat{\sigma}^2(\mathbf{x}|X_{N-1}).$$

Также в работе рассматривается несколько эвристических критериев адаптивного планирования эксперимента:

— Критерий uniform:

$$\mathbf{x}_N=\arg\max_{\mathbf{x}\in\mathbb{X}}\min_{\mathbf{v}\in X_{N-1}}d(\mathbf{x},\mathbf{v}),$$

где $d(\mathbf{x},\mathbf{v})$ - расстояние между точками $\mathbf{x}$ и $\mathbf{v}$ в некоторой метрике.

— Комбинация критериев ImseGain и Maximum Variance в виде мультипликативной формулы (ImseGain-Maximum Variance):

$$\mathbf{x}_N=\arg\min_{\mathbf{x}\in\mathbb{X}}\frac{\hat{\sigma}^2(\mathbf{u}|X_{N-1})}{|\mathbb{X}|}\int_{\mathbb{X}}\big(\hat{\sigma}^2(\mathbf{u}|X_{N-1}\cup\mathbf{x})-\hat{\sigma}^2(\mathbf{u}|X_{N-1})\big)\mathrm{d}\mathbf{u}=$$

$$=\arg\min_{\mathbf{x}\in\mathbb{X}}\frac{1}{|\mathbb{X}|}\int_{\mathbb{X}}\big[k(\mathbf{x},\mathbf{u})-k(\mathbf{x},X)^TK^{-1}k(\mathbf{u},X)\big]^2d\mathbf{u}.$$

Данный критерий также как и критерий ImseGain учитывает глобально поведение аппроксимации, но при этом более численно стабилен.

Экспериментальные результаты

В ходе экспериментов был использован большой набор тестовых функций, которые применяются для тестирования задач оптимизации [12, 13]. Всего тестирование проводилось на 10 различных функциях размерностей 3 и 4, для каждой из которых генерировалось 5 случайных начальных выборок размером (10 * размерность) точек. Процедура адаптивного планирования эксперимента проводилась для каждого из рассматриваемых критериев и для каждой из 5 начальных выборок. Причем, на каждой итерации в обучающую выборку добавлялась одна точка, выбранная в соответствии с рассматриваемым критерием. Всего добавлялось 50 точек. Результаты сравнивались по среднеквадратичной ошибке (1) на больших контрольных выборках из 10000 точек. Для оптимизации критериев используется алгоритм имитации отжига [11], который обладает хорошими глобальными свойствами и обеспечивает качественную оптимизацию критериев. Также для сравнения с критериями приведены результаты для случая, когда в выборку добавляются случайные точки из области дизайна. Для удобства результаты представлены в виде кривых Долан-Мора [14]. Чем выше кривая находится на графике, тем выше качество работы соответствующего алгоритма. Результаты представлены для аппроксимаций, построенных по выборке

с 50 адаптивно добавленными точками (см. рисунок 1). Заметим, что критерий ImseGain на значительной доле функций дает наилучший результат (значение кривой Долан-Мора в нуле), но при этом также на многих функциях значительно проигрывает, что вызвано численной нестабильностью этого критерия. Предложенный в данной работе критерий ImseGain-MaximumVariance позволяет избежать этого эффекта и показывает результат лучше, чем все остальные критерии.

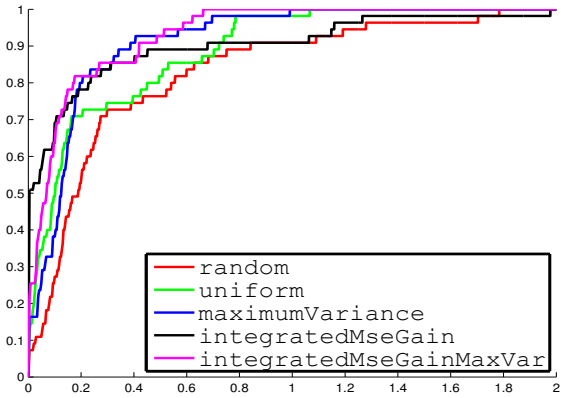

Рис. 1. Кривые Долан-Мора для тестовых функций, 50 добавленных точек.

Выводы

В работе предложен подход к задаче адаптивного планирования эксперимента в случае регрессии на основе гауссовских процессов, который позволяет сформулировать ее в виде задачи стохастического динамического программирования. Оптимальные на один шаг вперед решения этой задачи динамического программирования позволяют обосновать многие из известных в литературе критериев адаптивного планирования эксперимента. Также в работе предложен новый критерии ImseGain-MaximumVariance, который в простой форме комбинирует уже известные критерии. Экспериментальные результаты показали значительное преимущество адаптивных методов над случайным добавлением точек в обучающую выборку. При этом предложенный критерий показал более хорошие результаты по сравнению с известными критериями, а также оказался более численно стабильным.

Литература

[1] Бернштейн А.В., Бурнаев Е.В., Кулешов А.П. Интеллектуальный анализ данных в метамоделировании. // Труды 17 Всероссийского Семинара "Нейроинформатика и ее приложения к Анализу Данных", Красноярск, 2009 – с. 23-28.

[2] Forrester A., Sobester A., Keane A. Engineering Design via Surrogate Modelling. A Practical Guide. – Wiley, 2008. – 238 p.

[3] Giunta A., Watson L. T.A Comparison of Approximation Modeling Technique: Polynomial Versus Interpolating Models. // 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, Vol. 1, AIAA, Reston, VA, 1998 – pp. 392–404.

[4] Simpson T. W., Booker A. J., Ghosh S., Giunta A., Koch P. N., Yang, R. J. Approximation Methods in Multidisciplinary Analysis and Optimization: A Panel Discussion. // Structural and Multidisciplinary Optimization, Vol. 27, No. 5, 2004 – pp. 302–313.

[5] Batill S. M., Renaud J. E., Gu X., Modeling and Simulation Uncertainty in Multidisciplinary Design Optimization // AIAA Paper 2000-4803, Sept. 2000.

[6] Chen R.J.W., Sudjianto A. On sequential sampling for global metamodeling in engineering design // Proceedings of DETC'02. Montreal, Canada, September 29-October 2, 2002.

[7] Rasmussen C.E., Williams C.K.I. Gaussian Processes for Machine Learning. – the MIT Press, 2006.

[8] Burnaev E., Zaytsev A., Panov M., Prikhodko P. and Yanovich Yu. Modeling of nonstationary covariance function of Gaussian process on base of expansion in dictionary of nonlinear functions // In ITaS–2011, Gelendzhik, October 2–7 2011.

[9] Zimmerman D.L. Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction // Environmetrics, Volume 17, Issue 6, 2006 – pages 635–652.

[10] Gramacy R.B., Lee H.K.H Adaptive design and analysis of supercomputer experiments //arXiv:0805.4359v4

[11] Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P. Optimization by Simulated Annealing // Science, New Series, Vol. 220, No. 4598, 1983 – pp. 671-680.

[12] GDR MASCOT-NUM Toy Functions benchmark. `http://gdr-mascotnum.math.cnrs.fr/data2/benchmarks/jm.pdf`

[13] Lappeenranta University of Technology: evoluationary computation pages - the function testbed. `http://www.it.lut.fi/ip/evo/functions/functions.html`

[14] Dolan E. D., More J. J. Benchmarking optimization software with performance profiles // Mathematical Programmming, Ser. A 91, 2002 – pp. 201–213.

[15] Fedorov V.V. Theory of Optimal Experiments, Academic Press, 1972.

[16] Fedorov V.V. Design of spatial experiments: model fitting and prediction. Handbook of Statistics, volume 13. Elsevier, Amsterdam, 1996 – pp 515–553.

[17] Pukelsheim F. Optimal Design of Experiments. Wiley, New York, 1993.