

На правах рукописи



Приходько Павел Викторович

**Применение методов агрегации экспертов и регрессии на
основе гауссовских процессов для построения метамоделей**

Специальность: 05.13.17 — теоретические основы информатики

Автореферат диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2013

Работа выполнена в Институте проблем передачи информации им. А.А. Харкевича Российской Академии Наук (ИППИ РАН)

Научный руководитель:

кандидат физико-математических наук, доцент

Е.В. Бурнаев

Официальные оппоненты:

доктор физико-математических наук, профессор

В.В. Вьюгин,

*Институт проблем передачи и обработки информации
им. А.А. Харкевича, заведующий лабораторией*

кандидат физико-математических наук, профессор

А.Я. Червоненкис,

Институт проблем управления

им. В.А. Трапезникова РАН, ведущий научный сотрудник

Ведущая организация:

Институт проблем информатики Российской
академии наук (ИПИ РАН)

Защита состоится 2 декабря 2013 г. в 16.30 на заседании диссертационного совета Д 212.156.04 при Московском физико-техническом институте (государственном университете) по адресу: 141700, г. Долгопрудный, Московская обл., Институтский пер., д. 9, ауд. 204 Нового корпуса. С диссертацией можно ознакомиться в библиотеке Московского физико-технического института (государственного университета).

Автореферат разослан 25 октября 2013 г.

Ученый секретарь диссертационного совета Д 212.156.04

к.ф.-м.н.

Стрыгин Л. В.



Общая характеристика работы

Актуальность проблемы. Последнее десятилетие с развитием вычислительной техники при проектировании инженерных объектов возник спрос на построение метамоделей (суррогатных моделей)^{1,2} для анализа результатов экспериментов или сложных вычислительных кодов.

Спрос обусловлен следующей проблемой: инженеру требуется оптимизировать или исследовать поведение некоторой зависимости, являющейся функцией от ряда параметров³, однако, вычисление значения этой функции для новых значений параметров (в новой точке) затратно в смысле вычислительных ресурсов, времени или даже невозможно. Зачастую такое происходит, если для вычисления значения функции в новой точке требуется проводить натурный эксперимент или использовать “тяжелый” вычислительный код (например, обсчет модели конечных элементов даже на современных вычислительных мощностях может занимать несколько дней).

Для случаев, когда инженеру достаточно получить лишь приближенное решение задачи, одним из популярных подходов, позволяющих проанализировать зависимость, стало построение по имеющимся данным математической модели (метамоделю), которая, насколько это вообще возможно, имитирует поведение рассматриваемой зависимости, однако, при этом может быть очень быстро вычислена в произвольных точках. При таком подходе построенная метамоделю полностью заменяет собой (аппроксимирует) исходную зависимость во всех последующих расчетах (по этой причине метамоделю также иногда называют суррогатными моделями). Качество метамоделю обычно определяется тем, насколько она хорошо воспроизводит поведение исходной зависимости в смысле некоторой заданной метрики, например, среднеквадратичной разности между предсказаниями метамоделю и истинными значениями зависимости на некоторой репрезентативной выборке данных.

¹ Kuleshov A.P., Bernstein A.V., Burnaev E.V. *Adaptive models of complex systems based on data handling // Proceedings of the 3rd International Conference on Inductive Modelling, Kyiv, Ukraine, pp. 64-71, 2010.*

² Forrester A., Sobester A., Keane A. *Engineering Design via Surrogate Modelling. A Practical Guide. Wiley, 2008.*

³ Также в дальнейшем в том же смысле употребляются термины “признаки” и “входы”.

Довольно часто на практике исходная зависимость существенно нелинейна, характер ее поведения неизвестен, а в наличии имеется только выборка пар “параметры”—“значение зависимости”. В этом случае для построения метамоделей как правило применяются универсальные методы построения регрессионных моделей из статистики и машинного обучения, такие как: искусственные нейронные сети⁴ или регрессия на основе гауссовских процессов (кригинг)⁵. Соответствующие базовые метамоделю (далее будем называть их также просто моделями), построенные с помощью этих методов, не предполагают априори какой-то специальный вид аппроксимируемой зависимости, и их параметры могут быть настроены так, чтобы приблизить довольно широкий класс функций.

Построенная по данным базовая модель может не обеспечивать требуемую точность приближения, однако, в некоторых ситуациях качество аппроксимации можно повысить введением дополнительных алгоритмических надстроек над процессом построения моделей.

В работе рассматриваются такие надстройки для нескольких специальных случаев:

Недостаточная сложность модели. В ряде работ⁶ показывается, что в случае “простой” базовой модели, точность аппроксимации может быть существенно повышена, если итоговая модель определенным образом строится как линейная комбинация (ансамбль) некоторого количества базовых моделей.

Кроме того, в случае использования искусственных нейронных сетей (ИНС) в качестве базовой модели на практике не всегда возможно заранее оценить сложность зависимости и соответственно подобрать адекватную ей сложность (определяемую количеством узлов) ИНС, что приводит к необходимости итеративного выбора структуры модели в соответствии с некоторым дополнительным критерием. В этом смысле построение ансамблей может быть альтернативой прямому увеличению сложности ИНС.

Случай неоднородной выборки. Методы построения регрессионных моделей обычно неявно предполагают, что в выборке данных значе-

⁴ Хайкин С. *Нейронные сети: полный курс. Второе издание.* Вильямс, 2006.

⁵ Rasmussen C.E., Williams C.K.I. *Gaussian Processes for Machine Learning.* MIT Press, 2006.

⁶ N. Ueda and R. Nakano. *Generalization error of ensemble estimators, In Proceedings of International Conference on Neural Networks, pp. 90–95, 1996.*

ния входов равномерно заполняют пространство дизайна, а сама функция ведет себя достаточно однородно в том смысле, что она не содержит разрывы или области с качественно разным поведением (то есть для выбранного типа базовой модели можно задать такой набор параметров, что модель будет хорошо приближать рассматриваемую зависимость на всей области определения). Нарушение этого условия может приводить к проблемам при построении моделей, таким как низкая точность в определенных областях пространства дизайна, вырождение модели и т.п.

Для описания неоднородного поведения функции в работе рассматриваются следующие распространенные модели неоднородности:

- 1) данные порождены смесью нормальных распределений, у каждого элемента смеси имеется свой вид зависимости⁷,
- 2) данные описываются набором областей, разделенных линейными гиперплоскостями, причем в каждой области имеется свой вид зависимости.

Первая модель неоднородности позволяет эффективно описывать случаи, в которых обучающая выборка неравномерно заполняет пространство признаков (пространство дизайна); вторая модель подходит для описания задач, в которых аппроксимируемая зависимость имеет разрывы.

Избыточность в признаках. Набор признаков обычно может быть избыточен в одном из следующих смыслов:

- 1) Входные признаки могут быть скоррелированы,
- 2) Функция слабо зависит или не зависит вовсе от некоторых исходных признаков,
- 3) Функция зависит только от проекции входов на некоторое подпространство.

Отметим, что третий сценарий является наиболее общим и включает в себя предыдущие два.

⁷ Имеется в виду, что, если для элемента смеси построить базовую модель только по тем данным, которые порождены этим элементом смеси, то такая модель будет существенно точнее аппроксимировать рассматриваемый элемент смеси, чем модель, построенная по всей обучающей выборке.

Наличие избыточности в признаках при построении базовой модели может затруднять аппроксимацию и приводить к различным вырождениям или нестабильностям в алгоритме обучения.

Кроме того, снижение размерности данных, если при этом не теряется информация о зависимости, также эффективно уплотняет выборку (плотность точек в проекции на подпространство меньшей размерности получается выше, чем по всему пространству), что упрощает задачу аппроксимации.

В литературе присутствует множество процедур⁸, позволяющих найти проекцию входов на некоторое подпространство. Однако при использовании таких процедур как правило необходимо заранее указывать размерность сжатия, а также строить ядерные оценки зависимости (требующие подбора параметров, например, посредством кросс-проверки), которые хорошо работают лишь в случае малых размерностей. Указанные особенности делают эти методы малоприменимыми на практике.

Таким образом, **целью** данной работы является анализ существующих и разработка новых алгоритмов анализа данных, которые работают в сочетании с методом построения базовых моделей и позволяют специальным образом учитывать ситуации, когда базовая модель слишком проста для восстановления рассматриваемой зависимости, имеется избыточность во входных признаках, обучающая выборка неравномерна, и за счет этого повышать качество построенных моделей.

В соответствии с поставленной целью были определены следующие **задачи** исследования:

1. Определить специальные постановки задачи восстановления зависимости, в которых учет особенностей постановки может улучшить качество базовой модели.
2. Разработать методы построения моделей в случае, когда функция зависит лишь от некоторой проекции признаков на линейное подпространство.

⁸Yingcun Xia, Howell Tong, W. K. Li, Li-Xing Zhu. An adaptive estimation of dimension reduction space // *J. R. Statist. Soc. B* (2002) 64, Part 3, pp. 363–410.

3. Разработать методы агрегации базовых регрессионных моделей для случаев, когда

- а) выборка неоднородна, присутствуют выраженные кластеры точек и требуется выделить в пространстве признаков области, для каждой из которых будет строиться своя локальная модель,
- б) базовая модель слишком проста, чтобы получить аппроксимацию требуемого качества, поэтому для уточнения прогноза необходимо использовать ансамбль базовых моделей.

Общая методика исследования. Для решения поставленных задач в работе используются методы машинного обучения, статистики и статистической теории машинного обучения, языки Matlab и python для разработки прототипов предложенных методов. Также было проведено обширное экспериментальное исследование.

Научная новизна результатов, полученных в диссертации, заключается в том, что в ней предложен набор оригинальных процедур и исследованы их некоторые теоретические свойства:

- Разработана процедура построения метамоделей в случае, если функция зависит лишь от проекции признаков на некоторое неизвестное линейное подпространство, что позволяет существенно улучшить качество модели при выполнении этого предположения, а также дает возможность линейно снизить размерность данных. Кроме того, качество построенной модели становится устойчивым к повороту координатных осей пространства дизайна. Предлагается метод оценки размерности линейного подпространства, а также способ нахождения базиса в этом подпространстве.
- Разработана процедура БегБуст построения ансамблей базовых метамоделей. Получена верхняя граница на ошибку обобщающей способности процедуры.
- Разработана процедура построения метамоделей на основе кластеризации данных с помощью гауссовских смесей или разделения линейными гиперплоскостями, а также приводится пример успешного

применения процедуры в задаче построения высокоточной модели для оптимизации веса корпуса самолета.

Теоретическая и практическая значимость

Описанные в работе методы широко применяются для решения прикладных задач аэрокосмической отрасли в компании DATADVANCE. Кроме того, для метода построения ансамблей БегБуст получена верхняя граница на ошибку обобщающей способности. Разработанные при этом теоретические подходы можно использовать при анализе других алгоритмов построения ансамблей.

Основные результаты, выносимые на защиту:

1. Метод VEGA (Variable Extraction via Gradient Approximation) нахождения подпространства эффективной размерности для заданной функции по выборке. Вычислительно эффективный способ оценки матрицы ковариаций градиентов зависимости.
2. Метод БегБуст построения ансамблей базовых метамоделей.
3. Верхняя граница ошибки обобщающей способности метода БегБуст.
4. Метод построения метамоделей на основе кластеризации данных с помощью гауссовских смесей. Метод построения метамоделей на основе кластеризации данных через разделение линейными гиперплоскостями пространства дизайна. Вычислительно эффективный способ оценки параметров базовых моделей на основе оценки ошибки скользящего контроля для случая линейных экспертов.

Апробация работы

Результаты работы докладывались и обсуждались на следующих российских и международных конференциях:

- “Интеллектуализация обработки информации” (2012, Будва, Черногория),

- 1st International Conference on Composites Structures Dynamics (2012, Аркашон, Франция). Было сделано два доклада, один из которых пленарный,
- Uncertainty in Computer Models 2012 Conference (2012, Шеффилд, Англия),
- “Математические методы распознавания образов” (2011, Петрозаводск, Россия),
- “Информационные Технологии и Системы” (2009, Бекасово, Россия; 2010, 2011 Геленджик, Россия; 2012, Петрозаводск, Россия),
- Third International Workshop on Surrogate Modelling and Space Mapping For Engineering Optimization (2012, Рейкьявик, Исландия),
- 9th International Conference “Computer Data Analysis and Modeling: Complex Stochastic Data and Systems” (2010, Минск, Белоруссия),
- 52-я, 53-я, 54-я научные конференции Московского физико-технического института (2009-2012, Долгопрудный, Россия),
- Sixth conference “Mathematical Methods in Reliability” (2009, Москва, Россия).

Полученные в работе результаты также обсуждались на научных семинарах лаборатории структурных методов анализа данных в предсказательном моделировании МФТИ (2012, 2013) и на семинарах секторов 5 и 8.1 Института проблем передачи информации им А.А. Харкевича.

Описанные в работе методы были реализованы в программном продукте MACROS, разработанном компанией DATADVANCE, и используются при решении практических задач такими компаниями как Airbus, Eurocopter и др. В диссертации приводится несколько примеров прикладных задач, в которых применение предложенных процедур позволило существенно улучшить качество соответствующих метамоделей.

Публикации

По теме диссертации опубликовано 11 печатных работ, из них 4 работы — статьи в ведущих рецензируемых научных журналах, 7 работ в трудах ведущих российских и международных конференций.

Основные результаты представлены в работах [1], [2], [3] и [4].

В работах [1], [2] и [3] вклад автора был определяющим. В работе [4] автор отвечал за часть связанную с построением суррогатных моделей, в то время как S. Grihon отвечал за запуск оптимизации и валидацию результатов.

Во всех работах Бурнаев Е.В., как научный руководитель, отвечал за глобальную постановку задачи, кроме того автор хотел бы выразить ему благодарность за огромную помощь в написании и правке текстов статей.

Диссертационная работа была выполнена при поддержке лаборатории структурных методов анализа данных в предсказательном моделировании МФТИ, грант правительства РФ дог. 11.G34.31.0073.

Объем и структура работы

Диссертация состоит из введения, четырех глав, заключения и приложения. Полный объем диссертации составляет 121 страницу с 24 рисунками и 1 таблицей. Список литературы содержит 54 наименования.

Содержание работы

В *первой* главе описывается постановка задачи восстановления зависимости по данным, а также определяются такие понятия как базовая модель и обобщающая способность модели.

Пусть задана обучающая выборка $\mathbf{S}_m = (\mathbf{X}_m, \mathbf{Y}_m)$, $\mathbf{X}_m = \{\mathbf{x}^{(i)}, i = 1, \dots, m\}$, $\mathbf{Y}_m = \{\mathbf{y}^{(i)}, i = 1, \dots, m\}$, такая, что

- входные вектора (входы) $\mathbf{x}^{(i)} \in \mathbb{X} \subset \mathbf{R}^d$ (\mathbb{X} — произвольное компактное множество) порождаются независимым образом из некоторого неизвестного непрерывного распределения $P(\mathbf{x})$, $\mathbf{x} \in \mathbb{X}$,
- между выходным значением (выходом) $\mathbf{y}^{(i)} \in \mathbb{Y} \subset \mathbf{R}^1$ и входным

вектором $\mathbf{x}^{(i)} \in \mathbb{X}$ есть некоторая неизвестная функциональная зависимость $\mathbf{y}^{(i)} = f(\mathbf{x}^{(i)})$.

Задача восстановления неизвестной зависимости $\mathbf{y} = f(\mathbf{x})$ состоит в построении по обучающей выборке \mathbf{S}_m регрессионной зависимости (аппроксиматора) $\hat{f}(\mathbf{x}) = \hat{f}(\mathbf{x}|\mathbf{S}_m)$, такой, что её обобщающая способность, т.е. величина (риск) $er_{P,f}(\hat{f}) = \mathbb{E} \left(f(\mathbf{x}) - \hat{f}(\mathbf{x}) \right)^2 = \int_{\mathbb{X}} \left(f(\mathbf{x}) - \hat{f}(\mathbf{x}) \right)^2 dP(\mathbf{x})$ мала с вероятностью, близкой к единице. Здесь \mathbb{E} обозначает математическое ожидание по распределению $P(\mathbf{x})$.

На практике распределение $P(\mathbf{x})$ неизвестно, поэтому вместо $er_{P,f}(\hat{f})$ подсчитывают среднеквадратичную ошибку (эмпирический риск) $\hat{e}r_{\mathbf{S}_m}(\hat{f}) = \left\| f - \hat{f} \right\|_{\mathbf{x}_m}^2$, где выборочная норма $\|v\|_{\mathbf{x}_m}$ функции v определяется как $\|v\|_{\mathbf{x}_m} = \sqrt{\frac{1}{m} \sum_{i=1}^m (v(\mathbf{x}^{(i)}))^2}$.

Будем называть базовым аппроксиматором (базовой моделью) функцию $g = \text{Train}(\mathbf{S}_m, \mathbf{w})$, порождаемую некоторым методом обучения $\text{Train}(\cdot)$, применённым к обучающей выборке \mathbf{S}_m и, возможно, зависящим от некоторого набора гиперпараметров $\mathbf{w} \in \Omega$ (например, случайной инициализации начальных значений параметров базового аппроксиматора).

В качестве метода обучения могут быть использованы многие алгоритмы машинного обучения, такие как: линейная регрессия, регрессия на основе гауссовских процессов (кригинг), регрессия на основе опорных векторов, искусственные нейронные сети (ИНС), регрессионные деревья (CART).

Во *второй* главе рассматривается случай, когда базовая модель слишком проста для того, чтобы описать неизвестную зависимость с достаточной точностью.

Как вариант решения проблемы предлагается метод построения ансамблей регрессионных моделей БегБуст. На примере применения к искусственным нейронным сетям при некоторых дополнительных предположениях доказывается сходимость метода, а также проводится его сравнение с основными подходами к построению ансамблей, описанными в литературе.

В разделе **2.1** приводится постановка задачи построения ансамблей.

Под построением ансамбля базовых аппроксиматоров для решения задачи восстановления неизвестной зависимости на основе данных понимается способ получения аппроксиматора $\hat{f}(\mathbf{x})$ в виде некоторого функционала $\hat{f}(\mathbf{x}) = f_n(g_1(\mathbf{x}), \dots, g_n(\mathbf{x}))$, где $g_i(\mathbf{x})$, $i = 1, \dots, n$ являются базовыми аппроксиматорами из некоторого заданного класса G , построенными на основе данных \mathbf{S}_m . Рассматривается случай, когда ансамбль f_n линеен по g_1, \dots, g_n , т.е.

$$f_n = \sum_{i=1}^n \alpha_i g_i$$

для некоторых коэффициентов $\{\alpha_i\}_{i=1}^n$.

Таким образом, алгоритм построения ансамбля сводится к описанию того, как построить базовые модели g_1, \dots, g_n и задать веса $\{\alpha_i\}_{i=1}^n$.

В разделе 2.2 приводится описание основных методов построения ансамблей в литературе, а также полученные для них теоретические результаты.

В разделе 2.3 приводится описание предложенной методики.

Идея метода *БегБуст* состоит в том, чтобы сохранить форму предсказания ансамбля в виде усреднения предсказаний всех входящих в него аппроксиматоров и одновременно последовательно обучать базовые аппроксиматоры не на первоначальной выборке \mathbf{S}_m , а на некоторой разности выходных значений из выборки и предсказаний аппроксиматоров ансамбля, построенных на соответствующих предыдущих итерациях.

Допустим, что ансамбль после n шагов имеет вид $f_n = \frac{1}{n} \sum_{j=1}^n g_j(\mathbf{x})$, т.е. фактически является усреднением n базовых моделей. Необходимо строить модели g_j , $j = n+1, n+2, \dots$ последовательно, минимизируя эмпирический риск на обучающей выборке \mathbf{S}_m так, чтобы ансамбли f_{n+1}, f_{n+2}, \dots сохранили структуру усреднения.

Предложенный способ построения ансамбля позволяет добиться отрицательной скоррелированности каждой следующей базовой модели с предыдущими, что, как известно, улучшает обобщающую способность, при этом веса $\{\alpha_i\}_{i=1}^n$ в ансамбле не требуют подстройки, поскольку фиксируются равными $\frac{1}{n}$.

В разделе 2.4 приводятся основные теоретические результаты, полученные для метода *БегБуст*.

В определенных предположениях удается доказать, что ошибка $er_{P,f}(f_n)$ аппроксимации метода *БегБуст*, реализующего описанную выше идею, сходится к нулю.

Обозначим через

- G — некоторое подмножество функционального гильбертова пространства H с функциональной нормой $\|\cdot\|$;
- $co(G) = \{g : g = \sum_{i=1}^n \alpha_i g_i, \alpha_i \geq 0, \sum_{i=1}^n \alpha_i = 1, g_i \in G, n \geq 1\}$ — выпуклую оболочку множества G ;
- $\overline{co}(G)$ — замыкание выпуклой оболочки $co(G)$.

Положим

$$\varepsilon_k = \hat{er}_{\mathbf{S}_m^k}(g_k) - \inf_{g \in G} \hat{er}_{\mathbf{S}_m^k}(g), k = 1, 2, \dots,$$

$\varepsilon_n^{\max} = \max_{k=1, \dots, n} \{\varepsilon_k\}$, где $\mathbf{S}_m^k = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i),k}), i = 1, \dots, m\}$, $\mathbf{y}^{(i),k} = k \cdot \mathbf{y}^{(i)} - (k-1) \cdot f_{k-1}(\mathbf{x}^{(i)})$, $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \in \mathbf{S}_m$. В дальнейшем будут использоваться следующие предположения

- **A₁**. Подмножество G имеет вид

$$G = \{\alpha \psi(\cdot, \theta) \mid |\alpha| \leq B, \theta \in \Theta\} \subset L^2(\mathbb{X}, P),$$

где $\alpha \in \mathbf{R}^1$, $\psi(\mathbf{x}, \theta)$ — сигмоидальная функция активации, множество значений параметров $\theta \in \Theta \subset \mathbf{R}^{d+1}$ (Θ — произвольное компактное множество), $0 < B < \infty$ — заданная константа.

- **A₂**. Неизвестная функция f принадлежит $\overline{co}(G)$.
- **A₃**. Для некоторого $\varepsilon > 0$ при $k \geq 1$ выполнено, что

$$\varepsilon_k \leq \varepsilon.$$

Пусть $\Delta = B^2 - \frac{1}{m} \sum_{i=1}^m (\mathbf{y}^{(i)})^2 > 0$. При введенных предположениях **A₁** и **A₂** возможно доказать следующую лемму.

Лемма 1. *Для любой $f \in \overline{co}(G)$, фиксированной выборки \mathbf{S}_m , $h \in co(G)$ и $\alpha \in [0, 1]$ справедливо, что*

$$\min_{g \in G} \hat{er}_{\mathbf{S}_m}(\alpha h + (1-\alpha)g) \leq \alpha^2 \hat{er}_{\mathbf{S}_m}(h) + (1-\alpha)^2 \left(\max_{g \in G} \|g\|_{\mathbf{X}_m}^2 - \hat{er}_{\mathbf{S}_m}(0) \right).$$

Следствие. Для любой $f \in \overline{co}(G)$, фиксированной выборки \mathbf{S}_m , $h \in co(G)$ и $\alpha \in [0, 1]$ справедливо, что

$$\min_{g \in G} \hat{e}_{\mathbf{S}_m}(\alpha h + (1 - \alpha)g) \leq \alpha^2 \hat{e}_{\mathbf{S}_m}(h) + (1 - \alpha)^2 \Delta.$$

На основании приведенной леммы при выполнении предположений \mathbf{A}_1 и \mathbf{A}_2 может быть доказана следующая теорема.

Теорема 1. Для фиксированной выборки \mathbf{S}_m и любого $k = 1, 2, \dots$ выполнено, что

$$\hat{e}_{\mathbf{S}_m}(f_k) \leq \frac{\Delta + \varepsilon_k^{\max}}{k},$$

Следствие. Если на каждом шаге алгоритма БегБуст при построении базового аппроксиматора g_k достигается минимум ошибки $\hat{e}_{\mathbf{S}_m^k}(g_k)$, $k = 1, 2, \dots$, т.е. $\varepsilon_k = 0$, то ошибка $\hat{e}_{\mathbf{S}_m}(f_k)$ убывает как $\frac{\Delta}{k}$.

Данная теорема дает оценку сверху ошибки приближения построенной модели для произвольной выборки. Проведенные эксперименты показывают, что теорема выполняется на практике (рис. 1).

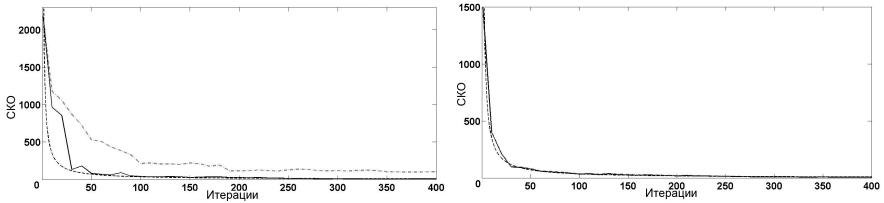


Рис. 1. Зависимость ошибок на обучающей (черная сплошная линия) и проверочной (серая штриховая линия) выборках от числа итераций для обучающей выборки объема $t = 100$ (слева) и $t = 1000$ (справа), функция Branin . Для сравнения приведена кривая $\frac{\Delta}{k}$ (черная пунктирная линия). СКО — средне-квадратичная ошибка.

Используя предыдущую теорему, также можно получить ограничение сверху на обобщающую способность:

Теорема 2. Пусть выполняются предположения \mathbf{A}_1 , \mathbf{A}_2 и \mathbf{A}_3 , тогда для любой $f \in \overline{co}(G)$, для всех $\gamma \in (0, 1)$ и $\rho > 0$, если алгоритм БегБуст применяется к выборке \mathbf{S}_m размера по крайней мере $t(\rho, \gamma) =$

$\frac{C_1}{\rho^2} \left(\ln \frac{4}{\gamma} + \frac{C_2}{\rho^2} \ln \frac{C_3}{\rho^5} \right)$, где $C_1 = 2^{12} B^4$, $C_2 = 5 \cdot 2^{14} (d+3) B^4$ и $C_3 = \frac{2^{21} \epsilon B^6}{(d+1)} C_2$, то с вероятностью не менее $1 - \gamma$ через $n = \lceil \frac{2(\Delta + \epsilon)}{\rho} \rceil$ шагов алгоритма БезБуст ошибка аппроксимации ансамбля f_n будет удовлетворять неравенству $er_{P,f}(f_n) < \rho$.

В разделе **2.5** приводится сравнение предложенного и представленных в литературе методов построения ансамблей на наборе искусственных задач, из которого видно, что при своей простоте предложенный подход демонстрирует хорошее качество работы.

В *третьей* главе рассматривается случай, когда во входных признаках модели содержится избыточность. Предполагается, что набор признаков избыточен в том смысле, что функция зависит только от проекции входов на некоторое линейное подпространство. Такое подпространство называется центральным средним подпространством регрессии. В литературе присутствует множество процедур, позволяющих оценивать эту проекцию, однако, как правило, при использовании этих процедур требуется заранее указывать размерность сжатия, а также строить ядерные оценки зависимости (с подбором параметров через кросс-проверку), что работает хорошо лишь в случае малых размерностей и делает эти методы малоприменимыми на практике.

По этой причине автором были предприняты попытки улучшить качество оценок проекции, результатом которых стало создание алгоритма VEGA (Variable Extraction via Gradient Approximation, Выделение Переменных Основанное на Аппроксимации Градиентов), позволяющего оценивать центральное среднее подпространство регрессии точнее, чем найденные автором в литературе подходы.

В разделе **3.1** приводится детальное описание алгоритма VEGA.

Пусть аппроксимируемая функция имеет следующую структуру

$$y = f(\mathbf{x}) = h(\mathbf{x}\mathbf{B}), \quad (1)$$

где $\mathbf{B} \in \mathbf{R}^{d \times p}$ ($p \leq d$) — искомая матрица, задающая подпространство сниженной размерности p , то есть, как и говорилось выше, функция f зависит только от проекции \mathbf{x} на некоторое подпространство, заданное матрицей \mathbf{B} .

Будем предполагать, что функция $f(\mathbf{x})$ является реализацией некоторого гауссовского процесса⁹, тогда структуру типа (1) можно достаточно удобно описать, предположив, что ковариационная функция гауссовского процесса имеет вид

$$k(\mathbf{x}, \mathbf{x}') = \sigma_0^2 \exp\left(-\frac{1}{2} \text{dist}(\mathbf{x}, \mathbf{x}')\right), \quad (2)$$

где расстояние Махаланобиса

$$\text{dist}(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')\mathbf{A}(\mathbf{x} - \mathbf{x}')^T = (\mathbf{x} - \mathbf{x}')\mathbf{B}\mathbf{A}\mathbf{B}^T(\mathbf{x} - \mathbf{x}')^T, \quad (3)$$

\mathbf{B} – ортогональная матрица размера $d \times p$, $p \leq d$, \mathbf{A} – диагональная матрица размера $p \times p$ с элементами $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ на диагонали. При такой параметризации матрица \mathbf{B} осуществляет поворот осей координат, а диагональная матрица \mathbf{A} регулирует ширину ядер вдоль новых повернутых осей.

Если матрица \mathbf{B} фиксирована и известна, то аппроксимацию \hat{f} можно искать в виде $\hat{f}(\mathbf{x}) = \hat{h}(\mathbf{x}\mathbf{B})$, где \hat{h} – апостериорное среднее значение гауссовского процесса h с функцией ковариации (2), в которой у расстояния Махаланобиса (3) $\mathbf{A} \equiv \mathbf{A}$, то есть нет поворота осей координат. В этом случае достаточно оценить (например, с помощью максимизации правдоподобия выборки \mathbf{S}_m) только p гиперпараметров ковариационной функции – элементы матрицы \mathbf{A} .

Если же матрица \mathbf{A} фиксирована и известна, то матрицу \mathbf{B} можно оценить из следующих соображений.

Пусть

- $\mathbf{M}_i = \left\{ \frac{\partial^2 k(\mathbf{x}, \mathbf{x})}{\partial x_j \partial x_l} \Big|_{\mathbf{x}=\mathbf{x}^{(i)}}, j, l = 1, \dots, d \right\} \in \mathbf{R}^{d \times d}$,
- $\mathbf{k}(\mathbf{x}) = \{k(\mathbf{x}, \mathbf{x}^{(i)}), i = 1, \dots, m\} \in \mathbf{R}^{1 \times m}$, $\mathbf{J}_i = \frac{\partial \mathbf{k}(\mathbf{x})}{\partial \mathbf{x}} \Big|_{\mathbf{x}=\mathbf{x}^{(i)}} \in \mathbf{R}^{d \times m}$,
- $\mathbf{K}_0 = \{k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}), i, j = 1, \dots, m\} \in \mathbf{R}^{m \times m}$,
- $\{\mathbf{\Gamma}_1, \dots, \mathbf{\Gamma}_m\} = \left\{ \frac{\partial f(\mathbf{x})}{\partial x_1} \Big|_{\mathbf{x}=\mathbf{x}^{(i)}}, \dots, \frac{\partial f(\mathbf{x})}{\partial x_d} \Big|_{\mathbf{x}=\mathbf{x}^{(i)}} \right\}_{i=1}^m \in \mathbf{R}^{m \times d}$ – неизвестные истинные значения градиентов $f(\mathbf{x})$ в точках выборки \mathbf{S}_m .

⁹ *Rasmussen C.E., Williams C.K.I. Gaussian Processes for Machine Learning. MIT Press, 2006.*

Тогда выполнены следующие утверждения.

Утверждение 1. Градиенты $f(\mathbf{x})$ в точках выборки \mathbf{S}_m имеют апостериорное распределение

$$\mathbf{\Gamma}_i \sim \mathcal{N}(\hat{\mathbf{\Gamma}}_i, \hat{\mathbf{\Sigma}}_i),$$

где $\hat{\mathbf{\Gamma}}_i = \mathbf{Y}_m^T \mathbf{K}_0^{-1} \mathbf{J}_i^T$, $\hat{\mathbf{\Sigma}}_i = \mathbf{M}_i - \mathbf{J}_i \mathbf{K}_0^{-1} \mathbf{J}_i^T$.

Утверждение 2. Пусть данные \mathbf{S}_m порождены гауссовским процессом с ковариационной функцией (2) и расстоянием Махаланобиса (3) с параметрами $\mathbf{A} = \mathbf{B}\mathbf{A}\mathbf{B}^T$, тогда

- $\hat{\mathbf{\Gamma}}_i = \mathbf{Y}_m^T \mathbf{K}_0^{-1} \text{diag}(\mathbf{k}(\mathbf{x}^{(i)})) (\mathbf{e}^T \mathbf{x}^{(i)} - \mathbf{X}_m) \mathbf{A}$, где \mathbf{e} — вектор из единиц размерности m ,
- $\hat{\mathbf{\Sigma}}_i = \sigma_0^2 \mathbf{A} - \mathbf{A} \tilde{\mathbf{\Sigma}}_i \mathbf{A}$,
где $\tilde{\mathbf{\Sigma}}_i = (\mathbf{e}^T \mathbf{x}^{(i)} - \mathbf{X}_m)^T \text{diag}(\mathbf{k}(\mathbf{x}^{(i)})) \mathbf{K}_0^{-1} \text{diag}(\mathbf{k}(\mathbf{x}^{(i)})) (\mathbf{e}^T \mathbf{x}^{(i)} - \mathbf{X}_m)$,
- $\Sigma = \mathbb{E}(\sum_{i=1}^m \mathbf{\Gamma}_i^T \mathbf{\Gamma}_i | \mathbf{S}_m) = \sum_{i=1}^m (\hat{\mathbf{\Gamma}}_i^T \hat{\mathbf{\Gamma}}_i + \hat{\mathbf{\Sigma}}_i)$, где \mathbb{E} — среднее значение по апостериорному распределению рассматриваемого гауссовского процесса.

Следствие. Пусть данные \mathbf{S}_m порождены гауссовским процессом с ковариационной функцией (2) и расстоянием Махаланобиса (3) с параметрами $\mathbf{A} = \mathbf{B}\mathbf{A}\mathbf{B}^T$, тогда градиенты $f(\mathbf{x})$ лежат в подпространстве, натянутом на столбцы матрицы \mathbf{B} , которое совпадает с подпространством, натянутом на собственные вектора апостериорной матрицы ковариаций градиентов Σ .

Кроме того эксперименты показывают, что качество итоговой оценки матрицы \mathbf{B} может быть заметно улучшено итеративным образом:

- пусть сделана $(j - 1)$ итерация и получены оценки $\{\mathbf{B}_k\}_{k=1}^{j-1}$,
- по новой выборке $\tilde{\mathbf{S}}_m$ с преобразованными входными векторами $\tilde{\mathbf{x}} = \mathbf{x} \prod_{k=1}^{j-1} \mathbf{B}_k$ строится аппроксимация $f(\mathbf{x})$ с помощью апостериорного среднего значения \hat{h} гауссовского процесса h с функцией ковариации (2), в которой у расстояния Махаланобиса (3) $\mathbf{A} \equiv \mathbf{A}$,

- на основе оценки матрицы \mathbf{A} , полученной при построении аппроксимации \hat{h} , и утверждения 2 оценивается апостериорная матрица ковариаций градиентов Σ^j ,
- матрица \mathbf{B}_j формируется из собственных векторов матрицы Σ^j .

В таком случае итоговая оценка матрицы \mathbf{B} получается как $\mathbf{B} = \prod_j \mathbf{B}_j$.

Оценив собственные числа матрицы Σ , можно отранжировать направления, задаваемые столбцами матрицы \mathbf{B} , по значимости: малость собственного числа означает, что вдоль соответствующего направления значение функции также меняется мало, что фактически позволяет построить процедуру линейного снижения размерности входных данных.

В разделе 3.2 приводится обзор методов оценки размерности центрального среднего подпространства.

В результате, как наиболее удобный алгоритм оценки размерности предлагается использовать выбор размерности p по значениям собственных чисел апостериорной матрицы ковариаций градиентов Σ , при этом пороговое значение оценивается либо с помощью бутстрепа, либо с помощью метода складного ножа (или аналогичного метода оценки стабильности собственных подпространств).

При прямом вычислении собственных чисел на каждой итерации метода складного ножа пришлось бы заново обращать матрицу ковариации (что имеет сложность $O(m^3)$ и является довольно затратной операцией для больших выборок). Можно сэкономить время и сделать обращение не на каждой итерации, а лишь один раз. Действительно, обозначим через $\mathbf{S}_{\sim i} = (\mathbf{X}_{\sim i}, \mathbf{Y}_{\sim i})$ выборку, которая получается из выборки \mathbf{S}_m выкидыванием i -го наблюдения, тогда верно следующее

Утверждение 3. Пусть данные \mathbf{S}_m порождены гауссовским процессом с ковариационной функцией (2) и расстоянием Махаланобиса (3) с параметрами $\mathbf{A} = \mathbf{B}\mathbf{B}^T$, тогда значения апостериорного среднего и апостериорной матрицы ковариаций градиентов $f(\mathbf{x})$ в точке $\mathbf{x}^{(i)}$ можно вычислить следующим образом

$$\hat{\Gamma}_{\sim i} = \mathbf{Y}_m^T \hat{\mathbf{K}}_i \text{diag} \left(\mathbf{k}(\mathbf{x}^{(i)}) \right) \left(\mathbf{e}^T \mathbf{x}^{(i)} - \mathbf{X}_m \right) \mathbf{A},$$

$$\hat{\Sigma}_{\sim i} = \sigma_0^2 \mathbf{A} - \mathbf{A} \tilde{\Sigma}_{\sim i} \mathbf{A},$$

где $\tilde{\Sigma}_{\sim i} = (\mathbf{e}^T \mathbf{x}^{(i)} - \mathbf{X}_m)^T \text{diag}(\mathbf{k}(\mathbf{x}^{(i)})) \hat{\mathbf{K}}_i \text{diag}(\mathbf{k}(\mathbf{x}^{(i)})) (\mathbf{e}^T \mathbf{x}^{(i)} - \mathbf{X}_m)$, матрица $\hat{\mathbf{K}}_i$ получается из матрицы $\tilde{\mathbf{K}}_i$ занулением i -ой строки и i -го столбца, $\tilde{\mathbf{K}}_i = \mathbf{K}_0^{-1} + \frac{(\mathbf{K}_0^{-1})_i^T (\mathbf{K}_0^{-1})_i}{(\mathbf{K}_0^{-1})_{ii}}$, здесь $(\mathbf{K}_0^{-1})_i$ — i -я строка матрицы.

В разделе 3.3 проводится обширное экспериментальное сравнение с другими методами эффективного снижения размерности из литературы. Демонстрируется, что метод позволяет обеспечить инвариантность качества аппроксимации к повороту координатных осей.

Кроме того, описывается решение задачи аппроксимации энергозатрат вертолета от параметров маневра, включающих 51 параметр: масса, моменты вращения, углы, высота и т.п. Эффективная размерность данных, оцененная с помощью предложенного статистического теста, равна 15. Сокращение числа параметров в ≈ 4 раза за счет выделения наиболее значимых направлений с помощью предложенного алгоритма позволило существенно уменьшить сложность задачи и при этом улучшить качество аппроксимации.

В *четвертой* главе обсуждается ситуация, когда выборка неоднородна, а аппроксимируемая зависимость может иметь разрывы.

В этих случаях оказывается эффективным проводить сначала предварительную кластеризацию данных, после чего в каждом кластере строить свою базовую модель (в литературе такой подход называется построением смеси экспертов).

Рассматриваются следующие распространенные модели неоднородности:

- 1) данные порождены смесью нормальных распределений,
- 2) данные описываются набором областей, разделенных линейными гиперплоскостями.

У каждого элемента смеси или в каждой области предполагается свой вид зависимости.

Отметим также, что, так как алгоритмы кластеризации обычно имеют меньшую вычислительную сложность по размеру выборки, чем методы построения базовых моделей, то предварительная кластеризация позволяет, кроме прочего, строить модели и в случае очень больших выборок,

поскольку каждая базовая модель может строиться только по тем точкам, которые принадлежат соответствующему кластеру.

В разделе 4.1 приводится обзор методов построения смесей экспертов.

В разделе 4.2 рассматривается ситуация, когда данные заполняют пространство неравномерно, образуя несвязные облака точек.

В этом случае удобно предположить, что данные были сгенерированы смесью нормальных распределений, то есть точки $(\mathbf{x}, \mathbf{y}) \in \mathbf{S}_m$ порождены в соответствии с распределением

$$\text{Law}(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{N_x} \alpha_k \mathcal{N}(\mu_k, \Theta_k),$$

где μ_k и Θ_k — среднее и матрица ковариации k -го нормального распределения смеси для выборки \mathbf{S}_m . Соответственно, если рассматривать только входной вектор \mathbf{x} , то

$$\text{Law}(\mathbf{x}) = \sum_{k=1}^{N_x} \alpha_k \mathcal{N}(\mu_{k,x}, \Theta_{k,x}),$$

где $\mu_{k,x}$ и $\Theta_{k,x}$ — подвектор средних значений вектора μ_k и подматрица ковариаций матрицы Θ_k , задающие безусловное распределение вектора входов $\mathbf{x} \in \mathbb{X}$.

Параметры гауссовской смеси $\{\mu_k, \Theta_k\}_{k=1}^{N_x}$ оцениваются с помощью EM-алгоритма и выделяют в \mathbb{X} области “однородности” $\cup_{k=1}^{N_x} \mathbb{X}^k$ следующим образом. Определим расстояние в метрике Махаланобиса от центра k -го кластера до точки \mathbf{x} в соответствии с формулой

$$d_k(\mathbf{x}) = (\mathbf{x} - \mu_{k,x}) \left(\Theta_{k,x} \right)^{-1} (\mathbf{x} - \mu_{k,x})^T,$$

тогда множество $\mathbb{X}^k = \{\mathbf{x} \in \mathbb{X} : d_k(\mathbf{x}) \leq \chi_{\alpha_1}^2(p)\}$, где $\chi_{\alpha_1}^2(p)$ является α_1 -квантилью распределения χ^2 с p степенями свободы. Локальные аппроксимационные модели $g_k \in G$ для каждой подобласти \mathbb{X}^k строятся по подвыборкам $\mathbf{S}^k = \{(\mathbf{x}, \mathbf{y}) \in \mathbf{S}_m : \mathbf{x} \in \mathbb{X}^k\}$, $k = 1, \dots, N_x$.

Степень вхождения точки \mathbf{x} в k -й кластер можно определить согласно формуле

$$w(k|\mathbf{x}) = \frac{1}{2} \left(\tanh \left(1 - 2 \frac{d_k(\mathbf{x}) - \chi_{\alpha_1}^2(p)}{\chi_{\alpha_2}^2(p) - \chi_{\alpha_1}^2(p)} \right) + 1 \right),$$

где $\chi_{\alpha_2}^2(p)$ и $\chi_{\alpha_1}^2(p)$ являются α_2 и α_1 квантилями распределения χ^2 с p степенями свободы соответственно (в экспериментах брались значения $\alpha_2 = 99\%$ и $\alpha_1 = 97\%$).

Таким образом классификатор, определяющий степень принадлежности точки \mathbf{x} к k -му кластеру, может строиться в соответствии с формулой

$$\hat{w}(k|\mathbf{x}) = \frac{w(k|\mathbf{x})}{\sum_r w(r|\mathbf{x})}. \quad (4)$$

Тогда итоговая метамодель имеет вид $\hat{f}(\mathbf{x}) = \sum_{k=1}^{N_x} \hat{w}(k|\mathbf{x})g_k(\mathbf{x})$.

Эксперименты показывают, что такой подход оказывается надежнее, чем подход, где в качестве весов (4) используются апостериорные вероятности $\hat{w}(k|\mathbf{x}) = P(k|\mathbf{x})$ точки \mathbf{x} принадлежать соответствующим кластерам¹⁰.

В разделе 4.3 рассматривается случай, когда функция f имеет разрывы вдоль линейных гиперплоскостей (“линейная” структура разрывов).

Один из способов моделировать такую функцию с разрывом в точке θ вдоль направления \mathbf{v} состоит в том, чтобы описать разрыв функцией-ступенькой $\sigma\left((\mathbf{x} - \theta)^T \mathbf{v}\right)$ (такой, например, как функция Хевисайда или логистическая функция), а по обе стороны от него задать гладкие базовые модели $g_1(\mathbf{x})$ и $g_2(\mathbf{x})$, определив итоговую модель как

$$\hat{f}(\mathbf{x}) = \sigma\left((\mathbf{x} - \theta)^T \mathbf{v}\right) g_1(\mathbf{x}) + \left(1 - \sigma\left((\mathbf{x} - \theta)^T \mathbf{v}\right)\right) g_2(\mathbf{x}),$$

или

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^I \omega_i(\mathbf{x}) g_i(\mathbf{x}), \quad (5)$$

где $I = 2$, $\sum_{i=1}^I \omega_i(\mathbf{x}) = 1$ и $\omega_i(\mathbf{x})$ — вес i -й модели в точке \mathbf{x} . В этом случае по одну сторону от разрыва функция будет вести себя как g_1 , а по другую как g_2 , “резкость” перехода будет контролироваться $\|\mathbf{v}\|$.

Для описания разрывов, имеющих более сложную, чем линейная, структуру, можно, сохранив форму модели (5), усложнить ее следующим образом. Поставим в соответствие модели \hat{f} бинарное дерево, на листьях

¹⁰Bettebghor D., Bartoli N., Grihon S., Morlier J., Samuelides M. Surrogate modeling approximation using a mixture of experts based on EM joint estimation, Structural and Multidisciplinary Optimization, pages 1–17, 2011, Springer.

которого находятся базовые модели $g_i, i = 1, \dots, I$, а каждая ветвь, выходящая из вершины J , имеет в точке \mathbf{x} вес $\phi_1^{(J)}(\mathbf{x}) = \sigma((\mathbf{x} - \theta_J) \mathbf{v}_J)$, если это левая ветвь, или $\phi_2^{(J)}(\mathbf{x}) = (1 - \sigma((\mathbf{x} - \theta_J) \mathbf{v}_J))$, если это правая ветвь. При вычислении $\hat{f}(\mathbf{x})$ по формуле (5) вес $\omega_i(\mathbf{x})$ может быть найден как произведение весов $\phi_l^{(M)}(\mathbf{x}), l = 1, 2$ по всем ветвям, ведущим от i -го листа к корню дерева.

Экспериментально показывается, что предложенная форма весов $\omega_i(\mathbf{x})$ позволяет моделировать функции с разрывами, имеющими более сложную, чем линейная, структуру.

Определим критерии качества, которые будут использоваться при настройке параметров модели.

Качество базовой модели на i -м листе будем оценивать как

$$E_{leaf}^i = \sum_{n=1}^m \left(\mathbf{y}^{(n)} - g_i(\mathbf{x}^{(n)}) \right)^2 \omega_n^i(\mathbf{x}^{(n)}),$$

где $\omega_n^i(\mathbf{x}^{(n)})$ – произведение весов $\phi_l^{(M)}(\mathbf{x}^{(n)}), l = 1, 2$ по всем ветвям, ведущим от i -го листа к корню дерева.

Качество подбора параметров θ_J, \mathbf{v}_J на ветвях, выходящих из вершины J , при фиксированных базовых моделях g_{J_1} и g_{J_2} будем оценивать как

$$E_{vertex}^J = \sum_{n=1}^m \left(\mathbf{y}^{(n)} - \phi_1^{(J)}(\mathbf{x}^{(n)}) g_{J_1}(\mathbf{x}^{(n)}) - \phi_2^{(J)}(\mathbf{x}^{(n)}) g_{J_2}(\mathbf{x}^{(n)}) \right)^2 \tilde{\omega}_n^J(\mathbf{x}^{(n)}),$$

где $\tilde{\omega}_n^J(\mathbf{x}^{(n)})$ – произведение весов $\phi_l^{(M)}(\mathbf{x}^{(n)}), l = 1, 2$ по всем ветвям, ведущим от J -й вершины к корню дерева (в корне дерева принимается, что $\tilde{\omega}_n^0(\mathbf{x}^{(n)}) = 1$).

Построение дерева производится итеративно и заключается в том, что на каждой итерации выбирается лист с наибольшим значением E_{leaf} и из него пускаются две ветви с новыми листьями (на первой итерации вместо листа с наибольшим E_{leaf} берется корень дерева).

Нахождение параметров \mathbf{v}_J и θ_J , весов ветвей, выходящих из вершины J , и соответствующих базовых моделей g_{J_1} и g_{J_2} на листах производится следующим образом:

- 1) \mathbf{v}_J и θ_J инициализируются некоторыми начальными значениями,
- 2) для текущих значений \mathbf{v}_J и θ_J через минимизацию $E_{leaf}^{J_1}$ и $E_{leaf}^{J_2}$ находятся g_{J_1} и g_{J_2} ,
- 3) для найденных g_{J_1} и g_{J_2} через минимизацию E_{vertex}^J подбираются параметры \mathbf{v}_J и θ_J ,
- 4) шаги 2 и 3 повторяются до сходимости оценок.

В работе проводится экспериментальное сравнение различных методов подбора параметров \mathbf{v}_J , θ_J и показывается, что приведенная выше методика обеспечивает наилучшее качество построения \hat{f} в смысле точности приближения f .

Отличие рассмотренной процедуры от встречающихся в литературе подобных подходов¹¹ состоит в том, что вместо того, чтобы на каждой вершине проводить оптимизацию только по точкам, лежащим по одну сторону от гиперплоскости, заданной параметрами \mathbf{v} и θ на вершине выше, вводятся веса и используются все точки выборки. В экспериментах показывается, что введение гладких весов значительно уменьшает артефакты в аппроксимации на разрывах и в целом повышает точность построенных моделей.

В случае, если в качестве g_i используются линейные модели, то для функционала E_{vertex}^J на каждом шаге оптимизации 3) можно выписать выражение для ошибки скользящего контроля $E_{vertex(loo)}^J$ (значение ошибки E_{vertex}^J , подсчитываемое в точках $\mathbf{x}^{(n)}$ обучающей выборки \mathbf{S}_m для фиксированных \mathbf{v}_J , θ_J и $\tilde{\omega}^J(\mathbf{x})$ при условии, что базовые модели g_{J_1} и g_{J_2} для каждого n строятся по обучающей выборке с исключенным n -м наблюдением).

Доказывается, что выполнено

Утверждение 4. Пусть $\mathbf{H}_i^J = \left(\tilde{\mathbf{X}}_m^T \mathbf{Q}_i^J \tilde{\mathbf{X}}_m \right)^{-1}$, где $\tilde{\mathbf{X}} = [\mathbf{X} \mathbf{e}^T]$, $\mathbf{Q}_i^J = \text{diag} \left(\phi_i^{(J)}(\mathbf{x}^{(1)}), \dots, \phi_i^{(J)}(\mathbf{x}^{(m)}) \right)$. Тогда ошибка скользящего контроля для

¹¹C. Hametner and S. Jakubek, *Neuro-fuzzy modelling using a logistic discriminant tree*, American Control Conference-2007 (ACC -07), pp. 864-869.

функционала E_{vertex}^J может быть записана как

$$E_{vertex(loo)}^J = \sum_{n=1}^m \left(\sum_{i=1}^2 \phi_i^{(J)}(\mathbf{x}^{(n)}) \frac{\mathbf{y}^{(n)} - g_{J_i}(\mathbf{x}^{(n)})}{1 - \phi_i^{(J)}(\mathbf{x}^{(n)}) (\tilde{\mathbf{x}}^{(n)})^T \mathbf{H}_i^J \tilde{\mathbf{x}}^{(n)}} \right)^2, \quad (6)$$

здесь $\tilde{\mathbf{x}} = (\mathbf{x}, 1)$, g_{J_i} — линейные модели, построенные по всей выборке $\tilde{\mathbf{X}}_m$ минимизацией E_{leaf}^i .

Использование $E_{vertex(loo)}^J$ вместо E_{vertex}^J при нахождении весов на новых ветвях может уменьшить переобучение модели.

На рисунке 2 показан пример работы метода при аппроксимации разрывной функции, в качестве базовых моделей использовались линейные модели.

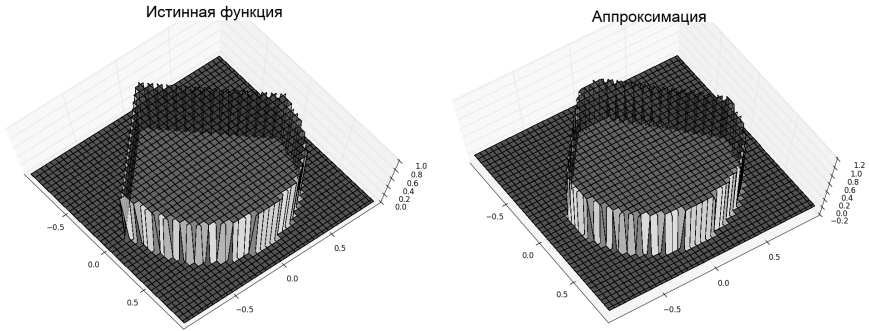


Рис. 2. Настоящая функция (слева), аппроксимация смесью линейных экспертов (справа)

В разделе 4.4 описывается сравнение предложенных методов с методами из литературы. Приводится пример успешного применения предложенной методологии к задаче построения суррогатных моделей ограничения в задаче оптимизации веса обшивки корпуса самолета.

В **заключении** сформулированы основные результаты диссертационной работы:

1. В случае, если неизвестная функция зависит не от всего пространства аргументов, а только от проекции на некоторое линейное подпространство, разработанный алгоритм VEGA позволяет найти это

подпространство, оценить его размерность, а также обеспечивает устойчивость точности построенной модели к повороту координатных осей в пространстве дизайна.

2. В случае, если базовая модель слишком проста для воспроизведения искомой зависимости, разработанный метод построения ансамблей базовых моделей БегБуст позволяет улучшить качество аппроксимации. При выполненных предположениях на зависимость для алгоритма БегБуст также получена верхняя граница ошибки обобщающей способности.
3. В случае, если приближаемая функция имеет разрывы, либо имеющаяся выборка неравномерно заполняет пространство дизайна, разработанные методы построения метамоделей на основе кластеризации данных с помощью гауссовских смесей или разделения линейными гиперплоскостями позволяют улучшить качество приближения.

Список публикаций

- 1) Бурнаев Е.В., Приходько П.В. Методология построения суррогатных моделей для аппроксимации пространственно неоднородных функций // Труды МФТИ. Т. 5, No. 4. 2013. С. 122-132.
- 2) Бурнаев Е.В., Ерофеев П.Д., Приходько П.В. Выделение главных направлений в задаче аппроксимации на основе гауссовских процессов // Труды МФТИ. Т. 5, No. 3. 2013. С. 24-35.
- 3) Бурнаев Е.В., Приходько П.В. Об одной методике построения ансамблей регрессионных моделей // Автоматика и телемеханика. 2013. Т. 10, С. 36-54.
- 4) S. Grihon, E. Burnaev, P. Prikhodko et al. Surrogate Modeling of Stability Constraints for Optimization of Composite Structures // Surrogate-Based Modeling and Optimization for Engineering applications / Eds. by S. Koziel, L. Leifsson. Springer, 2012. P. 359-391.

- 5) Burnaev E., Prikhodko P., Struzik A. Surrogate models for helicopter loads problems // Proceedings of 5th European Conference for Aerospace Sciences, 2013, 11 P.
- 6) Бурнаев Е.В., Приходько П.В. Эффективное снижение размерности на основе гауссовских процессов // Сборник докладов конференций ИОИ-9. 2012. С. 188-192.
- 7) Grihon S., Alestra S., Burnaev E., Prikhodko P. Surrogate Modelling of Bucking Analysis of approximation based on linear Expansions in parametric dictionaries // Proceedings of 1st International Conference on Composite Dynamics, Archanon, France, 2012.
- 8) Grihon S., Alestra S., Betebghor D., Burnaev E., Prikhodko P. Comparison of different techniques for surrogate modeling of stability constraints for composite structures // Proceedings of 1st International Conference on Composite Dynamics, Archanon, France, 2012.
- 9) Бурнаев Е.В., Приходько П.В. Теоретические свойства процедуры построения регрессионного ансамбля на основе беггинга и бустинга // Тр. конф. Информационные Технологии и Системы. 2011. С. 438-443.
- 10) Ерофеев П.Д., Приходько П.В. Эффективное снижение размерности на основе аппроксимации гауссовскими процессами и его применение при оптимизации крыла самолета // Труды 54-й научн. конф. МФТИ. Долгопрудный, ФУПМ, 2011. С. 82-83.
- 11) Burnaev. E.V., Belyaev M.G., Prikhodko P.V. Approximation of multidimensional dependency based on an expansion in the parametric functions from dictionary // Proc. of 9th International Conference "Computer Data Analysis and Modeling: Complex Stochastic Data and Systems". Vol. 2. 2010. pp. 223-226.

Подписано в печать 25.10.13

Тираж 100 экз.

Отпечатано: отдел оперативной полиграфии МАРХИ
107031, г. Москва, ул. Рождественка, 11/4, корп. 3

Тел: 8 (495) 621-15-66